이학박사 학위논문

정확한 서열정렬기법과 인메모리 핵심 유전자 데이터베이스 기반의 향상된 메타유전체 분류법

Application of Exact Alignments with an In-memory Core Gene Database for an Improved Metagenomic Taxonomic Classification

2017 년 8 월

서울대학교 대학원

협동과정 생물정보학

마우리치오

Ph.D. Dissertation in Science

Application of Exact Alignments with an In-memory Core Gene Database for an Improved Metagenomic Taxonomic Classification

Advisor: Jongsik Chun, Ph. D.

August 2020

Graduate School of Biological Sciences
Seoul National University
Bioinformatics Major

Mauricio Antonio Chalita Williams

Application of Exact Alignments with an In-memory Core Gene Database for an Improved Metagenomic Taxonomic Classification

Advisor: Jongsik Chun, Ph. D.

Submitting a Ph.D. Dissertation in Science

May 2020

Graduate School of Biological Sciences
Seoul National University
Bioinformatics Major

Mauricio Antonio Chalita Williams

Confirming the Ph.D. Dissertation written by
Mauricio Antonio Chalita Williams
June 2020

Chair BYEONG J. LEE (Seal)
Vice Chair JONGSIK CHUN (Seal)
Examiner HYESHIK CHANG (Seal)
Examiner Young Woon Lim (Seal)
Examiner Hana Yi (Seal)

ABSTRACT

Shotgun metagenomics is of great importance to understand the microbial community composition of a sample and the impact it has on its host. The proper identification and quantification of bacterial species is a key component of any microbiome research that is based on metagenomic samples.

In the last decade, several algorithms and databases have been developed, however the differences between references and the type of algorithm used for the classification makes the comparisons among themselves unfair and bias. The contents of the reference database, including genome sequences of type strains or reference genomes of uncultured species, have a great impact on the performance of the classification results of metagenomic samples.

Another significant factor on shotgun metagenomics is the classification speed as most current bioinformatic tools lack computational and memory optimization. Here, I propose several enhancements to a well-known method, exact match k-mer classification in order to increase the overall speed of a metagenomic classification. This method was further improved by the use of Up-to-date Bacterial Core Gene (UBCG) sequences to provide better method for a faster and accurate shotgun metagenomic profiling classification.

In order to prove the efficiency of our method, I built two UBCG-based reference databases: one containing UBCG sequences of valid named species, and the second one containing UBCG sequences of all valid named species and

genomospecies in the EzBioCloud database. Three datasets containing *Streptococcus* species were used to evaluate the improved method against the MetaPhlan2 tool which is the most widely used open-source shotgun metagenomic classifier: (i) synthetic metagenomic samples, (ii) clinical sputum samples from patients with chronic obstructive pulmonary disease (COPD), and (iii) clinical samples of a blood stream infection.

In this analysis, I demonstrated that UBCG sequences can be used as references for metagenomic classification, showing that they are easy to extract from genome sequences and accurate when predicting relative abundance. I also showed that the inclusion of genomospecies in the reference databases, significantly improves the classification accuracy of bacterial species within a metagenomic sample. Finally, I showed that while publicly available pipelines and databases are easily accessible, for accurate and reliable taxonomic classification, an updated database with proper taxonomic and genomic curation must be used.

The method devised in this work is then applied to profile the *Bacteroides* species in over 4,000 shotgun metagenomic samples, which is one of most abundant members of the human gut microbiome. This task cannot be accomplished using conventional tools such as MetaPhlAn2 due to the high processing time they require. The results in this study showed that *Bacteroides* is high abundant in human samples from urban areas while being low abundant in humans from rural areas, particularly African and South American tribes. Countries showed dominance for a specific *Bacteroides* species, but this could

also be explained by the type of study were the samples came from. Mice samples showed the most diversity of *Bacteroides*, this can be attributed by the number of bacterial references isolated from this organism. House cat and dog samples showed correlation between each other, this may be attributed to the similarities of their lifestyle and diet.

