

A vibrant, colorful illustration of various microorganisms, including bacteria, viruses, and fungi, scattered across the top and right sides of the page. The organisms are depicted in various shapes and colors, such as green, orange, purple, and pink, with some showing internal structures like nuclei and organelles.

**REPORT: MICROBIOME
LIFESTYLE TEST CREATION,
WITH THE GOAL OF OBESITY
TREATMENT**

NOVEMBER 2019

ADVISOR: PROFESSOR DOUTOR
FRANCISCO PINA MARTINS

SUPERVISOR: DOUTORA JOANA VAZ

AUTHOR: PEDRO MELO E SOUSA CASAL
RIBEIRO GONÇALVES

ACKNOWLEDGEMENTS

Thanking someone is primarily an act of gratitude, and in this paper I want to thank all that have contributed positively and assisted me in these past 3 years of academic journey.

Firstly, I want to thank my school, Instituto Politécnico de Setúbal and especially my teachers and colleagues that were always supportive and always endured my inquisitive nature.

To Centro de Medicina Laboratorial Germano de Sousa and its team, for taking me in, nurturing my professional growth, with a special focus on the genetics team, Professor Doutor. Germano de Sousa, Doutora Joana Vaz, Doutor José Leal, Genetics Specialist Ana Pereira, Sara Carapeta and Doutora Inês Sousa and my colleague Rodrigo Rente. I have learned so much this past year and I couldn't have without your guidance, support and friendship.

To Doutora Joana Vaz, Doutor José Leal for asking the tough questions and for offering guidance so I could present the best paper possible.

To Professor Doutor Francisco Pina Martins for the constructive feedback on this paper and for challenging me to keep on improve it.

And finally I consider this academic achievement a victory for both myself and my family. Especially to my wife for the support in these past 3 years. Thank you for believing in me and sharing my pains and little victories along the way!

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 2

ACRONYMS AND ABBREVIATIONS 4

1. ABSTRACT 5

2. INTRODUCTION 6

2.1 COMPANY INFORMATION 6

2.2 PROJECT OBJECTIVES 7

3. FUNDAMENTAL THEORY 8

3.1 MICROBIOTA AND THE MICROBIOME 8

3.2 WHY IS THE MICROBIOTA IMPORTANT? 9

3.2.1 HARNESSING THE POWER OF A HEALTHY MICROBIOTA 10

3.3 SYMBIOSIS VS DYSBIOSIS 12

3.3.1 MICROBIOME AND OBESITY 13

3.3.1.1 OBESITY AND ITS MICROBIOME INDICATORS 13

3.4 METAGENOMICS AND BIOINFORMATICS 15

3.4.1 NGS AND ITS DATA 15

3.4.2 16S rRNA GENE 17

3.5 RELEVANT FILE FORMATS 19

3.6 PROGRAMMING LANGUAGES 21

4. PRODUCT DEVELOPMENT 23

4.1 FINDING THE RIGHT TOOLS 23

4.2 PHASE 1 30

4.3 PHASE 2 33

4.4 PHASE 3 36

4.5 PHASE 4 41

5. DISCUSSION 42

6. RESULTS 44

7. CONCLUSIONS 45

8. LITERATURE CITED 46

9. APPENDICES 49

ACRONYMS AND ABBREVIATIONS

CMLGS – Centro de Medicina Laboratorial Germano de Sousa

AG – American Gut Project

NCBI – National Center for Biotechnology Information

PGM – Personal Genome Machine

BP – Base Pair

SIBO – Small Intestine Bacterial Overgrowth

IBD – Inflammatory Bowel Disease

DB – Database

GG – GreenGenes

RDP – Ribosomal database project

LCA – Lowest Common Ancestor

FMT – Fecal Microbiota Transplantation

GI – Gastrointestinal

OTU – Operational Taxonomic Unit

PGM – Ion Personal Genome Machine

NGS – Next-Generation Sequencing

rRNA – Ribosomal ribonucleic acid

SAM – Sequence Alignment Map

BAM – Binary Alignment Map

BIOM – Biological Observation Matrix

JSON – JavaScript Object Notation

PNG – Portable Network Graphics

PDF – Portable Document Format

SED – Stream Editor

1. ABSTRACT

This paper details the work developed by myself at Centro de Medicina Laboratorial Germano de Sousa (CMLGS) in the genetics department, related to the fields of metagenomics and microbiota. The main goal was to develop the bioinformatics pipeline component of a new clinical test focused on the gut microbiota and obesity. This clinical test will help a doctor/nutricionist in finding actionable paths of diet for possible obesity treatments.

The work developed was divided between researching, testing and implementing bioinformatics solutions as well as understanding the metrics and terms necessary to reach the main goal.

As a result CMLGS now has a market-ready clinical microbiome test. It is called **GUTHEALTH**.

KEYWORDS

Bacteria, Microbiome, Microbiota, 16S rRNA, Diet, Obesity

2. INTRODUCTION

2.1 COMPANY INFORMATION

The Centro de Medicina Laboratorial Germano de Sousa group specializes in laboratory medicine and has more than 40 years of experience in clinical analysis. It currently comprises more than 450 collection places and 15 laboratories across Portugal and it performs over 11 million laboratory tests per year. Its headquarters and the department of Genetics are located in Telheiras where the investigation and production of this test took place.

The department of Genetics answers the growing need in the areas of Genetic and Genomic related diagnoses. It comprises of areas Cytogenetics/Molecular Cytogenetics, Molecular Genetics and Genetic Biochemistry, presenting a complete offer of genetic tests for the most diverse areas of Clinical specialty.

Currently the goal of creating “Lifestyle” health products such as the one presented in this paper is one of the priorities of the CMLGS group. A brand called Lifestyle Genomics was created to accommodate such goal.

To provide scientific support in the context of microbiome analysis, partnership between the CMLGS group and the NOVA Medical School – Faculty of Medical Sciences of the Nova University of Lisbon. Namely the list of Genus/Species of bacteria to be analyzed and included and the conclusions related to obesity, were the responsibility of the research group led by Doutora Conceição Calhau.

2.2 PROJECT OBJECTIVES

The main objective was to create a marketable sequencing-based clinical microbiome test related to obesity without the need of human intervention in the processes of analysis and reporting. The project was divided into 4 phases:

Phase 1 included: **1)** Finding in the data the right bacteria that would be identified with the used Next Generation Sequencing (NGS) methods; **2)** Detecting actionable bacteria through diet; **3)** Implementing the right metrics that could assist health professionals in recommending healthier diets to the patients; **4)** Creating a database of taxonomic sources that would be able to properly identify bacteria present in the patient's with our microbiome test.

Phase 2 included: **1)** Finding publicly available datasets that would serve as populational references; **2)** To create a pipeline that would be able to extract metrics from the compiled reference samples.

Phase 3 was devoted to further develop the software and metagenomic pipelines available. This included: **1)** Creating our own pipeline that to the metrics and charts to be used; **2)** Designing the test itself (i.e, how it would be presented to our customers) and finally **3)** Creating the pipeline that would analyse the data and produce the report automatically without human intervention and creating the PDF report.

Phase 4 was related to finding the right metrics and statistic tests that would allow for sample profiling. We needed to identify whether a sample more closely resembled an obese gut microbiota or a normal gut microbiota.

3. FUNDAMENTAL THEORY

3.1 MICROBIOTA AND THE MICROBIOME

The terms Microbiota and Microbiome were often interchangeably used. While the core concept is similar, microbiota and microbiome are fundamentally different definitions^[1].

A microbiota is an ecological community of apathogenic (commensal and symbiotic) and pathogenic microorganisms found in and on all multicellular organisms studied to date. A microbiota includes bacteria, archaea, protists, fungi and viruses. It has been found to be a factor in immune response and hormonal and metabolic equilibrium of their host^[2].

A microbiome as established by ^[1], refers to the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions. Humans, plants, and other animals all have microbiomes; these can be generalized to their entire organism, or broken down into specific microbiomes for different locations on them.

Microbiota are specific to each organism and the diversity in microbiomes between individuals is huge, and even within a person there can be extensive variation in their microbiome makeup. For humans, there are a number of specific and separate microbiomes present. From skin to lungs to the gastrointestinal tract, all of these specific microbiomes make up a unique microbiome for each human.

Each individual has a unique microbial composition that is influenced by the types of bacteria acquired through maternal vertical transmission, genetic composition of the individual, diet, use of medications, intestinal infections, stress and day-to-day interactions.

3.2 WHY IS THE MICROBIOTA IMPORTANT?

Trillions of microbes exist inside an individual's intestines and on his skin. A study^[3] from 2018 extrapolated that there are more bacterial cells in your body than human cells. There are roughly 40 trillion bacterial cells in one's and only 30 trillion human cells. Combined, these microbes may weigh as much as 1-2 kg, which is roughly the weight of the brain. Together, they function as an extra organ in our bodies and play a role in an individual's health^[4].

Microbes begin to affect our bodies the moment we are born as we first exposed to them when we pass through our mother's birth canal. But our knowledge of this subject is expanding as new evidence suggests that babies may come in contact with some microbes while inside the womb^[5].

As we grow, our gut microbiota begins to diversify. Higher microbiota diversity is considered good for one's health^[6].

Some examples of how the microbiota affects our bodies include:

- Some of the bacteria that first begin to grow inside babies' intestines are called *Lactobacilli*. They can degrade lactose and use challenging to digest substrates such as milk glycans^[7].
- Certain bacteria digest fiber, producing short-chain fatty acids^[7], which are important for gut health. Fiber may help prevent weight gain, diabetes, heart disease and the risk of cancer^{[8] [9]}.
- The gut microbiota also helps in shaping how our immune system works^[10]. By interacting with immune cells, the gut microbiome can affect how our body responds to infection.

-
- The gut microbiota may also affect the central nervous system^[11], which controls brain function and this relationship has been linked to Autism^[12].

In summary, the data gathered so far is pointing to the major importance of gut microbiota for human health. Thus understanding the composition, the fluctuations associated with health and disease and how to manipulate the gut microbiota is becoming a main goal in human health.

This serves as motivation as to why we need to create microbiota related clinical tests. These need to be able to characterize microbiota complex relationships with the host. This is why a test such as GUTHEALTH is necessary and the reason behind its development.

3.2.1 HARNESSING THE POWER OF A HEALTHY MICROBIOTA

Fecal Matter Transplant (FMT) is an innovative investigational treatment that has been used to resolve infections caused by recurrent *C. difficile* that does not respond to antibiotics. During an FMT, a fecal preparation from healthy stool donor is transplanted into the colon of the patient and works by repopulating the patient's microbiota with diverse microorganisms, that competitively exclude *C. Difficile* and restore the symbiotic relationship of the microbiota^[7].

However this treatment has inherent risks. For example if the donor carries within its microbiota pathogenic and resistant bacteria and without proper screening, the transplant is made. Its effects can be deadly as shown by the medical case below:

A 73 year old patient, participating in a clinical trial involving fecal transplants, died after developing sepsis due the presence of an antibiotic-

resistant strain of E. Coli in his blood. Other participants developed similar symptoms but their infection responded to the antibiotics while this particular patient did not^[13].

Following this episode FDA implemented microbiome analysis for multidrug-resistant organisms in all FMT procedures.

FMT shows that manipulation of the microbiota can be a viable treatment option. This episode emphasizes the need for reliable microbiota clinical tests such as GUTHEALTH.

3.3 SYMBIOSIS VS DYSBIOSIS

Symbiosis is any type of a close and long-term biological interaction between two different biological organisms. As defined by the article ^[14] “In the human microbiome literature, the definition of symbiosis ranges from a commensalistic relationship, wherein the interaction is decidedly beneficial for one of the partners (the host), to mutualistic, involving beneficial outcomes for all organisms involved.”

In the context of the microbiota, Dysbiosis is any perturbation of the normal content that could disrupt the symbiotic relationship between the host and associated microbes^[15]. This can originate a disease state on its host.

Exposure to aggressive factors such as stress, excessive consumption of alcohol, smoking, antibiotic use, unhealthy diet and even sleep deprivation may cause imbalance of the microbiota (e.g. increased pathogenic bacteria versus a decrease in commensal bacteria). This disruption can result in diseases^[15], such as inflammatory bowel disease and other gastrointestinal (GI) disorders, including gastritis, peptic ulcer disease, irritable bowel syndrome, gastric and colon cancer and other systemic diseases such as obesity^[7].

The medical community is becoming more and more aware of the importance^[16] of these imbalances for human health and consequently therefore having clinical tests designed and available to recognize Dysbiosis is crucial to treating the conditions outlined above. GUTHEALTH was created for this purpose.

3.3.1 MICROBIOME AND OBESITY

Obesity is becoming worldwide epidemic given its rapid growth. Obesity and obesity-related metabolic disorders are characterized by specific alterations in the composition and function of the human gut microbiome^[17]. Mechanistic studies have indicated that the gastrointestinal microbiota can influence energy consumption and generation. As a factor influencing energy utilization from the diet and as a factor that influences host genes that regulate energy expenditure and storage^[18].

