ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO EN INGENIERÍA DE LA SALUD

MÉTODOS Y FLUJOS DE TRABAJO PARA EL ANÁLISIS TAXONÓMICO DE DATOS METAGENÓMICOS

METHODS AND WORKFLOWS FOR THE TAXONOMIC ANALYSIS OF METAGENOMIC DATA

Realizado por

Pablo Rodríguez Brazzarola

Tutorizado por

Oswaldo Trelles Salazar

Departamento

Arquitectura de Computadores

UNIVERSIDAD DE MÁLAGA

MÁLAGA, JUNIO 2018

Fecha defensa:

El Secretario del Tribunal

RESUMEN

Todas las plantas y animales tienen comunidades microbianas estrechamente asociadas que hacen que los nutrientes, metales y vitaminas necesarios estén disponibles para su huésped, contribuyendo esencialmente a la vida en la Tierra. El campo inherentemente complejo que tiene como objetivo comprender las contribuciones de estas microbiotas a la biósfera se conoce como metagenómica. Uno de los principales objetivos en este campo de investigación es determinar la composición de los organismos presentes en una muestra ambiental. Para ello, se han desarrollado diversas herramientas, la mayoría de ellas basadas en los resultados de búsqueda de similitud obtenidos al comparar un conjunto de secuencias biológicas contra una base de datos. Aunque el campo ha avanzado significativamente desde su inicio, todavía hay otros asuntos por resolver como tratar con variantes genómicas y detectar secuencias repetidas que podrían pertenecer a diferentes especies en una mezcla de organismos desiguales y desconocidos. Los distintos enfoques al analizar una muestra de metagenoma dan lugar a preguntarse si el análisis de una muestra con lecturas (fragmentos cortos de ADN producto de procedimientos de secuenciación) proporciona una mayor comprensión del metagenoma que con contigs (lecturas superpuestas que se han ensamblado juntas). El ensamblaje produce fragmentos genómicos más grandes, pero conlleva el riesgo de producir contigs a partir de lecturas de diferentes organismos. Por otro lado, las lecturas son más cortas y por ello su significación estadística es más difícil de evaluar, pero son más numerosas. En este proyecto, evaluamos y comparamos la calidad de cada una de estas alternativas para establecer el enfoque de datos que proporciona los mejores resultados en términos de informar la abundancia relativa de especies dentro de una muestra. Para validar los resultados, generamos conjuntos de datos de lectura sintéticos que pertenecen a organismos previamente identificados manteniendo las distribuciones de abundancia relativa. Posteriormente, los ensamblamos en un conjunto de contigs y realizamos un análisis taxonómico con ambos enfoques. Debido a que podemos rastrear el origen de las colecciones de lecturas, también se puede medir la calidad de estas asignaciones con un conjunto de herramientas desarrolladas para demostrar que el análisis con lecturas proporciona una representación más confiable de las especies en una muestra que usando los contigs, especialmente en casos que presentan una alta variabilidad genómica. Esperamos que las herramientas desarrolladas contribuyan a mejores soluciones en metagenómica y que brinden apoyo a los investigadores que trabajan en dicho campo.

Palabras clave: flujo de trabajo, metagenómica, asignación taxonómica; análisis de secuenciación; comparación metagenómica, ensamblaje de ADN.

ABSTRACT

All plants and animals have closely associated microbial communities that make necessary nutrients, metals, and vitamins available to their host, essentially contributing to all life on Earth. The inherently complex field that aims to understand the contributions of these microbiotas to the biosphere is known as metagenomics. One of the primary goals in this research field is to determine the composition of organisms present in an environmental sample. In order to do so, diverse tools have been developed, most of them based on the similarity search results obtained from comparing a set of biological sequences against a database. Although the field has advanced significantly since its beginning, there still are affairs to solve such as dealing with genomic variants and detecting repeated sequences that could belong to different species in a mixture of uneven and unknown representation of organisms in the sample. The distinct approaches when analyzing a metagenome sample give rise to the question of whether analyzing a sample with reads (short fragments of DNA product of sequencing procedures) provides further understanding of the metagenome than with contigs (overlapping reads that have been assembled together). The assembly yields larger genomic fragments but bears the risk of producing contigs from reads of different organisms. On the other hand, reads are shorter and therefore their statistical significance is harder to asses, but there is a larger number of them. In this project, we assess and compare the quality of each of these alternatives to establish the data-approach that provides the best results in terms of reporting the relative abundance of species within a sample. To validate the results, we generate synthetic read datasets that belong to previously identified organisms maintaining the relative abundance distributions. Afterwards, we assemble these into a set of contigs and perform a taxonomic analysis on both approaches. Since we can trace the origin of the reads collections we are able to measure the quality of these assignments with a set of developed tools in order to demonstrate that analyzing with reads provide a more trustworthy representation of the species in a sample than using contigs, especially in cases that present a high genomic variability. We expect the developed tools will contribute for better solutions in metagenomics providing support to researchers working in such field.

Keywords: workflow; metagenomics; taxonomic assignment; sequencing analysis; metagenome comparison, DNA assembly.

TABLE OF CONTENTS

Executive Summary

Metagenomics is a field that aims to study an uncultured biological sample taken directly from its original environment. This area of research presents many more challenges than traditional genomics, such as the uneven and unknown abundance of species and the fact that not all species will be completely represented by the reads generated from the sequencing experiment.

One of the main goals in this field is to analyze the composition of species within a sample. This study is known as a taxonomic analysis and multiple approaches have been designed for this purpose. Two of the most common ones are (1) using reads (generated from the sequencing experiment) or (2) using contigs (obtained by assembling the reads). Even though the goal is the same, each approach provides different results, and to the best of our knowledge, there is no study addressing such difference. Therefore there is a need to assess and compare the quality of these taxonomic assignments in order to obtain the best possible results in metagenomic taxonomic analysis.

One of the problems that arise when attempting to compare a taxonomic assignment from a real metagenomic sample is the fact that the real relative abundance of species is unknown. To solve this problem we have prepared a software that generates a metagenomic synthetic dataset of reads from a selection of genomes and specifying the abundance of reads per genome. Afterwards, these reads are assembled into contigs.

In this project, we perform the taxonomic analysis of a metagenomic sample with the reads and contigs approach. This is executed in order to obtain several indicators by applying a set of developed software tools. that measure the quality of the analysis, enabling a comparison between these different approaches. To facilitate the use of these tools, an automatic pipeline has been made available.

Lastly we present two use cases that apply the developed software tools with the intent of validating and consolidating an appropriate procedure to obtain the best possible results

when performing a metagenomic taxonomic analysis, whether it is with the reads or with the contigs approach.

In addition, this project has been developed under the group “Bioinformatics and Information Technologies Laboratory” (BITLAB), part of the Departamento de Arquitectura de Computadores, Universidad de Málaga and presented in the 6th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018).

CHAPTER 1.
INTRODUCTION

1.1 Motivation

The deoxyribonucleic acid, or DNA, is the hereditary material in almost all organisms[1], and it stores information as a code made up of four chemical bases (nucleobases): adenine (A), cytosine (C), guanine (G), and thymine (T). The order of this bases determines the information available for building and maintaining an organism. Genomics is the interdisciplinary field of science that studies whole genomes of organisms, and a lot of the experiments in this field begin by determining the content and order of such nucleobases from the genetic material of an organism, also known as DNA sequencing. This procedure can be performed by different methods in which each one portrays its advantages and disadvantages. Nonetheless, the most typical sequencing experiments consist on fragmenting the genome into smaller molecules known as reads. A set of overlapping reads is referred to as a contig (See Figure 1). The first sequencing technologies were developed in 1977 by Sanger et al. [2] from Cambridge University awarded a Nobel Prize in chemistry 1980. These were very expensive and time consuming, but their discovery opened the door to study the genetic code and inspired researchers to develop faster and more efficient sequencing technologies.

Figure 1. From sequencing reads to contigs.

Recently, a drastical reduction of time and cost per sequencing experiment has taken place, dropping from 10,000$ at the beginning of this century down to a few cents in less than 20 years, due to major breakthroughs in sequencing technologies that have occurred in the last

decades [3]. These techniques produce a huge amount of data overcoming the main barrier during the early Genomic Era which was the data generation problem. Biologists now face a torrent of data that has paved the way towards the analysis of numerous unknown biological communities and the research of pioneering scientific areas such as metagenomics (beyond genomes).