This study shows the importance of having a great number of samples for any given metagenomic analysis, and even though, we have profiled thousands of samples, more might be needed in the future. The method proposed in this thesis demonstrates that core genes are reliable reference sequences for shotgun metagenomics. Their implementation as reference sequences in metagenomic databases improves the accuracy of the abundance prediction of any given sample. Additionally, with the use of a k-mer approach, this method's running time outperforms the most popular shotgun metagenomic tools.

The work presented in this thesis aims to help microbial research by providing faster and accurate metagenomic taxonomic predictions. Finally, with the ability of updating a metagenomic database with ease, will help researchers to obtain the most up-to-date results to find potential diagnosis or treatments for diseases associated to human microbial communities.

Keyword: Metagenome, Shotgun, K-mer, Exact match, *Streptococcus*, *Bacteroides*, Core Genes, Sequence classification.

Student Number: 2014-31487

TABLE OF CONTENTS

ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
rRNA	ribosomal RNA
NGS	Next-generation sequencing
DDH	DNA-DNA Hybridization
ANI	Average nucleotide identity
NCBI	National Center for Biotechnology Information
PCR	Polymerase chain reaction
CDS	Coding DNA sequence
BLAST	Basic Local Alignment Search Tool
UBCG	Up-to-date Bacterial Core Genes
AWS	Amazon Web Services
OTU	Operational taxonomic unit
CG	Comparative genomics
WGS	Whole genome sequencing

RAM	Random Access Memory
ORF	Open reading frame
POG	Pairwise orthologs group
SNV	Single nucleotide variant
SNV	Single nucleotide polymorphism
TNF	Tetra-nucleotide frequency
SRG	Species-specific reference genome
GI	Genomic island
DPG	Differentially presented genes
KEGG	Kyoto Encyclopedia of Genes and Genomes
UPGM	Unweighted pair group method with arithmetic mean
POM	Pairwise orthologs matrix
IJSEM	International Journal of Systematic and Evolutionary Microbiology

LIST OF FIGURES

LIST OF TABLES

Chapter 1.

General Introduction

1.1. Introduction to metagenomics

Metagenomics is the study of microorganisms in their natural living environment, which also includes the complex microbial communities in which they exist. It also analyses the genomic composition of an entire organism and each of the microorganisms that co-exist within their community. Instead of considering each microorganism independent from each other, metagenomics assumes dependency for each microorganism within its community.

The field of metagenomics is relatively new because microorganisms have been traditionally studied and analyzed through a classic laboratory setting, which assumes that the microorganism is independent from its host. Therefore, knowledge of microorganisms within their environment is scarce.

Metagenomic studies where limited prior the development of Next Generation Sequencing (NGS) technologies, which provides the capabilities to profile entire microbial communities from complex samples in order to discover new organisms and explore the dynamic nature of microbial populations. There are currently two common methods for environmental metagenomics: 16S rRNA sequencing and shotgun metagenomics.

1.2. 16S rRNA sequencing

16S rRNA gene sequencing is commonly used for identification, classification and quantitation of microorganisms within complex biological communities such as environmental samples and gut samples. The 16S rRNA gene is a highly conserved component of the transcriptional machinery of all DNA-based life forms [1], and because of this, is highly suited as a target gene for sequencing DNA in samples containing hundreds of different microorganisms.

Universal PCR primers can be designed to target conserved regions of 16S making it possible to amplify the gene in a wide range of different microorganisms from a single sample. Conveniently, the 16S rRNA gene consists of both conserved and variable regions (Figure 1). While the conserved region makes universal amplification possible, sequencing the variable regions allows discrimination between specific different microorganisms such as bacteria, archaea and microbial eukarya. Identification of viruses requires metagenomic sequencing (the direct sequencing of the total DNA extracted from a microbial community) due to their lack of the phylogenetic marker gene 16S.