3.3.1.1 OBESITY AND ITS MICROBIOME INDICATORS

In order to design GUTHEALTH it was important to understand the metrics that characterized an obese microbiota profile and how we could measure them. These are the global measures that are now a part a GUTHEALTH test:

- ~ Richness is a measure for the total number of the species in a community. A study showed that individuals with a low bacterial richness (23% of the population) are characterized by more marked overall adiposity when compared with high bacterial richness individuals^[19].
- ~ Evenness is the measure of uniformity of abundance between species in a community. One study found statistical support for decreased evenness amongst obese individuals^[20].
- ~ The diversity index (Alpha diversity) takes into consideration the how many species are present (richness) and their overall representation relative to other species (evenness). Typical values are generally between 1.5 and 3.5 in most ecological studies, and the index is rarely greater than 4. Cross-sectional studies have shown lower microbiota diversity in obese subjects compared to lean controls^[21].

~ Firmicutes/Bacteroidetes ratio has been consistently demonstrated by numerous studies that it is increased in obese people compared to lean people, and tend to decrease with weight loss ^[18] ^[21].

3.4 METAGENOMICS AND BIOINFORMATICS

Research of the microbiota is mainly supported by NGS techniques and its data. The scientific field that encompasses this research is called Metagenomics. It is defined as the direct genetic analysis of genomes contained with an environmental sample^[22].

This wouldn't be possible with Bioinformatics which is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data.

The data generated by metagenomics studies are both enormous and inherently noisy, containing fragmented data. Collecting, curating, and extracting useful biological information from datasets of this size represent significant computational challenges for researchers^[22]. This is where a bioinformatician comes in.

3.4.1 NGS AND ITS DATA

The NGS instrument used by CMLGS is from Thermofisher. The Ion Personal Genome Machine (PGM) System combines semiconductor sequencing technology with natural biochemistry to directly translate chemical information into digital data. The system leverages direct, real-time sequencing detection, providing sequencing results typically in 3-7 hours^[23]. The PGM can provide sequence data in two read-lengths: 200bp and 400bp.

Concerning metagenomics PGM sequencing accelerates and simplifies its research by using whole-genome or targeted sequencing of the bacterial 16S rRNA gene. 16S rRNA sequencing was the one selected by our lab to implement our microbiome test. Not only 16S rRNA sequencing is the most widely accepted approach by the medical community but it also has a reduced cost and

processing time and finally, due to its wide acceptance by the medical community, there is easy access to reference samples that can be used as control samples in the test.

3.4.2 16S rRNA GENE

The use of 16S rRNA gene to study metagenomics, bacterial phylogeny and taxonomy has been considered the workhorse of scientific community [24] and, more recently, of the private companies offering microbiome testing. The main reasons are:

- They are omnipresent in all prokaryotic species, ribosomes can't translate mRNA without their 16S rRNA component, so all prokaryotic species have it [24].
- Our ability to sequence it at a lower cost and to recognize bacteria down to the Genus/Species level [24].
- The function of the 16S rRNA gene has not changed over time and is therefore a conserved gene which facilitates using it across studies and over time. This also means it is possible to construct a tree of life linking together all known bacteria [24].
- Finally, is one of the most well-studied and characterized genes used in metagenomics, the phylogenetic trees are well developed and taxonomic information is readily available in a variety of databases.

The 16s rRNA gene is comprised of conserved and variable regions. As shown by figure 1.



CONSERVED REGIONS: unspecific applications

VARIABLE REGIONS: group or species-specific applications

Figure 1 - Representation of the 16s rRNA gene. Digital Image. The Ishaq Lab. May 8 2016.
<https://sueishaqlab.org/tag/16s-rrna/>

The conserved regions allow primers to be designed to target all bacteria, but they can amplify the 16s gene through a variable sequence of bases in which its differences allow the identification of the organism to the Genus/Species level. However, it is important to note that not all regions of the 16s rRNA gene are equally good at differentiating between different taxon[25].

16S rRNA DATABASES

An important step when using the 16S rRNA gene for bacteria classification is having a means of comparison with agreed upon classified reads[24]. The main free-to-use available databases are named Greengenes, Ribosomal Database Project (RDP), Silva, Open Tree of life Taxonomy (OTT) and National Center for Biotechnology Information (NCBI)[26]. Not all available databases have the same information or the same level of detail as shown by *figure 2*.

Taxonomy	Type	No. of nodes	Lowest rank	Latest release
SILVA	Manual	12,117	Genus	Sep 2016
RDP	Semi	6,128	Genus	Sep 2016
Greengenes	Automatic	3,093	Species	May 2013
NCBI	Manual	1,522,150	Species	Today ^a
OTT	Automatic	2,627,066	Species	Sep 2016

Figure 2 - Overview of 5 databases mentioned and their taxonomic classifications (table taken from "SILVA, RDP, Greengenes, NCBI and OTT-how do these taxonomies compare?")

3.5 RELEVANT FILE FORMATS

A file format is the layout of a file in terms of how the data within the file is organized. It is important to understand the concepts behind each format so we can replicate previous work and have our own available for replication in standard usable formats. I will describe the formats used in this report below and the reasons behind using them.

SAM & BAM

SAM (Sequence Alignment Map) is a text-based format for storing biological sequences aligned to a reference sequence. It is widely used for storing data, such as nucleotide sequences, generated by next generation sequencing technologies, and the standard has been broadened to include unmapped sequences. The binary equivalent of a SAM file is a BAM (Binary Alignment Map) file, which stores the same data in a compressed binary representation^[27].

FASTQ

This is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity^[28].

BIOM

It is a Biological Observation Matrix Data. The BIOM file format is designed to be a general-use format for representing biological sample by observation contingency tables^[29].

PHYLOSEQ

It is a R object that provides a set of classes and tools to facilitate the import, storage, analysis, and graphical display of microbiome census data. Currently, phyloseq uses 4 core data classes. They are the OTU abundance

table (otu_table), a table of sample data (sample_data); a table of taxonomic descriptors (taxonomyTable) and a phylogenetic tree ("phylo"-class) ^[30].

JSON

JSON (JavaScript Object Notation) ^[31] is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. is built on two structures:

- A collection of name/value pairs. In various languages, this is realized as an object, record, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an array, vector, list, or sequence.

PNG

PNG (Portable Network Graphics) is a raster-graphics file-format that supports lossless data compression ^[32].

3.6 PROGRAMMING LANGUAGES

Programming is important to automate, collect, manage, calculate, analyze the processing of data and information accurately. When it comes to deciding which programming languages used it is important to consider the field of work and the tools and libraries available. The discipline of Bioinformatics cannot be done without the use of programming languages. In this report the main languages used were R and PYTHON with some minor use of BASH/SHELL scripts and JAVA as “wrapper” languages for the command line menus and processing of files in bulk. I will describe the programming languages used in this report below.

R

It is a language and environment for statistical computing and graphics. It provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering and others) and graphical techniques, and is highly extensible^[33].

PYTHON

It is a widely used general-purpose, high level programming language. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code^[34].

JAVA

It is a general-purpose, concurrent, strongly typed, class-based object-oriented language. It is normally compiled to the bytecode instruction set and binary format defined in the Java Virtual Machine Specification^[35].

BASH/SHELL

A shell script is a computer program designed to be run by the Unix shell, a command-line interpreter. Bash (Bourne Again Shell) is a type of interpreter that processes shell commands. The main purpose of a BASH shell

is to allow users to interact effectively with the system through the command line^[36].

4. PRODUCT DEVELOPMENT

In order for CMLGS to bring this test to market I had to ensure that I researched the available and appropriate tools, always with the focus of creating a product, that would generate its analysis and reporting itself automatically. While I mentioned four separate phases previously, each with their own tasks, all of them were performed in parallel of each other. This allowed fine tuning to be done quicker but also the manifestation of new and improved ideas.

In this section I've added snippets of the scripts used and created however, I wasn't able to provide all the code written as it is intellectual property of CMLGS.

4.1 FINDING THE RIGHT TOOLS

Considering the time constraints to take this test to market coupled with fact that I was the sole bioinformatician available I had to simplify the development process as much as possible while making sure I maintained the quality of the product. It was crucial for the timely success of this project that I used the right tools.

My main requirements for choosing bioinformatics tools were: **1)** Being user-friendly and easy to implement; **2)** A proven track record amongst the bioinformatician community; **3)** Easy to maintain; **4)** And if possible coded in a language that required the least amount of installation of script dependencies as possible.

BAM FILE CONVERSION

The raw sequencing data from PGM instrument is in BAM format. This meant converting these into a more usable format. In this case I chose the FASTQ format because it is commonly used by other tools and had quality metrics that could be used to perform quality control.

I tested Picard^[37] and Bamtofastq from BEDTOOLS^[38]. Both met my requirements but because Picard was based on JAVA I decided to use Bamtofastq. It meant using less dependencies. This is a tool from BEDTOOLS which is a fast, flexible toolset for a wide-range of genomics analysis tasks. BAMTOFASTQ is a conversion utility for extracting FASTQ records from sequence alignments in BAM format. Command invocation as shown in *figure 3*.

```
$ bedtools bamtofastq [OPTIONS] -i $BAM -fq $FASTQ

[-i]      = BAM input file.
[-fq]     = FASTQ output file.
```

Figure 3 - BAMTOFASTQ STANDARD USAGE. NOTE: OPTIONS DISPLAYED ARE THE ONES USED ON THIS REPORT.

FASTQ QUALITY CONTROL

High-throughput sequencing means a probability of sequencing errors and chimeras in the process. This means performing quality control on the its data.

I tested Trimmomatic^[39], Cudadapt^[40] and Fastq_Quality_Filter. All tools met my requirements and I chose Fastq_Quality_Filter because I had used it before and it worked quickly and efficiently. This is a tool from the FASTX-TOOLKIT^[41] which is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing. FASTQ Quality Filter removes low-quality sequences from FASTQ files. Command invocation as shown in *figure 4*.

```
$ fastq_quality_filter [-q N] [-p N] [-i INFILE] [-o OUTFILE]

[-q N]      = Minimum quality score to keep.
[-p N]      = Minimum percent of bases that must have [-q] quality.
[-i INFILE] = FASTA/Q input file. default is STDIN.
[-o OUTFILE] = FASTA/Q output file. default is STDOUT.
```

Figure 4 - FASTQ_QUALITY_TRIMMER USAGE. NOTE: OPTIONS DISPLAYED ARE THE ONES USED ON THIS REPORT.

CONVERSION OF SINGLE TO PAIRED FASTQ FORMAT

The PGM instrument sequences in both directions (3' & 5'). This meant converting the single raw read file into paired format.

I tested using a set of bash commands and a tool called Reformat^[42]. I chose the Reformat tool because it was more commonly used by the bioinformatics community. It is a tool from BBT00LS which is a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data.

Reformat is designed for generic streaming read-processing tasks that have low memory or computational demands. In this case it was used to separate paired reads. Command invocation as shown in *figure 5*.

```
$ reformat.sh in=reads.fq out1=read1.fq out2=read2.fq

[in]      = FASTA/Q input file.
[out1]    = FASTA/Q output file with forward reads.
[out2]    = FASTA/Q output file with reverse reads.
```

Figure 5 - REFORMAT USAGE; NOTE: OPTIONS DISPLAYED ARE THE ONES USED ON THIS REPORT.

METAGENOMIC CLASSIFICATION

Metagenomic classification tools match sequences against a database of microbial genomes to identify the taxon of each sequence. There are two different approaches to this problem, clustering-first or assignment first classification^[43]. Approaches outlined in *figure 6*.