The goal of metagenomics is to study microbial communities, also known as microbiotas, in their natural environment, without requiring to aisle and cultivate the species that make up such community. This field brings a profound transformation in multiple fields, such as: biology, medicine, ecology, agriculture, and biotechnology [4]. Despite these benefits, metagenomic sequence data presents several challenges. For instance, most communities are so diverse that most genomes are not utterly represented by reads. The difficulty of performing direct comparisons through sequence alignment is even greater due to distinct reads from the same gene that may not overlap. However, when they do overlap it is not always noticeable whether they are from the same or different genomes, making the the sequence assembly much more challenging. Additionally, its bioinformatic analysis is more complicated when dealing with poor quality reads, detecting repeated sequences from similar organisms, and genomic variants or species that have not yet been sequenced within a sample in which the representation of organisms is uneven and unidentified [5].

A primary objective in metagenomics is portray the organisms present in an environmental sample, known as a taxonomic assignment. A correct classification of the species within a sample enables a further insight about several issues such as: the microbial ecosystems models used to describe and predict community-based microbial processes, changes, and sustainability; the global scale descriptions of the role of the human microbiome in different health states in individuals and populations; and the exploitation of the remarkably versatile and diverse biosynthetic capacities of microbial communities to generate beneficial industrial, health, and food products.

Tools such as MEGAN [6], FANTOM [7], MG-RAST [8] or META-GECKO [9] perform a taxonomic analysis with reads and are also prepared to work with contigs, since each approach has advantages and disadvantages. Analyzing contigs provide larger genomic

fragments, nevertheless this entails a risk of generating chimeric contigs due to the heterogeneity of the sample. On the other hand, with reads this risk is non-existent, however the analysis is affected by several factors such as the quality and length of the sequences, thus may generate matches with low statistical significance. Moreover, there can be almost identical reads that belong to similar organisms within the sample that make it almost impossible to know the origin of such sequence. Nevertheless, overlapping these reads together into contigs may provide helpful insight about the metagenomic sample.

The main contributions of this project are a set of tools that performs a metagenomic taxonomic analysis, then evaluates the quality of the taxa assigned to the metagenomic sample. Afterwards, it establishes statistical differences between reads and contigs in order to provide a better judgement to properly identify the correct taxa distribution in a metagenomic sample. Additionally, it provides a workflow that employs the previous tools to propose suggestions on how to perform an optimal taxonomic analysis of a metagenomic sample, either with reads or with contigs

1.2 Objectives

The main goals of this project are the following:

- Determine the best data-approach to perform a metagenomic taxonomic analysis, with reads or contigs:
    - We will assess and compare the quality of each of these alternatives in order to consolidate an appropriate, standard procedure to obtain the best possible results when these analysis are carried out.
    - Apply the scientific method with two use cases in order to validate the comparison results.
- Design a workflow in order to:
    - Generates synthetic datasets of metagenomic reads in which the abundance of species is known.
    - Assemble the generated reads into contig.
    - Map using the reads and using the contigs against the same reference database.

- Perform a taxonomic analysis for each approach.
- Measure the quality of each approach with a set of implemented tools in order to obtain valid comparisons.

CHAPTER 2.
STATE OF THE ART

In this chapter we will discuss two most common approaches when performing a taxonomic analysis in metagenomics, with reads and with contigs. This will be followed by a section were we examine the main challenges for these alternatives. Finally we will talk about the most common metagenome analysis packages. However, before we start introducing procedures, concepts and software, we should briefly address what is a metagenome and the goal of a metagenomic taxonomic analysis.

Metagenomic differs from traditional microbiology because a metagenome is an uncultured sample directly recovered from its original environment, meaning that the sample is not cultivated in a laboratory and there is no need to design specific primers as in traditional microbiology. From a scientist's point of view, a metagenome might be a collection of unknown species that interact in some way that it is interesting to research. In this sense, to determine the organisms present in an environmental sample is known as a metagenomic taxonomic analysis. While the goal is the same, there are different approaches to perform it. In the following section we describe the two most typical ones.

2.1 Metagenomic Taxonomic Analysis Approaches

The first step when performing a taxonomic analysis with reads is to obtain the data from a sequencing experiment from a metagenomic sample. The following sections detail the two compared approaches in this project.

2.1.1 Reads Approach

The reads are mapped against a reference database of a collection of genomes. Afterwards, a taxonomic rank is specified and the taxonomic assignment is performed with such mapped reads. This generates a report of the species present in such sample (See Figure 1).

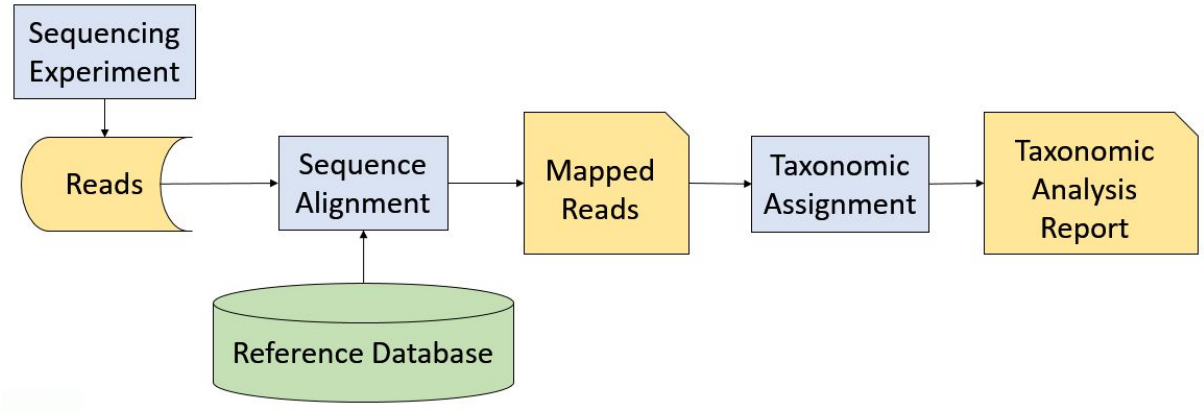

Figure 2. Workflow of the reads approach when performing a metagenomic taxonomic analysis.

2.1.2 Contigs Approach

This approach is very similar to the one with reads, yet it requires an assembly prior to the alignment. After the reads are generated from a sequencing experiment, they are assembled into contigs. Afterwards, the sequence alignment against a reference database, of genomes from different species or stains, is performed and, lastly, the taxonomic assignment is executed to obtain the taxonomic analysis report (See Figure 2).

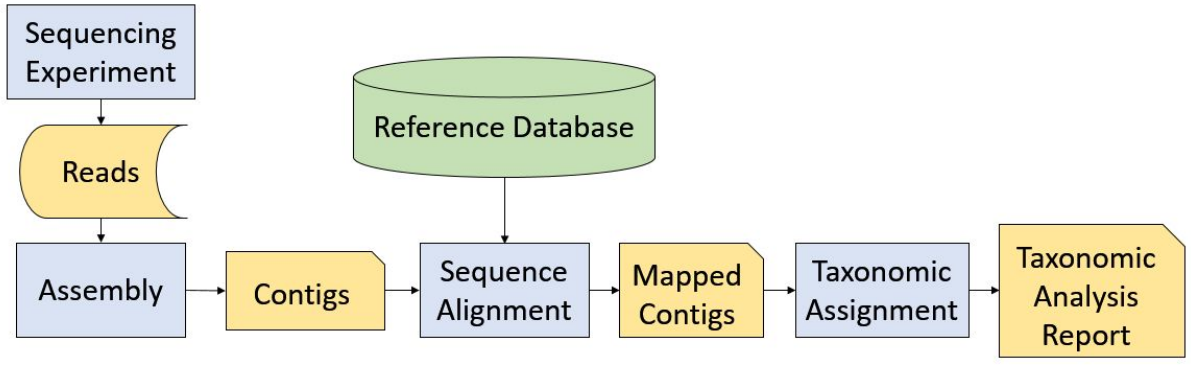

Figure 3. Workflow of the reads approach when performing a metagenomic taxonomic analysis.

2.2 Challenges in Metagenomic Taxonomic Analysis

Independent to the approach, there are issues that arise in metagenomics that do not come up in traditional genomics. For instance, a metagenomic sample will present an uneven and unknown distribution of species. This implies that it is not possible to measure the accuracy of a taxonomic assignment from a real metagenomic sample since the relative abundance of species is uncharted. Another issue is that most communities are very diverse, therefore most genomes are not completely represented by the reads.