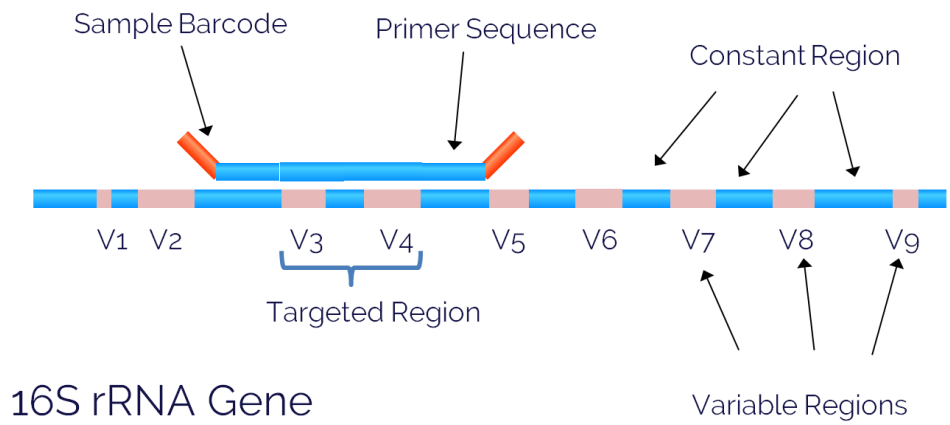


Figure 1. 16S rRNA gene sequence showcasing nine variable regions, making this gene an ideal target as a marker gene.

1.3. Shotgun metagenomic sequencing

Shotgun metagenomic sequencing allows researchers to comprehensively sample all genes in all organisms present in a given community. It enables researchers to evaluate microorganism diversity and detect the abundance in various environments. Shotgun metagenomics also provides a means to study unculturable microorganisms that are otherwise difficult or impossible to analyze.

In shotgun metagenomics, DNA is again extracted from all cells in a community. But, instead of targeting a specific genomic locus for amplification, all DNA is subsequently sheared into tiny fragments that are independently sequenced. This results in DNA sequences (i.e., reads) that align to various genomic locations for the genomes present in the sample, including non-microbes. Some of these reads will be sampled from taxonomically informative genomic loci (e.g., 16S), and others will be sampled from coding sequences that provide insight into the biological functions encoded in the genome.

There are several steps involved in a sequencing based metagenomics project. These include DNA extraction, library preparation and sequencing.

1.3.1. History

Shotgun sequencing is the method used to sequence the human genome by Craig Venter at Celera Genomics. The first method of DNA sequencing, the chain termination method or Sanger sequencing, is limited to a maximum DNA chain length of about 1,000 base pairs. On the other hand, shotgun sequencing

increases the total amount of DNA that can be sequenced. It is more of a strategy than a distinct method.

A shotgun approach was first used for early sequencing of small genomes like cauliflower mosaic virus. Later, shotgun methods were adapted (with the development of powerful computer algorithms) for sequencing and reassembling large genomes, most notably the human genome.

In the late 1990s, Craig Venter adapted the shotgun approach to large genomes. In that method, the DNA is randomly broken into many small pieces, cloned into a bacterial host, and sequenced using chain termination. Multiple rounds of fragmentation and sequencing are carried out, creating overlapping sequences. Powerful computer algorithms are then used to reassemble the sequence. Venter first developed his shotgun sequencing method while working on the bacterial species *Haemophilus influenzae* at the National Institutes of Health (NIH) in the US. The project took four months, compared to thirteen years researchers spent sequencing *E. coli* using Sanger sequencing, and ten years for yeast organisms. The alternative at the time was to create a low-resolution map of the genome first, and then perform a calculation of the minimum number of fragments needed to sequence the entire genome. The genome was then broken up randomly into fragments and the fragments cloned into bacterial hosts. Based on the map, the cloned fragments were assembled into a scaffold, or tiling path, that theoretically covers the entire sequence, and those fragments were sequenced.

Shotgun sequencing was a more direct alternative, but required a great deal more computing power, pushing the limits of processors available at the time.