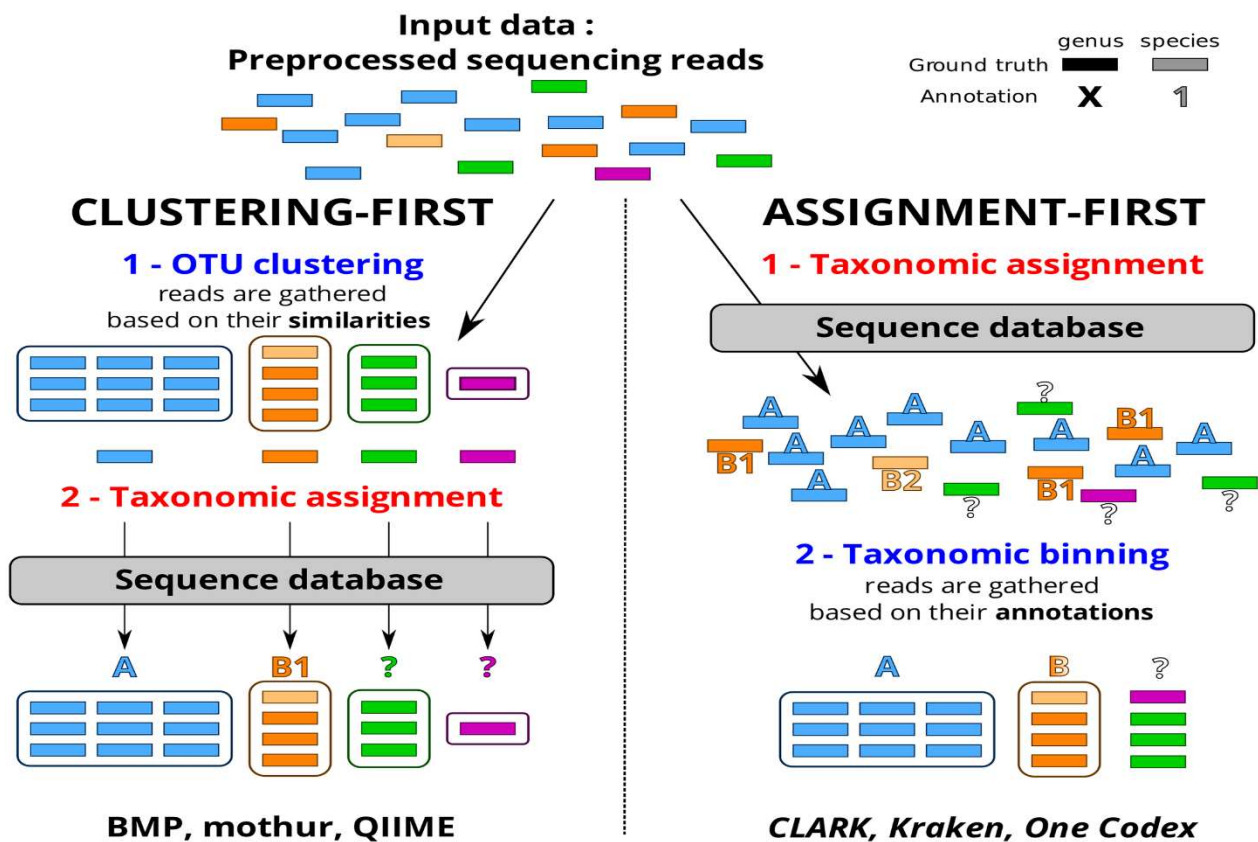


Figure 6 - *DISTINCTIONS BETWEEN CLUSTERING-FIRST AND ASSIGNMENT-FIRST APPROACHES.* (image taken from "Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics")

Each approach has its own merits and issues. OTU-clustering means that reads are gathered into OTUs (Operation Taxonomic Unit) based on their sequence similarities. This allows for the discrimination of unclassified reads but can also be consuming on the available resources such as time, CPU usage or RAM depending on the amount of input data^[43]. Assignment-first was faster and sometimes lighter on resources but could be more prone to chimera errors.

Ultimately I decided that assignment-first was the viable choice more specifically Kraken2, for time and resource constraints but also because the

product involved knowing beforehand the bacteria involved so I didn't need a *de-novo* approach.

Kraken2^[44] tool it is a taxonomic sequence classifier that assigns taxonomic labels to DNA sequences. Kraken examines the k-mers within a query sequence and uses the information within those k-mers to query a database. That database maps k-mers to the lowest common ancestor (LCA) of all genomes known to contain a given k-mer. DB build command invocation as shown in *figure 7* and Classification command invocation as shown in *figure 8*.

For targeted 16S sequencing projects Kraken2 provides support for building databases from three publicly available 16S databases, Greengenes, RDP and SILVA.

Kraken 2's standard sample report format is tab-delimited with one line per taxon. The fields of the output, from left-to-right, are as follows:

1. Percentage of fragments covered by the clade rooted at this taxon.
2. Number of fragments covered by the clade rooted at this taxon.
3. Number of fragments assigned directly to this taxon.
4. A rank code, indicating (U)nclassified, (R)oot, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies.
5. NCBI taxonomic ID number.
6. Indented scientific name.

```
$ kraken2-build --db $DBNAME --special $TYPE
```

```
 [--db]           = Name of the DB to be created.
```

```
 [--special]      = Type of DB to be created (greengenes, rdp, silva)
```

Figure 7 - KRAKEN2 BUILD USAGE FOR 16S DB TYPES

```
$ kraken2 --threads $THREADS --db $DBNAME --paired R1.fastq R2.fastq --
confidence NUM --report FILE.report
```

```
 [--db]           = Name of the DB to be used.
 [--threads]      = NUM switch to use multiple threads.
 [--confidence]   = Confidence threshold used [0-1].
 [--paired]       = NUM switch to use multiple threads.
 [--report]       = Name of the report file created.
```

Figure 8 - KRAKEN2 USAGE. NOTE: OPTIONS DISPLAYED ARE THE ONES USED IN THIS REPORT.

METAGENOMIC ABUNDANCE ESTIMATION

In the creation of this product it was also crucial to get the correct abundance information for each taxa. Therefore, and because I chose to use Kraken2, I decided to use BRACKEN^[45] (Bayesian Reestimation of Abundance with Kraken).

It is a highly accurate statistical method that computes the abundance of species in DNA sequences from a metagenomics sample. It uses the taxonomy labels assigned by Kraken to estimate the number of reads originating from each species present in a sample. It produces accurate species and genus-level abundance estimates even when a sample contains two or more near-identical species. Build command invocation as shown in *figure 9* and Classification command invocation as shown in *figure 10*.

```
$ bracken-build -d $DBNAME -t $THREADS -k $KMER_LEN -l $READ_LEN
```

```
 [-d]           = Name of the KRAKEN2 DB to be used.
 [-t]           = NUM switch to use multiple threads.
 [-k]           = Length of kmer used to build the database. [default: 35]
 [-l]           = Ideal length of reads in your sample.
```

Figure 9 - BRACKEN BUILD USAGE FOR KRAKEN2

```
$ bracken -d $DBNAME -t $THREADS -i $KRAKEN2.report -r $NUM -o $NAME.bracken
```

```
[-d] = Name of the KRAKEN2 DB to be used.  
[-t] = NUM switch to use multiple threads.  
[-i] = Input KRAKEN2 report  
[-r] = Ideal length of reads in your sample.  
[-o] = Name of the output file to be generated.
```

Figure 10 - BRACKEN USAGE ; NOTE: OPTIONS DISPLAYED ARE THE ONES USED IN THIS REPORT.

CONVERSION FROM KRAKEN TO PHYLOSEQ

Due to the nature of the report I had to extract statistical metrics from the data, both from sample and reference populations. I decided to use the R object Phyloseq. This would allow me to use ready functions that were peer tested. For this I had to convert the report output from Kraken2+Bracken to a biom format so it could be imported as a Phyloseq format. This was also helpful considering that I could store all population kraken report observations in one single biom file which would help the speed of the pipeline.

For this step the KRAKEN-BIOM^[46] algorithm was implemented. This takes as input, one or more files output from the kraken-report tool. Each file is parsed and the counts for each OTU are recorded, along with database ID (e.g. NCBI), and lineage. The extracted data are then stored in a BIOM table where each count is linked to the Sample and OTU it belongs to. Sample IDs are extracted from the input filenames. Command invocation as shown in *figure 11*.

```
$ kraken-biom $KRAKEN.report -o $SAMPLE.biom --fmt hdf5
```

```
[-o] = Name of output biom file used.  
[-fmt] = format in which the biom file coded. [hdf5, json,tsv]
```

Figure 11 - KRAKEN-BIOM USAGE

4.2 PHASE 1

In order to identify the bacteria most interesting for medical intervention through our test we have established the partnership between CMLGS and NOVA previously described. This team work resulted in the compilation of a list of bacteria, presented in the *table 1*.

Table 1 - TABLE WITH CHOSEN BACTERIA

NAME	TAXONOMIC RANK	POSSIBLY COMMENSAL / PATHOGENIC
Bacteroides	Genus	Possibly Commensal
Bilophila	Genus	Possibly Commensal
Blautia	Genus	Possibly Commensal
Butyrivibrio	Genus	Possibly Commensal
Lactobacillus	Genus	Possibly Commensal
Prevotella	Genus	Possibly Commensal
Roseburia	Genus	Possibly Commensal
Ruminococcus	Genus	Possibly Commensal
Escherichia coli	Species	Possibly Commensal
Akkermansia muciniphila	Species	Possibly Commensal
Salmonella enterica	Species	Pathogenic
Campylobacter	Genus	Pathogenic
Clostridium difficile	Species	Pathogenic
Shigella	Genus	Pathogenic
Vibrio Cholerae	Species	Pathogenic

Due to the nature of the report I had to find these bacteria in our samples but also compare its abundance with one or more populations.

As a result of these requirements the task of finding the right classification database was a priority and considering that the lowest taxonomic rank necessary was Species I tested Greengenes and NCBI.

To perform these tests, I used a control sample from ZymoBIOMICS Microbial Community DNA Standard^[47]. This enabled me to test the effectiveness and accuracy of each database, because I had access to the composition and abundance of bacteria from this control sample. Neither of the tested databases proved to be satisfactory.

During this testing phase I encountered some concerns with these databases: **1)** Greengenes (version dated 05/13) lacked the ability to track *Escherichia coli*; **2)** NCBI meant that I had to create my own database of 16s rRNA sequences and I needed to use a peer curated database.

I chose to update Greengenes to a later version (dated 08/13) and adding more high quality sequences to that database. I focused on adding sequences from NCBI from the list of bacteria mentioned earlier.

To create the classification database, I used the pipeline outlined in the diagram from *figure 12*:

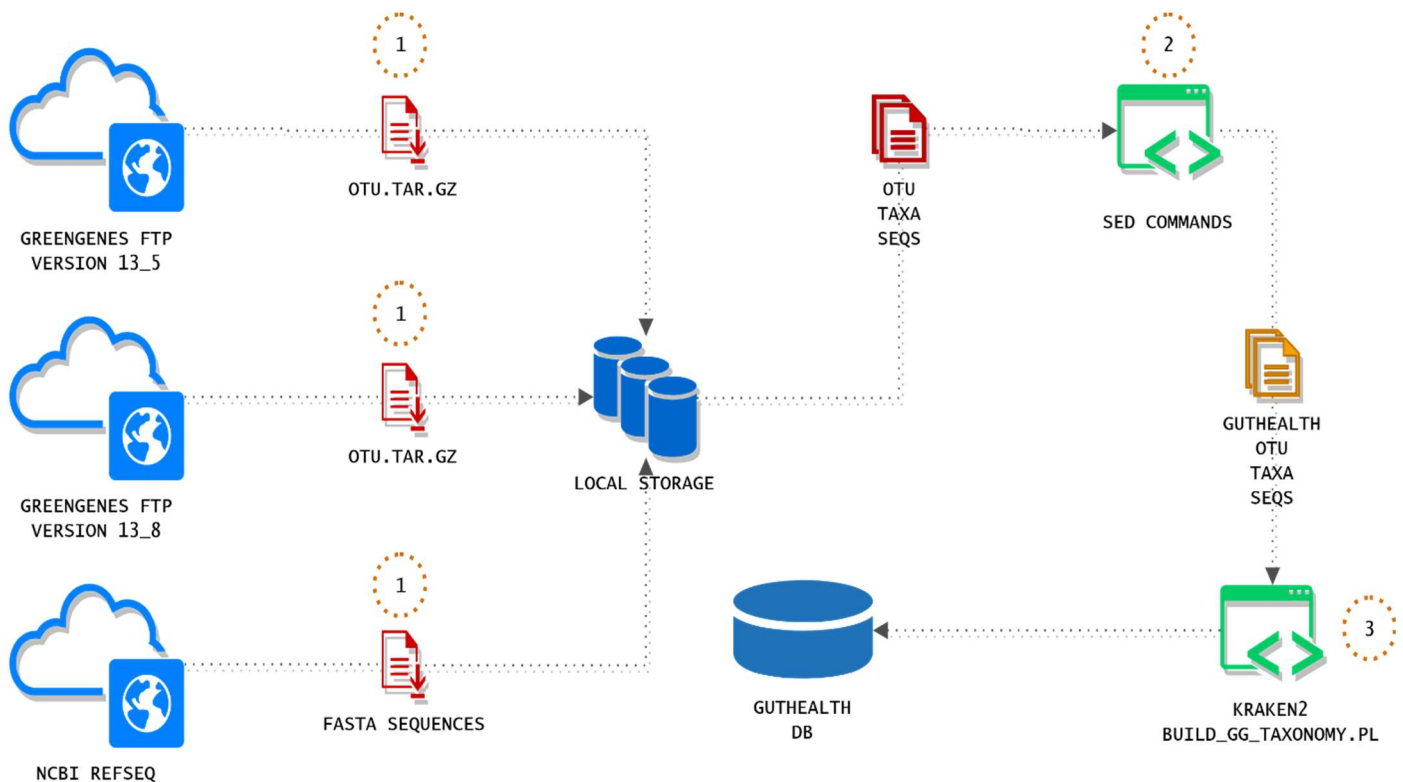


Figure 12 - GUTHEALTH DB CREATION PIPELINE

(1) Downloaded the necessary files from GG and NCBI from:

~ ftp://greengenes.microbio.me/greengenes_release/gg_13_5/

~ <https://www.ncbi.nlm.nih.gov/nucleotide>

-
- (2) Used SED (stream editor) commands to perform text transformations in order to merge all the data from these sources. This was necessary because the NCBI taxonomic ranks didn't follow the same format as the GG.
 - (3) Kraken2 requires 2 files to create a database: **1)** One fasta compressed file with the taxonomic IDs and sequences; **2)** A text file with the taxonomic IDs and names.

