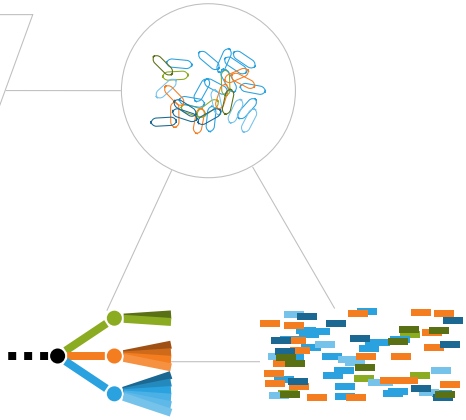
Bioinformatics Master Degree 2.2 - Year 2014 / 2016
Science and Technology Faculty
Rouen University
Normandy University

A two-year apprenticeship thesis

Cheese Ecosystems Metagenomics



Explorations & improvements around a bioinformatics tool

Charlie PAUVERT

Supervisors:
Anne-Laure ABRAHAM, Research Engineer
and Pierre RENAULT, Research Director

Team:
Food and Commensal Bacteria Team (BAC)
Research Unit MICALIS - UMR 1319
The French National Institute for Agricultural Research - AgroParisTech

Remerciements

Je tiens à adresser ci-après mes remerciements aux personnes qui ont permis l'aboutissement de ce rapport, de près ou de loin, et qui ont contribué de façon plus ou moins directe à mes travaux depuis Septembre.

à Nicolas VERGNE, qui a su lever des doutes et répondre à mes anxiétés (parfois infondées) lors de nos rendez-vous à Rouen.

À l'équipe "Bactéries Alimentaires et Commensales" (BAC) pour son accueil agréable ainsi qu'aux collègues de l'étage. Leur patience à m'entendre parler de cuisine régulièrement est louable et ils seront remerciés à mon pot de départ !

À Hugo DEVILLERS, intermittent du bureau (et futur représentant Bodum), pour nos discussions geeks et ses retours d'expérience académiques, autour d'un café bien sûr.

À Thibaut GUIRIMAND, pour ses blagues derrière moi, littéralement, et sans qui les relations entre "collègues" seraient différentes.

À Mathieu ALMEIDA, pour ses suggestions intéressantes et les folles conversations à chacune de ses visites au laboratoire.

À Christine DELORME et Éric GUÉDON pour leurs questions pertinentes lors de mes présentations.

à Sophie SCHBATH sans qui mes premières modélisations n'auraient pas dépassé un stade embryonnaire.

Aux personnes rencontrées à JOBIM, face au poster et surtout ailleurs, pour les discussions re-dynamisantes.

À la communauté du logiciel libre, sans qui de nombreux outils performants n'existeraient pas pour tous.

À mon "bro" de trail, grâce à qui j'ai pu lever la tête du guidon et m'aérer les méninges en forêt cette année.

À Pierre RENAULT, pour ces fascinantes histoires sur la diversité de ses chers micro-organismes et les discussions scientifiques qui suivent.

À Mahendra MARIADASSOU pour ses conseils avisés et ses mots justes, toujours, malgré mes questions bêtes. Sa clarté de formulation des problèmes et l'élégance de ses solutions est stimulante.

À Anne-Laure ABRAHAM, dont j'ai parfois maudit le stylo correcteur, mais en appréciant toujours la justesse de ses remarques et son art de la reformulation claire. Sa patience quotidienne face à mes digressions et bavardages est désormais légendaire. Lorsque je serais à moitié aussi bien organisé qu'elle, ce sera un accomplissement personnel.

Ces relecteurs méritent aussi un gâteau au vu du harcélement à des heures indues dont j'ai pu faire preuve.

Aux personnes qui m'entourent et avec qui je partage stress et déboires tout autant qu'excitation et succès.

Aux optimistes, sans qui on aurait tout arrêté sans doute, mais grâce à qui l'on continue.

À toi, lecteur lectrice, qui est au début de l'histoire.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Food-Microbiome projects

We spend a fair amount of time eating, or thinking about it. Daily subjects are naturally studied and food –as a complex product– is investigated at several scales. Multiple factors drive food-related studies: from scientific curiosity to yield and quality improvements. Fermented foods –like cheese, sausage and beer– provides a level of diversity and complexity worth studying given their world-wide spread.

1.1.1 Why?

Industrial partners and academics are both interested by insights into cheese ecosystems using metagenomics. Their common interests were crystallised into two consecutive projects: (1) Food-Microbiome project and (2) Food-Microbiome Transfert project

Food-Microbiome project

The Food-Microbiome project aims to deeply understand cheese ecosystems given next-generation sequencing technologies potential (Renault 2009). It is coordinated by Pierre RENAULT (PhD,Research Director) and funded from 2009 to 2013 by the ANR –*The French National Research Agency*. It intends to provide a proof-of-concept and preliminary results concerning the use of metagenomics for cheese ecosystems exploration.

Food-Microbiome Transfert project

The second project is funded by industrial partners from 2015 to 2018. It is divided into two parts: (1) provide a convenient metagenomics analysis tool and (2) characterise dairy strains genomics and functional features. Metagenomics data analysis will be facilitated through an online interface where public and user-provided genomes could be used. A comprehensive database and a website are being developed to this end and are targeting industrial partners and scientific collaborators.

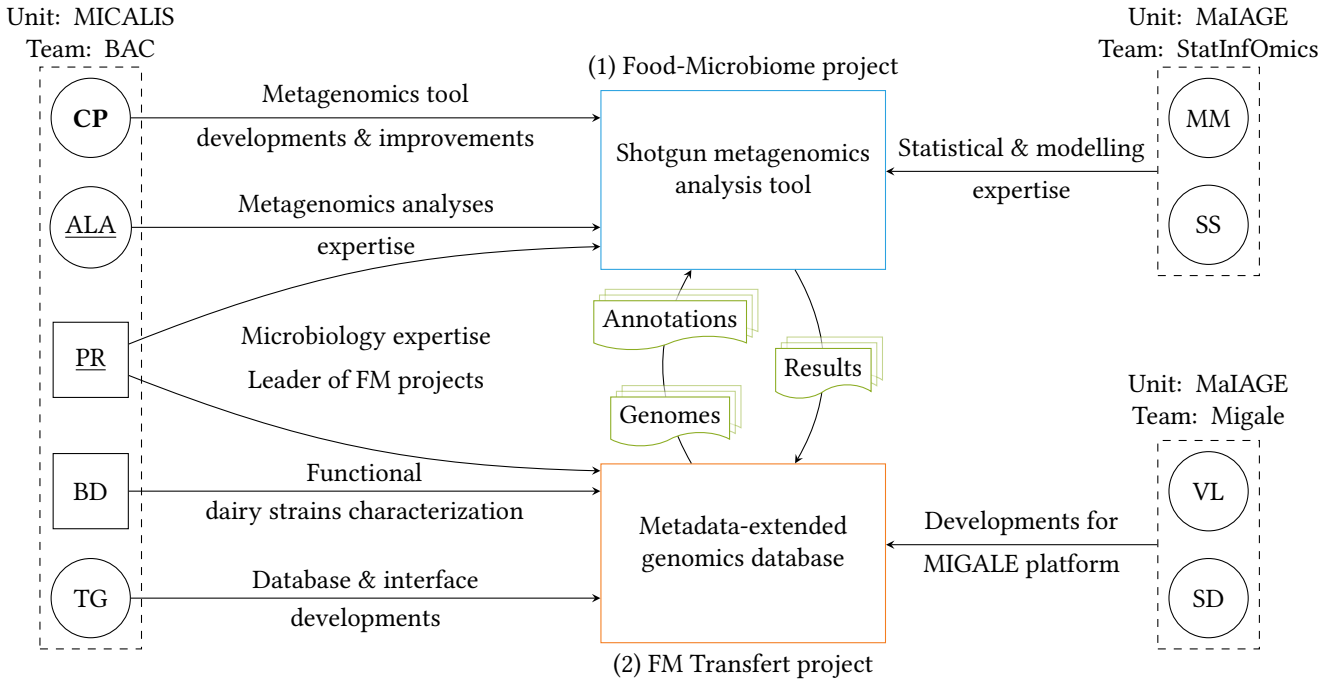


Figure 1.1: Actors and teams in Food-Microbiome projects. Round nodes depicts bioinformaticians or statisticians whereas square nodes depicts microbiologists. Underline names are my supervisors. Initials are detailed in the text (§1.1.2).

Wet-lab experiments are planned in order to unravel technological abilities of unknown strains and improve genomics annotations.

1.1.2 Who?

Food-related studies are often led by INRA –*The French National Institute for Agricultural Research*. INRA is an EPST –*Scientific and Technological Public Structure*– which yields scientific knowledge concerning food, agriculture and environment since 1946. It is organised into well-defined scientific divisions or *department* –e.g., MICA for *Microbiology and the Food Chain*– which gather several research *units* –e.g., MICALIS for *Food microbiology for health*. INRA is spread across the country in *centres* and scientists work locally in *teams* that are part of units.

Several actors and contributors to these projects

These Food-Microbiome projects gather dairy products experts –mainly microbiologists– from 3 INRA centres: Aurillac (F-15), Grignon (F-78) and Jouy-en-Josas (F-78). Industrial partners are brought together by the CNIEL –*The French Dairy Inter-branch Organisation*– and are involved in these two projects. Academics participants and their corresponding expertise and affiliations are presented together in Figure 1.1. Name abbreviations used in this figure are reported with INRA teams description in the following paragraphs.

Food and Commensal Bacteria Team BAC Team is led by Pierre RENAULT (PhD, Research Director, abbrev: PR) and belongs to MICALIS research unit. Part of the team is working on Food-Microbiome projects and are listed below. Anne-Laure ABRAHAM (PhD, Research Engineer, abbrev: ALA), Thibaut GUIRIMAND (Engineer, abbrev: TG), and myself Charlie PAUVERT (abbrev: CP) constitute the development bioinformatics team. Bedis DRIDI (Post-doc, abbrev: BD) is implied in wet lab experiments.

Applied Mathematics and Computer Science, from Genomes to the Environment MaIAGE research unit harbours several collaborators. In the StatInfOmics team –*Bioinformatics and Statistics for omics data*- Mahendra MARIADASSOU (PhD, Research Associate, abbrev: MM) and Sophie SCHBATH (PhD, Research Director, abbrev: SS) kindly provided their statistical expertise for modelling issues. Valentin LOUX (Engineer, abbrev: VL) and Sandra DEROZIER (Engineer, abbrev: SD) from the MIGALE platform provided database development expertise.

Personal involvement in both projects

Pierre RENAULT and Anne-Laure ABRAHAM were my supervisors during my two-year apprenticeship in this team. During the first project –Food-Microbiome project– I was mainly involved in development, improvements and tests. I was more taking part of know-how transfer and integration during the second one.

1.2 Cheese ecosystems particularities

1.2.1 Micro-organisms and definition of species

Micro-organisms unite bacteria, yeasts, moulds and viruses. These micro-organisms are usually classified using taxonomic ranks as illustrated in 1.2.

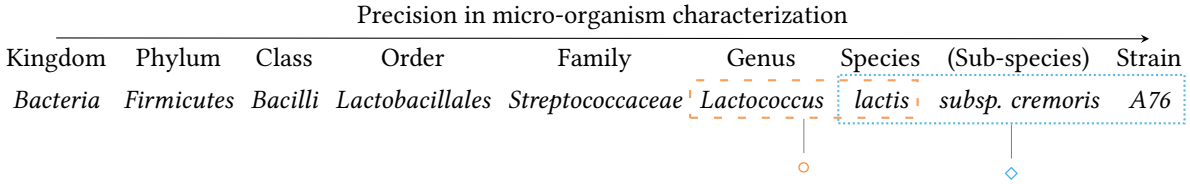


Figure 1.2: Illustration of taxonomic ranks. Current limits of taxonomic identification in several metagenomics sequencing strategies are described by symbols: diamonds (◇) depicts shotgun sequencing technology whereas circles (○) depicts amplicon sequencing strategies.

However whole-genome sequencing blurred these ranks –for instance, some taxonomic ranks were reevaluated after sequencing the corresponding organism– and these taxonomic ranks were thought

to be augmented by genomics information (Chun and Rainey 2014). DNA-DNA hybridisation and its *in silico* counterpart are proposed to delineate species limits (Richter and Rosselló-Móra 2009; Varghese et al. 2015). For example, it is commonly stated that strains from a common species have between 95 to 96% genome identity. For convenience I will use the standard classification when referring to micro-organisms.

1.2.2 Cheese ecosystems

From milk to cheeses

Cheese-making starts with a matrix composed of milk more or less transformed –raw milk or pasteurised. A cocktail of enzymes –rennet– triggers the milk state transition from liquid to solid, the resulting product being called curd (Button and Dutton 2012). The curd will then undergo many transformations and technological processes to be manufactured into a consumable cheese.

The great diversity of cheeses comes from variations in these highlighted processes. Experts proposed a classification in order to unify terms (Almena-Aliste and Mietton 2014). Cheese diversity also stem from the wide range of micro-organisms involved.

Micro-organisms sources shapes cheese ecosystems

Micro-organisms inoculations are among these transformations. Proteins, fatty acids and water –composing the milk– combined with tailored conditions provide a suitable media for microbial growth (Monnet et al. 2015). Micro-organisms may come from two main sources: starter cultures and environment (F. Irlinger et al. 2015). Starter cultures can be distinguished into two categories: manufactured and undefined. The first one consists of few selected strains –usually with interesting technological features– gathered in a mixture applied on the curd. The second can be much more complex and results from subcultures of either other starters or previous dairy products –cheese rind communities or yogourt for instance.

While a manufactured starter composition can be trivial and documented, unravel microbial composition in undefined starters is a challenge. Biotic and abiotic factors are hence involved in cheese-making and can greatly impact cheese quality and microbial composition (F. Irlinger et al. 2015; Monnet et al. 2015). These ecosystems are described as a *good* model: both simple and complex (Wolfe et al. 2014).

Why simple?