1.3.2. Sample extraction

A reproducible method to extract DNA from microbial communities is essential for surveying and whole genome metagenomic analysis. Isolation and extraction must yield high quality nucleic acid for subsequent library preparation and sequencing. Sampling variation can have an effect on comparisons, and abundance measurements. This introduces several challenges as some samples must be delivered anaerobically. Exposure to oxygen or freezing can change the dynamic composition of a given microbial community. For example, freezing, thawing and subsequent bead-beating can affect the cell wall of Gram-positive bacteria, and introduce artifacts compared to extraction performed on fresh samples.

If the target community is associated with a host, e.g. human or plant, then physical fractionation or selective lysis can be employed to ensure host DNA is kept to a minimum. Host material can also be removed during bioinformatics filtering and mapping. Regardless of the approach used, it’s important to remember that extraction and isolation methods can introduce bias in terms of microbial diversity, yield and fragment lengths. It’s highly recommended that the exact same extraction method be used when comparing samples.

1.3.3. Library preparation

One of the biggest considerations for library preparation of environmental samples for shotgun metagenomic sequencing has to do with amplification. Certain types of samples (water, swabs) yield small amounts of DNA, necessitating amplification during library preparation. Amplification by PCR can over amplify certain fragments over others confounding abundance and microbial diversity measurements. Often the user does not have a choice when faced with low inputs of DNA. Minimizing variability, constructing libraries together to reduce batch effects and keeping library preparation steps as consistent as possible between samples is good practice.

1.3.4. Sequencing

Shotgun metagenomic sequencing is unique in the sense that you're trying to sequence a large diverse pool of microbes, each with a different genome size, often mixed with host DNA. Current sequencing technologies offer a wide variety of read lengths and outputs. Illumina sequencing technology offers short reads, 2x250 or 2x300 bp but generates high sequencing depth. Longer reads are preferred as they overcome short contigs and other difficulties during assembly. However, instruments that offer longer reads, e.g. PacBio and Oxford Nanopore are accompanied with higher error rates, lower sequencing depth and higher costs. PacBio error rates can be reduced using circular consensus sequencing (CCS) which involves repeat sequencing of a circular template and generation of a DNA insert consensus. High quality 500-4000 bp can be generated with >99% Q20 accuracy.

1.4. Shotgun metagenomic classification

Shotgun metagenomic data can be relatively complex and large, complicating its bioinformatic analysis. For example, it can be difficult to determine the genome from which a read was derived. Additionally, most communities are so diverse that most genomes are not completely represented by reads. Also, metagenomic analysis tends to require a large volume of data to identify meaningful results because of the vast amount of genomic information being sampled [2]. This requirement can pose computational problems. In order to solve this issue, several heuristics have been proposed, being the most popular homology and k-mer matching due to their accuracy and speed [3].

1.4.1. Homology-based approaches

Homology-based approached are those algorithms that use sequence alignment (global or local) between a reference (database) and the input query (one or multiple reads). MetaPhlAn2 (Metagenomic Phylogenetic Analysis) [4] is a tool that profiles microbial communities and estimates the relative abundance of microbial cells by mapping reads against a reduced set of clade-specific marker sequences that are computationally preselected from coding sequences that identify specific microbial clades at the species level or higher taxonomic levels and cover all of the main functional categories. It compares each metagenomic read from a sample to this marker catalog to identify high confidence matches. The catalog contains only ~4% of sequenced microbial genes, and each read of interest has at most one match due to the markers' uniqueness. The classifier

normalizes the total number of reads in each clade by the nucleotide length of its markers and provides the relative abundance of each taxonomic unit, taking into account any markers specific to subclades. Microbial reads belonging to clades with no available sequenced genomes are reported as an ‘unclassified’ subclade of the closest ancestor for which there is available sequence data.

MEGAN (Metagenome Analyzer) [5] allows analysis of large metagenomic data sets. In a pre-processing step, the set of DNA reads (or contigs) is compared against databases of known sequences using a comparison tool such as BLAST. MEGAN is then used to estimate the taxonomical content of the data set, using the NCBI taxonomy to summarize and order the results. The program uses a simple algorithm that assigns each read to the lowest common ancestor (LCA) of the set of taxa that it hit in the comparison. As a result, species-specific sequences are assigned to taxa near the leaves of the NCBI tree, whereas widely conserved sequences are assigned to high-order taxa closer to the root.