I updated the GG database creation PERL script of Kraken2 to accept my local versions of "gg.fasta.gz" and "gg_taxonomy.txt.gz", instead of procuring these files online and lastly created the classification database by using the build command with Kraken2 as shown in *figure 9*. I called this database **GUTHEALTH**. Snippet of Kraken2 perl script as shown in *figure 13*.

```
mkdir -p "$KRAKEN2_DB_NAME"
pushd "$KRAKEN2_DB_NAME"
mkdir -p data taxonomy library
pushd data
cp /home/ophiomicsp/Desktop/BD/gg_13_8_taxonomy.txt| .
cp /home/ophiomicsp/Desktop/BD/gg_13_8.fasta .

build_gg_taxonomy.pl gg_13_8_taxonomy.txt
popd
mv data/names.dmp data/nodes.dmp taxonomy/
mv data/seqid2taxid.map .
mv data/gg_13_8.fasta library/gg.fna
popd

# kraken2-build --add-to-library full_bact.fasta --db $KRAKEN2_DB_NAME
kraken2-build --db $KRAKEN2_DB_NAME --build --threads $KRAKEN2_THREAD_CT
```

Figure 13 - SNIPPET OF KRAKEN CREATE DB GUTHEALTH SCRIPT

4.3 PHASE 2

This phase included the discovery and usage of curated metagenomics data to be used as our reference population. From this data it was possible to extract population related metrics from, and be able to use those metrics as comparison with the sample in our test. I decided to use samples from the American Gut Project^[48] (AG). The decision to use these samples was based on the following:

- ~ Previous knowledge of working with this database as part of the Bioinformatics Laboratory project.
- ~ The fact that its sequencing data was the same as the one chosen by this project, the 16s rRNA gene.
- ~ The availability of the data. It is located in a free repository <https://www.ebi.ac.uk/ena> under the project code **PRJEB11419**.
- ~ Its metadata was abundant which allowed me to fine tune and filter the samples based on strict rules.

To create the population database, I used the pipeline outlined in the diagram in *figure 14*:

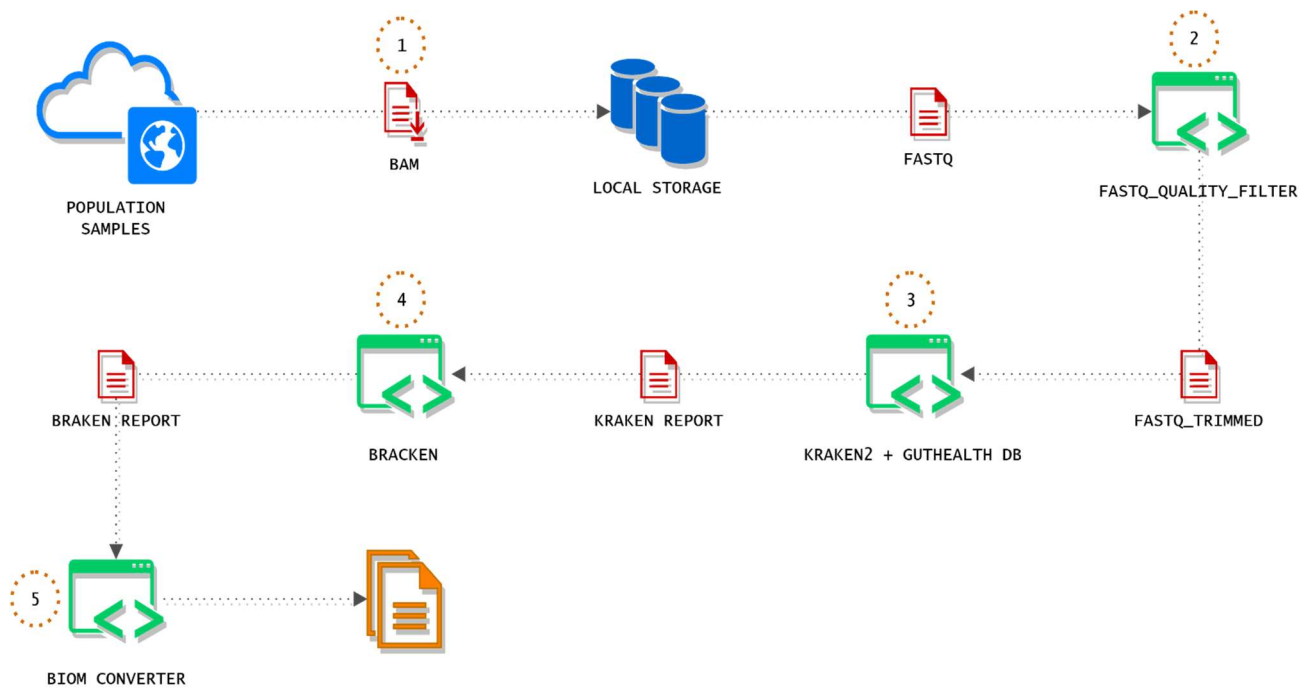


Figure 14 - POPULATION BIOM FILE PIPELINE

(1) I downloaded each sample individually using a python script that I designed called **import_AG_data.py**. This script uses the AG metadata to extract each sample id and then it would download both the sample metadata and its fastq file. This script was created during the Bioinformatics Laboratory project and adapted to this one. It uses the PYTHON libraries present in *table 2*.

Table 2 - TABLE WITH PYTHON LIBRARIES USED IN IMPORT_AG_DATA.PY

LIBRARY NAME	FUNCTION
REQUESTS	Ability To Perform HTTP Requests
PANDAS	High-Performance, Easy-To-Use Data Structures And Data Analysis Tools

The project's metadata was extensive so I was able to filter and separate samples. Since the main objective of this test is obesity treatment I decided to use the following filters and created the populations described in *table 3*.

Table 3 - *ALL SAMPLES THAT HAD DISEASES SUCH AS IBS, IBD, DIABETES WERE FILTERED OUT

POPULATION NAME	FILTERS USED
NORMAL	18.5 < BMI values < 24.9
OVERWEIGHT	25 < BMI values < 29.9
OBESE	30 < BMI values
UNDERWEIGHT	18.5 > BMI values

(2) Once I had all samples files downloaded I proceed to perform quality control using **fastq_quality_filter** as described in *figure 4*.

(3) Using the fastq filtered files after step number 2 I classified each one using Kraken2 as described in *figure 8*.

(4) Having each individual kraken report I proceed to use Bracken as described in *figure 10*.

(5) And lastly I used a python script that I designed called **bracken_to_biom.py** to convert each bracken report individually to a biom

format and append each to a population.biom file. A snippet of the script is present in *figure 15*. It uses the PYTHON libraries as shown in *table 4*.

Table 4 - TABLE WITH THE PYTHON LIBRARIES OF BRACKEN_TO_BIOM.PY

LIBRARY NAME	FUNCTION
ARGPARSE	Makes It Easy To Write User-Friendly Command-Line Interfaces.
SUBPROCESS	Gives The Developer The Ability To Start Processes Or Programs From A Python Script
BIOM	Provides Rich Table Objects To Support Use Of The BIOM File Format
QIIME	Collection Of Python Code And Scripts For Performing Microbiome Analysis. In this case I used a function called write_biom_table

```

directory = '/media/ophiomicsp/DATA/Ophiomics/AGDBs/AG_Overweight_NOF/'

if os.path.isdir(directory):
    for filename in os.listdir(directory):
        if filename.endswith("_bracken.report"):
            cmd_line = 'kraken-biom ' + os.path.join(directory, filename) + ' --output_fp ' + os.path.join(
                directory, filename) + '.biom ' + '--fmt hdf5'
            subprocess.Popen(cmd_line, shell=True)
        else:
            print ("Directory does not exist")

list_files = [filename for filename in os.listdir(directory) if filename.endswith('.biom')]
master_table = load_table(os.path.join(directory, list_files[0]))
count = 0
if os.path.isdir(directory):
    for filename in os.listdir(directory):
        if filename.endswith(".biom") and filename != list_files[0]:
            try:
                f = load_table(os.path.join(directory, filename))
                master_table = master_table.merge(f)
                count += 1
                print count
            except:
                print filename

write_biom_table(master_table, os.path.join(directory, 'AG_Overweight_NOF.biom'))

```

Figure 15 - SNIPPET OF BRACKEN_TO_BIOM SCRIPT

4.4 PHASE 3

This phase was used to create our own analysis pipeline and automatize it. This would allow us to automatically generate the metrics and charts to be used according to an established design of the test into a pdf file. This meant integrating all the scripts into one continuous and uninterrupted pipeline that took writing more than 2500 lines of code in BASH scripting, PYTHON AND R. Diagram outlined in *figure 16*.

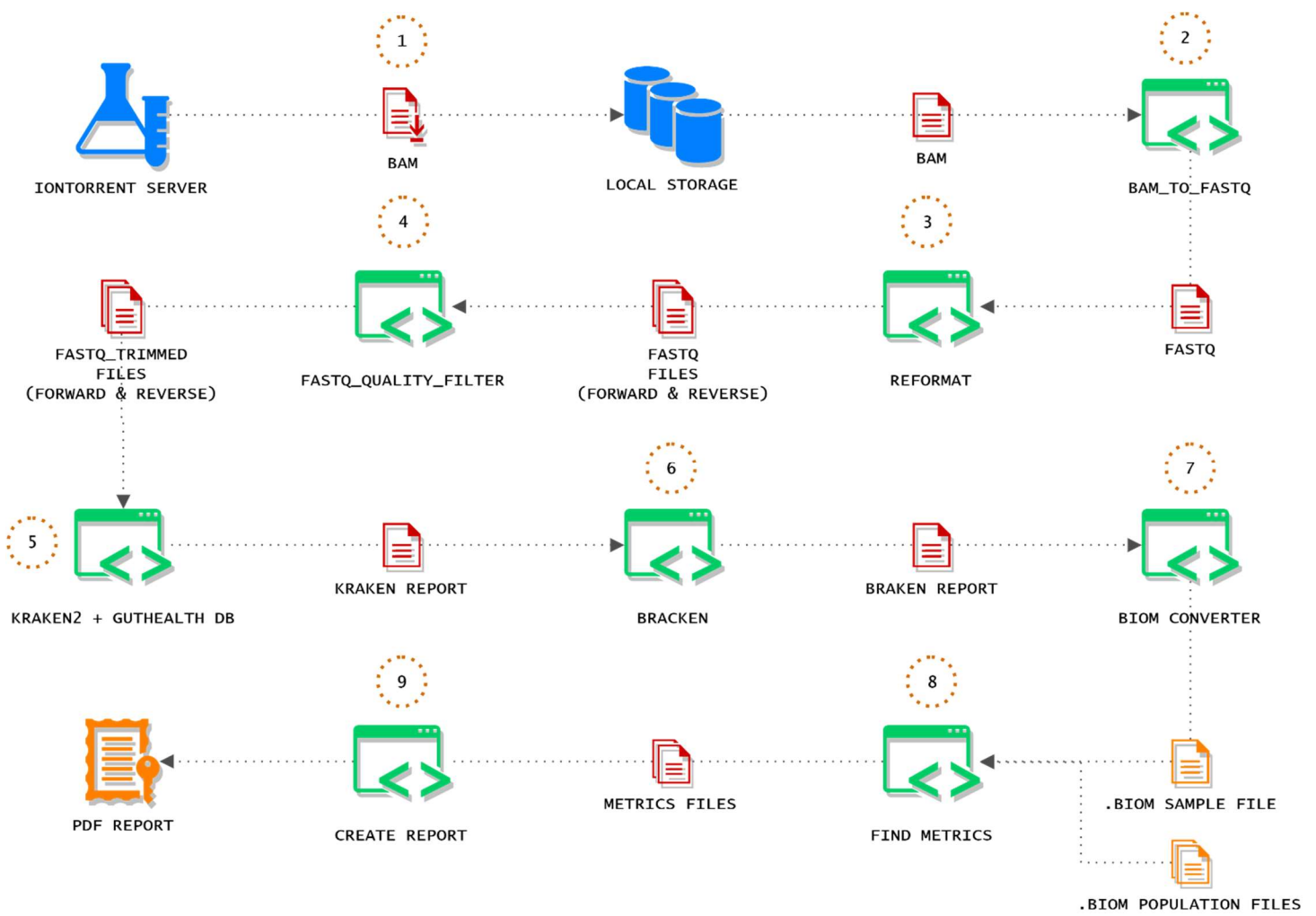


Figure 16 - REPORT CREATION PIPELINE

(1) First I need to extract the file from the Ion Torrent Servers. At the moment this is the only step that does require human action. It requires access to the PGM Ion Torrent interface and click on the BAM file name. Example shown in *figure 17*.