Cheese is usually referred as a less complex environment than soil or gut. They harbour a relative small number of distinct species compared to others ecosystems as well as a better rate of cultivable organisms (Bourdichon et al. 2012). Less than 200 species are estimated in cheese (*ibib.*) compared to

more than a thousand in aforementioned ecosystems (J. Qin et al. 2010). Cheese also benefits from many studies in literature[1], being an efficient milk storage technique for 7,000 years (Salque et al. 2013).

However recent methods –e.g., high-throughput technologies– shed light on a microbial diversity in what was thought to be well-described ecosystems (Quigley et al. 2012). These methods will be reviewed later in section 1.3.

Why complex ?

Cheese cannot be reduce to an association of milk and micro-organisms. The environment influence is significant –e.g., others sources of micro-organisms and their interactions– shape microbial composition of cheeses. So cheese ecosystem exploration is not so simple because this ecosystem is influenced by several factors. Moreover, little is known about low abundant micro-organisms in cheese ecosystems, despite rich media and supposed easy culture. Finally, cheese ecosystems are an important reservoir of microbial strains given their interactions with the environment. The diversity of strains in cheese ecosystems remains to be unraveled.

Reference genomes are not always available

Many dairy strains genomes recently enriched genomics database in the scope of Food-Microbiome project (Almeida et al. 2014). Despite these recent –and less recent– contributions, not all reference genomes are available. Some species do not have any genome representation in databases.

Apart from this limiting case, two others cases can be challenging. First, a representative genome exists in databases but the genome sampled in the ecosystem differ. For example, a genome of the same species is available but not the same strain –both sharing 90 to 99% identity. Second, several genomes –e.g., a catalog of strains– are available and one of them exists in the ecosystem. For example, recognise a specific strain in a cheese community after an artificial starter inoculation. These cases are illustrated in figure 1.3.

1.2.3 Why study cheese ecosystems?

Cheese ecosystems harbour an important diversity of micro-organisms (Montel et al. 2014). Diversity stem from the necessary functional redundancy in a community for the ecosystem to be reliable against environmental changes (Konopka 2009). For example, these microbial communities enable resistance to pathogens such as phages or others unwanted bacteria like *Salmonella* (Ortolani et al. 2010; Callon et al. 2014).

[1] A Pubmed search with “cheese” keyword yields 9,177 entries as of June, 2016.

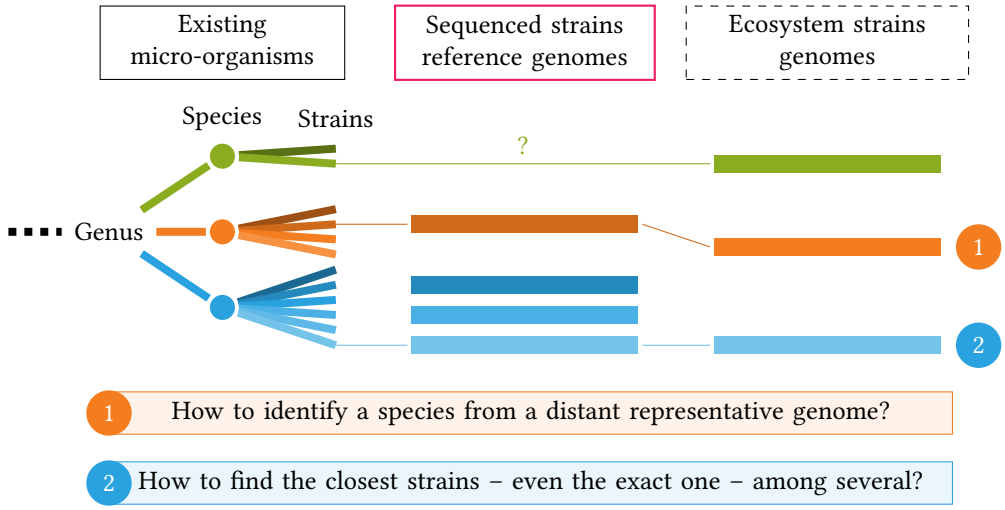


Figure 1.3: Discrepancies between reference genomes available and strains genomes from ecosystems. Several cases are illustrated: (1) when a representative reference genome exists and (2) when a catalog of sequenced strains is available.

Moreover, strains in cheese ecosystems can demonstrate specific metabolic pathways and organoleptic compounds synthesis. Their characterisation will provide leverage for food industry improvements and sustainable product quality. Hence, academics and industrial partners are exploring cheese ecosystems to unravel their relative complexity using a wide panel of tools and methods.

1.3 State-of-art tools for ecosystem exploration

1.3.1 Methods

From culture to metagenomics

The canonical approach to ecosystem exploration used to be culture of microbial-organisms. First, they were isolated from their ecosystem and then cultivated under proper conditions. Meeting these conditions and finding a convenient media could be a struggle –or worse quite unfeasible– depending on the strains and the ecosystem studied. Media is not an issue in cheese ecosystem studies, however low-abundant microbial-organisms or interactions-dependant micro-organisms would be out-of-scope.

Culture-independent approaches were then developed. The capture and analysis of ecosystem genomics information –or *metagenomics*– provides insights into genomes from micro-organisms bypassing cultivation related issues. High-throughput sequencing technologies fast development enables multiple methods and innovative strategies to blossom. These recent approaches provide a new perspective for ecosystem exploration and are outlined in figure 1.4.

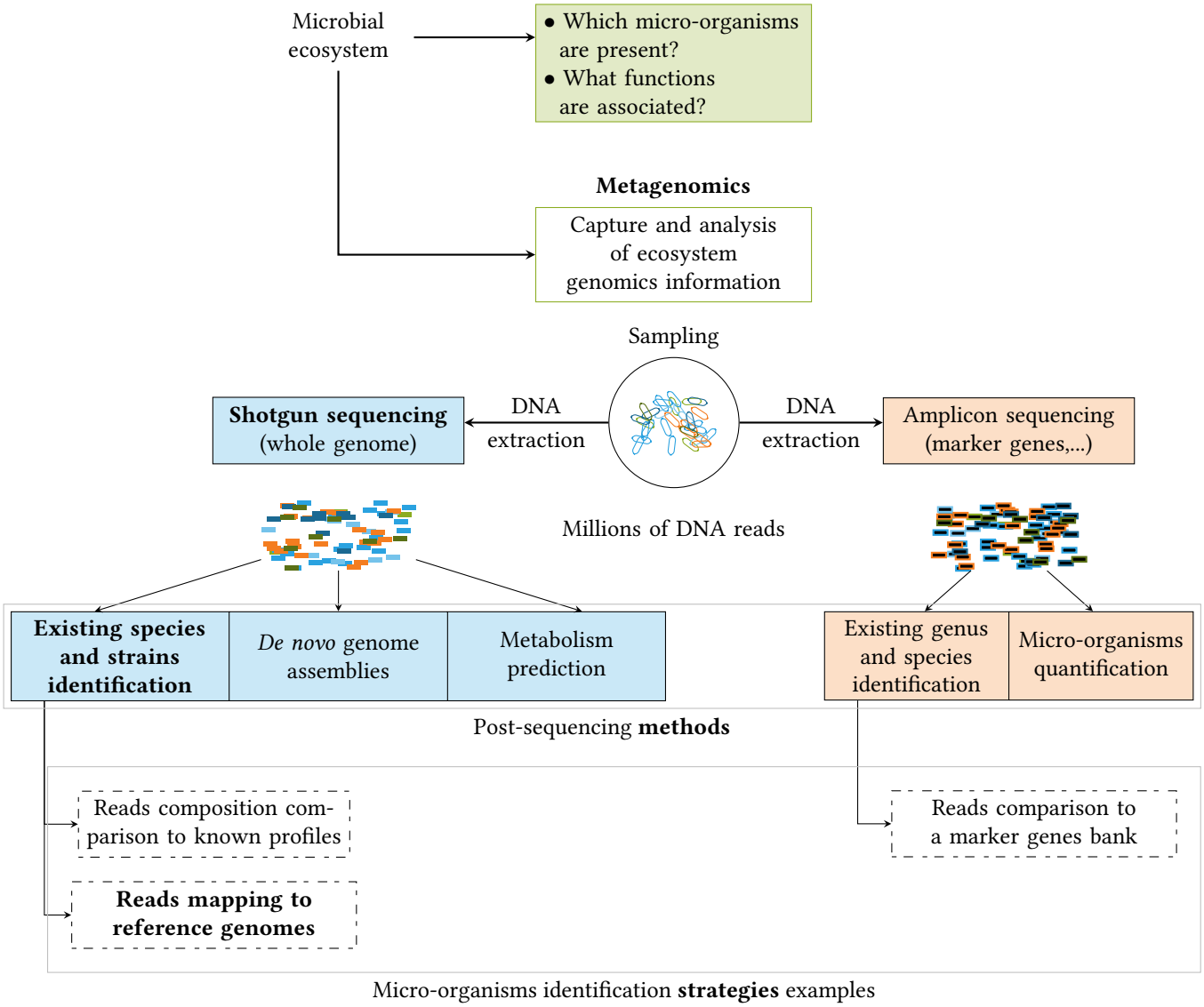


Figure 1.4: Overview of microbial communities exploration using metagenomics methods. Microbial community DNA is extracted after sampling a microbial ecosystem (circle). This biological material can undergo two types of sequencing: (1) amplicon sequencing (right pane in orange) and (2) shotgun sequencing (left pane in blue). In short the first can provide microbial identification down to species or genus level and relative quantification whereas the latter grant access to potential functions and a more precise taxonomic identification. Aforementioned aims are reached after post-sequencing methods and corresponding bioinformatics strategies. Bold items describe technologies and methods used in our projects.

Different questions and corresponding sequencing technologies

An ecosystem sample in a metagenomics study can be sequenced through two major technologies –amplicon or shotgun– depending on questions asked. Both are reviewed in the following paragraphs.

Amplicon sequencing Only one or few marker genes are sequenced in this approach. The preferred marker –or barcodes– for prokaryotes is the ribosomal small sub-unit DNA –termed 16S rDNA– and its counterpart in eukaryotes 18S rDNA. Another common barcode for eukaryotes especially fungi is the *internal transcribed spacer* –ITS. Their sequences harbour several conversed regions next to more variable zones. Conserved regions from multiples organisms will be used to generate PCR primers. These primers will then amplify variables regions that are specific to a species or genus. Generic markers can be used to assess overall diversity in exploratory studies (Ciccarelli et al. 2006). The choice of 16S regions to amplify –or marker genes– can also depends on prior knowledge concerning existing micro-organisms in the environment and can be tailored in consequence (Kumar et al. 2011).

Shotgun sequencing DNA extracted from samples is sheared into fragments. Fragment are randomly sequenced which yield millions of DNA reads (Sharpton 2014). These reads embody coding DNA sequences –CDS– which provide insights into potential functions harboured in the studied microbial community (Venter 2004).

Application to cheese ecosystems Both sequencing technologies were used in cheese ecosystem exploration. Delbes and al. relied on amplicon sequencing combined with culture experiments for validation to provide a microbial overview of raw milk cheeses (2007). Ercolini reviewed the use of shotgun sequencing specifically in food microbiology (see 2013). Wolfe and al. exploited both amplicon and shotgun metagenomics sequencing to cheese ecosystems (see 2014). Dugat-bony and al. decided to combine the latter with metatranscriptomics and biochemicals analyses (see 2015).

1.3.2 Post-sequencing strategies

High-throughput sequencing yields millions of DNA reads. This technology provides an ecosystem overview. However, the link between DNA read and organism is lost in the process of metagenomics sequencing –in contrast to the sequencing of one individual. Innovative strategies needs to be applied –or developed– in order to circumvent this information loss. Accordingly, these strategies will achieve ecosystem information retrieval using metagenomics –e.g., micro-organism identification. Some methods and strategies are summarised in figure 1.4 and detailed below.

How to deal with millions of DNA sequences? The information embodied in these reads enables several applications: (1) micro-organisms identification to distinct taxonomic levels or (2) micro-organisms quantification, (3) new genomes assembly and (4) metabolism prediction (Sharpton 2014).

Micro-organisms identification Several metagenomics reads analysis strategies are available to this end (see bottom frame 1.4). The first strategy is *de novo* genome assembly where reads are merged provided their similar ends in order to yield longer continuous fragments (Namiki et al. 2012). Accessing reads composition –in tetranucleotides or k-mers– is another strategy to bin reads. It relies on previous reads composition profiles associated with known organisms (Wood and Salzberg 2014; McHardy et al. 2007). Otherwise, reads can be aligned –or mapped– to public and user-made references (Ahn, Chai, and Pan 2015; Lindner and Renard 2015). Typically 16S rDNA reads will be compared to a bank –e.g., SILVA– whereas reads from shotgun sequencing will be aligned to several reference genomes or marker genes. This approach yields perfect matches as well as inexact matches.

Genome assembly is mentioned here due to potential identification of new species. However, specific conditions regarding input data for assembly makes this strategy quite distinct from others listed here.

Strain level identification is a challenge in metagenomics. Few approaches were available at the beginning of my apprenticeship. However several research teams showed a growing interest towards this challenge. Therefore many publications and tools are recently available but we did not have the time to review them thoroughly (Piro, Lindner, and Renard 2016; Truong et al. 2015; Luo et al. 2015; Cleary et al. 2015).