There is also the noise that can be generated during the sequencing experiment due to artifacts, bad quality reads or sequencing errors. The informatic analysis is much more complex when dealing with repeated sequences from similar organism; and detecting genomic variants or species that have not yet been sequenced within a sample.

Additionally, assembly errors must be taken into account when performing the contigs approach. For example, distinct reads that belong to the same gene or genome may not overlap, and if they do it is not always noticeable whether they are from the same or different genomes (See Figure 3). There is also the possibility of generating contigs from overlapping inter-species reads, also known as chimeric contigs (See Figure 4).

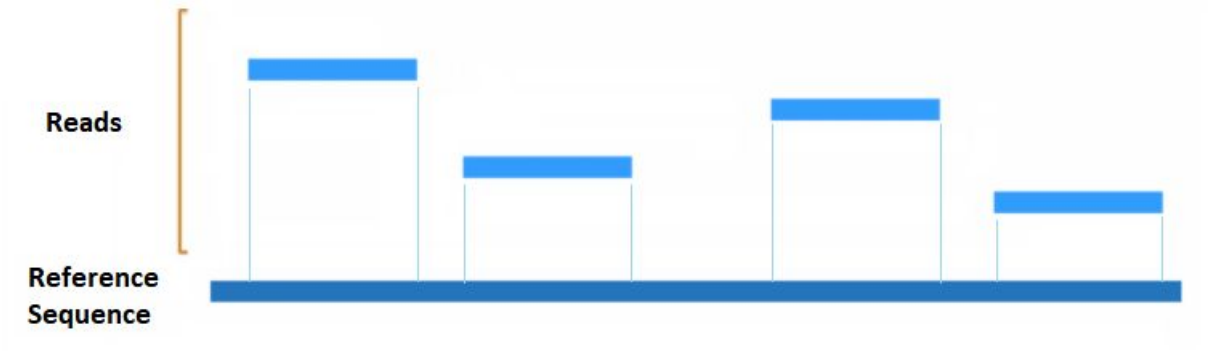

Figure 4. Reads from the same gene that do not overlap.

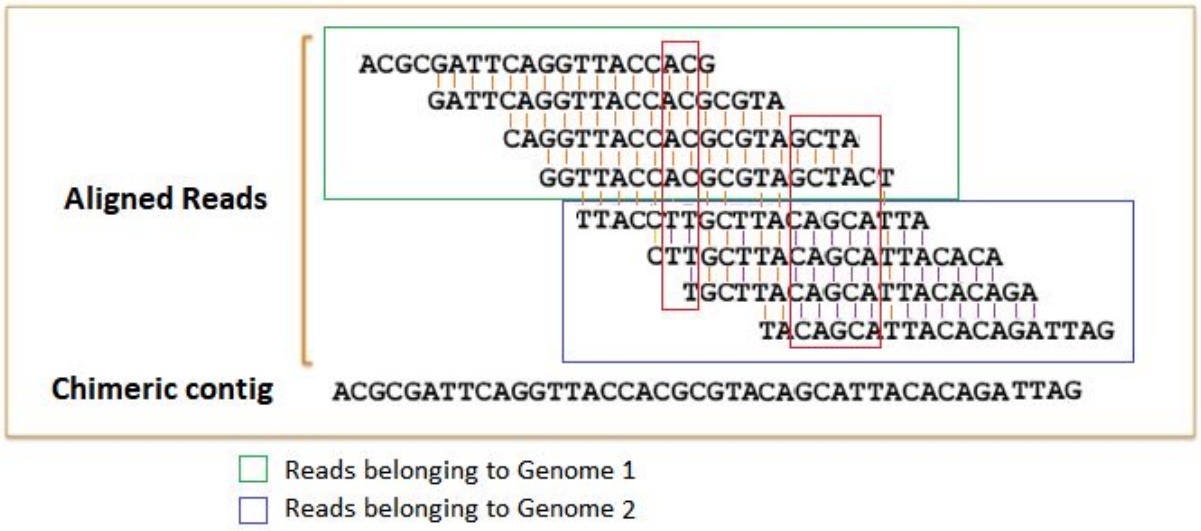

Figure 5. Reads from different genomes assembled into a chimeric contig.

2.3 Metagenome Analysis Packages

The interpretation of metagenomics data is important for understanding the ecosystem functioning and assessing differences between different environmental samples. The following section present some of most popular tools used to explore metagenomic data in taxonomic and functional context.

2.3.1 MEGAN

MEGAN (MEtagGenome ANalyzer) is a very easy to use, comprehensive microbiome analysis tool. It can be applied to analyze metagenomic (DNA), metatranscriptomic (RNA), peptide sequences and amplicon data (16S rRNA). The installation is very simple and straightforward. The Community Edition of MEGAN is free software that contains all features required to perform analysis of microbiome samples. The Ultimate Edition is built on top of the free edition, however it provides extra features, a command-line interface, and a set of command-line tools to customize classification schemes and mapping files used for the program. The community webpage is very active and provides support for both editions.

The aim of MEGAN is to provide a tool for studying the taxonomic content of a set of DNA reads, generally from a metagenomic project. As a preprocessing step, a sequence

alignment of all reads against a reference database is required to produce an input file for the program. This software facilitates an interactive exploration of the NCBI taxonomy which consists of over one million taxa.

The main application of the program is to parse and analyze the results of an alignment of a set of nucleotide sequences against one or more reference databases. The typical programs for this alignment are BLASTN [10], BLASTX [11] or similar tools such as DIAMOND [12] to compare against genome specific databases. The results of such analysis is a taxonomic profiling of the sample from which the sequences were collected. MEGAN provides different algorithms to assign each sequence a taxon on some level in the NCBI [13] hierarchy, based on their hits to known sequences recorded in the alignment file.

This software also provides a functional analysis using a number of different classification systems, but for the Community Edition only an early 2011 version of KEGG [14] is available. The Ultimate Edition contains an up-to-date version of KEGG.

2.3.2 FANTOM

FANTOM (Functional ANd Taxonomic analysis Of Metagenomes) is a software for the analysis of quantitative metagenomics data. This tool allows for an exploratory and comparative analysis of metagenomics data integrated with metadata information and biological databases. The software is implemented in Python, therefore is platform independent.

2.3.3 MG-RAST

MG-RAST (MetaGenomic Rapid Annotation using Subsystems Technology) is an automated platform that has served as a public resource of annotation and analysis of metagenomic sequence data, providing a repository for over 150,000 datasets (over 60 tera-base-pairs) with more than 23,000 publicly available.

This server allows users to upload raw metagenomic sequence data in FASTQ or FASTA format. Assessments of sequence quality and annotation with respect to multiple reference databases are performed automatically with minimal input from the user. Post-annotation analysis and visualization are also possible directly through the web interface or with R packages that utilize the MG-RAST API to easily download data from any stage in the MG-RAST processing pipeline. This tool provides support for shotgun and amplicon metagenomic samples, as well as metatranscriptomes.

2.3.4 META-GECKO

A software framework developed by Perez-Wohlfeil et al. that provides different mapping alternatives against reference databases, mapping reads over unannotated regions of genomes. Moreover, it provides evidence of the species present in metagenomics by mapping reads to specific regions of genomes. In addition, this workflow is an open platform composed of an expandable set of separate modules, which enables an easy incorporation of new processing tools.

CHAPTER 3.
ANALYSIS AND DESIGN

In this section we will explain the the requisites that should be accomplished in order that the tools used and developed for this project work properly. Moreover, the developed software tools will be briefly described.

3.1 Analysis of general requisites

In this section, the specifications regarding the software used in this project are described:

1. Platform:
    a. This project was designed under the Ubuntu 16.04.4 LTS, however it is supported in other UNIX environments.
2. Base requirements:
    a. UNIX system shell
    b. Python 3.5.2
    c. GNU Compiler Collection (gcc)
    d. R 3.3.4
    e. pdflatex (For PDF report)
3. Third-party software:
    a. Grinder [15] (Version 0.5.4)
    b. MEGAHIT [16] (Version v1.0.5)
    c. MEGANv6 (Version 6.10.13)
    d. BLASTN (Version 2.7.1)

A detailed explanation of the procedure to install the required third-party software is available in the github (https://github.com/pabrodbra/RACKit/) repository of this project.