Barcode Name	Sample	Output	%>=Q20	Reads	Mean Read Length	Read Length Histogram	BAM
No barcode	E2575-p7	32.6M	20.1M	408254	80 bp		BAM BAI

Figure 17 - ION TORRENT INTERFACE

(2) (3) Conversion of the BAM file to FASTQ format using BAMTOFAST and into its forward and reverse reads counterparts. Snippet of the script as shown in *figure 18*.

```
bamtofq(){
echo "BAM to FASTQ"
echo
bamToFastq -i "${folder_id}${sample_id}.bam" -fq "${folder_id}${sample_id}.fastq"
reformat.sh in="${folder_id}${sample_id}.fastq" out1="${folder_id}${sample_id}_R1.fastq" out2="${folder_id}${sample_id}_R2.fastq"
echo
echo "BAMTOFASTQ DONE && FORWARD AND REVERSE FASTQ FILES CREATED!"
}
```

Figure 18 - SNIPPET OF BAM TO FASTQ & ONE TO TWO FASTQ SCRIPT

(3) Perform quality control on both files. Snippet of the script as shown in *figure 19*.

```
quality(){
echo "Quality Check"
fastq_quality_filter -Q 33 -q 20 -p 75 -i "${folder_id}${sample_id}_R1.fastq" -o "${folder_id}${sample_id}_tq_R1.fastq"
fastq_quality_filter -Q 33 -q 20 -p 75 -i "${folder_id}${sample_id}_R2.fastq" -o "${folder_id}${sample_id}_tq_R2.fastq"
echo
echo "QUALITY CHECK DONE!"
}
```

Figure 19 - - SNIPPET OF QUALITY CONTROL SCRIPT

(4) (6) Classify the sample using Kraken2 + Bracken combination. Snippet of the script as shown in *figure 20*.

```
krakbrak(){
echo "Kraken + Bracken"
kraken2 --threads 12 --db ${guthealth} --confidence ${threshold_num} --paired "${folder_id}${sample_id}_tq_R1.fastq" "${folder_id}${sample_id}_tq_R2.f
${folder_id}${sample_id}.report"
echo
echo "KRAKEN ANALYSIS DONE!"
sleep 2
bracken -d ${guthealth} -i "${folder_id}${sample_id}.report" -r 200 -t 5 -o "${folder_id}${sample_id}.bracken"
echo
echo "BRACKEN ANALYSIS DONE!"
echo
sleep 2
echo "KRAKEN+BRACKEN DONE!"
}
```

Figure 20 – SNIPPET OF CLASSIFICATION SCRIPT

(5) Convert the kraken report to biom format using KRAKEN-BIOM. Snippet of the script as shown in *figure 21*.

```
brakbiom(){
echo "Bracken to BIOM"
echo
kraken-biom "${folder_id}${sample_id}_bracken.report" -o "${folder_id}${sample_id}.biom" --fmt hdf5
sleep 2
echo "CONVERSION DONE!"
}
```

Figure 21 – SNIPPET OF BIOM CONVERSION SCRIPT

(7) To extract the population and sample metrics from the population and sample .biom files I chose to use R and created a script called **meta_metrics_2.R**. It generates all values and graphs to be used in the pdf report. Snippet of the script as shown in *figure 22*.

```
rmetrics(){
eval "$(conda shell.bash hook)"
conda activate guthealth2

cd metrics/src
echo "Report Metrics"
Rscript meta_metrics_2.R -s "${sample_id}" -b "${folder_id}${sample_id}.biom"
echo
sleep 2
echo "REPORT METRICS DONE!"
}
```

Figure 22 – SNIPPET OF META METRICS CREATION SCRIPT

It uses the R libraries shown in *table 5*.

Table 5 - TABLE WITH THE PYTHON LIBRARIES OF META_METRICS.R

LIBRARY NAME	FUNCTION
GETOPT	Parsing Unix Command Line Options
BIOMFORMAT	Interfacing With The BIOM Format
PHYLOSEQ	Set Of Classes And Tools To Facilitate The Import, Storage, Analysis, And Graphical Display Of Microbiome Census Data
APE	Functions For Reading, Writing, Plotting, And Manipulating Phylogenetic Trees, Analyses Of Comparative Data In A Phylogenetic Framework
DPLYR	Grammar Of Data Manipulation
GGALT	New Geometries, Coordinate Systems, Statistical Transformations, Scales And Fonts For 'Ggplot2'
TIDYVERSE	Meant To Load Ggplot2
EBMC	Ensemble-Based Methods For Class Imbalance Problem (For Machine Learning)
CARET	Set Of Functions That Attempt To Streamline The Process For Creating Predictive Models.
COWPLOT	Add-On To Ggplot And Functions That Make It Easy To Annotate Plots And Or Mix Plots With Images.
GRIDEXTRA	Arrange Multiple Graphs In A Grid-Like Format
MAGICK	Open-Source Image Processing
GGIMAGE	Ggplot2 Equivalent Of Image
RETICULATE	Comprehensive Set Of Tools For Interoperability Between Python And R
METACODER	Tools For Parsing, Manipulating, And Graphing Taxonomic Abundance Data
JSONLITE	JSON Parser

The data created is exported in JSON format and its related graphs are in PNG format.

- (9) To create the PDF report I resorted in creating a Python script that I called **meta_report.py**. Snippet of the script as shown in *figure 23*.

```

creport(){
conda activate base
cd ..
cd ..
cd report
echo "Create Report"
python3 -m src.meta_report -i "${sample_id}" -f "${folder_id}"
echo
sleep 2
echo "FINAL REPORT CREATED!"
pause
}

```

Figure 23 - SNIPPET OF META REPORT CREATION SCRIPT

It uses the PYTHON libraries in table 6:

Table 6 - TABLE WITH THE PYTHON LIBRARIES OF META_REPORT.PY

LIBRARY NAME	FUNCTION
ARGPARSE	Makes It Easy To Write User-Friendly Command-Line Interfaces.
PANDAS	Easy-To-Use Data Structures And Data Analysis Tools
JSON	JSON Encoder And Decoder
REPORTLAB	An Open Source Python Library For Generating Pdfs And Graphics.

4.5 PHASE 4

This phase was crucial as our microbiome test had to be able to correctly profile a sample and assign it to its specific group (Underweight, Normal, Overweight or Obese).

This proved to be elusive, given that one or more individual factors might point out to one group while other factors might favor another group. To mitigate this, I researched Machine Learning algorithms for classification, and based my decision on one article evaluation of these methods^[49] and a pipeline already created for microbiome data^[50].

Considering the time constraints and available pipelines, I decided to use a RandomForest algorithm that took into account all abundances and created a predictive model. It had an error rate of 29% and it was implemented.

However, this error rate of 29% meant that there was room for improvement and during my research on Machine Learning algorithms I learned about dealing with unbalanced datasets^[51].

The population datasets used were uneven in size. (i.e, the group Normal has over 1400 samples whereas the group Obese has a little over 130 samples.)

I was able to eliminate the bias that came with unbalanced datasets by implementing an oversampling technique called SMOTE (Synthetic Minority Over-Sampling Technique). I created a new RandomForest prediction algorithm that included all abundances and also the metrics for Diversity, Richness and Evenness. This new algorithm had over 97% accuracy rate for all population types.

For this I used the R libraries called ebmc^[52] (Ensemble-Based Methods for Class Imbalance Problem) and caret^[53] (Classification and Regression Training). This algorithm is present in the **meta_metrics_2.R** script and it is implemented.

5. DISCUSSION

During the development and implementation of our test, we came across several issues that for now could not be solved with our currently implemented bioinformatics methods. While such issues were to be expected and do not represent a major drawback and thus allowed the product to be commercialized, it will soon require our attention and further implementations to solve them. These issues are:

ESCHERICHIA COLI

We noticed that E-coli was not present in the analyzed microbiomes despite the fact that, even being present in low abundance, it is a regular organism that lives in the human gut. After a deep investigation we found that the current usage of 16S rRNA gene sequencing methods to profile the microbiome can lead to underrepresentation, when doing a metagenomics sequencing run. This meant removing this taxon from the reported list until we can sequence it by using alternative methods.

SHIGELLA OR NOT SHIGELLA

We also came across 2 samples that were over representing Shigella. Taking a closer look at the classified data, and after using BLASTN of the sequences, I discovered that the sequences were being mislabeled. Deleting these mislabeled IDs solved this particular problem.

GRAPH LITERACY

During our interactions with potential or actual customers we became aware that understanding the data behind a graph might not always be present with our customers. We felt the need to change the way data was presented. We changed from the representation in *figure 24* to *figure 25*.

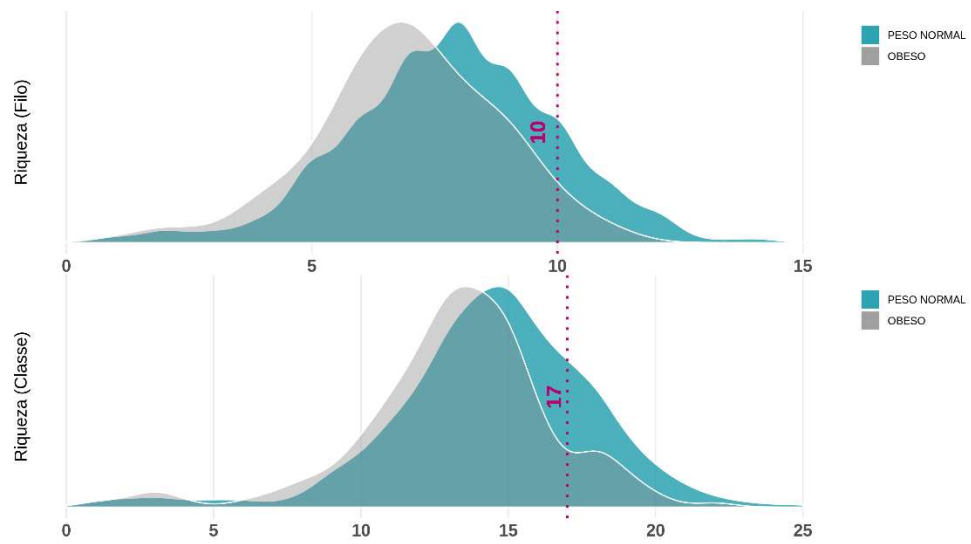


FIGURE 24 – VERSION 1 OF RICHNESS REPRESENTATION OF PHYLUM AND CLASS RANKS

Riqueza ao nível taxonómico: FILO



Riqueza ao nível taxonómico: CLASSE



FIGURE 25 – VERSION 2 OF RICHNESS REPRESENTATION OF PHYLUM AND CLASS RANKS

6. RESULTS

As a result of this project and of the work carried out by our laboratory, the microbiome test is now a finished product. It has 2 main groups (global and individual metrics) and currently it has 4 sections.

Its main formula is metrics comparison between the sample, its assigned group (by our classification algorithm) and the Normal reference population. (NOTE. If the assigned group is Normal, then the second reference group will be Obese)

The sections are:

SUMMARY

In this section we characterize the main global metrics, we identify any individual metric that isn't within normal parameters and, if we have found any pathogenic bacteria in an individual's sample we flag it.

GLOBAL

In this section we expand on the global metrics. These are Microbiome Diversity, Richness, Firmicutes/Bacteroidetes Ratio and a global Phylogenetic tree.

INDIVIDUAL

In this section we provide insights on the individual bacteria previously chosen for this test and its abundance within the sample.

BACTERIA INVENTORY

In this section we provide a table listing all bacteria found in the sample as well as its abundance.

I have attached a sample GUTHEALTH report in the appendix section of this report.

7. CONCLUSIONS

The main objective of creating a market-ready sequencing-based clinical microbiome test, related to obesity without the need of human intervention in the processes of analysis and reporting was achieved. The product is now being commercialized by CMLGS actively.

I have found the field of Metagenomics and Microbiota to be flourishing with opportunities of research making Bioinformatics indispensable and, by association, the role of the bioinformatician.

My role in this project evolved from a bioinformatics developer and intern to product manager where I'm able to provide insights based on the data.