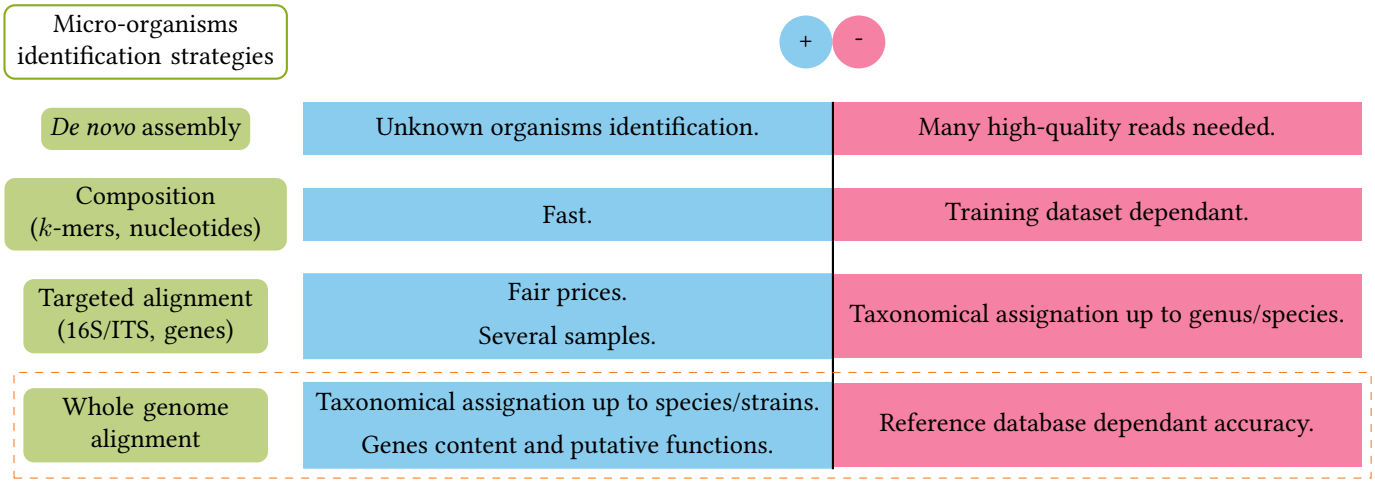


Figure 1.5: Comparison of post-sequencing strategies. Four types of strategies are summarised here and described in the text. Majors pros –respectively cons– are presented on the left side –respectively right– of the central edge. The dashed frame highlight the strategy used in our projects.

Taxonomic identification and ecosystem exploration Taxonomic identification level depends on specific interests concerning the studied ecosystem. Depending on ecosystem complexity, species level identification provides ecosystem global features, or generic micro-organisms characteristics. However only strain level identification –more accurate than previously– can shed light on environmental adaptation or distinguish specific metabolic pathways for instance. Network analysis can

complement ecosystem exploration up to genus or species level and thus provide an extended view of the ecosystem (Parente et al. 2016).

1.4 Aims

Microbial communities are (1) ubiquitous –e.g., soil, plants, fermented foods, etc.– and (2) necessary for host health or manufacturing processes. Their study highly benefited from high-throughput technologies such as metagenomics sequencing. However, metagenomics data analysis comes with methodological and computational challenges. Among them, strain level identification remains for a time ideal in metagenomics. Recently this challenge shifts from hypothetical to achievable thanks to the scientific and industrial community efforts. However, at the time unraveling cheese ecosystems was a challenge. In the scope of Food-Microbiome projects, we rely on existing reference genomes to (1) identify micro-organisms present in the ecosystem –if possible to the strain level– and (2) characterise low-abundant micro-organisms.

I have focused my research on a in-house metagenomics analysis tool –named GeDI– during my bioinformatics apprenticeship. This two-year work can be crystallised into four axes –tackled or soon-to-be.

The primary goal was to develop a method to automatically identify species and strains in the ecosystem. To this end, I relied on shotgun metagenomics reads alignments on a set of reference genomes. However, specific cases leads to the exploration of several *improvements* in terms of scientific methods. For instance, reference genomes are likely to differ from the strains actually present in the ecosystem, hence micro-organism identification has to be tailored. I have also worked on computational features improvements such as compatibility by using standard bioinformatics files. This conversion leads to space usage and speed improvements.

Recently, I was involved in GeDI *integration* with the metadata extended genomics database developed in the Food-Microbiome Transfert project and others tools. Finally, I have drafted results yields by GeDI *applications* to simulated and real datasets. For illustration purpose, I will present how one strain is represented in 9 cheese samples.

2 Materials and methods

2.1 Computing facilities

Personal computers The institute computer I am working on runs under Microsoft Windows 7 64bits. Its features –Intel Xeon 2.66Ghz and 8Go RAM– support a virtual machine set up. Through VirtualBox I can work under a suitable operating system for my needs: Linux Debian "wheezy" 7.8 (stable). From times to times, I use my personal laptop for writings during commutes or remote working.

Clusters farms Some greedy simulations or tasks requires supplementary computational resources. I have been granted access to two INRA clusters farms to meet these needs. The first one –`migale`– located in Jouy-en-Josas harbours 580 nodes and a large choice of bioinformatics tools. The second one –`genotoul`– in Toulouse benefits from a larger storage space and has more than 5,000 nodes.

2.2 Professional practice

2.2.1 Technology and literature monitoring

Literature monitoring I usually browse state-of-the art journals –e.g., *Bioinformatics*, *Plos Computational Biology*– outlines in order to quickly get a glance at new articles. Word of mouth and informal discussions are far from outdated and relevant to pin point interesting papers. This network is defined at several scales: (i) at office level with Anne-Laure et Pierre, (ii) with team members, (iii) through collaborations and others informals interactions within the institute –e.g., young researchers association, bioinformaticians network– and finally (iv) within master degrees colleagues.

We set up and promote a wiki –within the institute– dedicated to biologists and bioinformaticians from the MICALIS research unit. Articles and tips are shared through this platform with colleagues working on similar topics. A shared bibliography was set up especially with MetaGenoPolis.[1]

[1]Industrial and academics consortium tied with INRA working on human gut microbiota using metagenomics.

My bibliography is managed with Zotero (v4.0). This tool provide remote synchronisation between the several terminals I use –lab computer and personal computer. Articles of interests are hence gathered both at the lab or at the university.

To circumvent BibLaTeX-related issues –such as non-unique citations keys or specials characters– I have been using Better BibLaTeX Zotero extension which tackle previous issues and provide automatic bibliography exports easily.

Technology watch Many resources exists to keep up with bioinformatics and biostatistics related activities. Mailing lists from SFBI –*French Bioinformatics Society*– or AFEM –*French Microbial Ecology Association*– or more specific to the institute are examples of resources. I usually browse SFBI "bioinformations", bioinfo-fr.net website, or R-bloggers for general information and stackoverflow.com forum –and its sub-components– to answer technical issues.

Meetings, work groups and seminars Science without regular communication with peers would not be. It is one way to learn new methods and share feedbacks. Adapting your speech to a public with different scientifc background is a good opportunity to synthetise your research topic. To this end, I had attended monthly bioinformatics meetings organized in the institute.

I was also involved in an INRA working group belonging to the PEPI network –*Experience and good practices sharing in Computer Science*. Bioinformaticians and statisticians gathered in this group are interested by amplicon metagenomics analysis –especially 16S rDNA. I could compare these approaches and biases outlined with my project. I participated to the beta test of the training course organized by this working group. These meetings provide a new perspective to my project and in the long term to metagenomics analysis.

Conferences are a common way of keeping yourself up to date in a field as broad as bioinformatics. I attended three national and international conferences such as JOBIM –*Open Days in Biology, Computer Science and Mathematics*– or ECCB –*European Conference on Computational Biology* throughout my apprenticeship. I also attended a yearly workshop RCAM –*Recent Computational Advances in Metagenomics*– where I could discuss with experts of the field.

2.2.2 Good practices and tracability

The `vim` editor is my favorite tool for every writing –scripts, summaries, reports–. I enjoy its advanced functions, ubiquity and flexibility.

I use `git` the version management system almost every day for code related projects or not. It enables proper work organisation –even when you are the only contributor. Most of my repositories are stored on the migale server. But I created a private GitHub repository to manage versions of this thesis. It also provides online storage and availability from anywhere.

Besides I use the R software mostly through the IDE Rstudio. It provides documentation, history associated with an editor and a R console (v3.1). Work and files can be conveniently splitted in projects. Moreover, Rstudio is associated with `knitR` to do litterate programming. I daily use this technology to maintain reproductibility and tracability in my own analyses. This R package dynamically generate high-quality reports in LaTeX. I rely on this package for summaries of on going work as well as personal documentation or more elaborate reports.

This modular and adaptation code philosophy with knitR is completed by illustrations with TikZ. Figures are generated in several possible filetypes and can easily fit in many communication support –thesis, presentation, poster.

2.2.3 Research communication

Previous tools enable easy communication with peers. I talk with my supervisors on a daily basis. I communicate in team meetings and annually in conference.

I share the office with my supervisor, consequently she can easily monitor my work. Regularly we organise planned meetings before and after going to Rouen. These periods are concluded by summaries to my tutor.

Last year, I presented twice ongoing work to the team, hence benefiting from advices in biology or scientific methods. In February 2016, I presented my work on the mixture model to the monthly bioinformatics meeting of the institute despite negative results.

Conferences broaden the scope of communication and diffusion. I had the chance to present my work to every conference I attended (ECCB-JOBIM 2014, JOBIM 2015 and 2016). This provides an easy and comprehensive way to talk with scientists and welcome ideas and suggestions. I was awarded the young scientist best poster award last summer.

2.3 Tools

Short reads alignment on reference genome Bowtie (v. 0.12) (Langmead et al. 2009) is a reads aligner software meant to assess efficiently this issue (Schbath et al. 2012). Index creation is done before the alignment and we use the default parameters of `bowtie-build`. We align here short reads of 35 bases with up to 3 mismatches included. We used the following parameters: `-a -best -strata -M 1` which means that an alignment is outputed for each reads aligned once, if more, one alignment is randomly choose from the those with less mismatches. We use this version of `bowtie` because cheese samples were sequenced with SOLiD technology and `bowtie` does not support this technology anymore.

Synthetic metagenomics samples creation Grinder (v. 0.5.3) (Angly et al. 2012) sample one or several FASTA files to create a FASTQ file with artificial reads. It was used to simulate metagenomics shotgun sequencing from a microbial communinity sample, in our case it yields an important number of short DNA reads of 35 bases. Grinder was used because community diversity can be controlled: either manually with an abundance textfile (with `-abundance_file`), or either from a statistical distribution (with `-abundance_model exponential 0.5`) with a fixed parameter but ranks are randomly sampled. Variable genome length are taken into account and sequencing error models are available for an accurate simulatation.

In between genomes identity percentage computation La proximité des génomes utilisés dans le jeu de données *Streptococcus* a été évaluée avec Genome closeness used in the *Streptococcus* dataset was assessed by Gegenees (v2.2) (Ågren et al. 2012). A sliding window of 200 bases every 100 bases yields fragments for every genomes. All-versus-all BLAST strategy is applied to these fragments and Average Nucleotide Identity is returned.

Bioinformatics toolbox `SAMtools` (1.3.1) were used in the last version of our tool. I relied on `BEDtools` (2.17.0) to provide fast and efficient alignment file and gene annotation intersection. These toolbox were either called from our tool with a Python library `pysam` (v0.14.1) or through the Python `os` module.

Data mining and exploratory analyses Data mining was mostly supported by the GNU toolbox –e.g. `awk, xargs, sort, sed`–, the R software (v3.1) (R Core Team 2014) and its IDE Rstudio. Plots are generated with `knitR` and the `ggplot2` package (v.1.0.0) (Wickham 2009).

Distribution fitting I relied on the R package `fitdistrplus` (v1.0-4)(Delignette-Muller and Dutang 2015) to fit several distributions to characterise densities presented in figure 3.2. Maximum likelihood method is used to choose between multiple fit.

A metagenomics analysis tool: GeDI

Our tool is developed in Python (v.2.7.5). Its is a wrapper for the short read mapper `bowtie 1`. The principle of the former version is described in 2.1. Besides index construction and reads alignment that are part of the mapper, several specific task are implemented in GeDI: for example CDS filtering and few computations and verifications. However as of the former version, it was only using tab separated files and in-house files. Recently, an important release has seen the conversion to standard bioinfomatics filetypes: GFF for annotations, BAM for alignments, BED for regions to be excluded etc. This new version is depicted in figure 2.2. Output files are now GFF files for each genomes with

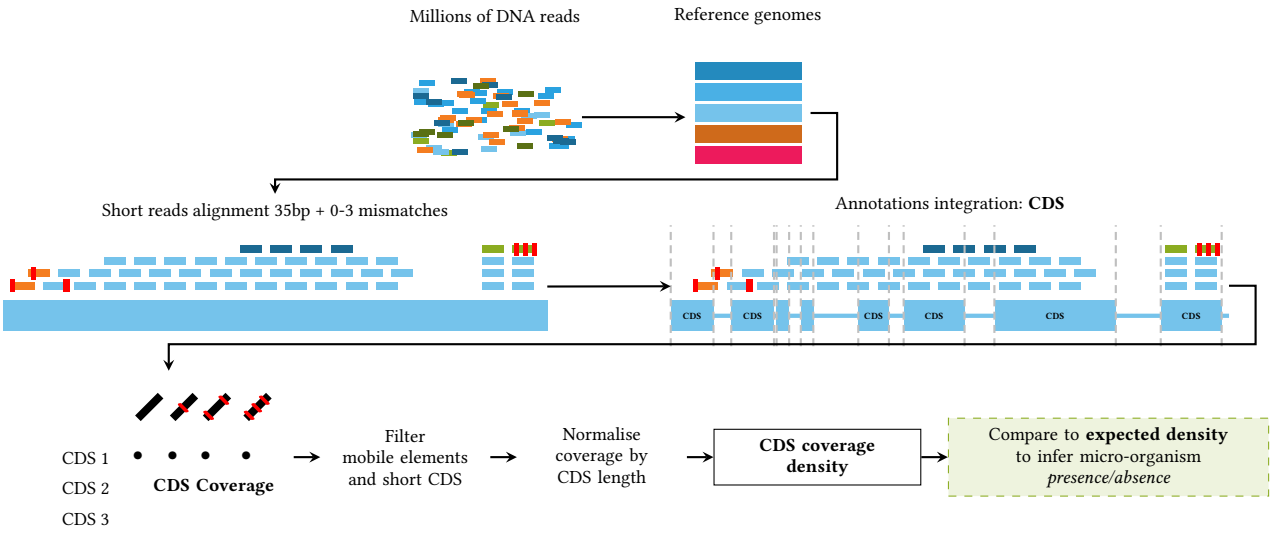


Figure 2.1: Previous data flow: GeDI

additional flags stating whether the CDS was filtered or not; its coverage total and with mismatches (see figure 2.3 for details}.