GOTTCHA (Genomic Origins Through Taxonomic CHAllenge) [6] is a semi-automated metagenomic community-profiling tool that is able to provide accurate community composition profiles at multiple taxonomic levels with reliable abundance estimates. It automatically eliminates genomic regions that generate the majority of false-positive signals in existing tools by analyzing the distribution and depth of coverage of only the unique fraction of each reference genome—the unique genome—to identify the true community composition and accurate relative abundance of members of the community. GOTTCHA uses empirically-derived

coverage limits, supported by machine-learning approaches, to set the limits of detection.

PhyloSift [7] is a method for phylogenetic analysis of metagenomic samples and for comparison of community structure among multiple related samples. It leverages phylogenetic models of molecular evolution to provide high resolution detection of organisms in a metagenome. It's based on well known statistical phylogenetic models, is amenable to Bayesian hypothesis testing, and uses name-independent and OTU-free analyses to provide higher resolution about microbial community assemblages (versus methods that rely on taxonomy or OTUs). Additionally, it proposes a set of 37 “elite” marker gene families that have largely congruent phylogenetic histories, thus improving the limit of detection for rare organisms in microbial communities.

1.4.2. Exact match K-mer approaches

CLARK (CLAssifier based on Reduced K-mers) [8]is a method based on a supervised sequence classification using discriminative k-mers. Considering two distinct specific classification problems (1) the taxonomic classification of metagenomic reads to known bacterial genomes, and (2) the assignment of BAC clones and transcript to chromosome arms/centromeres (in the absence of a finished assembly for the reference genome). CLARK offers two modes of execution. The first mode outputs for each object the number of hits against all the targets and the confidence score of the assignment (which is a number 0.5–1.0). The second mode employs sampling to reduce the number the target-specific k-

mers for classification, and outputs assignments without any detailed statistics so that the output size is significantly reduced.

The Kraken sequence classification algorithm [9] uses exact alignment of k-mers in order to classify a metagenome sample. To classify a sequence, each k-mer in the sequence is mapped to the lowest common ancestor (LCA) of the genomes that contain that k-mer in a database. The taxa associated with the sequence's k-mers, as well as the taxa's ancestors, form a pruned subtree of the general taxonomy tree, which is used for classification. In the classification tree, each node has a weight equal to the number of k-mers in the sequence associated with the node's taxon. Each root-to-leaf (RTL) path in the classification tree is scored by adding all weights in the path, and the maximal RTL path in the classification tree is the classification path (nodes highlighted in yellow). The leaf of this classification path (the orange, leftmost leaf in the classification tree) is the classification used for the query sequence.

Chapter 2.

An exact match k-mer algorithm

2.1. An exact match k-mer classification approach

2.1.1. Definition of the problem

We define the computational problem of finding the classification of multiple strings *S* in a metagenomic sample. Given a string *S*, we are often interested in counting the number of occurrences in *S* of every substring of length *k*. These length-*k* substrings are called *k*-mers and the problem of determining the number of their occurrences is called *k*-mer counting. *S* is either one DNA sequence or a concatenation of many DNA sequences where $\Sigma = \{A, C, G, T\}$.

2.1.2. Building a k-mer reference database

In order to build a k-mer database, two steps are required. First, k-mer counting must be performed, which is the process where all the possible k-mers from the reference database are generated. Finally, for each k-mer generated, a taxon must be associated, in some cases this taxon can be a bacterial species, or a higher taxa; this step is called k-mer mapping.