I also want to point out that I understood the importance of being amongst a great team of professionals. A bioinformatician's work does not depend only on himself but also on the quality of the data and, by association, on the quality of the team around him.

Lastly, reflecting on my own progress during this project, I can safely say that today I'm much a more accomplished programmer, capable of implementing my own bioinformatics pipelines, and that my understanding of the role of a Bioinformatician and the field of Metagenomics is much greater.

Since this is a field that I wish to pursue, having the opportunity to work on this project meant a great deal to me.

8. LITERATURE CITED

- [1] J. R. Marchesi and J. Ravel, 'The vocabulary of microbiome research: a proposal', *Microbiome*, vol. 3, no. 1, pp. 31, s40168-015-0094-5, Dec. 2015.
- [2] The NIH HMP Working Group *et al.*, 'The NIH Human Microbiome Project', *Genome Research*, vol. 19, no. 12, pp. 2317-2323, Dec. 2009.
- [3] R. Sender, S. Fuchs, and R. Milo, 'Revised Estimates for the Number of Human and Bacteria Cells in the Body', *PLoS Biol*, vol. 14, no. 8, p. e1002533, Aug. 2016.
- [4] 'Why the Gut Microbiome Is Crucial for Your Health', *Healthline*. [Online]. Available: <https://www.healthline.com/nutrition/gut-microbiome-and-health>. [Accessed: 25-Nov-2019].
- [5] K. Aagaard, J. Ma, K. M. Antony, R. Ganu, J. Petrosino, and J. Versalovic, 'The Placenta Harbors a Unique Microbiome', *Science Translational Medicine*, vol. 6, no. 237, pp. 237ra65-237ra65, May 2014.
- [6] The Human Microbiome Project Consortium, 'Structure, function and diversity of the healthy human microbiome', *Nature*, vol. 486, no. 7402, pp. 207-214, Jun. 2012.
- [7] F. Godoy-Vitorino, 'Human microbial ecology and the rising new medicine', *Ann. Transl. Med.*, vol. 7, no. 14, pp. 342-342, Jul. 2019.
- [8] J. Slavin, 'Fiber and Prebiotics: Mechanisms and Health Benefits', *Nutrients*, vol. 5, no. 4, pp. 1417-1435, Apr. 2013.
- [9] D. Ríos-Covián, P. Ruas-Madiedo, A. Margolles, M. Gueimonde, C. G. de los Reyes-Gavilán, and N. Salazar, 'Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health', *Front. Microbiol.*, vol. 7, Feb. 2016.
- [10] A. L. Kau, P. P. Ahern, N. W. Griffin, A. L. Goodman, and J. I. Gordon, 'Human nutrition, the gut microbiome and the immune system', *Nature*, vol. 474, no. 7351, pp. 327-336, Jun. 2011.
- [11] J. F. Cryan and T. G. Dinan, 'Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour', *Nat Rev Neurosci*, vol. 13, no. 10, pp. 701-712, Oct. 2012.
- [12] J. G. Mulle, W. G. Sharp, and J. F. Cubells, 'The Gut Microbiome: A New Frontier in Autism Research', *Curr Psychiatry Rep*, vol. 15, no. 2, p. 337, Feb. 2013.
- [13] M. W.-S. W. 24 days ago Health, 'How a Man's Fecal Transplant Turned Fatal', *livescience.com*. [Online]. Available: <https://www.livescience.com/fecal-transplant-death.html>. [Accessed: 25-Nov-2019].
- [14] E. A. Eloe-Fadrosh and D. A. Rasko, 'The Human Microbiome: From Symbiosis to Pathogenesis', *Annu. Rev. Med.*, vol. 64, no. 1, pp. 145-163, Jan. 2013.
- [15] M. Levy, A. A. Kolodziejczyk, C. A. Thaiss, and E. Elinav, 'Dysbiosis and the immune system', *Nat Rev Immunol*, vol. 17, no. 4, pp. 219-232, Apr. 2017.
- [16] D. Hadrach, 'Microbiome Research Is Becoming the Key to Better Understanding Health and Nutrition', *Front. Genet.*, vol. 9, p. 212, Jun. 2018.
- [17] E. Patterson *et al.*, 'Gut microbiota, obesity and diabetes', *Postgrad Med J*, vol. 92, no. 1087, pp. 286-300, May 2016.
- [18] R. Jumpertz *et al.*, 'Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans', *The American Journal of Clinical Nutrition*, vol. 94, no. 1, pp. 58-65, Jul. 2011.
- [19] MetaHIT consortium *et al.*, 'Richness of human gut microbiome correlates with metabolic markers', *Nature*, vol. 500, no. 7464, pp. 541-546, Aug. 2013.
- [20] M. A. Sze and P. D. Schloss, 'Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome', *mBio*, vol. 7, no. 4, pp. e01018-16, /mbio/7/4/e01018-16.atom, Sep. 2016.

-
- [21] O. Castaner *et al.*, 'The Gut Microbiome Profile in Obesity: A Systematic Review', *International Journal of Endocrinology*, vol. 2018, pp. 1–9, 2018.
- [22] T. Thomas, J. Gilbert, and F. Meyer, 'Metagenomics - a guide from sampling to data analysis', *Microb Informatics Exp*, vol. 2, no. 1, p. 3, Dec. 2012.
- [23] 'Ion Personal Genome Machine™ (PGM™) System'. [Online]. Available: <http://www.thermofisher.com/order/catalog/product/4462921>. [Accessed: 26-Nov-2019].
- [24] J. M. Janda and S. L. Abbott, '16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls', *Journal of Clinical Microbiology*, vol. 45, no. 9, pp. 2761–2764, Sep. 2007.
- [25] T. Větrovský and P. Baldrian, 'The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses', *PLoS ONE*, vol. 8, no. 2, p. e57923, Feb. 2013.
- [26] M. Balvočiūtė and D. H. Huson, 'SILVA, RDP, Greengenes, NCBI and OTT – how do these taxonomies compare?', *BMC Genomics*, vol. 18, no. S2, p. 114, Mar. 2017.
- [27] H. Li *et al.*, 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.
- [28] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Research*, vol. 38, no. 6, pp. 1767–1771, Apr. 2010.
- [29] D. McDonald *et al.*, 'The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome', *GigaSci*, vol. 1, no. 1, p. 7, Dec. 2012.
- [30] P. J. McMurdie and S. Holmes, 'phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data', *PLoS ONE*, vol. 8, no. 4, p. e61217, Apr. 2013.
- [31] 'JSON'. [Online]. Available: <https://json.org/>. [Accessed: 26-Nov-2019].
- [32] 'PNG (Portable Network Graphic) Definition'. [Online]. Available: <https://techterms.com/definition/png>. [Accessed: 26-Nov-2019].
- [33] 'R: The R Project for Statistical Computing'. [Online]. Available: <https://www.r-project.org/>. [Accessed: 26-Nov-2019].
- [34] 'Welcome to Python.org', *Python.org*. [Online]. Available: <https://www.python.org/about/>. [Accessed: 26-Nov-2019].
- [35] 'What is Java? - Definition from Techopedia'. [Online]. Available: <https://www.techopedia.com/definition/3927/java>. [Accessed: 26-Nov-2019].
- [36] 'What is a Shell Script? - Definition from Techopedia'. [Online]. Available: <https://www.techopedia.com/definition/9341/shell-script>. [Accessed: 26-Nov-2019].
- [37] 'Picard Tools - By Broad Institute'. [Online]. Available: <https://broadinstitute.github.io/picard/>. [Accessed: 26-Nov-2019].
- [38] 'bamtofastq - bedtools 2.29.0 documentation'. [Online]. Available: <https://bedtools.readthedocs.io/en/latest/content/tools/bamtofastq.html>. [Accessed: 26-Nov-2019].
- [39] 'USADELLAB.org - Trimmomatic: A flexible read trimming tool for Illumina NGS data'. [Online]. Available: <http://www.usadellab.org/cms/?page=trimmomatic>. [Accessed: 26-Nov-2019].
- [40] 'User guide - Cutadapt 2.7 documentation'. [Online]. Available: <https://cutadapt.readthedocs.io/en/stable/guide.html>. [Accessed: 26-Nov-2019].
- [41] 'FASTX-Toolkit - Command Line Usage'. [Online]. Available: http://hannonlab.cshl.edu/fastx_toolkit/commandline.html. [Accessed: 26-Nov-2019].
- [42] 'Reformat Guide', *DOE Joint Genome Institute*. [Online]. Available: <https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/reformat-guide/>. [Accessed: 26-Nov-2019].

-
- [43] L. Siegwald, H. Touzet, Y. Lemoine, D. Hot, C. Audebert, and S. Caboche, 'Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics', *PLoS ONE*, vol. 12, no. 1, p. e0169563, Jan. 2017.
- [44] 'Kraken2'. [Online]. Available: <https://ccb.jhu.edu/software/kraken2/>. [Accessed: 26-Nov-2019].
- [45] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg, 'Bracken: estimating species abundance in metagenomics data', *PeerJ Computer Science*, vol. 3, p. e104, Jan. 2017.
- [46] S. Dabdoub, *smdabdoub/kraken-biom*. 2019.
- [47] 'ZymoBIOMICS Microbial Community Standards', ZYMO RESEARCH. [Online]. Available: <https://www.zymoresearch.com/collections/zymbiomics-microbial-community-standards>. [Accessed: 26-Nov-2019].
- [48] 'American Gut Project'. [Online]. Available: <https://microbio.me/AmericanGut/introduction/>. [Accessed: 26-Nov-2019].
- [49] A. Statnikov *et al.*, 'A comprehensive evaluation of multicategory classification methods for microbiomic data', *Microbiome*, vol. 1, no. 1, p. 11, Dec. 2013.
- [50] M. Berry, *michberr/randomforest-microbe*. 2019.
- [51] Shuo Wang and Xin Yao, 'Multiclass Imbalance Problems: Analysis and Potential Solutions', *IEEE Trans. Syst., Man, Cybern. B*, vol. 42, no. 4, pp. 1119-1130, Aug. 2012.
- [52] 'ebmc: Ensemble-Based Methods for Class Imbalance Problem version 1.0.0 from CRAN'. [Online]. Available: <https://rdrr.io/cran/ebmc/>. [Accessed: 26-Nov-2019].
- [53] M. Kuhn, *The caret Package*.

9. APPENDICES

TABLE & FIGURE INDEX

Figure 1 - Representation of the 16s rRNA gene. Digital Image. The Ishaq Lab. May 8 2016. https://sueishaqlab.org/tag/16s-rrna/	17
Figure 2 - Overview of 5 databases mentioned and their taxonomic classifications (table taken from “SILVA, RDP, Greengenes, NCBI and OTT–how do these taxonomies compare?”)....	18
Figure 3 - BAMTOFASTQ STANDARD USAGE. NOTE: OPTIONS DISPLAYED ARE THE ONES USED ON THIS REPORT.....	24
Figure 4 - FASTQ_QUALITY_TRIMMER USAGE. NOTE: OPTIONS DISPLAYED ARE THE ONES USED ON THIS REPORT.....	24
Figure 5 - REFORMAT USAGE; NOTE: OPTIONS DISPLAYED ARE THE ONES USED ON THIS REPORT....	25
Figure 6 - DISTINCTIONS BETWEEN CLUSTERING–FIRST AND ASSIGNMENT–FIRST APPROACHES. (image taken from “Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics”).....	26
Figure 7 - KRAKEN2 BUILD USAGE FOR 16S DB TYPES.....	27
Figure 8 - KRAKEN2 USAGE. NOTE: OPTIONS DISPLAYED ARE THE ONES USED IN THIS REPORT.....	28
Figure 9 - BRACKEN BUILD USAGE FOR KRAKEN2.....	28
Figure 10 - BRACKEN USAGE ; NOTE: OPTIONS DISPLAYED ARE THE ONES USED IN THIS REPORT...	29
Figure 11 - KRAKEN-BIOM USAGE.....	29
Figure 12 - GUTHEALTH DB CREATION PIPELINE.....	31
Figure 13 - SNIPPET OF KRAKEN CREATE DB GUTHEALTH SCRIPT.....	32
Figure 14 - POPULATION BIOM FILE PIPELINE.....	33
Figure 15 - SNIPPET OF BRACKEN_TO_BIOM SCRIPT.....	35
Figure 16 - REPORT CREATION PIPELINE.....	36
Figure 17 - ION TORRENT INTERFACE.....	37
Figure 18 - SNIPPET OF BAM TO FASTQ & ONE TO TWO FASTQ SCRIPT.....	37
Figure 19 - - SNIPPET OF QUALITY CONTROL SCRIPT.....	37
Figure 20 - SNIPPET OF CLASSIFICATION SCRIPT.....	38
Figure 21 - SNIPPET OF BIOM CONVERSION SCRIPT.....	38
Figure 22 - SNIPPET OF META METRICS CREATION SCRIPT.....	38
Figure 23 - SNIPPET OF META REPORT CREATION SCRIPT.....	40
FIGURE 24 - VERSION 1 OF RICHNESS REPRESENTATION OF PHYLUM AND CLASS RANKS.....	43
FIGURE 25 - VERSION 2 OF RICHNESS REPRESENTATION OF PHYLUM AND CLASS RANKS.....	43
Table 1 - TABLE WITH CHOSEN BACTERIA.....	30
Table 2 - TABLE WITH PYTHON LIBRARIES USED IN IMPORT_AG_DATA.PY.....	34
Table 3 - *ALL SAMPLES THAT HAD DISEASES SUCH AS IBS, IBD, DIABETES WERE FILTERED OUT..	34
Table 4 - TABLE WITH THE PYTHON LIBRARIES OF BRACKEN_TO_BIOM.PY.....	35
Table 5 - TABLE WITH THE PYTHON LIBRARIES OF META_METRICS.R.....	39
Table 6 - TABLE WITH THE PYTHON LIBRARIES OF META_REPORT.PY.....	40

IDENTIFICAÇÃO DO UTENTE

Nome:

Data de Nascimento:

Género:

Idade:

AMOSTRA

Identificação da amostra:

Tipo de amostra: Fezes

Data de receção:

Data de emissão: 14-11-2019

REQUISITANTE

Nome:

Médico

Nutricionista

Outro

Sumário dos Resultados

Composição

- Uma análise estatística da composição global desta amostra identificou-a como mais próxima de **NORMOPONDERAL**, assemelhando-se a uma população de indivíduos saudáveis de referência.