2.3.1 Mixture model of distributions

Reads alignment on a reference genome is not trivial in metagenomics. Aligned reads can stem from one or *multiples* organisms. These contributors genomes are the ecosystem genomes. They would be sequenced and they harbour genomics regions –of at least 35 bases long– similar to the reference genome. These ecosystem genomes can be: (1) exactly the same as the reference, or (2) a close strain or close species genome, or (4) a genome that share only these short aligned regions –e.g., mobile elements.

Contributors genomes are gathered in *classes* –or clusters in literature– based on closeness to the reference. Each closeness class is represented by a distinct distribution. The observed distribution is modeled as a mixture of these distributions. The posterior cluster probability of one CDS will be a proxy of an organism contribution to the alignment. We describe here the construction of a mixture model.

Reference genomes are independently considered. Let X_i a random variable describing the aligned reads number per base –or coverage– of a genome CDS i. X_i is defined on $[0; +\infty[$. We suppose $x_1, \ldots, x_n$ values of n random variables noted $X_1, \ldots, X_n$, where n is the genome total number of CDS. Contributor genome closeness class with the reference is noted c and $c \in \mathcal{C} = \{0, 1, 2, 3\}$. Let $Z_{ic} = (Z_{i0}, \ldots, Z_{i3})$ a latent variable stating the contributor genome closeness class whose reads were aligned on CDS i. Z_{ic} equals 1 if reads aligned on CDS i stem from –a genome– closeness class c, and 0 if not. With Z_i we model the heterogeous origin of reads aligned on CDS i. The observed coverage density is modeled with a mixture model of distributions: each closeness class is a mixture

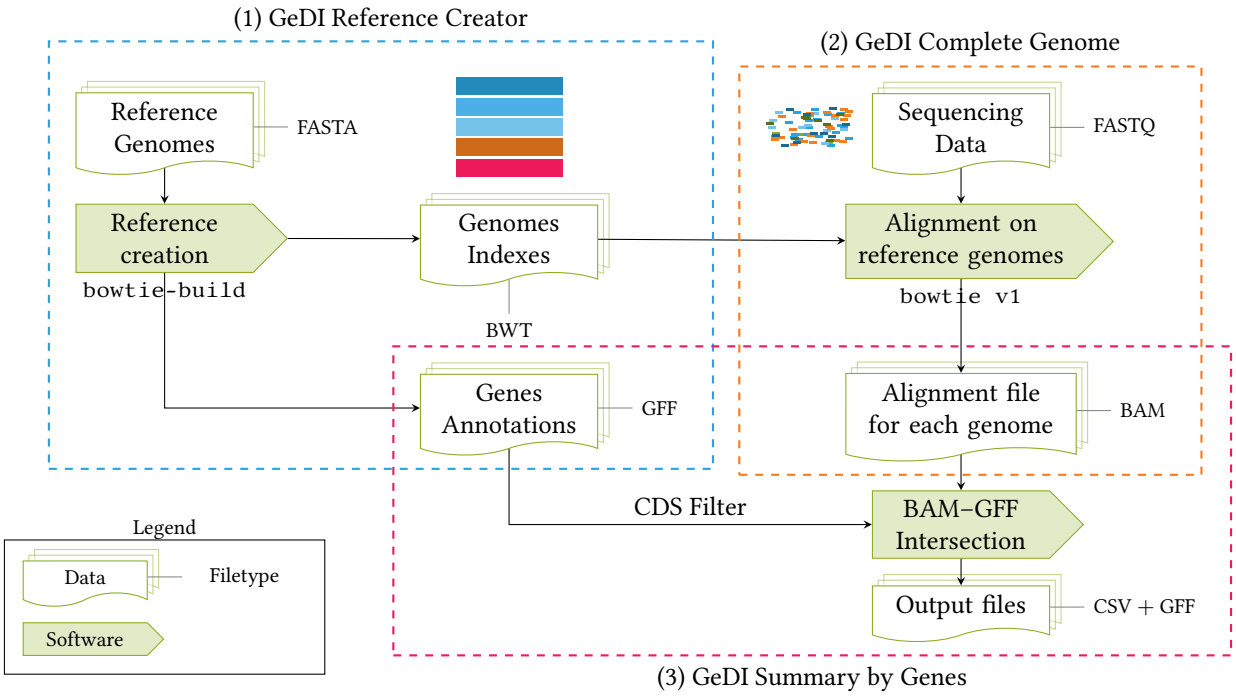


Figure 2.2: Overview of our metagenomics analysis tool: GeDI. It consists of three Python (v2.7.5) modules interconnected. Each module is framed here with dashed lines and encompass data and software.

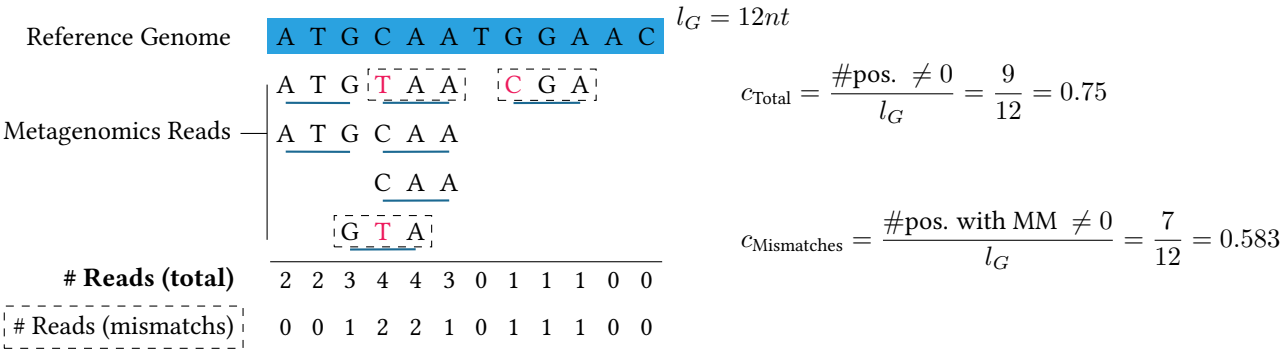


Figure 2.3: Genome coverage computation after reads alignement. In our metagenomics analysis tool –GeDI– we discriminate (1) coverage (c_{Total}) yield by the number of reference genome positions where at least one read is mapped divided by genome length (l_G) and (2) coverage ($c_{\text{Mismatchs}}$) concerning the same metric only for reads aligned with one to three mismatches compared to the reference.

component. We then estimate mixture model parameters.

Contribution The contribution of a given distribution to the model matches the ratio of class c in the mixture. This contribution is noted $\pi_c = \mathbb{P}(Z_{ic} = 1)$.[2] CDS are considered independent so the contribution of class c for the CDS i is equal for each $i \in [1; n]$. Then $\sum_{c \in \mathcal{C}} \pi_c = 1$.

Distribution The distribution probability of class c is the distribution of X_i knowing that reads stem from –a genome belonging to– class c. It is noted $f_c\,(x) = \mathbb{P}(X_i = x | Z_{ic} = 1)$.

Model Hence the mixture model is stated as follow:

$$f(x, \Theta) = \sum_{c \,\in\, \mathcal{C}} \pi_c \; f_c\,(x) \begin{cases} x & \text{Observed CDS coverage} \\ \Theta & \text{Model parameters} \\ \mathcal{C} & \text{Closeness classes} \\ \pi_c & \text{Distribution contribution to the model} \\ f_c\,(x) & \text{Identified zero-inflated distribution of class } c \end{cases}$$

In our case, we choose the following parametric forms for every distribution f_c: 3 Gaussian and one log-normal. However, these distributions are zero-inflated. In other words, chance to draw a null value from this distribution will be augmented. This mathematical adaptation enables the inclusion of CDS without any aligned reads in the model. To this end, we introduce ρ_c the ratio of CDS uncovered by reads from class c. Hence we have:

- For class $c \in \{0, 1, 2\}$:

$$f_c\,(x) = \underbrace{\rho_c \mathbf{1}_{\{x=0\}}}_{\text{Uncovered CDS ratio}} + \overbrace{(1-\rho_c)\mathbf{1}_{\{x\neq 0\}} \; \underbrace{\Phi(x; \mu_c, \sigma_c^2)}_{\text{Gaussian density}}}^{\text{Covered CDS ratio}}$$

- For class $c = 3$:

$$f_c\,(x) = \rho_c \mathbf{1}_{\{x=0\}} + (1-\rho_c)\mathbf{1}_{\{x\neq 0\}} \; \underbrace{\frac{1}{x}\Phi(\ln(x); \mu_c, \sigma_c^2)}_{\text{log-normal density}}$$

Parameters estimation We have to estimate model parameters Θ, which encompass: (1) contributions π_c, (2) ratio ρ_c for each class c, and (3) f_c distribution parameters θ_c for each class c.

These are estimated using maximum likelihood method computed on every CDS of the reference genome. To this end, we use the Expectation-Maximization algorithm (Dempster, Laird, and Rubin

[2] These quantities can also be found in the literature under the term "mixing weights" and noted α_c.

1977).

Expectation-Maximization algorithm This algorithm consists of iteratively increasing the model log-likelihood. In others words, the E step gives the conditional probabilities that data –CDS coverage– was drawn from every class parametric distribution f_c given model parameters Θ. The next step consists in the estimation of optimal values of model parameters Θ based on previous conditional probabilities. To this end, we use explicit parameters formulas for Gaussians in order to estimate with respect to the maximum likelihood criterion. This iterative algorithm stops either when Θ values converged or until an upper bound of steps. With the help of Mahendra MARIADASSOU, I implemented a version of this mixture model in R.

Output files I decided to rely on RDS files –R binary data filetype– to store all models estimated for the sake of reusability and interoperability. Reports are automatically generated with a table for every BIC values. The most likely model is selected with respect to the BIC criterion and its start and final parameters are outlined. The observed CDS density is plotted and on top the subsequent distributions are drawn from the selected model. Total and non-zeros densities are plotted for exploratory purposes.

2.4 Data

Micro-organisms genomes used I worked with up-to-date Genbank files thanks to the servers `migale` and `genotoul`. Our tool GeDI is based on these flat files to generate its reference genome catalog. However, somes genomes come from local database in the institute where NCBI genomes meets internal contributions. This database contains the recent genomes of the Collective genomes project (Almeida et al. 2014).

Annotation files Standard files for gene annotation –GFF3– were fetched from either from public databases like NCBI or from within the institute databases. In the absence of such file type, I sometimes used a BioPerl script to convert GenBank files to a pair of FASTA file and GFF file.

***Streptococcus* training datasets** I simulated a shotgun sequencing for each *Streptococcus* strains listed in the following table. These reads are then aligned separetely on *one* reference genome: the strain *Streptococcus salivarius* JIM8777 using the former version of GeDI–see 2.2

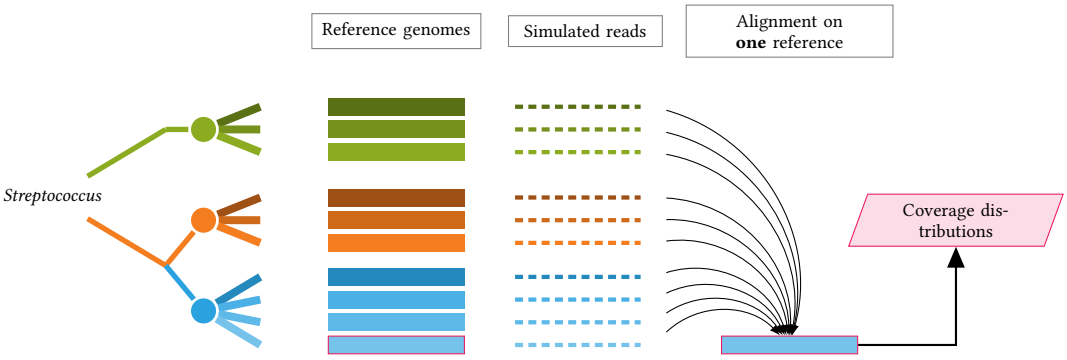


Figure 2.4: Training dataset overview.

Table 2.1: *Streptococcus* dataset composition and distance to the reference strain (*Streptococcus salivarius* JIM8777). Closeness classes are based on ANI *Average Nucleotide Identity* computed with Gegenees (Ågren et al. 2012).

Strains	**ANI** (%)	**Distance**	**Closeness**
Streptococcus salivarius JIM8777	100.0	Exact strain	0
Streptococcus salivarius NCTC 8618	89.63	Close strain	1
Streptococcus salivarius CCHSS3	88.57	Close strain	1
Streptococcus salivarius K12	86.01	Close strain	1
Streptococcus vestibularis ATCC 49124	80.01	Sub-species	2
Streptococcus thermophilus JIM 8232	74.68	Sub-species	2
Streptococcus salivarius PS4	73.74	Sub-species	2
Streptococcus agalactiae NEM316	67.14	Close species	3
Streptococcus infantarius ATCC BAA-102	62.82	Close species	3
Streptococcus mutans UA159	59.86	Distant species	3

Similarly a *Lactococcus* dataset was build –termed *composition dataset*– in which specific strains were iteratively mixed to create datasets of growing complexity.

Mock Community The Mock Community is a micro-organism cocktails of 22 strains gathered along a staggered abundance. Shotgun metagenomics sequencing was performed on this community and yields 8 millions of reads –available at this address hmpdacc.org/HMMC/ or on the NCBI website under the accession number: `PRJNA48475`.

3 Results

"Which micro-organisms exist in the studied ecosystem?"

We are trying to answer this question through the proxy of reference genomes and metagenomics reads alignments. Hundreds of genomes extracted from dairy products are currently available in genomics databases (Almeida et al. 2014). In the scope of Food-Microbiome projects, we rely on existing reference genomes to (1) identify micro-organisms present in the ecosystem –if possible to the *strain* level– and (2) characterise low-abundant organisms.

In this chapter, I will focus on my two-year work around a metagenomics analysis tool –`GeDI`– through four main axes. I will present (1) several approaches explored to improve the tool, followed by (2) a comparison with similar tools and (3) how this module was integrated with others, to conclude with (4) results from its application on simulated and real datasets.

3.1 Scientific and computational improvements of GeDI

Two kind of improvements are highlighted concerning this tool. First, scientific improvements through an exploration of micro-organisms taxonomical classification. Second, computational improvements that covers software performance as well as proper documentation.