2.1.2.1. K-mer counting

To store and count our k-mers into memory, we utilize a hash table, which is defined as an array of (*key*, *value*) pairs. When applied to *k*-mer counting, *key* is the sequence of the *k*-mer, and *value* is the number of times that *k*-mer occurs. The position in the hash table of a given *key* is determined by a hashing function

hash. In the hash table, if M is the length of the table, the i-th possible location for a given mer m is defined as:

$$\text{Pos}(m,i) = \text{hash}(m) + \text{reprobe}(i)) \bmod M.$$

During the creation of the hashmap, the length n of the hash table is maintained as the power of two, $M = 2^{\ell}$ for some ℓ, and the key representing the k-mer is encoded as an integer in the set $Uk = [0, 4k - 1]$. The function hash is a function mapping Uk into $[0, M - 1]$.

When a new k-mer is added to the hash table, it's stored in pos(m, 0), and if that position is already filled with a different key, we try pos(m, 1) and so on up to some limit. A quadratic reprobing is used and it's defined as reprobe(i) = $i(i + 1)/2$. To allow concurrent update operations on the hash table, a lock-free hash approach with open addressing (Purcell and Harris, 2005) is implemented. A lock-free hash table uses the compare and swap (CAS) assembly instruction that is present in all modern multi-core CPUs. The CAS instruction updates the value at a memory location provided that the memory location has not been modified by another thread. The CAS algorithm is defined as:

```
function cas (p : pointer to int, old : int, new : int) returns bool {
        if *p ≠ old {
                return false
        }
        *p ← new
        return true
}
```

2.1.2.2. K-mer mapping

Mapping of *k*-mers to taxa is performed by querying a pre-computed hashmap structure described in the previous section. The process begins with the selection of a library of genomic sequences also known as reference sequences. Secondly, the 4-byte spaces used in the hashmap to store the *k*-mer counts are instead used to store the taxonomic ID numbers of the *k*-mers' lowest common ancestor (LCA) values. For each sequence, the taxon associated with it is used to set the stored LCA values of all *k*-mers in the sequence. As sequences are processed, if a *k*-mer from a sequence has had its LCA value previously set, then the LCA of the stored value and the current sequence's taxon is calculated and that LCA is stored for the *k*-mer.

2.1.3. Classification of a metagenomic read

In order to classify a DNA sequence *S*, we collect all *k*-mers within that sequence into a set, denoted as K(S). We then map each k-mer in K(S) to the lowest common ancestor (LCA) taxon of all genomes that contain that k-mer. These LCA taxa and their ancestors in the taxonomy tree form what we term the classification tree, a pruned subtree that is used to classify *S*. Each node in the classification tree is weighted with the number of *k*-mers in *K*(*S*) that mapped to the taxon associated with that node. Then, each root-to-leaf (RTL) path in the classification tree is scored by calculating the sum of all node weights along the path. The maximum scoring RTL path in the classification tree is the *classification path*, and *S* is assigned the label corresponding to its leaf (if there are multiple

maximally scoring paths, the LCA of all those paths' leaves is selected). This allows the algorithm to consider each *k*-mer within a sequence as a separate piece of evidence, and then attempt to resolve any conflicting evidence if necessary. For an appropriate choice of *k*, most *k*-mers will map uniquely to a single species, greatly simplifying the classification process. Sequences for which none of the *k*-mers in $K(S)$ are found in any genome are left unclassified by the algorithm. Figure 2 demonstrates an example of 4 reference sequences being added to a HashMap using the LCA methods; as each sequence is added, some k-mers originally added to a leaf node, will end up being tagged to an internal node.

The use of RTL path scoring in the classification tree is necessary due to the inevitable differences between the sequences to be classified and the sequences present in any library of genomes. Such differences can, even for large values of *k*, result in a *k*-mer that is present in the library but associated with a species far removed from the true source species. By scoring the various RTL paths in the classification tree, we can compensate for these differences and correctly classify sequences even when a small minority of *k*-mers in a sequence indicate that the sequence should be assigned an incorrect taxonomic label.

Adding the k-mers to the HashMap using LCA method:

Figure 2. Example of the process of reference k-mer processing using the LCA method.