3.2 Workflow design

The workflow has been designed with the intent of analyzing the levels of concordance between the reads and the contigs they assemble and retrieve reliable comparison results (concordance levels and comparison metrics are detailed in the next chapter). One of the requirements to establish a valid comparison is to know beforehand the relative abundance of species in a sample. This is achieved by selecting a set of genomes and the abundance distribution so that Grinder creates a synthetic reads dataset.

Such dataset is generated in the FASTQ format, therefore it is pipelined to a BASH script that converts this file into a FASTA format. Once formatted, the produced multi-fasta file of synthetic reads are assembled into contigs using MEGAHIT, an assembler developed for large and complex metagenomic Next Generation Sequencing (NGS) reads. Afterwards, both sets of sequences (reads and contigs) are mapped against a reference database to acquire the possible species that each sequence came from. These results are then fed to MEGAN, a microbiome analysis tool that applies the Last Common Ancestor (LCA) algorithm to assign each sequence to a taxa. Finally, the retrieved information from previous steps is processed by the developed toolkit in order to generate a set of results that assesses the quality of the taxa assigned to the reads and contigs and provides statistical insight about such results (See figure 6).

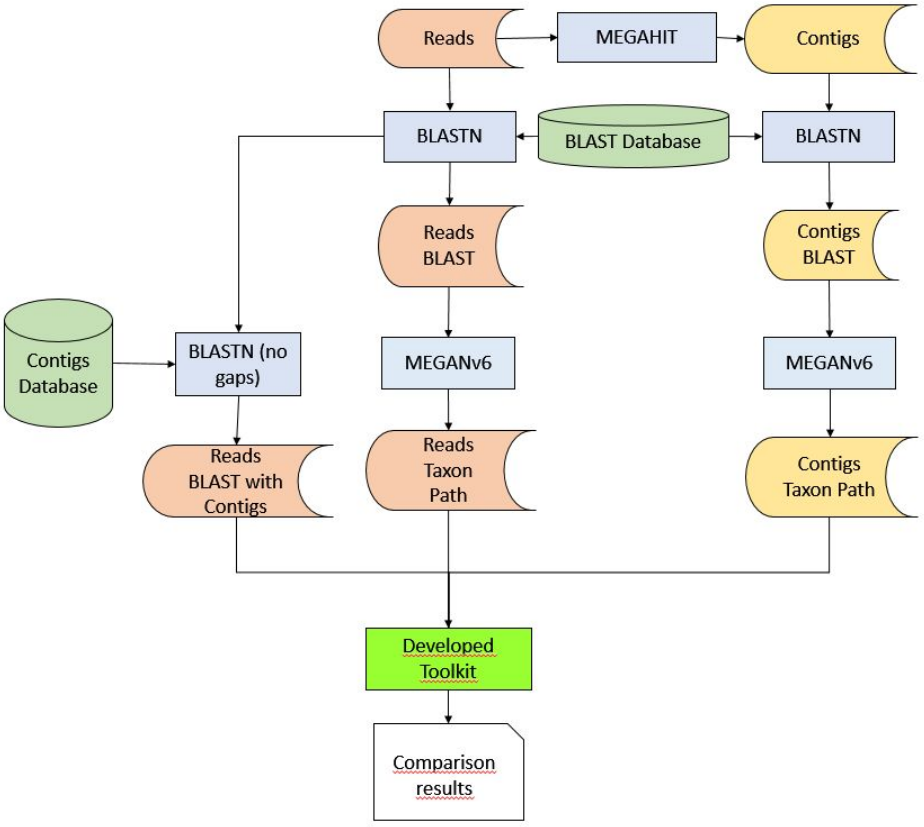

Figure 6. Read assembly and Taxonomic Analysis Comparison workflow. Red: File Generated from Reads. Yellow: File generated from Contigs. Blue: Third-Party software. Light green: Proprietary software. Green: Databases.

3.3 Developed software tools and scripts

The developed proprietary software or scripts are listed below accompanied by a brief explanation.

- **Data preprocessing:**
  - **FASTQ to FASTA converter:** A script programmed in BASH that receives a FASTQ file as input, converts it to a FASTA format and outputs such file. FASTQ and FASTA are two different files to store sequencing information. The main difference is FASTQ contains additional information on the quality of each base (nucleotide) and the certainty of the lecture.

- **BLAST Result Parser:** The developed workflow interconnects several modules that extract different information from BLAST results. Because of this we have designed a pipeline composed of a small program developed in C and some BASH scripts that requires as input the result of a BLAST comparison in its typical format and parses its information into a tab delimited file to reduce its size and facilitate its processing for the other tools.

- **Analysis Tools:**
  - **Reads Against Contigs Python Toolkit:** These are a set of tools developed in Python that requires different inputs from other tools applied in this workflow in order to qualify the concordance levels and quantify the quality of the taxonomic assignments from two different approaches. This is performed by pipelining the tools as described below:
    - One tool gathers the reads that were assembled into each contig from a parsed BLAST result of the reads against the nucleotide sequence of contigs, used as a reference database. This generates two dictionaries: one that associates the ID of the reads to the ID of the assembled contig; and another one that correlates the ID of the contig with all the reads that assembled it.
    - The second software first loads the taxonomic assignments of the reads approach and contigs approach. Then, it generates the dictionary of contigs associated to the read that assembles it. Afterwards, it retrieves the inconsistencies between the assignment of each of the relationships in the dictionary based on a specified taxonomic rank. This tool outputs the ID of the sequences for each of the inconsistencies found, classifies them (this classification is described in Chapter 4), and specifies the taxonomic rank in which these inconsistencies are resolved.
    - This software generates a confusion matrix for each genome in the reference database (multiple 1 vs All comparisons) and populates it from the Parsed Blast Result generated from the toolkit. Afterwards it calculates different statistical measurements such as accuracy, sensitivity, specificity, precision and fallout (detailed description in Chapter 4).

- **UniseqDBCoverage:** This software developed in C, receives the reference database and the primary sequence alignment results. Afterwards, from the best matches in the mapping, the width coverage of the database mapping and the average of top scoring matches is calculated (detailed description in the Chapter 4).
- **Reads Against Contigs R Script:** This R script retrieves all the results from the previously described developed tools and the MEGAN taxonomic analysis results, in order to generate a report with the plots that can be analyzed by the researcher to compare different approaches.

CHAPTER 4.
METHODS AND IMPLEMENTATION

The definitions, procedures and algorithms employed to compare reads and contigs when analyzing a metagenomic sample are describe in this section. To achieve a reasonable comparison we have defined a set of conditions that describe the taxonomic concordance on a specific taxonomic rank between each read and the contig (handled as one sequence) it assembles.

4.1 General Definitions

Let R be the set composed by the Reads. Let C be the set composed by the Contigs. Each Read can only be assembled into one Contig.

$$R = \{r_1, r_2, \ldots, r_n\} \wedge |C| < |R|$$

$$r_i \in R \wedge c_i \in C$$

$$c_i \subseteq R \wedge c_1 \cap c_2 \cap \ldots \cap c_n = \emptyset$$

Let S be the set of composed Reads and Contigs sequences. Let T be the set composed by the Taxa in a taxonomic rank and None.

$$S = \{R, C\} \wedge s \in S \wedge t \in T$$

$$Taxon\,(s) \rightarrow t$$

In order to detect chimeric contigs, we have defined the following levels of concordance for a contig and the read that assembles to classify them:

- **Consistency (C):** Both, read and contig, have the same taxon assigned or were not assigned at all.

$$Taxon\,(Read) = Taxon\,(Contig)$$

- **Weak Inconsistency (WI):** Either the read or the contig has been assigned to a taxon while the other one was not assigned to any. These relationships are classified based on which sequence was unassigned. It will be a **Weak Inconsistency by Read** (WIR) granted that the read does not match to a taxon in a specific taxonomic rank. However,

it will be classified as a **Weak Inconsistency by Contig** (WIC) if the contig was the unassigned sequence.

$$Taxon\,(Read) = None \wedge Taxon\,(Contig) \neq None$$

or

$$Taxon\,(Read) = X \wedge Taxon\,(Contig) \neq None$$

- **Strong Inconsistency (SI):** Both sequences, read and contig, are assigned to a taxon in the same taxonomic rank, but to different taxa. In the case that either the read or the contig is not assigned it will be classified as a WI.

$$Taxon\,(Read) \neq Taxon\,(Contig) \neq None$$

Having settled the previous definitions, the subsequent sections provide a detailed description of the internal functioning of the workflow.