We want an automatic criterion to infer a micro-organism presence or absence from its alignment data. To this end, we have focused on several modeling approaches –summarised in Figure 3.1– throughout my work in BAC team.

3.1.1 Modeling approaches

Our method is based on metagenomics reads alignment onto a set of reference genomes. It is followed by an goodness of fit analysis between observed and predicted genome coverage. The predicted coverage stem from a model assuming that the considered micro-organism genome exists in the ecosystem.

However, ecosystem genomes usually differ from corresponding genomes in databases. Hence this constraint harden reads alignment analysis in metagenomics, and subsequent modeling efforts. We

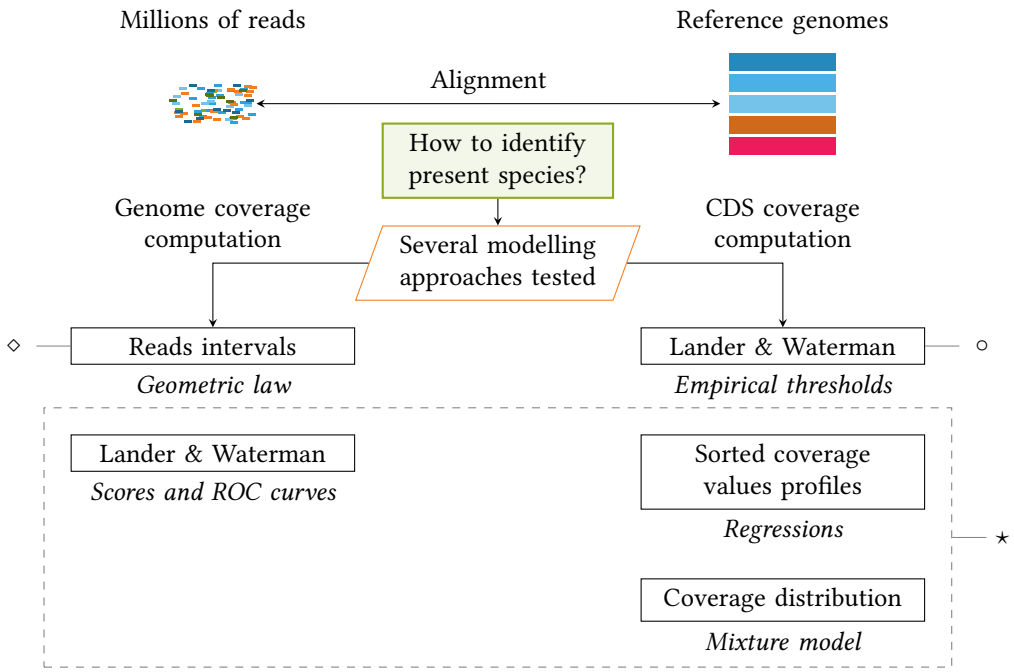


Figure 3.1: Previous modeling approaches summary embedded in GeDI. These different approaches were tested at different period depicted as symbols: before my internship (○), during my first year of master degree internship (◇) and during my two-year apprenticeship (⋆).

need to be able (1) to identify a strain if its reference genome is available and (2) find the closest species when it is not the case. We struggled to define a model that provide both accuracy and flexibility. I have worked to explore, test and precise several modeling approaches. Approaches principles are briefly reviewed below.

Previous modeling approaches

Homogeneity Lander and Waterman stated that reads distribution is homogeneous on a genome as long as: (1) sequencing is a random process and (2) reads stem from the considered genome (see 1988). This work is relevant in metagenomics for a sufficient reads number. In this case, testing the reads distribution homogeneity is a proxy for testing presence or absence of a genome. This hypothesis was tested by comparing observed and expected genome coverage –under the homogeneity hypothesis– before and during my internship. However, thresholds used to compare distributions were empirical and we could not explicitly validated this approach. Nevertheless, it yields interesting results as long as repeated regions were not too numerous and the reference genome was close to strains in the dataset.

Reads position intervals I then explored several leads in order to improve automatic taxonomic identification. Simulated datasets enable reads distributions exploration and modeling. Intervals between every two successive reads aligned is supposed to follow a geometric law. This approach

was meant to target genome with a low coverage, that corresponds to low abundant micro-organisms. However, this model was too strict and could only be applied if reads were aligned to the same genome reads stem from. For example, reads from a strain A aligned to the reference genome A would yield a reads interval distribution suitable to pass the test. However, it was not the case if these same reads were aligned to a close strain of genome A –say 98% identity.

Scores and ROC curves In order to circumvent previously identified limits, I have also worked on threshold determination. I relied on simulated datasets to infer decision rules regarding a micro-organism presence or absence from alignment data. For instance reads from one strain were aligned to more or less distant genomes. Simple datasets –like the latter– yield interesting results from this approach. However, it did not scales to metagenomics data where reads stem from multiples contributors.

Previous approaches were based on genome coverage. While bypassing potential annotation bias, these methods were impacted by both inter-species and intra-species genomics variation noises –e.g., pseudo-genes, mobile elements, etc. In 2015, I started going back to a previous approach in the team: focused on CDS –*Coding DNA Sequences*.

Exploratory approach to micro-organism identification

A CDS-centred approach has several advantages. First, these coding sequences are much more conserved. Then, CDS can be strain-specific and yield organism functions –inferred for instance by protein sequences prediction. Moreover, reference genomes sheared in multiple contigs could also be used. Minimal annotation of hypothetical CDS is automatically done nowadays. Hence even draft genomes could provide information on the studied metagenome.

CDS coverage densities We tested this approach with simulated datasets. We define a strain as a *reference* here: *Streptococcus salivarius* JIM8777. Shotgun metagenomics reads are simulated from this reference genome and nine other genomes more or less distant to the reference. These additional genomes can be broken down into 3 closeness classes: close strains, sub-species and close species –described in figure 2.4. The ten simulated dataset –or ten FASTQ files– are then *independently* processed by our tool GeDI –see former version 2.1. Reads aligned in CDS are used to compute CDS coverage. These coverage are then normalised by CDS length and presented in the figure 3.2.

In the left pane of the figure 3.2, the first density –in blue– represents how CDS coverage values are distributed when the same genome –the exact strain– is used both for simulating reads and aligning them on this reference genome. The latter CDS coverage density follows a Gaussian distribution with a mean μ and variance σ^2. Therefore, a majority of CDS are covered around the mean –here around 0.04 reads per base in a CDS. On both sides of this bell curve stands CDS that are either much

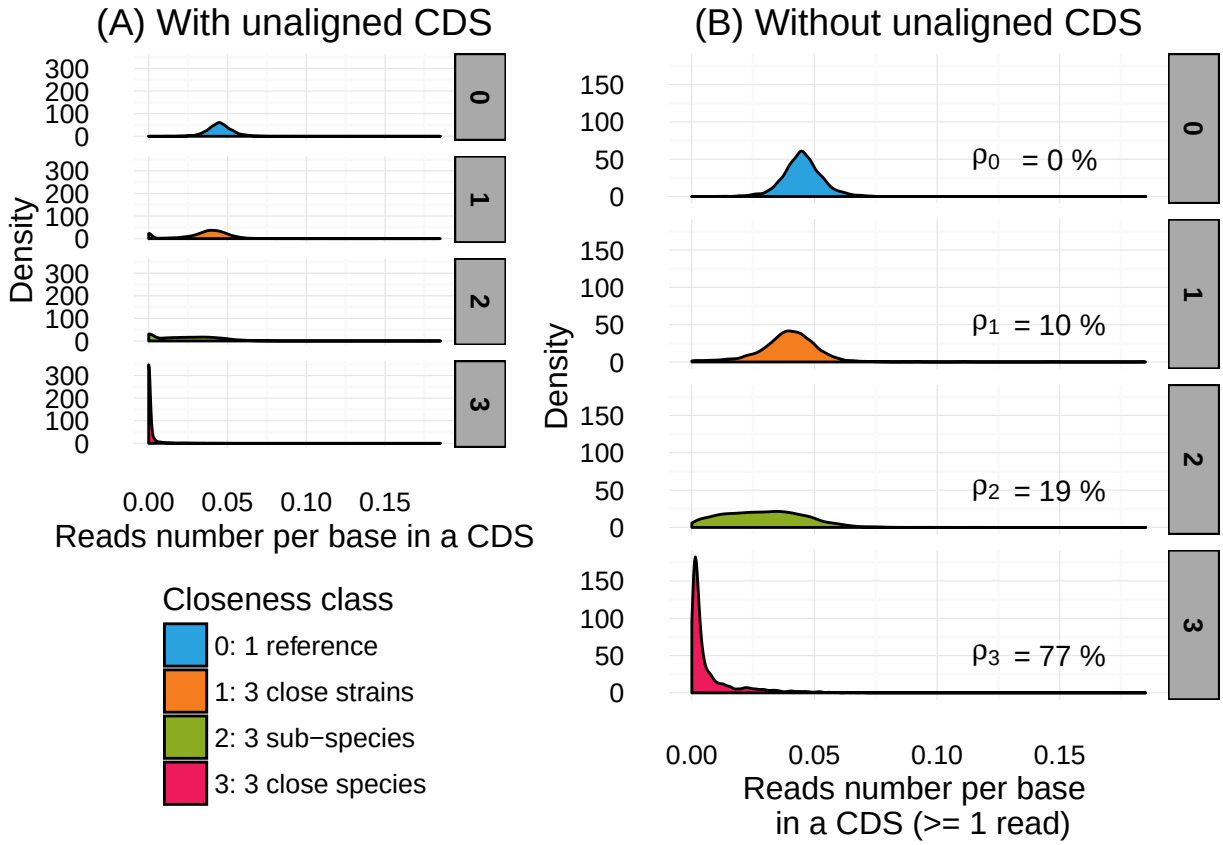


Figure 3.2: Training dataset CDS coverage densities depending on closeness classes. Reads counts are normalised by each CDS length of the reference *Streptococcus salivarius* JIM8777. (A) Densities of these values are presented on the left pane. (B) CDS without any reads aligned are excluded from densities estimation on the right pane. Densities are showed by closeness class. Hence, the first one only includes one genome –the reference strain–, others classes encompass three genomes each. Unaligned CDS ratios are showed by closeness class despite being excluded from densities. This ratio is noted ρ_c for a class c where $c \in [0; 3]$ and is a parameter in the mixture model.

more covered than the mean coverage or very less covered. Their low density indicate that there are few CDS well covered and few CDS highly covered. No CDS are left uncovered when enough reads are aligned on the exact reference. Hence, there is no difference between the blue curves on pane (A) and (B) in figure 3.2.

The second curve represents CDS coverage yielded by previous dataset alignment on a close strain. This curve is slightly shifted to the left, indicating a lesser mean coverage. But most notably, some CDS of the reference genome are not covered at all by reads from a close strain –it is indicated by the bump on the left side of the plot. Interestingly, once null coverages are removed –see the right pane of the figure 3.2– the latter density looks like the exact strain coverage density. In general, the less close the genome, the more CDS uncovered, shifting densities to the left.

The last density –closeness class 3– possess a representative density, breaking with previous Gaussian-like densities. Only a few CDS are covered, hence an important quantity of null values shape this density. The few CDS covered match housekeeping genes covered by a few reads.

CDS coverage densities crystallise closeness information CDS coverage densities seems to be a good proxy for the closeness between a reference genome and reads aligned. They appear (1) to be sensitive enough so that close strains reads aligned to the reference yield both a Gaussian density –see closeness class 0 and 1– and (2) to be specific enough to express distinct signal between closeness class through the ρ_c parameter. The unique shape of the more distant class –closeness class 3– density ensure to model noisy contributions to the alignment such as reads from any genomes. This non specific noise signal will be used to feature reference genomes absent from the ecosystem.

A mixture model to unravel contributors genomes Among available reference genomes, we want to be able to distinguish (1) exact strains that yields the metagenomics reads –if available, (2) close strains or species, and (3) absent genomes.

Metagenomics sequencing is a proxy to infer the presence of micro-organisms. Unfortunately the link between DNA reads and corresponding micro-organism is lost contrary to a single genome sequencing. Reads aligned on a reference genome may originate from multiple contributors. I had worked on a mixture model of distributions in order to bypass this information loss. An overall idea of the approach developed is summarised in figure 3.3. From the CDS coverage density yielded after reads alignment, the mixture model of distribution is meant to split this observed signal into contributions of defined distributions. In our case the distributions used are based on the four densities in figure 3.2.

This approach comes down to multiple parameters estimation. Providing proper parameters estimation, we plan to infer the presence or absence of a micro-organism based on estimated contributions. Especially the contribution π_0 of the distribution modeling the closeness 0 is likely to illustrate whether the reference genome is close to the strain in the ecosystem.