2.1.3.1. K-mer search

Because *k*-mer's are usually queried immediately after looking for an adjacent *k*-mer, and because adjacent *k*-mers share a substantial amount of sequence, we utilize the minimizer concept to group similar *k*-mers together. We define the canonical representation of a DNA sequence *S* as the lexicographically smaller of *S* and the reverse complement of *S*. To determine a *k*-mer's minimizer of length *M*, we consider the canonical representation of all *M*-mers in the *k*-mer, and select the lexicographically smallest of those *M*-mers as the *k*-mer's minimizer. In practice, adjacent *k*-mers will often have the same minimizer.

To search in the hashmap efficiently, all *k*-mers are stored consecutively, and are sorted in lexicographical order of their canonical representations. A query for a *k*-mer *R* can then be processed by looking up in an index the positions in the database where the *k*-mers with *R*'s minimizer would be stored, and then performing a binary search within that region. Because adjacent *k*-mers often have the same minimizer, the search range is often the same between two consecutive queries, and the search in the first query can often bring data into the CPU cache that will be used in the second query. By allowing memory accesses in subsequent queries to access data in the CPU cache instead of RAM, this strategy makes subsequent queries much faster than they would otherwise be. The index containing the offsets of each group of *k*-mers in the database requires 8×4^M bytes.

2.1.3.2. Scoring a metagenomic read

The approach used to score a metagenome read is for the user to specify a threshold score in the [0,1] interval; which represents a confidence value that represents how close the metagenome read is to the reference database, where 1 means that all *k*-mers from the read belong to the given taxon. If the user gives a threshold and the read doesn't exceeds the threshold, it's labeled as unclassified.

A sequence label's score is a fraction C/Q, where C is the number of *k*-mers mapped to LCA values in the clade rooted at the label, and Q is the number of *k*-mers in the sequence that lack an ambiguous nucleotide (i.e., they were queried against the database).

2.1.4. Calculating the metagenome profile

A metagenomic profile is a report that contains the total counts of reads for a given taxon that are labeled as classified while maintaining the minimum confidence threshold given by the user. This report represents the abundance of reads per taxon within a metagenomic sample.

2.1.4.1. Normalization for LCA-assigned reads

When classifying raw sequence reads, many reads correspond to identical regions between two or more genomes. The algorithm solves this problem by labeling the sequence with the lowest common ancestor (LCA) of all species that share that sequence. Bracken (Bayesian Reestimation of Abundance after Classification with KrakEN) [10], estimates species abundances in metagenomics samples by probabilistically re-distributing reads in the taxonomic tree. Reads assigned to nodes above the species level are distributed down to the species nodes, while reads assigned at the strain level are re-distributed upward to their parent species. In order to re-assign reads classified at higher-level nodes in the taxonomy, we need to compute a probabilistic estimate of the number of reads that should be distributed to the species below that node.

Reallocating reads from a genus-level node in the taxonomy to each genome below it can be accomplished using Bayes' theorem, if the appropriate probabilities can be computed. Let $P(S_i)$ be the probability that a read in the sample belongs to genome S_i, $P(G_j)$ be the probability that a read is classified by Kraken at the genus level G_j, and $P(G_j|S_i)$ be the probability that a read from genome S_i is classified by the algorithm as the parent genus G_j. Then the probability that a read classified at genus G_j belongs to the genome S_i can be expressed as:

$$\mathrm{P(Si|Gj)} = \frac{\mathrm{P(Gj|Si)P(Si)}}{\mathrm{P(Gj)}}$$

2.1.4.2. Normalization for cell count relative abundance

Abundance from any given species given by the algorithm must be normalized. Without any normalization, the abundance profile will represent the total DNA abundance. However, in order to calculate cell count relative abundance, we must use the length of the reference (gene length, genome length etc) and it can be calculated as:

$$\text{Cell count abundance} = \frac{\text{Predicted DNA abundance}}{Reference\ length * ploidy}$$

2.2. RAM memory usage

Access and execution to the hash map database requires many random accesses to a very large data structure. To obtain maximal speed, these accesses need to be made as quickly as possible. This means that the database must be in physical memory during execution. To overcome this, we created a memory resident filesystem (ram disk).