De todos os **organismos** investigados que podem ser modificados pela dieta ^[1], apenas os seguintes apresentam frequência fora do intervalo de valores de uma população de indivíduos saudáveis de referência:

- A presença de *Roseburia* encontra-se em valores aumentados/diminuídos relativamente a uma população normal de referência.
- A presença de *Butyrivibrio* encontra-se em valores aumentados/diminuídos relativamente a uma população normal de referência.
- A presença de *Bilophila* encontra-se em valores aumentados/diminuídos relativamente a uma população normal de referência.
- A presença de *Akkermansia muciniphila* encontra-se em valores aumentados/diminuídos relativamente a uma população normal de referência.

Todos os outros géneros e espécies analisados encontram-se em frequências semelhantes a uma população de indivíduos saudáveis de referência.

- Não foram identificados na amostra os seguintes organismos patogénicos investigados:
Salmonella enterica, Campylobacter, Clostridium difficile, Shigella, Vibrio Cholerae

Medidas Globais

- O teste GUT HEALTH V1.0 indica que o microbioma intestinal representado pela amostra analisada tem um índice de **diversidade** de 2.88 e não se assemelha à população de indivíduos saudáveis de referência.
- A **riqueza** de grupos biológicos (filó e classe) assemelha-se à população de indivíduos saudáveis de referência.
- O **Rácio Firmicutes/Bacteroidetes** de **1.051** assemelha-se à população de indivíduos saudáveis de referência.



O teste Gut Health(TM) resulta de uma colaboração entre o Centro de Medicina Laboratorial Germano de Sousa, através da sua marca Lifestyle Genomics, e o grupo de "Nutrição e Metabolismo" da Faculdade de Ciências Médicas da Universidade NOVA de Lisboa. As primeiras desenvolvem e implementam a componente laboratorial e bioinformática do teste, assegurando o cumprimento de todos os critérios de qualidade, segurança e rigor do resultado, enquanto a segunda desenvolve a interpretação clínica dos resultados.

Análises Globais

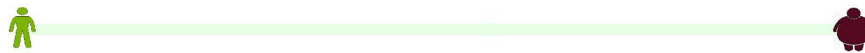


Diversidade

A disbiose é uma alteração da composição e função do microbioma intestinal que pode estar na base de diversas patologias como as doenças inflamatórias intestinais e a obesidade. A expansão de patobiontes (microrganismos comensais que podem causar patologias quando a sua proliferação está descontrolada), o decréscimo das bactérias benéficas e a perda de diversidade (abundância e riqueza de espécies) são as características mais comuns de um estado de disbiose.

A medida de Diversidade usada aqui é o Índice de Shannon.

A análise desta amostra revela um índice de diversidade de **2.88**. Este encontra-se mais distante do intervalo de valores de uma população de indivíduos saudáveis de referência.

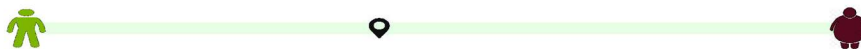


Riqueza

A riqueza de diferentes microorganismos no microbioma intestinal indica o potencial que este tem para se adaptar a diferentes condições. Um desequilíbrio do microbioma, com dominância de espécies com efeitos nocivos ou perda de espécies com efeitos protectores, é frequentemente acompanhado de uma redução da Riqueza de espécies.

A análise desta amostra revelou **11** filós e **27** classes diferentes, o que indica uma riqueza de **filós** e de **classes** mais próxima do intervalo de valores de uma população de indivíduos saudáveis de referência.

Riqueza ao nível taxonómico: FILO



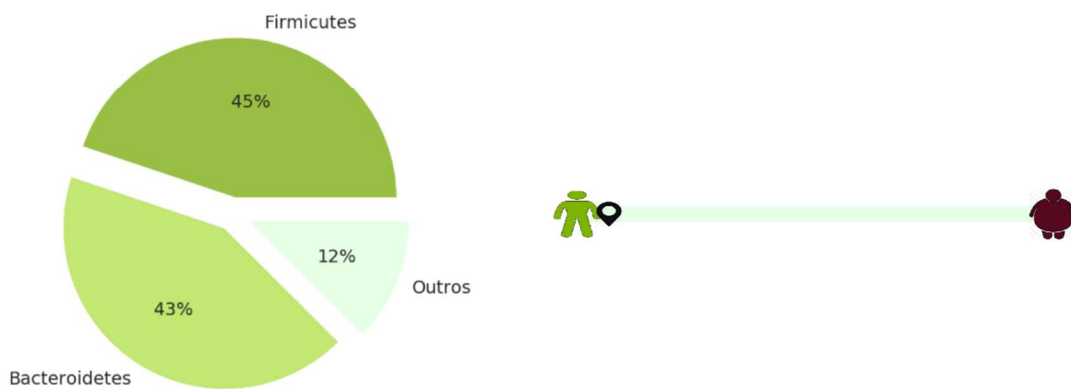
Riqueza ao nível taxonómico: CLASSE



✓ **Rácio Firmicutes/Bacteroidetes**

O microbioma intestinal é constituído maioritariamente por bactérias pertencentes aos filos Firmicutes e Bacteroidetes. O rácio entre Firmicutes e Bacteroidetes está aumentado na obesidade.

Nesta amostra o valor determinado foi de **1.051**. Este encontra-se mais próximo do intervalo de valores de uma população de indivíduos saudáveis de referência.

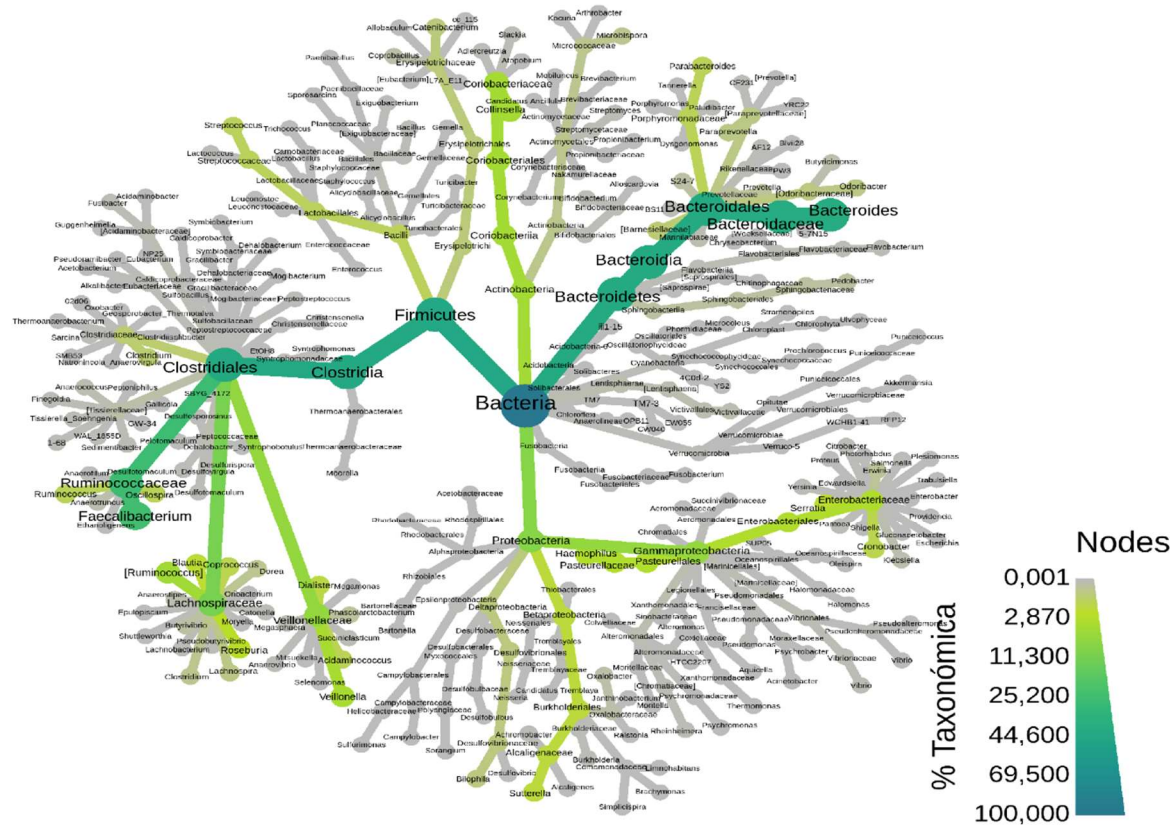


Géneros e Espécies investigadas

Encontraram-se 42 géneros e 35 espécies bacterianas diferentes nesta amostra, representadas globalmente na árvore filogenética em baixo.

Descrevem-se em baixo a abundância de géneros e espécies bacterianas escolhidos para integrar este relatório. Estes organismos foram seleccionados pela existência de evidência científica clara, relevância clínica e cuja abundância seja manipulável através de intervenção alimentar.

Para cada género e espécie apresenta-se a sua relevância fisiológica e patológica e compara-se a abundância do microbioma representado por esta amostra com a abundância numa população de indivíduos saudáveis de referência. Os gráficos seguintes representam o resultado da amostra no contexto da distribuição de valores dessa mesma população.



Visão global dos taxa bacterianos encontrados nesta amostra. A árvore representada ilustra a diversidade de géneros identificados na análise. A cor e dimensão relativa representa a frequência (ver legenda ao lado).



Uma análise estatística da composição global desta amostra identificou-a como mais próxima de **NORMOPONDERAL**, assemelhando-se a uma população de indivíduos saudáveis de referência.



Análise detalhada por grupo taxonómico

✔ *Blautia* (Gram-positivo; Anaeróbio estrito)

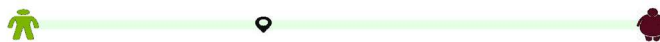
O género de *Blautia* é importante para a assimilação de nutrientes. Este encontra-se aumentado na obesidade^[3].



A presença de *Blautia* corresponde a **1.71%** da amostra. A amostra encontra-se próxima da população norma de referência

✔ *Bifidobacterium* (Gram-positivo; Anaeróbio estrito)

O género de *Bifidobacterium* produz ácidos gordos de cadeia curta, melhora a barreira intestinal e diminui os níveis de LPS (do inglês, Lipopolysaccharide) no intestino. Este género está diminuído na obesidade^[2] e uma dieta rica em fibra estimula o crescimento destas bactérias^[19].



A presença de *Bifidobacterium* corresponde a **0.01%** da amostra. A amostra encontra-se próxima da população norma de referência

✔ *Bacteroides* (Gram-negativo; Anaeróbio estrito)

O género de *Bacteroides* ativa as células T CD4 + e a sua presença está aumentada na doença inflamatória intestinal^[9,10]. Uma dieta rica em gordura saturada e em proteína de origem animal está associada a um aumento destas bactérias^[19].



A presença de *Bacteroides* corresponde a **40.3%** da amostra. A amostra encontra-se próxima da população norma de referência

✘ *Roseburia* (Gram-variável; Anaeróbio estrito)

O género de *Roseburia* produz ácidos gordos de cadeia curta e está diminuído na doença inflamatória intestinal^[7].