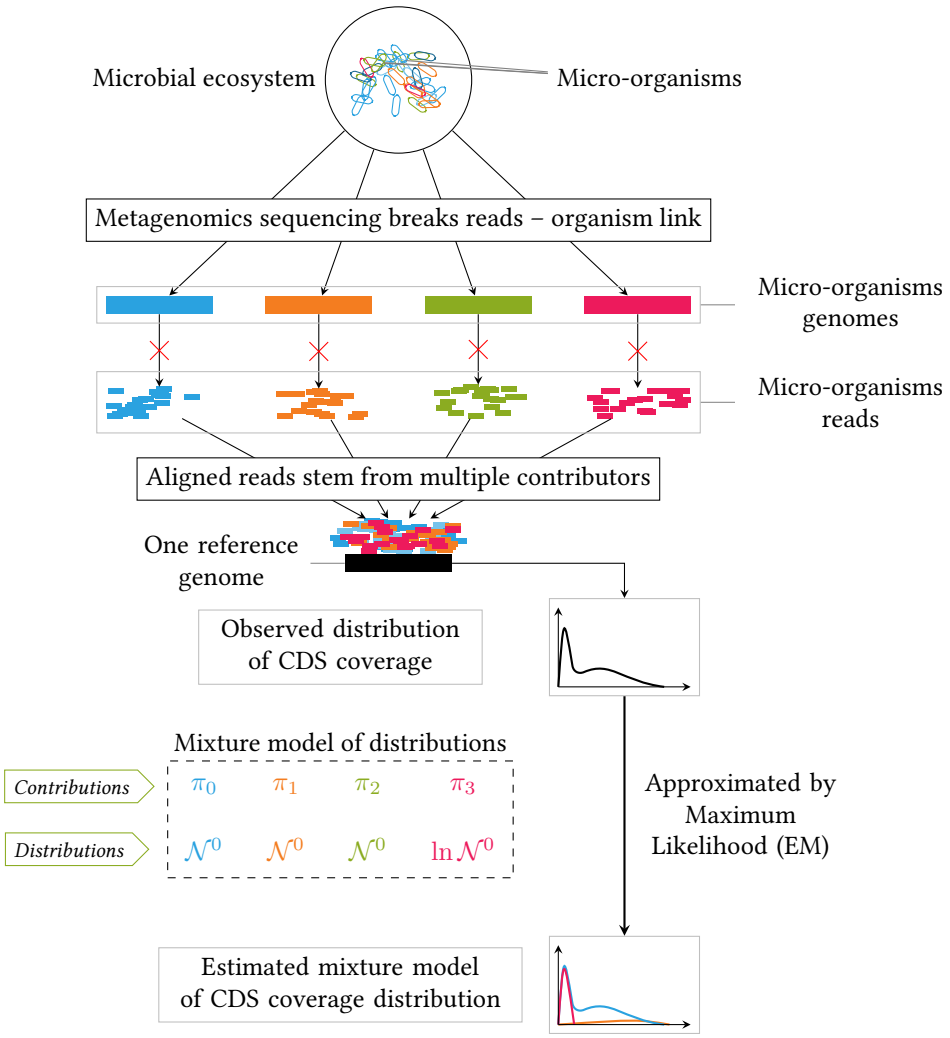


Figure 3.3: Contributors genomes issues and mixture model principles.

3.1.2 Modeling exploration results

For parameter estimation, I rely on the canonical maximum likelihood method with the Expectation-Maximization algorithm (Dempster, Laird, and Rubin 1977). I then applied this mixture model to simulated datasets and at the same time worked to counterbalance some model and method limitations. Results of these joint explorations are described below.

Mixture model limits and solutions

Number of classes Four classes of distribution were relevant given the training dataset I designed. However, there might be some cases where this number is either (1) too high, hence many superficial parameters needs to be estimated or (2) too low, thus the observed distribution will not be described properly. To this end, I down-sized the model considering that four classes of closeness was the upper bound. I hence removed iteratively intermediate Gaussian distributions building a simpler model of two and three distributions. The Gaussian distribution is meant to capture the information embodied in reads aligned from a close –at best exact– strain. The log-normal distribution intend to fit noise or distant genomes reads alignment. I decided to use the BIC –*Bayesian Information Criteria*– to choose between these models. It provides strong penalties towards parameters numbers thus enabling parsimonious model selection. The relevance of a single log-normal distribution test was examined, especially for one specific case with simulated data. However, the presence of only one organism is highly unrealistic in metagenomics, hence a minimum of 2 distributions was settled.

Aligned reads number and start values The EM algorithm –used for parameters estimation– provides local optimum and hence is known to be sensitive to start values. In addition, the mixture model was first designed for a fixed number of reads. I decided to treat the algorithmic and methodologic limit –respectively the first and second mentioned above– at the same time. Therefore, I explored the influence of aligned reads number on the mixture model distributions. The idea was to show the expected resolution loss between the closeness classes when the aligned reads number decreased. In the meantime, it provides first grasps on model behaviour towards low-abundant micro-organisms –which are obviously yielding few reads.

Reads number is an input parameter when crafting simulated datasets. However, it is different from the effective aligned reads number on the reference. Actually only the latter is known with real datasets, so our predictions needs to rely on this effective number rather than the input parameter. This information is provided on the x-axis of figure 3.4, where its variation is highlighted with the ratio of aligned CDS in the reference genome. Despite stating the obvious, we see that the more close the genome is to the reference, the more reads effectively aligned. Aligned CDS ratio shows a plateau for each class as long as reads number increases. This reflects the fraction of CDS shared by the reference genome and the genome from which reads where sequenced.

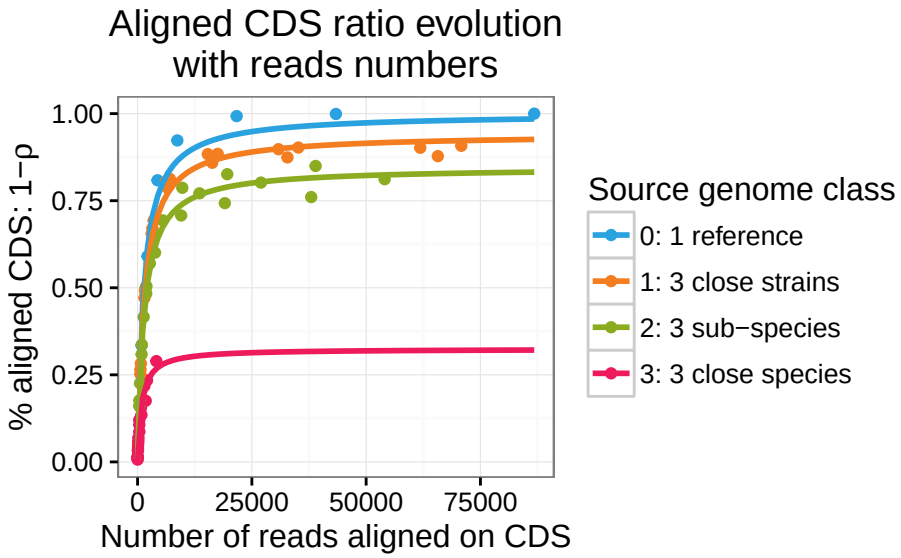


Figure 3.4: Aligned reads number influence on one parameter: ρ or aligned CDS ratio. Each dot represents an alignment metric on a genome–the ratio of CDS with at least one read aligned– function of the effective number of reads aligned. Data stem from the *Streptococcus* dataset. Reference genomes are classified on their closeness to the reference strain we choose throughout this study –see figure 2.4. Non-linear regressions for each class are printed on plain lines. Each parameters –with respect to Michaelis-Menten formula– yields significant estimation.

The aforementioned tendency could be captured with non linear regression. The fit for each class was based on Michaelis-Menten formula and each parameters –such as V_m or K yields significant estimations. For the parameters of mixture model pure distributions –gaussians and log-normal–, I observed a linear tendency between aligned reads number and the mean μ or the variance σ^2 –yet with different coefficient.

The strategy set up to circumvent start values issues will encompass both prior knowledge –in the form of regression models– and a wider space parameter exploration –in the form of random sampling. Hence, the following initialisation: (1) contributions π_c are drawn from a uniform random law, (2) ratios ρ_c are initialised with respect to the number of reads aligned based on non linear regression models above, and finally (3) distribution parameters –such as Gaussian and log-normal– are initially set up both by predicting from linear regressions and by sampling in four times the prediction interval. This approach yields several mixture models ranked with respect to the BIC criterion, hence selecting the most likely.

Several models are tested and the space of parameters is well explored. I had also written several R scripts to explore alternative models generated. A report is automatically created giving an overview of model estimation. However, parameters estimations seems far from the primary aims of identifying micro-organisms. I then worked on decisions rules.

Decision rules concerning the mixture model output

Composition dataset In order to bridge the gap from contribution estimations to a boolean vector of presence / absence of micro-organisms, I explored decisions rules from the mixture model. I designed a dataset termed *composition dataset* different from the training dataset with *Streptococcus* but with a similar approach. This dataset actually consists of 3 subdatasets of growing complexity –from 1 the simpler to 3 the more complex.

The first dataset encompass reads from a selected reference genome – here *Lactococcus lactis subsp. cremoris* A76– mixed with reads from a *noise* genome. This distant genome needs to be close enough from the reference to provide reads that can be aligned and distant enough from the reference to be able to distinguish them. I choosed *Lactococcus raffinolactis* JIM2957 genome to be the *noise* genome. I expected this data to illustrate strain identification when the reference genome is available.

The second dataset encompass reads stem from the reference genome, the noise genome and an addtional close strain –*Lactococcus lactis subsp. cremoris* MG1363 from closeness class 1. Similarly the third dataset –noted 3– comprise the latter dataset with a supplementary genome to stem the reads from. The sub species *Lactococcus lactis subsp. lactis* Il1403 is the one added.

From these composition of genomes, I simulated three datasets mimicking shotgun metagenomics sequencing at two different coverages: High Coverage ($1\times$) and Low Coverage ($0.1\times$). They are respectively noted HC and LC. Their genome composition is the same and is outlined in figure 3.5 where genomes used in the simulation are depicted with triangles and circles. The 3 datasets at two coverages are hence aligned to 10 reference genomes distributed like the *Streptococcus* earlier: 1 reference genome, 3 close strains, 3 sub-species and 3 close species or distant genomes.

Decision rules After reads alignement on each genome, parameters are estimated for all differents models presented above and a model is selected thanks to the BIC criterion. The mixture component $c = 0$ and its estimated parameters will be used for decisions rules. We expect this component to capture the information of the closest strain. Hence if the exact strain is present, the signal will be strong to support this information. At first, I outlined two rules to decide whether the exact reference genome had contributed to the alignment observed. Closeness class 0 contribution to the entire mixture model needed to be important –say $\pi_0 > 0.3$– and the number of CDS uncovered should be marginal for a fair number of reads aligned –say ρ_c < 0.01. These thresholds stem from prior design and local observations. As the approach was still developed, I settled for the previous values and assess them onto a dataset in order to refine them if needed.

Mixture model results

Preliminary results on simulated datasets The first observation we can state from figure 3.5 is the following: in Low Coverage datasets no genome was considered present with the given thresholds.

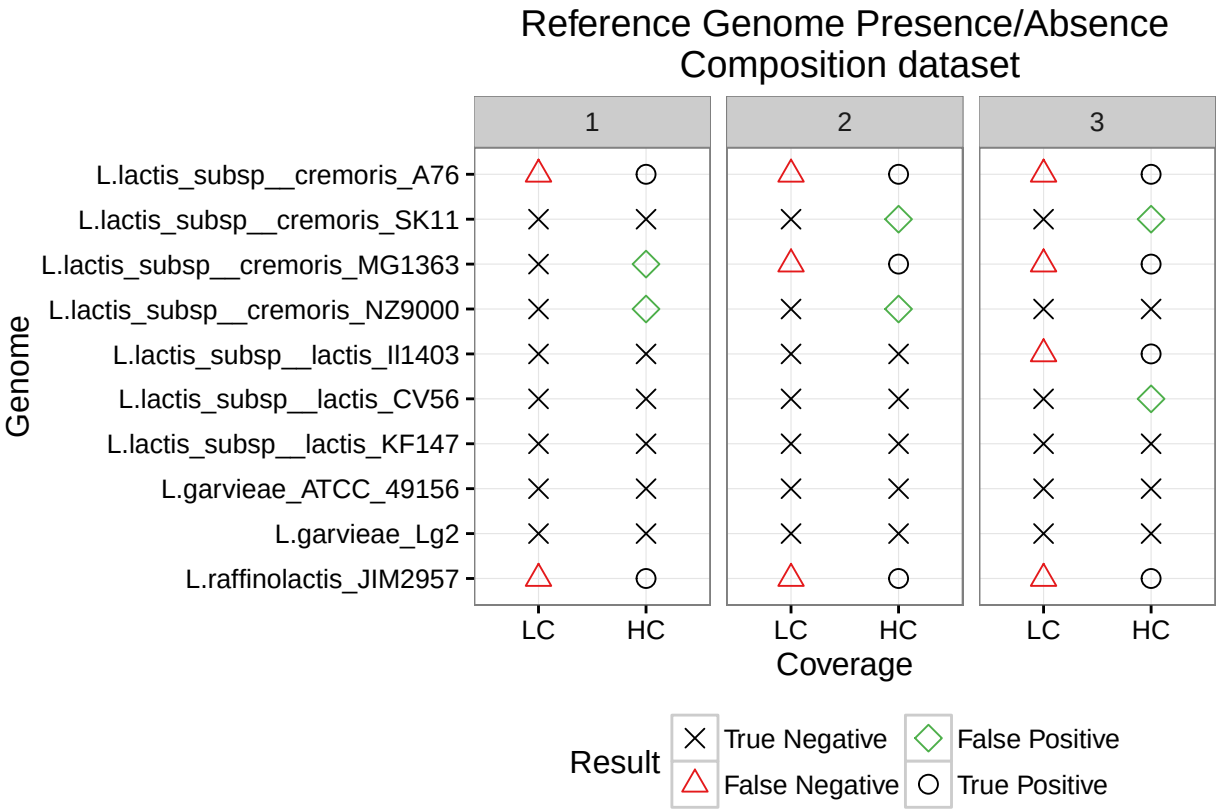


Figure 3.5: Composition dataset outcomes after mixture model estimation. On the y-axis are the 10 reference genomes onto which reads alignment was done. Some of them were chosen to simulate reads from their genome sequence. The x-axis consists of the two coverages used in the simulation. Labeled facets separate results for each dataset. Results consists in a combination of truth vectors: one stating the presence/absence according to simulated datasets and the other one yields by the outcome of decision rules mentioned above. Symbols highlights the four possible values of the combination. Triangles and circles shapes depict genomes used to simulate data.

This high rate of false negative occurs despite an important ratio of covered CDS –data not shown. Exact strains were retrieved only in High Coverage experiment. However, false positives were also retrieved. They matches to close strains which is not specifically a wrong outcome.

The effect of threshold was assessed by encapsulate this plot into a Shiny application in order to interactively explore decision rules. However, no suitable pattern emerge from this exploration and there was no interesting threshold combination.

Others datasets

Verification dataset I then design a *verification* dataset in order to assess the following questions: (1) how to determine a genome absence in a dataset despite an important number of reads aligned? and (2) could we identify the exact strain if few reads from this strain genome were aligned in addition to this important noise signal? It was constructed with the reference genome mentioned above with combination of coverages $-0.01 - 0.1\times$ to $10 - 100\times$. The results are not shown here but the

conclusion is similar to earlier, this mixture model yields proper results only when the coverage is important –that is better than $0.1\times$. This lower bound in disappointing as this coverage values is considered high in metagenomics.