4.2 Detecting differences between taxonomic analysis approaches: Reads and Contigs

The developed toolkit has been designed for comparing the results obtained after performing a primary sequence comparison and a biological taxonomic analysis between reads and contigs. The output information provided by this tool is composed by:

- **The associations for each contig and the reads that assemble it:** The associations between the reads and contigs are extracted from best alignments of the BLASTN output obtained by performing a DNA primary sequence alignment between them. Other comparison tools can be used by adding an specific parser. This result is processed to obtain two collections of the relationships between the reads and the contigs: one in which the reads are assigned to the contig that it assembles; the other where the contigs are partnered with the group of reads used to assemble it.
- **Concordance of the taxa assigned between the reads and the contig assembled:** Firstly, the identifier of all the sequences that have been assigned to a taxon in the selected biological classification rank are extracted from the MEGAN results. Afterwards, this information is used to classify the previously obtained associations between reads and contigs based on the concordance level of the taxon assigned to a contig and the reads that assembles it.

- **Coverage of the reference database:** The amount of base pairs that were aligned to the database obtained from the results after executing the BLASTN with each set of sequences is compared to the number of base pairs in such reference database to obtain the following metrics: total coverage of the database for each set of sequences; total coverage of the database that the reads and contig map together.
- **Ratios of highest scoring matching species per sequence in a metagenomic dataset:** The average of top scoring matches resulting from the sequence alignment against the reference database is calculated for each of the datasets. Afterwards, the calculated ratios are compared to decide which set of sequences provides less variable matches. An appropriate approach should report a lower ratio of top matches, yet a higher coverage of the reference database.
- **Confusion matrix and performance metrics from the taxonomic classification:** This measurement can only be calculated when the original genome from which each read was generated. For the contigs, it is impossible to know the original genome due to the possibility of chimeric contigs. Therefore we define the correct genome of each contig as the one to which the majority of the reads that assemble it belong to. For each genome in the reference database we calculate a confusion matrix (multiple One vs All binary classification). The numbers that populate each matrix are calculated from: the genome of the confusion matrix, the original genome of the sequence, and the genome such sequence mapped. The measured instances are the following:

  - True positives (TP): The current confusion matrix belongs to the genome that the sequence belongs to. The sequence mapped to its original genome.

  - **False positives (FP):** The current confusion matrix belongs to the genome that is not the one that the sequence belongs to. The sequence mapped to such genome.

  - False negatives (FN): The current confusion matrix belongs to the genome that the sequence belongs to, however such sequence did not map to its original genome.

  - True negatives (TN): The current confusion matrix belongs to the genome that is not the one that the sequence belongs to. The sequence did not match to the current genome.

  After populating each confusion matrix, these are averaged together to obtain a final confusion matrix results. From this results, the following statistical performance metrics are calculated:

- Accuracy (ACC): Ability to properly differentiate between the correct mappings.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

- Sensitivity (TPR): Ability to determine a sequence only maps to its original genome.

$$TPR = \frac{TP}{TP+FN}$$

- Specificity (TNR): Ability to ascertain that a sequence does not map to a genome that it does not belong to.

$$TNR = \frac{TN}{TN+FP}$$

- Precision (PPV): Probability that a sequence to maps to the correct genome.

$$PPV = \frac{TP}{TP+FP}$$

- Fallout (FPR): Probability that a sequence matches to an inappropriate genome.

$$FPR = \frac{FP}{TN+FP}$$

CHAPTER 5.
RESULTS AND DISCUSSION

Two use cases have been design with the intent of applying the described workflows and obtain valid comparison results. In both, the metagenomic reads dataset must fulfill the condition of knowing beforehand the origin of each read because this enables us to assess the quality of the taxonomic assignment and to establish whether it is better to perform a taxonomic analysis with reads or with contigs.

The two use cases are the detailed in the section below (See Figure 7). The relative abundance of species for both metagenomic datasets is represented in the Figure 8:

- **Fully synthetic dataset/use case (FSD):** The gastrointestinal tract genomes provided by the Human Microbiome Project (HMP) [17] were fed to the synthetic data generation workflow. An equitably number of reads are generated from each genome in such database in order to obtain a mixed sample of reads from different species. The total number of reads is 521,334 with and average length of 391, from which the length of the 97.42% is over 300 nucleotides. These reads represent a 7,27% of the nucleobases from HMP database.
- **Semi-synthetic dataset/use case (SSD):** After analyzing the study *"Comparative metagenomic, phylogenetic and physiological analyses of microbial communities across nitrogen gradients"* [18], a set of genomes that represent each of the classes were selected. This selection of genomes were fed to the synthetic data generation workflow to generate a set of reads proportional to the class relative abundance specified in such article. The remaining percentage of the metagenomic sample (9%) was obtained by generating a set of random reads that followed the nucleotide distribution from the rest of the dataset. In order to provide a soil sequencing framework, these genomes were selected from the soil microbial genomes in the RefSoil [19] database. The total number of reads is 499,991 with an average length of 250, from which a 100% of them have a length of over 200 nucleobases. These reads represent a 2,89% of the reference database.

The Figure 8 represents the original relative abundance of both datasets, presented in a logarithmic scale from the most abundant specie to the least. For the FSD we can observe the 221 species present in the HMP dataset and that, although the number of reads for each sequence was uniformly distributed, there are some species with a higher percentage of sequences. This phenomenon happens because some of the species present in the HMP database have multiple strains sequenced, therefore more reads will be generated for such species. The SSD follows the relative abundance from real metagenomic samples as detailed in the referenced paper and represents 21 species.

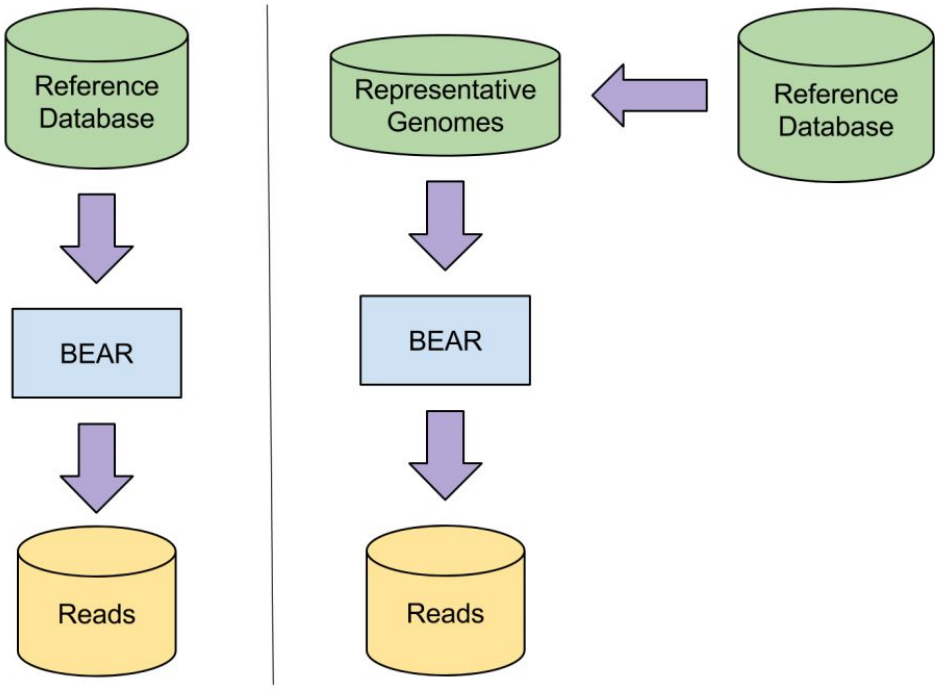

Figure 7. On the left: Generation of fully synthetic reads. On the right: Generation of semi synthetic reads.