A presença de *Roseburia* corresponde a **2.25%** da amostra. A amostra revela valores aumentados/diminuídos relativamente a uma população normal de referência

✘ **Butyrivibrio** (Gram-positivo; Anaeróbio estrito)

O género de *Butyrivibrio* produz butirato. Este impede o ganho de peso^[4].



A presença de *Butyrivibrio* corresponde a **0.07%** da amostra. A amostra revela valores aumentados/diminuídos relativamente a uma população normal de referência

✔ **Prevotella** (Gram-negativo; Anaeróbio estrito)

O género de *Prevotella* produz ácidos gordos de cadeia curta e está associado a um melhor perfil cardiometabólico (colesterol LDL mais baixo)^[8]. Uma dieta rica em fibra estimula o crescimento destas bactérias^[19].



A presença de *Prevotella* corresponde a **0.1%** da amostra. A amostra encontra-se próxima da população norma de referência

✔ **Lactobacillus** (Gram-positivo; Anaeróbio facultativo)

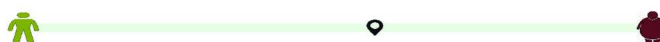
O género de *Lactobacillus* produz ácidos gordos de cadeia curta e têm atividade anti-inflamatória e anti-carcinogénica. Estes atenuam a doença inflamatória intestinal^[6]. E estão presentes nos alimentos fermentados como o iogurte, queijo e kefir^[19].



A presença de *Lactobacillus* corresponde a **< 0.01%** da amostra. A amostra encontra-se próxima da população norma de referência

✘ **Bilophila** (Gram-negativo; Anaeróbio estrito)

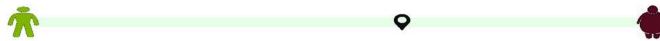
O género de *Bilophila* é resistente aos ácidos biliares e têm atividade pró-inflamatória. A *Bilophila wadsworthia* está associada à colite e à colecistite^[13]. Uma dieta rica em gordura saturada e em proteína de origem animal está associada a um aumento destas bactérias^[19].



A presença de *Bilophila* corresponde a **0.35%** da amostra. A amostra revela valores aumentados/diminuídos relativamente a uma população normal de referência

✘ ***Akkermansia muciniphila*** (Gram-negativo; Anaeróbio estrito)

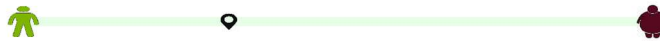
A espécie de *Akkermansia muciniphila* tem atividade anti-inflamatória. Esta está diminuída na doença inflamatória intestinal e na obesidade^[15].



A presença de *Akkermansia muciniphila* corresponde a < 0.01% da amostra. A amostra revela valores aumentados/diminuídos relativamente a uma população normal de referência

✔ ***Faecalibacterium Prausnitzii*** (Gram-positivo; Anaeróbio estrito)

A espécie de *Faecalibacterium Prausnitzii* produz ácidos gordos de cadeia curta e têm atividade anti-inflamatória. Esta está diminuída na doença inflamatória intestinal e na obesidade^[5].



A presença de *Faecalibacterium prausnitzii* corresponde a 24.41% da amostra. A amostra encontra-se próxima da população norma de referência

Metodologia e Limitações

O teste foi realizado de acordo com todas as recomendações de qualidade em vigor, em laboratório licenciado para Análises Clínicas, incluindo Patologia Molecular e para Genética Médica, podendo assim ser alvo de interpretação em contexto clínico.

Foi utilizada como população comparativa com a amostra presente neste teste, uma base de dados de 1342 amostras. As mesmas foram selecionadas do projecto American Gut (www.americangut.org) e filtradas de modo a apresentarem valores saudáveis e não saudáveis de referência de acordo com o seu IMC (Índice de Massa Corporal). Cada amostra foi caracterizada pelo conjunto de sequências de 16s rRNA da região V4 e a sua taxonomia foi identificada com acesso a uma base de dados revista '16S rRNA' + 'Greengenes 13_8' nomeada de GutHealth_DB.

Para efeitos de classificação da amostra como NORMOPONDERAL/NÃO NORMOPONDERAL foi utilizado um algoritmo de *machine learning* de classificação (Random Forests)^[18], tendo como base os valores saudáveis e não saudáveis de referência descritos em cima.

Foram selecionados para esta análise detalhada apenas organismos que podem ser modificados pela dieta^[1]:

Ao nível taxonómico de Género são: *Bacteroides*, *Bifidobacterium*, *Bilophila*, *Blautia*, *Butyrivibrio*, *Lactobacillus*, *Prevotella*, *Ruminococcus*, *Roseburia*.

Ao nível taxonómico de Espécie são: *Akkermansia muciniphila*, *Faecalibacterium prausnitzii*, *Escherichia Coli*.

A amostra foi sequenciada por NGS (do inglês, Next Generation Sequencing), utilizando a plataforma IonTorrent (Ion 16STM Metagenomics Kit). A taxonomia foi identificada com o programa Kraken^[16] e melhorada com o programa Bracken^[17] com recurso à base de dados 'GutHealth_DB'.

No GUT HEALTH reportam-se frequências de organismos na amostra presente, bem como algumas métricas derivadas destas. **Recomenda-se que discuta o significado destes resultados com o seu médico assistente, nutricionista ou outro profissional de saúde que requisitou este teste.**

Filtramos a amostra para excluir organismos com número de *reads* inferior ou igual a 10, pelo que a ausência de identificação de um organismo na presente amostra, não pode ser interpretada como um negativo pois este pode estar presente com valores inferiores aos limites de detecção da técnica; não estar correctamente representado nas bases de dados de referência; ou outros.

Géneros e Espécies encontradas

Apresentam-se em baixo todos os 42 géneros e 35 espécies que estão representadas por frequências superiores a 0,01%.

Organismos (Género)	%	Organismos (Espécie)	%
<i>Bacteroides</i>	40.3	<i>Faecalibacterium prausnitzii</i>	24.41
<i>Faecalibacterium</i>	24.41	<i>Bacteroides plebeius</i>	13.07
[<i>Ruminococcus</i>]	4.56	<i>Bacteroides fragilis</i>	9.6
<i>Collinsella</i>	4.14	<i>Bacteroides caccae</i>	8.74
<i>Veillonella</i>	3.8	<i>Collinsella aerofaciens</i>	4.14
<i>Haemophilus</i>	2.94	<i>Bacteroides ovatus</i>	4.07
<i>Roseburia</i>	2.25	<i>Bacteroides uniformis</i>	3.16
<i>Serratia</i>	1.86	<i>Haemophilus parainfluenzae</i>	2.94
<i>Blautia</i>	1.71	[<i>Ruminococcus</i>] <i>gnavus</i>	2.77
<i>Sutterella</i>	1.5	<i>Veillonella dispar</i>	2.51
<i>Dialister</i>	1.34	<i>Roseburia faecis</i>	2.25
<i>Cronobacter</i>	1.11	<i>Serratia marcescens</i>	1.86
<i>Coprococcus</i>	1.07	[<i>Ruminococcus</i>] <i>torques</i>	1.8
<i>Acidaminococcus</i>	0.96	<i>Bacteroides eggerthii</i>	1.45
<i>Oscillospira</i>	0.95	<i>Veillonella parvula</i>	1.29
<i>Parabacteroides</i>	0.94	<i>Blautia obeum</i>	1.24
<i>Ruminococcus</i>	0.91	<i>Cronobacter sakazakii</i>	1.11
<i>Streptococcus</i>	0.85	<i>Coprococcus eutactus</i>	1.07
<i>Clostridium</i>	0.52	<i>Parabacteroides distasonis</i>	0.94
<i>Catenibacterium</i>	0.37	<i>Streptococcus luteciae</i>	0.85
<i>Bilophila</i>	0.35	<i>Ruminococcus bromii</i>	0.79
<i>Paraprevotella</i>	0.31	<i>Clostridium perfringens</i>	0.52
<i>Odoribacter</i>	0.31	<i>Blautia producta</i>	0.47
<i>Lachnospira</i>	0.29	<i>Bilophila sp.</i>	0.35
<i>Microbispora</i>	0.2	<i>Microbispora rosea</i>	0.2
<i>Clostridium</i>	0.2	<i>Clostridium piliforme</i>	0.19
<i>Dorea</i>	0.19	<i>Dorea formicigenerans</i>	0.19
<i>Erwinia</i>	0.13	<i>Bacteroides barnesiae</i>	0.15
<i>Phascolarctobacterium</i>	0.1	<i>Ruminococcus flavefaciens</i>	0.11
<i>Prevotella</i>	0.1	<i>Pedobacter cryoconitis</i>	0.09
<i>Pedobacter</i>	0.09	<i>Bacteroides coprophilus</i>	0.07
<i>Butyrivibrio</i>	0.07	<i>Prevotella copri</i>	0.06
<i>Coprobacillus</i>	0.04	<i>Coprobacillus cateniformis</i>	0.04
<i>Oxalobacter</i>	0.04	<i>Oxalobacter formigenes</i>	0.04

Organismos (Género)	%
<i>Butyricimonas</i>	0.03
<i>Klebsiella</i>	0.03
<i>Anaerotruncus</i>	0.02
<i>Anaerostipes</i>	0.02
<i>Anaerococcus</i>	0.02
<i>Megasphaera</i>	0.01
<i>Streptomyces</i>	0.01
<i>Fusibacter</i>	0.01

Organismos (Espécie)	%
<i>Prevotella nigrescens</i>	0.03

Referências Bibliográficas

[1] - Singh, R. K. et al. Influence of diet on the gut microbiome and implications for human health. *J Transl Med* 15, 73, doi:10.1186/s12967-017-1175-y (2017).

[3] - Kasai, C. et al. Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterol* 15, 100, doi:10.1186/s12876-015-0330-2 (2015).

[5] - Miquel, S. et al. Faecalibacterium prausnitzii and human intestinal health. *Curr Opin Microbiol* 16, 255-261, doi:10.1016/j.mib.2013.06.003 (2013).

[7] - Eloe-Fadrosh, E. A. et al. Functional dynamics of the gut microbiome in elderly people during probiotic consumption. *MBio* 6, doi:10.1128/mBio.00231-15 (2015).

[9] - Lucke, K., Miehle, S., Jacobs, E. & Schuppler, M. Prevalence of *Bacteroides* and *Prevotella* spp. in ulcerative colitis. *J Med Microbiol* 55, 617-624, doi:10.1099/jmm.0.46198-0 (2006).

[11] - Mishra, S. & Imlay, J. A. An anaerobic bacterium, *Bacteroides thetaiotaomicron*, uses a consortium of enzymes to scavenge hydrogen peroxide. *Mol Microbiol* 90, 1356-1371, doi:10.1111/mmi.12438 (2013).

[13] - Baron, E. J. *Bilophila wadsworthemehantethia*: a unique Gram-negative anaerobic rod. *Anaerobe* 3, 83-86, doi:10.1006/anae.1997.0075 (1997).

[15] - van Passel, M. W. et al. The genome of *Akkermansia muciniphila*, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One* 6, e16876, doi: 10.1371/journal.pone.0016876 (2011).

[17] - Jennifer Lu, Florian P Breitwieser, Peter Thielen, Steven L Salzberg. Bracken: Estimating species abundance in metagenomics data. doi: 10.7717/peerj-cs.104 (2017).

[19] - Dong TS, Gupta A: Influence of Early Life, Diet, and the Environment on the Microbiome. *Clin Gastroenterol Hepatol* 2019;17:231–242.

[2] - Schwartz, A. et al. Microbiota and SCFA in lean and overweight healthy subjects. *Obesity (Silver Spring)* 18, 190-195, doi:10.1038/oby.2009.167 (2010).

[4] - Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* 500, 541-546, doi:10.1038/nature12506 (2013).

[6] - Venturi, A. et al. Impact on the composition of the faecal flora by a new probiotic preparation: preliminary data on maintenance treatment of patients with ulcerative colitis. *Aliment Pharmacol Ther* 13, 1103-1108 (1999).

[8] - de Moraes, A. C. et al. Enterotype May Drive the Dietary-Associated Cardiometabolic Risk Factors. *Front Cell Infect Microbiol* 7, 47, doi:10.3389/fcimb.2017.00047 (2017).

[10] - Prindiville, T. P. et al. *Bacteroides fragilis* enterotoxin gene sequences in patients with inflammatory bowel disease. *Emerg Infect Dis* 6, 171-174, doi:10.3201/eid0602.000210 (2000).

[12] - Haro, C. et al. The gut microbial community in metabolic syndrome patients is modified by diet. *J Nutr Biochem* 27, 27-31, doi:10.1016/j.jnutbio.2015.08.011 (2016).

[14] - Darfeuille-Michaud, A. et al. High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease. *Gastroenterology* 127, 412-421 (2004).

[16] - Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. doi: 10.1186/gb-2014-15-3-r46 (2014).

[18] - Leo Breiman. Random Forests. doi: 10.1023/A:1010933404324.