Mock Community I also ran the mixture model on a published dataset: the Mock Community. It is the real sequencing experiment conducted on a known composition microbe cocktail. This metagenomics dataset was supposed to be a trivial test but unfortunately it yields too many false negatives again: around 50% of species are not retrieved with previous thresholds. Some false negatives could be discarded after some models modifications. For instance, label switching is the possible inversion of components in a mixture model. Mainly due to the iterative process, it is known to cause issues in classification. To circumvent this issue, I set up a sorting strategy on gaussian components. It consists in labelling to 0 the first component yields by a double sort: (1) ascending sort of ρ_c, then (2) descending sort of π_c. This sorting strategy relies on the prior that the closest strain contributing to the alignment will have few CDS uncovered –small ρ_0– and could be the major contributor of the component –high π_0. I am aware that this strategy may prevent the discovery of low -abundant micro-organisms. Although these modifications decrease the false negative rate, it was still too important.

A more relevant mock community? I had planned to test this model on a mock dataset designed by Dugat-Bony et al (see 2015). This microbial community dataset is relevant to our projects because it is associated with cheese ecosystems and it has been sequenced with SOLiD technology. Time constraints prevented this dataset to be tested as well as aforementioned negative results on the Mock Community. In spite of an almost perfect match of this dataset to the scope of our project, I did not used the mixture model approach on it yet.

3.2 Integration

3.2.1 Integration with the Food-Microbiome Transfert database

In the scope of the Food-Microbiome Transfert project, a metadata-extended genomics database has been build thanks to a bioinformatician engineer in the team since 8 months. The interaction between our metagenomics analysis tool GeDI and this database was first described in figure 1.1. Actually, GeDI fetch genomes and annotations information from flat files in directories. This system is totally relevant and based on standard files such as FASTA and Genbank files. However, in the scope of the Food-Microbiome project, we planned to fetch these information directly from the database.

3.2.2 Integration with the Galaxy platform

A tool is relevant only if is it used. We planned to be able to run GeDI from the web interface in order to promote the use of our metagenomics analysis tool. Therefore, we worked to integrate GeDI to the Galaxy platform thus taking advantage of Galaxy API to run softwares. The Galaxy platform is a web-based tool providing (1) easy access to command line tools and (2) data history and workflow management. It is hosted on the same location of the migale server. A bioinformatician engineer in the team attended a training to integrate new tools into the Galaxy platform. Our tool's integration was his exercise throughout this course. This training was an opportunity for me to craft small datasets and documents to provide him material for tests. Remarks and conclusions after this formation were taken into account to properly adapt our tool to some requirements, filetypes for instance.

3.2.3 Integration with other tools and software performance

In addition to scientific improvements covered above, I assess our tool performance and modify it in order to enhance them. I was mainly working on converting in-house files generated by GeDI into standard filetype in bioinformatics –such as BAM and GFF. We expected a decrease in file size due to the compression of alignment data with BAM files. Likewise, we expected an increase in software speed due to the transition to highly efficient tools such as samtools instead of home-made scripts. These expectations were met with a 4-fold decrease in running time and 7-fold decrease in disk space.

Standard files ease the exporting to other tools. For example, industrial partners and academics will be able to export metagenomics reads alignment on their strain of interest, and visualize data with Tablet (Milne et al. 2013) or process it for parallel downstream analyses.

3.3 Application

This new version of GeDI was applied to 9 cheeses samples in order to illustrate the tools ability. I wanted to highlight the importance of mismatches through the draft analysis of the representation of one strain in these samples. The following cheeses were used. Note that the surface of the cheese was sampled for every cheese. Each sample was sequenced with SOLiD technology.

I used a strain reference genome *Psychrobacter aquimaris* sequenced in the collective genome project (Almeida et al. 2014). It consists of 126 contigs and a total of 2856 CDS and 2495 post filter.

Roquefort, Salers and Toscanello metagenomes harbours similar coverage profiles. Half the genome is covered and less than 15% covered only by reads with mismatches. For some cheese like Epoisses and Munster especially, the majority of genome coverage is contained in coverage yielded by reads

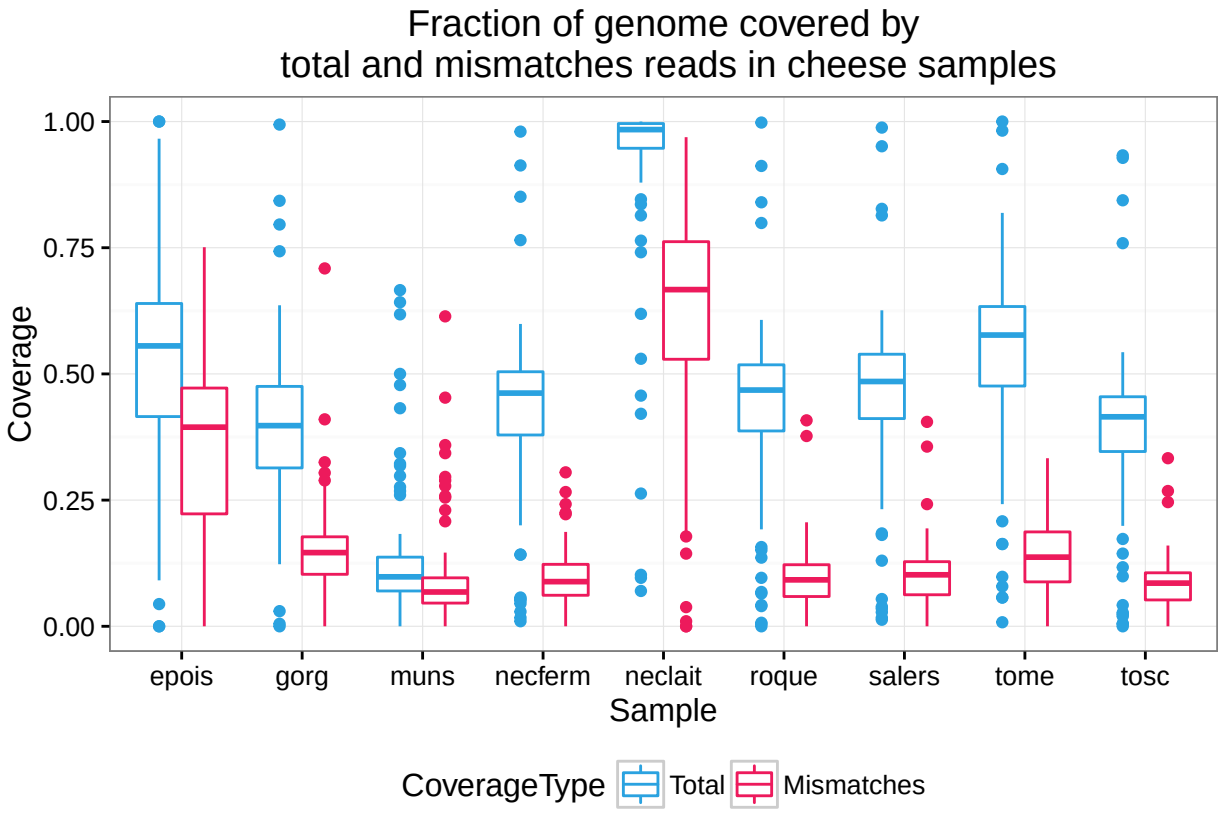


Figure 3.6: Strain representation in cheese samples: an overview

with mismatches. This preliminary observations provide an illustration of the type of data GeDI could output.

Table 3.1: Cheese samples overview after alignment to *Psychrobacter aquimaris*.

Cheese Name	Legend	# Reads aligned
Epoisses	`epois`	436,056
Gorgonzola	`gorg`	77,697
Munster	`muns`	139,804
Saint-Nectaire fermier	`necferm`	65,914
Saint-Nectaire laitier	`neclait`	4008,074
Roquefort	`roque`	68,080
Salers	`salers`	72,226
Tome des Bauges	`tome`	102,503
Toscanello	`tosc`	54,974

4 Conclusion and prospects

4.1 Conclusion

4.1.1 Food microbiology context

Cheese flora is a microbial community. Like other microbial communities, it benefits from recent advances in metagenomics. On one hand, technological advances with high-throughput sequencing provided insights into these communities. On the other hand, computational developments enabled better data analysis with the emergence of new algorithms or tools. Cheese ecosystem benefits from multiple sources of micro-organisms –from starters to environment. Strains in cheese ecosystems possess interesting technological interests –e.g., organoleptic compounds synthesis, phage resistance. Hence, deciphering the microbial diversity of cheeses communities is relevant.

The Food-Microbiome project aims to deeply understand cheese ecosystems given next-generation sequencing technologies potential. In the scope of these two projects, we rely on existing reference genomes to (1) identify micro-organisms present in the ecosystem –if possible to the strain level– and (2) characterise low-abundant micro-organisms.

Throughout two years of apprenticeship, I have focused my research on a in-house metagenomics analysis tool –named GeDI. I was mainly interested in the development of a method to automatically identify species and strains in the ecosystem. To this end, I relied on shotgun metagenomics reads alignments on a set of reference genomes. However, specific cases leads to the exploration of several *improvements* in terms of scientific methods. For instance, reference genomes are likely to differ from the strains actually present in the ecosystem, hence micro-organism identification has to be tailored.

4.1.2 Exploratory approach for microbial strains identification

I have been trying to take into account the discrepancy between between reference genomes and genomes from the environment since my apprenticeship start. Previous approaches were only able to detect when the exact strain was present both in the ecosystem and as a reference genome onto which reads are mapped.

Mixture model of distribution Reads aligned on a reference genome may originate from multiple contributors, hardening the task of microbial identification. Last year, I tried to tackle this issue through a new angle: CDS coverage densities. I worked on a proof of concept based on a mixture model and thus designed simulated datasets in order to document this idea. The mixture model of distribution is meant to split observed CDS coverage density into contributions of defined distributions. Therefore, mixture contributions are proxies of micro-organisms contributions to an alignment to a reference genome.

Limits identified and tackled During the mixture development, I stumble upon some limits –algorithm or method related–, I managed to circumvent most of them. I extended the initial scope of four components to build three mixture models –from two to four components. The relevant model is chosen with respect to the BIC criterion.

The EM algorithm is known to be sensitive to start values. Moreover, given sequencing depth disparities in metagenomics, I wanted to document reads number influence on the distributions. It appears that both limits could be addressed at the same time. Therefore, I set up a strategy combining (1) prior knowledge on reads number through regressions and (2) randomness from sampling in order to provide multiples relevant start values.

Label switching is the possible inversion of components distribution in the iterative process. If conclusions are drawn from a specific component, this common issue faced by mixture model was addressed by a sorting strategy agreeing with the model.

However, the model struggle with low coverage and yields too many false negatives. It performs worse than the first approach with Lander and Waterman. On the Mock Community dataset, model selection is sometimes irrelevant despite being the more likely.

Model prospects Altering thresholds could be a lead to decrease the false negative rate. However, this approach was meant to be accurate and will not be suitable for a large screening followed by refiments. I also consider adding other critera in the decision rule such as the number of components for instance. But I foresee a difficult biological conclusion to this approach. Lastly I suggested to switch from a boolean decision to an index of presence or closeness. But this implies to let the user defined the threshold, and no consensus could emerge. However, metagenomics tools often only advice the user on thresholds to apply.

Despite exploring deeply the limits of this new approach, no suitable pattern arise from its application results. While this model provides an extended view of metagenomics reads alignment, it will not be used in our tool to routinely identify micro-organisms in metagenomes.

4.1.3 Tool integration

Our tool can be fully integrated into the Galaxy platform thanks to the engineer training. Moreover, our tool now complies requirements stated by the platform. These requirements accelerated the expected conversion of our tool to standard bioinformatics output files.

This spring cleaning leads to the removal of temporary files, and the use of compiled toolbox instead of home made scripts. Hence, these changes strongly reduced disk space usage and increase the software speed.

4.2 Prospects

4.2.1 Future of GeDI

Several approaches were evaluated to infer the presence or absence of micro-organisms in an ecosystem. CDS coverage holds relevant information to address this task. While continuing on-going developments in this direction, we continue to be aware of recent advances in metagenomics.

Comparison with other tools Recently the challenge of strain retrieval in metagenomics shifts from hypothetical to achievable thanks to the scientific and industrial community efforts. Therefore many publications and tools are recently available and could pave the way to new analyses and exploration (Piro, Lindner, and Renard 2016; Truong et al. 2015; Luo et al. 2015; Cleary et al. 2015).

Two tools were published last year addressing similar challenges as our project. A framework to compare our tool with Sigma (Ahn, Chai, and Pan 2015) and MicrobeGPS Lindner and Renard (2015)]is already defined. We plan to use the Mock Community dataset with these tools and assess whether the same strain was foundby the three tools.

Future improvements I was working recently on implementing variant calling strategies to strains in high abundance –and hence with high coverage– in metagenomes. We are also starting to think how to decipher strains cocktails in ecosystems, mostly by choosing accordingly the catalog of reference genomes and then carefully set alignment parameters to get the most out of DNA reads information. Besides, our approach can only scale to hundreds of genomes so we plan to use a faster approach to screen for candidate genomes with Kraken (Wood and Salzberg 2014). Then, we will apply our method on this subset to identify micro-organisms.

4.2.2 Next with Food-Microbiome Transfert and others projects

The aim of the Food-Microbiome Transfert is to provide a functional web based interface transparently running our metagenomics analysis tool. It will be used by industrials partners and academics to gain insights into food related ecosystems. Specific trainings will be provided by the Migale platform team given their experience in Galaxy trainings and their involvement in the Food-Microbiome Transfert project.

For now GeDI does not rely on the metadata concerning cheese ecosystem stored in the Food-Microbiome Transfert database. However, in the near future, including these metadata could enhance visualisation and provide a new perspective on cheese ecosystems.

Our team is involved in several projects, where the metagenomics analysis tool presented here could be use. For instance, the 1350 cheeses project in collaboration with the CNIEL. This project led by Françoise IRLINGER (INRA-Grignon) was submitted again in order to gain insights into cheeses with controlled origin –AOP.