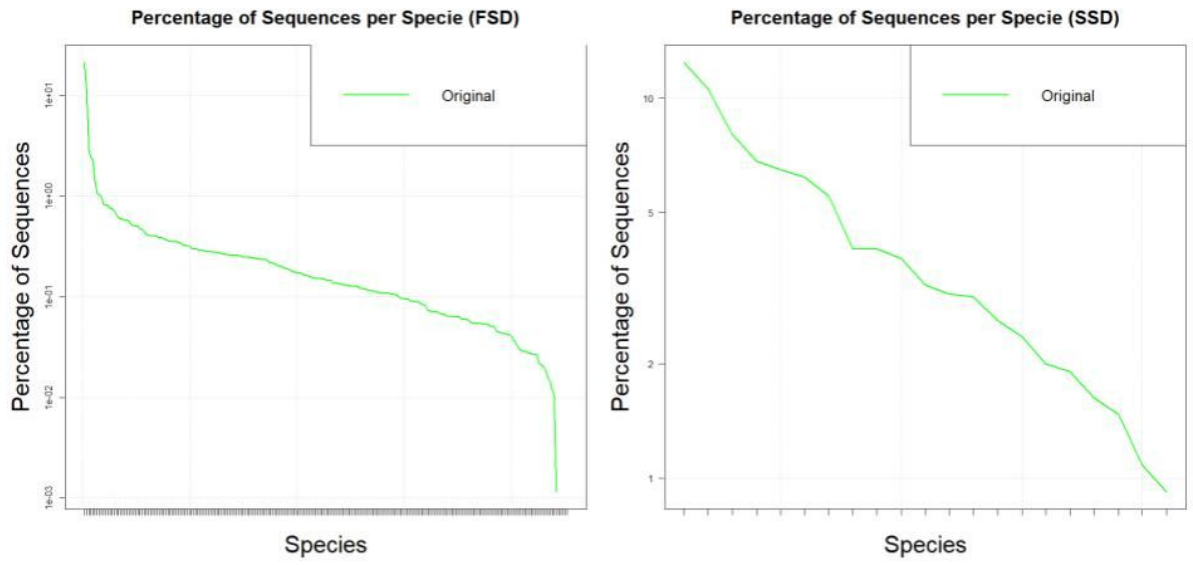

Figure 8. Relative abundance of species for the: (left) fully synthetic dataset (FSD) and (right) semi synthetic dataset (SSD).

The workflow, depicted in the Methods and Implementation section has been applied to each of the use cases. For each use case, the generated output from the developed tools are interpreted to obtain the following results:

5.1 Comparison with the Original Relative Abundance of Species

The relative abundance of species obtained by performing a taxonomic assignment with the reads and contigs is compared with the original dataset in Figure 9 in a logarithmic scale for the percentages. For the FSD, both reads and contigs seem to have differences when compared to the original dataset. However, it is not noticeable which one is more similar to the authentic dataset. This is not the case for the SSD since the reads present and almost identical relative abundance of species in comparison to the original, while on the other hand the contigs clearly have noticeable differences.

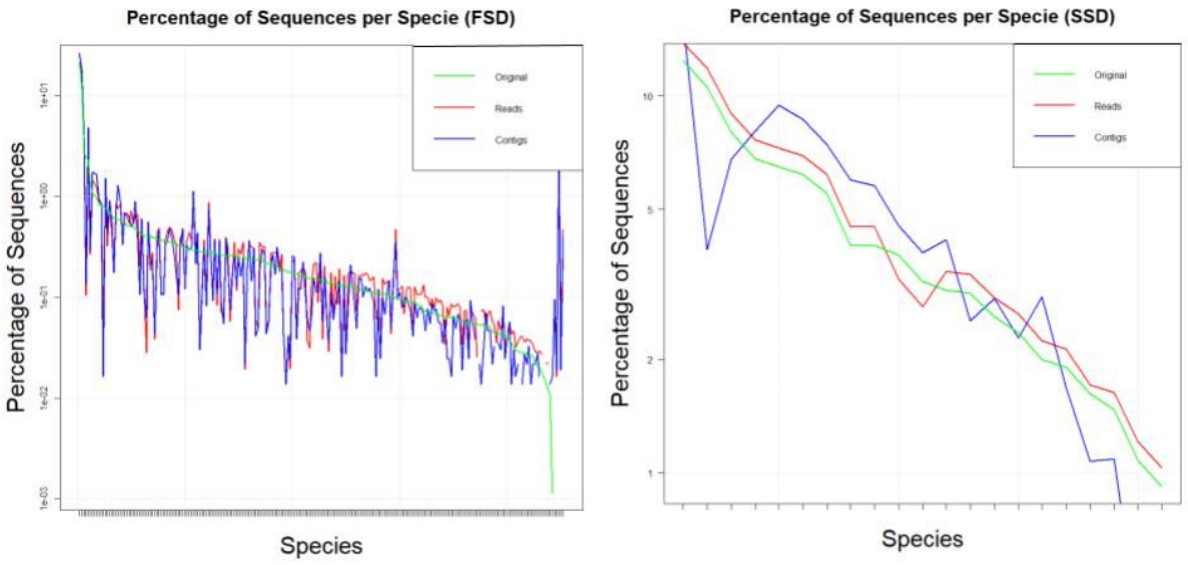

Figure 10. Relative abundance of species in a metagenomic original dataset (green), reads (red) and contigs (blue) for the: (left) fully synthetic dataset and (right) semi synthetic dataset.

The plots from Figure 10 are aesthetically pleasing and describe the results of the taxonomic analysis for each approach. However, we can observe that for the FSD, the numerous amount of species complicate the analysis of these results. However, for the SSD, the relative abundance for the reads is clearly much more similar to the original one than the contigs. we must calculate a measurement that allows us to establish the difference between the relative abundance estimated by each approach to properly compare them. To achieve that, we calculate the Root Mean Square Error as described in the following section.

5.2 Root Mean Square Error (RMSE) after the Taxonomical Analysis

The RMSE is calculated for both reads and contigs approach using the original dataset as reference. A lower RMSE implies that the taxonomic analysis obtained from such approach, describes with more precision the ideal abundance of species in the metagenomic sample (Table 1).

Dataset	RMSE for FSD	RMSE for SSD
Reads	0.3187	0.4031

Contigs	0.3858	4.2534

Table 1. Root Mean Squared Error of the assignment of species for reads and contigs compared to the original dataset for both use cases.

From Table 1, we can observe that the reads provide a lower RMSE than the contigs in both use cases. From these results we can clear up the confusion about which approach provides more insight about the proper taxonomic assignment of species. In Figure 10 it was very hard to depict which approach was better for the FSD, however the RMSE suggest that the reads provide a better taxonomic assignment than the contigs. Moreover, this measurement confirm that the reads provide a more precise report of the species within the SSD.

5.3 Inconsistencies Found

A concordance level is established to each of the associations between each read and the contig it assembles. Identifying the types of of inconsistencies aids us at the moment of determining the reason behind the RMSE. If there are more weak inconsistencies at the species taxonomic rank, then most of the reads or contigs involved were assigned to a taxon in a higher and less specific taxonomic rank. the detected inconsistencies and the percentage of relationships they represent are shown in the Table 2.

Type of Inconsistency	**Found on FSD (%)**	**Found on SSD (%)**
Weak Inconsistency by Read	21,393 (4.10)	4,003 (0.80)
Weak Inconsistency by Contig	24,183 (4.64)	1,622 (0.32)
Hard Inconsistency	4,464 (0.84)	2,231 (0.45)

Table 2. Number of inconsistencies found for each use case at the species taxonomic rank.

5.4 Inconsistency Resolution

The previously found inconsistencies can always be solved by selecting a higher taxonomic rank, since it covers a broader range of taxa that a sequence can be assigned. For both use cases, the sequences belong to bacterias, therefore the discrepancy between the assignment of a contig and the read that assembles it will always be sorted out in the taxonomic rank "Domain". This can be appreciated in Figure 10.