References

Ahn, Tae-Hyuk, Juanjuan Chai, and Chongle Pan. 2015. "Sigma: Strain-Level Inference of Genomes from Metagenomic Analysis for Biosurveillance." *Bioinformatics* 31 (2): 170–77. doi:10.1093/bioinformatics/btu641.

Almeida, Mathieu, Agnès Hébert, Anne-Laure Abraham, Simon Rasmussen, Christophe Monnet, Nicolas Pons, Céline Delbès, et al. 2014. "Construction of a Dairy Microbial Genome Catalog Opens New Perspectives for the Metagenomic Analysis of Dairy Fermented Products." *BMC Genomics* 15: 1101. doi:10.1186/1471-2164-15-1101.

Almena-Aliste, Montserrat, and Bernard Mietton. 2014. "Cheese Classification, Characterization, and Categorization: A Global Perspective." *Microbiology Spectrum* 2 (1): CM–0003–2012. doi:10.1128/microbiolspec.CM-0003-2012.

Angly, Florent E, Dana Willner, Forest Rohwer, Philip Hugenholtz, and Gene W Tyson. 2012. "Grinder: A Versatile Amplicon and Shotgun Sequence Simulator." *Nucleic Acids Research* 40 (12): e94. doi:10.1093/nar/gks251.

Ågren, Joakim, Anders Sundström, Therese Håfström, and Bo Segerman. 2012. "Gegenees: Fragmented Alignment of Multiple Genomes for Determining Phylogenomic Distances and Genetic Signatures Unique for Specified Target Groups." Edited by Niyaz Ahmed. *PLoS ONE* 7 (6): e39107. doi:10.1371/journal.pone.0039107.

Bourdichon, François, Serge Casaregola, Choreh Farrokh, Jens C. Frisvad, Monica L. Gerds, Walter P. Hammes, James Harnett, et al. 2012. "Food Fermentations: Microorganisms with Technological Beneficial Use." *International Journal of Food Microbiology* 154 (3): 87–97. doi:10.1016/j.ijfoodmicro.2011.12.030.

Button, Julie E., and Rachel J. Dutton. 2012. "Cheese Microbes." *Current Biology* 22 (15): R587–R589. doi:10.1016/j.cub.2012.06.014.

Callon, Cécile, Emilie Retureau, Robert Didienne, and Marie-Christine Montel. 2014. "Microbial Biodiversity in Cheese Consortia and Comparative Listeria Growth on Surfaces of Uncooked Pressed Cheeses." *International Journal of Food Microbiology* 174 (March): 98–109.

doi:10.1016/j.ijfoodmicro.2014.01.003.

Chun, Jongsik, and Fred A. Rainey. 2014. "Integrating Genomics into the Taxonomy and Systematics of the Bacteria and Archaea." *International Journal of Systematic and Evolutionary Microbiology* 64 (Pt 2): 316–24. doi:10.1099/ijs.0.054171-0.

Ciccarelli, Francesca D., Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. 2006. "Toward Automatic Reconstruction of a Highly Resolved Tree of Life." *Science* 311 (5765): 1283–7. doi:10.1126/science.1123061.

Cleary, Brian, Ilana Lauren Brito, Katherine Huang, Dirk Gevers, Terrance Shea, Sarah Young, and Eric J Alm. 2015. "Detection of Low-Abundance Bacterial Strains in Metagenomic Datasets by Eigengenome Partitioning." *Nature Biotechnology* 33 (10): 1053–60. doi:10.1038/nbt.3329.

Delbes, C., L. Ali-Mandjee, and M.-C. Montel. 2007. "Monitoring Bacterial Communities in Raw Milk and Cheese by Culture-Dependent and -Independent 16S RRNA Gene-Based Analyses." *Applied and Environmental Microbiology* 73 (6): 1882–91. doi:10.1128/AEM.01716-06.

Delignette-Muller, Marie Laure, and Christophe Dutang. 2015. "Fitdistrplus: An R Package for Fitting Distributions." *Journal of Statistical Software* 64 (4): 1–34.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Dugat-Bony, Eric, Cécile Straub, Aurélie Teissandier, Djamila Onésime, Valentin Loux, Christophe Monnet, Françoise Irlinger, et al. 2015. "Overview of a Surface-Ripened Cheese Community Functioning by Meta-Omics Analyses." *PLOS ONE* 10 (4): e0124360. doi:10.1371/journal.pone.0124360.

Ercolini, Danilo. 2013. "High-Throughput Sequencing and Metagenomics: Moving Forward in the Culture-Independent Analysis of Food Microbial Ecology." *Applied and Environmental Microbiology* 79 (10): 3148–55. doi:10.1128/AEM.00256-13.

Irlinger, F., S. Layec, S. Helinck, and E. Dugat-Bony. 2015. "Cheese Rind Microbial Communities: Diversity, Composition and Origin." *FEMS Microbiology Letters* 362 (2): 1–11. doi:10.1093/femsle/fnu015.

Konopka, Allan. 2009. "What Is Microbial Community Ecology?" *The ISME Journal* 3 (11): 1223–30. doi:10.1038/ismej.2009.88.

Kumar, Purnima S., Michael R. Brooker, Scot E. Dowd, and Terry Camerlengo. 2011. "Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing." *PloS One* 6 (6): e20956. doi:10.1371/journal.pone.0020956.

Lander, E. S., and M. S. Waterman. 1988. "Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis." *Genomics* 2 (3): 231–39.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25.

doi:10.1186/gb-2009-10-3-r25.

Lindner, Martin S., and Bernhard Y. Renard. 2015. "Metagenomic Profiling of Known and Unknown Microbes with MicrobeGPS." *PLoS ONE* 10 (2): e0117711. doi:10.1371/journal.pone.0117711.

Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J Xavier, and Dirk Gevers. 2015. "ConStrains Identifies Microbial Strains in Metagenomic Datasets." *Nature Biotechnology* 33 (10): 1045–52. doi:10.1038/nbt.3319.

McHardy, Alice Carolyn, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. 2007. "Accurate Phylogenetic Classification of Variable-Length DNA Fragments." *Nature Methods* 4 (1): 63–72. doi:10.1038/nmeth976.

Milne, Iain, Gordon Stephen, Micha Bayer, Peter J. A. Cock, Leighton Pritchard, Linda Cardle, Paul D. Shaw, and David Marshall. 2013. "Using Tablet for Visual Exploration of Second-Generation Sequencing Data." *Briefings in Bioinformatics* 14 (2): 193–202. doi:10.1093/bib/bbs012.

Monnet, Christophe, Sophie Landaud, Pascal Bonnarme, and Dominique Swennen. 2015. "Growth and Adaptation of Microorganisms on the Cheese Surface." *FEMS Microbiology Letters* 362 (1): 1–9. doi:10.1093/femsle/fnu025.

Montel, Marie-Christine, Solange Buchin, Adrien Mallet, Céline Delbes-Paus, Dominique A. Vuitton, Nathalie Desmasures, and Françoise Berthier. 2014. "Traditional Cheeses: Rich and Diverse Microbiota with Associated Benefits." *International Journal of Food Microbiology* 177 (May): 136–54. doi:10.1016/j.ijfoodmicro.2014.02.019.

Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. "MetaVelvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads." *Nucleic Acids Research* 40 (20): e155. doi:10.1093/nar/gks678.

Ortolani, Maria Beatriz Tassinari, Anderson Keizo Yamazi, Paula Mendonça Moraes, Gabriela Nogueira Viçosa, and Luís Augusto Nero. 2010. "Microbiological Quality and Safety of Raw Milk and Soft Cheese and Detection of Autochthonous Lactic Acid Bacteria with Antagonistic Activity Against Listeria Monocytogenes , *Salmonella* Spp., and *Staphylococcus Aureus*." *Foodborne Pathogens and Disease* 7 (2): 175–80. doi:10.1089/fpd.2009.0390.

Parente, Eugenio, Luca Cocolin, Francesca De Filippis, Teresa Zotta, Ilario Ferrocino, Orla O'Sullivan, Erasmo Neviani, Maria De Angelis, Paul D. Cotter, and Danilo Ercolini. 2016. "FoodMicrobionet: A Database for the Visualisation and Exploration of Food Bacterial Communities Based on Network Analysis." *International Journal of Food Microbiology* 219 (February): 28–37. doi:10.1016/j.ijfoodmicro.2015.12.001.

Piro, Vitor C., Martin S. Lindner, and Bernhard Y. Renard. 2016. "DUDes: A Top-down Taxonomic Profiler for Metagenomics." *Bioinformatics*, March, btw150. doi:10.1093/bioinformatics/btw150.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf,

Chaysavanh Manichanh, Trine Nielsen, et al. 2010. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing." *Nature* 464 (7285). doi:10.1038/nature08821.

Quigley, Lisa, Orla O'Sullivan, Tom P. Beresford, R. Paul Ross, Gerald F. Fitzgerald, and Paul D. Cotter. 2012. "High-Throughput Sequencing for Detection of Subpopulations of Bacteria Not Previously Associated with Artisanal Cheeses." *Applied and Environmental Microbiology* 78 (16): 5717–23. doi:10.1128/AEM.00918-12.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Renault, Pierre. 2009. *ANR Food-Microbiomes : Outils Innovants Pour Favoriser, En Toute Sécurité, L'utilisation Des Microorganismes Naturels Dans L'alimentation.* Agence Nationale de la Recherche.

Richter, Michael, and Ramon Rosselló-Móra. 2009. "Shifting the Genomic Gold Standard for the Prokaryotic Species Definition." *Proceedings of the National Academy of Sciences of the United States of America* 106 (45): 19126–31. doi:10.1073/pnas.0906412106.

Salque, Mélanie, Peter I. Bogucki, Joanna Pyzel, Iwona Sobkowiak-Tabaka, Ryszard Grygiel, Marzena Szmyt, and Richard P. Evershed. 2013. "Earliest Evidence for Cheese Making in the Sixth Millennium Bc in Northern Europe." *Nature* 493 (7433): 522–25. doi:10.1038/nature11698.

Schbath, Sophie, Véronique Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, and Jean-François Gibrat. 2012. "Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis." *J. Comput. Biol.* 19 (6): 796–813. doi:10.1089/cmb.2012.0022.

Sharpton, Thomas J. 2014. "An Introduction to the Analysis of Shotgun Metagenomic Data." *Plant Genetics and Genomics* 5: 209. doi:10.3389/fpls.2014.00209.

Truong, Duy Tin, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 12 (10): 902–3. doi:10.1038/nmeth.3589.

Varghese, Neha J., Supratim Mukherjee, Natalia Ivanova, Konstantinos T. Konstantinidis, Kostas Mavrommatis, Nikos C. Kyrpides, and Amrita Pati. 2015. "Microbial Species Delineation Using Whole Genome Sequences." *Nucleic Acids Research* 43 (14): 6761–71. doi:10.1093/nar/gkv657.

Venter, J. C. 2004. "Environmental Genome Shotgun Sequencing of the Sargasso Sea." *Science* 304 (5667): 66–74. doi:10.1126/science.1093857.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer New York.

Wolfe, Benjamin E., Julie E. Button, Marcela Santarelli, and Rachel J. Dutton. 2014. "Cheese Rind Communities Provide Tractable Systems for In Situ and In Vitro Studies of Microbial Diversity." *Cell* 158 (2): 422–33. doi:10.1016/j.cell.2014.05.041.

Wood, Derrick E, and Steven L Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classifica-

tion Using Exact Alignments." *Genome Biology* 15 (3): R46. doi:10.1186/gb-2014-15-3-r46.

Cheese ecosystems metagenomics
Exploration & improvements around a bioinformatics tool

Cheese complex flora –composed of dairy micro-organisms– is not completely and exactly known in most cheeses. Further understandings of these ecosystems needs a better characterization of the cheese flora and a precise taxonomic identification. Hundreds of genomes extracted from dairy products are currently available in genomics databases. But the key points to tackle are low abundant species and identification up to the strain level. A tool was developed in the team to address these issues using shotgun metagenomics sequencing data.

I have focused my research around this tool for two years. I mostly worked on scientific improvements: deeply exploring new leads to enhance ecosystem taxonomic identification. Notably I designed an approach taking into account the discrepancy between reference genomes and genomes from the environment. However, it did not yields expected results despite a thorough examination of model limits.

I have also worked on computational features improvements such as compatibility by using standard bioinformatics files. This conversion leads to space usage and speed improvements.
I am involved in the integration of this tool with a metadata-extended genomics database.

Keywords metagenomics, microbial community, simulated datasets.
Technical keywords short reads alignement, CDS coverage, mixture model.

Métagénomique des écosystèmes fromagers
Améliorations et explorations d'un outil bioinformatique

La composition exacte des flores fromagères n'est pas connue dans la plupart des fromages. Afin d'approfondir les connaissances de ces écosystèmes, une meilleure caractérisation de la flore fromagère et une assignation taxonomique précise sont nécessaires. Quelques centaines de génomes issus de produits laitiers sont désormais disponibles dans les banques de données. Cependant, la faible abondance de certaines espèces et l'identification jusqu'à la souche restent des défis. Dans l'équipe, un outil a été développé pour les aborder à partir de données de séquençage métagénomique global aléatoire.

Durant deux ans, j'ai focalisé mes travaux autour de cet outil. J'ai principalement travaillé à des perfectionnements telles que l'exploration de nouvelles pistes pour améliorer l'identification de micro-organismes. J'ai développé une approche qui implique la discordance entre génomes de références et génomes de l'environnement. Malgré une exploration exhaustive des limites du modèles, cette méthode n'offre cependant pas les résultats escomptés.

J'ai également travaillé à améliorer les caractéristiques de l'outil telles que l'inter-opérabilité en utilisant des standards en bioinformatique. Cette conversion a permis de limiter l'espace disque utilisé et de diminuer le temps d'éxecution. J'ai également été impliqué dans l'intégration de cet outil avec une base de données génomique enrichies en métadonnées.

Mots-clés métagénomique, microbiome, alignement de courtes lectures.
Mots-clés techniques données simulées, couvertures des CDS, modèle de mélange.