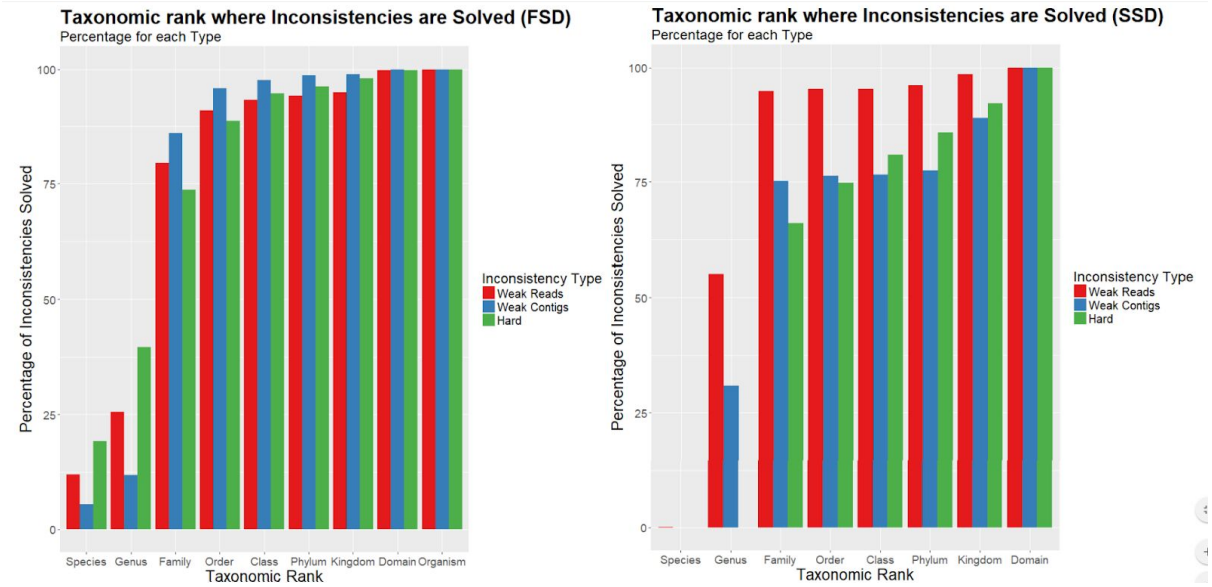

Figure 10. Percentage of inconsistencies solved at different taxonomic ranks. In both use cases, over 50% of the inconsistencies are resolved if the desired taxonomic group to analyze is the family. On the left: inconsistency resolution for the fully synthetic dataset. On the right: inconsistency resolution for the semi synthetic dataset

The heterogeneity of the samples make it so a noticeable amount of the contigs are chimeric. This confirms that the inconsistencies arise due to the intrinsic difficulty of the assembly process. These results suggest that the reads associated to a SI are used to assemble chimeric contigs. Moreover, we can observe that these chimeric contigs are more often generated from reads from reads that belong to different organisms within the same family taxonomy (over 50% of the SI).

5.5 Coverage and Mapping Comparison against the Reference Database

For each use case, the ratio of top scoring matches after performing the primary sequence alignment against the reference database and the percentage of nucleotides covered by the full set of sequences is described in the Table 3.

Measurement	**FSD**		**SSD**	
	Reads	Contigs	Reads	Contigs
Ratio of Matches per Sequence	7.05	7.50	4.52	8.38
% Coverage of Database	21.21	7.16	5.59	3.37
Common % within Use Case	6.42		3.03	
Common Coverage % against Contigs	89.66	Not Applicable	89.91	Not Applicable

Table 3. Mapping and coverage comparison between reads and contigs for each use case.

In both of the use cases, the reads obtain a lower average of top scoring matches than the contigs. This tends to happen due to the assembly noise generated by forming contigs from reads that belong to different species. Moreover, it is noteworthy that over 85% of the nucleotides covered by contigs are also covered by reads, yet reads cover a wider range of the database. This means that reads provide more information that may be of interest depending on the goal of the metagenomic experiment.

5.6 Confusion Matrices and Performance Metrics based on the Correct Assessment of a Taxon for each Sequence

Both use cases fulfil the prerequisite to calculate this measurements, which is to know the genome to which each read was originated from. The results of the statistical performance matrices are described in the Table 4. The best approach is the one that provides: a higher accuracy, sensitivity, specificity and precision; and a lower fallout.

Dataset		**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Fallout**
SSD	Reads	19.04 %	1.06 %	99.49 %	90.26 %	0.51 %
	Contigs	12.68 %	0.67 %	99.20 %	85.44 %	0.80 %
FSD	Reads	13,43 %	$2{,}29 \times 10^{-6}$ %	99,99 %	13,43 %	$9{,}52 \times 10^{-5}$ %
	Contigs	8,81 %	$1{,}42 \times 10^{-6}$ %	99,98 %	8,81 %	$15{,}28 \times 10^{-5}$ %

Table 4. Confusion matrices for species average and performance metrics

A sequence can map to multiple genomes from the reference database with the same identity, similarity, length and e-value. This occurs because the reads could have originated of a region which is very similar within different genomes. For instance, different strains of the same species have an almost identical genome. Moreover, if the read was originated from an orthologous gene region, different species with the same common ancestor may share that nucleotide sequence. Because of this fact, the sensitivity and specificity are very extreme.

The explanation for these values is that, for each confusion matrix (one for each genome in the reference database), one sequence can only map to to the original genome (TP) once, however multiple matches for one sequence cause multiple FP. Furthermore, the number of TN becomes extremely high in comparison to the other due to the fact that one sequence be considered as a TN for each sequence in the reference database that it does not match to, as long as it does not belong from it. Even with very restrictive coverage, similarity and e-value thresholds, almost all the sequences match to a genome, therefore the number of FN is extremely low.

Almost all the sequences are mapped at least once, and since every map is considered a positive, we suggest that the most reliable measurement are the one that evaluate the positive rate, such as precision (true positive rate) and fallout (false positive rate). As observed from the obtained results, the reads approach obtain better measurements in comparison to the contigs approach.

CHAPTER 6.
CONCLUSIONS

A metagenomic taxonomic assignment aims to determine the composition of organism present in an uncultured biological sample taken directly from its original environment. The analysis of metagenomes present more challenges than traditional genomics, such as: the uneven and unknown abundance of species, and the fact that not all the species will be completely represented by the reads from the sequencing experiment. Hence, it requires more accurate, refined and computationally expensive methods. However, it provides an in-depth and unbiased method of obtaining genomic information, whereas traditional microbiology presents and inherent bias since culture methods can only confirm the presence of microorganisms that can grow on the selected media.

As it was mentioned, the main goal of this project is to determine which approach was more appropriate when performing a metagenomic taxonomic analysis, using either reads or using contigs. In order to do so, we first have designed, implemented and applied a workflow (RACKit) to obtain and validate the results by applying the scientific method. Such workflow was developed to: generate synthetic datasets composed of a user specified abundance of species; assemble them into contigs; map using such reads and contigs against the same reference database; perform a taxonomic analysis applying a last common ancestor algorithm for each approach; and calculate several indicators from the previous steps which enable a valid comparison between both alternatives.

The developed workflow RACKit was executed for two different use cases (fully synthetic and semi-synthetic datasets). Both analysis suggest that the reads approach provide a more precise assignment of taxa and a relative abundance of species resembles to a larger extent to the one that belongs to the original metagenomic sample than using the contigs approach. Such outcome suggests that the best data-approach to obtain a more accurate metagenomic taxonomic analysis are obtained with the reads approach.

These results conjecture that the contigs approach presents challenges during the assembly process due to several reasons. To begin with, the quality of the assembly will vary strongly on the length and quality of the reads. Another issue is that the number of contigs will vary based on their length and how many reads are used to assemble such contig. Likewise, a noticeable amount of reads that belong to different species are put together into chimeric contigs as a result of the great heterogeneity of species in a metagenomic sample. At the moment of assigning a taxon to each sequence, these previously mentioned issues have a negative impact for the contig approach because each contig is handled as one sequence although it was formed by many reads, misrepresenting the original sample.

In conclusion, we expect that the existing modern tools and algorithms to solve problems in metagenomics will provide support to researchers working in the metagenomics field. However, these tools present shortcomings which are difficult to solve due to the intrinsic complexity of analyzing a metagenomic sample. Therefore, it is pertinent to properly identify and address such drawbacks to develop upgraded tools in the future to obtain a better understanding about the contributions of microbiotas to the health of the planet, their roles in human health, and the consequences of human activities towards the biosphere. Accordingly, the results obtained in this project suggest that the metagenomic assembly is a very challenging process caused by several issues that arise in this field, and that the reads approach provides further understanding of a metagenomic sample than the contigs approach in this current day and age.

In addition, this study has been presented at the 6th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018)[20] and has been selected to be extended and submitted to BMC Bioinformatics, a high impact factor journal, as a research article. It is currently under inspection at the time of writing these conclusions.

6.1 Conclusiones

Un análisis taxonómico en metagenómica tiene como objetivo determinar la composición de los organismos presentes en una muestra biológica no cultivada tomada directamente de su medio natural. El análisis de los metagenomas presenta más desafíos que

la genómica tradicional, tales como: la abundancia desconocida y desigual de especies, y el hecho de que no todas las especies estarán completamente representadas por las lecturas del experimento de secuenciación. Por lo tanto, requiere métodos más precisos, refinados y computacionalmente costosos. Sin embargo, proporciona un método profundo e imparcial para obtener información genómica, mientras que la microbiología tradicional presenta un sesgo inherente, ya que los métodos de cultivo solo pueden confirmar la presencia de microorganismos que pueden crecer en los medios seleccionados.

Como se mencionó anteriormente, el objetivo principal de este proyecto es determinar qué enfoque es más apropiado cuando se realiza un análisis taxonómico en metagenómica, usando lecturas o usando contigs. Para hacerlo, primero hemos diseñado, implementado y aplicado un flujo de trabajo (RACKit) para obtener y validar los resultados aplicando el método científico. Tal flujo de trabajo se desarrolló para: generar conjuntos de datos sintéticos compuestos de una abundancia de especies especificada por el usuario; ensamblarlos en contigs; mapear usando tales lecturas y contigs contra la misma base de datos de referencia; realizar un análisis taxonómico aplicando un algoritmo del ancestro común más bajo para cada enfoque; y calcular varios indicadores de los pasos anteriores que permiten una comparación válida entre ambas alternativas.

El flujo de trabajo desarrollado RACKit se ejecutó para dos casos de uso diferentes (conjuntos de datos totalmente sintéticos y semisintéticos). Ambos análisis sugieren que el enfoque de lectura proporciona una asignación más precisa de los taxones y una abundancia relativa de especies se parece en mayor medida a la que pertenece a la muestra metagenómica original que utilizando el enfoque de contigs. Tal resultado sugiere que el mejor enfoque de datos para obtener un análisis taxonómico metagenómico más preciso se obtiene con el enfoque de lecturas.

Estos resultados conjeturan que el enfoque de contigs presenta desafíos durante el proceso de ensamblaje debido a varias razones. Para empezar, la calidad del conjunto variará considerablemente según la longitud y la calidad de las lecturas. Otro problema es que el número de contigs variará en función de su longitud y la cantidad de lecturas que se utilizan para ensamblar dicho contig. Del mismo modo, una notable cantidad de lecturas que

pertenecen a diferentes especies se juntan en contigs quiméricos como resultado de la gran heterogeneidad de especies en una muestra metagenómica. En el momento de asignar un taxón a cada secuencia, estos problemas mencionados anteriormente tienen un impacto negativo para el enfoque de contig porque cada contig se maneja como una secuencia, aunque se formó por muchas lecturas, tergiversando la representación de la muestra original.

En conclusión, esperamos que las herramientas y algoritmos modernos existentes para resolver problemas en metagenómica brinden apoyo a los investigadores que trabajan en el campo de la metagenómica. Sin embargo, estas herramientas presentan deficiencias que son difíciles de resolver debido a la complejidad intrínseca del análisis de una muestra metagenómica. Por lo tanto, es pertinente identificar y abordar adecuadamente tales inconvenientes para desarrollar herramientas mejoradas en el futuro a fin de comprender mejor las contribuciones de las microbiotas a la salud del planeta, sus funciones en la salud humana y las consecuencias de las actividades humanas hacia la biósfera. En consecuencia, los resultados obtenidos en este proyecto sugieren que, a día de hoy, el ensamblaje metagenómico es un proceso muy desafiante causado por varios problemas que surgen en este campo, y que el enfoque de lectura proporciona una mayor comprensión de una muestra metagenómica que el enfoque contigs.

Además, este estudio fue presentado en la 6ta Conferencia Internacional de Trabajo sobre Bioinformática e Ingeniería Biomédica (IWBBIO 2018) y ha sido seleccionado para ser extendido y presentado a BMC Bioinformatics, una revista de alto impacto, como artículo de investigación. Actualmente está bajo inspección en el momento de escribir estas conclusiones.

6.2 Ongoing work

In terms of future work, the toolkit is being applied to compare the quality of different metagenomic assembly tools and to compare the quality of the assembly using different parameters. Likewise, adjusting the presented workflow to compare the functional analysis between the reads and contigs approach would be very interesting to establish a proper methodology when analyzing metagenomic samples. Moreover, the comparison between the reads and contigs approach will be carried out with more use cases in order to establish a

more statistically significant comparison. In such comparisons, different parameters will be employed during the reads assembly, mapping against the reference database, and during the taxonomic assignment.

6.3 Acknowledgments

I wish to express my sincere gratitude to my advisor, Dr. Oswaldo Trelles, for having confidence in my skills and giving me the opportunity to be part of his research team. I also wish to thank all my colleagues at the Bitlab team, specially to Esteban Pérez for all the support and guidance.

Last but not least, I want to thank all my family for the unconditional support they have given me throughout the years, specially to my parents, Luis and Luisa, and sister, Lucía, for teaching me to always keep a positive mindset and that best things in life are the people you love, the places you've seen, and the memories you've made along the way.

CHAPTER 7.
BIBLIOGRAPHY

1. Genetics Home Reference, (2018) What is DNA? *https://ghr.nlm.nih.gov/primer/basics/dna.*
2. M Kchouk, JF Gibrat, M Elloumi, (2017) "Generations of Sequencing Technologies: From First to Next Generation", Biol Med (Aligarh), vol. 9, pp. 395.
3. National Human Genome Research Institute, The Cost of Sequencing a Human Genome (2016) *https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/*.
4. National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. (2007, January) Washington (DC): National Academies Press (US) Why Metagenomics? *https://www.ncbi.nlm.nih.gov/books/NBK54011/.*
5. Sharpton, T. J. (2014) An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 209. http://doi.org/10.3389/fpls.2014.00209.
6. Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007) MEGAN analysis of metagenomic data. Genome Research, 17(3), 377386. http://doi.org/10.1101/gr.5969107.
7. Sanli, K., Karlsson, F. H., Nookaew, I. & Nielsen, J. (2013) FANTOM: Functional and taxonomic analysis of metagenomes. *BMC Bioinformatics* 14, 38, doi: 10.1186/1471-2105-14-38.
8. Keegan K.P., Glass E.M., Meyer F. (2016) MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. In: Martin F., Uroz S. (eds) Microbial Environmental Genomics (MEG). *Methods in Molecular Biology*, vol 1399. Humana Press, New York, NY.
9. Pérez-Wohlfeil E., Arjona-Medina J.A, Torreno O., Ulzurrun E., Trelles O. (2017) Computational workflow for the fine-grained analysis of metagenomic samples. *BMC Genomics* 17(8): 802.

10. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSIBLAST: a new generation of protein database search programs." *Nucleic Acids Res*. 25:33893402
11. Gish, W. & States, D.J. (1993) "Identification of protein coding regions by database similarity search." *Nature Genet*. 3:266272.
12. Buchfink et al, Fast and sensitive protein alignment using DIAMOND, *Nature Methods*, 2015, 12:5960.
13. NCBI, Resource Coordinators. (2013) "Database resources of the National Center for Biotechnology Information." *Nucleic acids research* 41.Database issue: D8.
14. Kanehisa, Minoru, and Susumu Goto. (2000) "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research* 28.1: 2730.
15. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic acids research* 40: e94–e94
16. Li, D., Liu, C-M., Luo, R., Sadakane, K., & Lam, T-W. (2015) MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, doi: 10.1093/bioinformatics/btv033.
17. The NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Guyer, M. (2009) The NIH Human Microbiome Project. *Genome Research*, 19(12), 23172323. http://doi.org/10.1101/gr.096651.109.
18. Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME Journal*, 6(5), 10071017. http://doi.org/10.1038/ismej.2011.159.
19. Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., Howe, A. (2017) Strategies to improve reference databases for soil microbiomes. The ISME Journal, 11(4), 829834. http://doi.org/10.1038/ismej.2016.168.
20. Rodríguez-Brazzarola P., Pérez-Wohlfeil E., Díaz-del-Pino S., Holthausen R., Trelles O. (2018) Analyzing the Differences Between Reads and Contigs When Performing a Taxonomic Assignment Comparison in Metagenomics. *Bioinformatics and Biomedical Engineering. IWBBIO 2018. Lecture Notes in Computer Science*, vol 10813. Springer, Cham.