The RAM disk driver is a way to use main system memory as a block device. It is required for initrd, an initial filesystem used if you need to load modules in order to access the root filesystem. It can also be used for a temporary filesystem, since the contents are erased on reboot. The RAM disk dynamically grows as more space is required. It does this by using RAM from the buffer cache. The driver marks the buffers it is using as dirty so that the VM subsystem does not try to reclaim them later.

Ramfs is a very simple filesystem that exports Linux's disk caching mechanisms (the page cache and dentry cache) as a dynamically resizable RAM-based filesystem.

During the beginning of the process, the database will be loaded from local storage media to the RAM disk, and it will remain there until the user powers off the computer or deletes the RAM disk from the process. Once the database is in memory, random accesses can be performed to the database during any type of classification process on the database. Figure 3 shows a diagram showing the bi-directional flow between the process and the RAM memory. Access to the local storage is limited to the reading of the RAW data in order to maintain constant reading speed. Temporary files generated during the process are also stored on RAM disk, this in order to facilitate the removal and to avoid constant local media write up, making the reading of the RAW data slower.

2.3. Quality Control

2.3.1. Read Trimming

Read trimming is performed in-process as shown on Figure 4. Two pointers will indicate the region of the read that fulfils the minimum quality requirements and allow the k-mers align only within that region. Traditionally, this step is called read trimming, and it involves a separate process where a RAW fastq file is loaded and each read is trimmed and rewritten in a separate file; however, this process is

quite inefficient. By doing this in-process, it allows us to only align k-mers in high quality areas of the read without the need of a separate process.

2.3.2. Host read removal

Removal of reads from unwanted organisms is usually common on metagenomics. Removal of host reads, PhiX and sequencing adaptors in some cases is a mandatory step in order to avoid miss-classifications or to speed up the process. However, removal of reads from an unwanted organism is a slow process, which usually involves the mapping of the reads to the reference genome of the unwanted organism.

Here, in order to avoid this slow process, as shown on Figure 5, by creating all the k-mers from the human genome, sequencing adaptors and PhiX, and removing the overlapping k-mers from our database, we prevent any read from those organisms to be classified. If a read from those organisms is contained in a sample, it will simply be ignored.

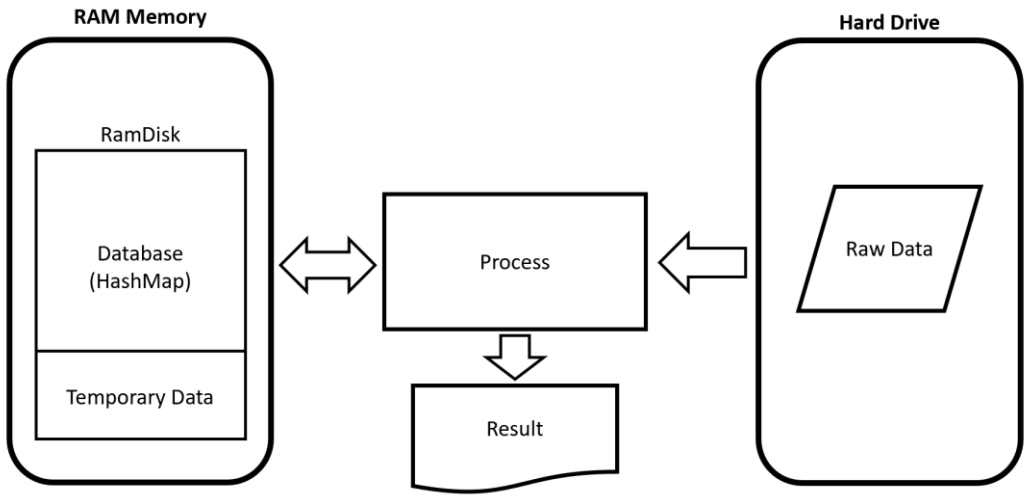


Figure 3. Diagram showing the flow of information during the classification process.

Figure 4. Low quality regions of a read are ignored in-process using two pointers.

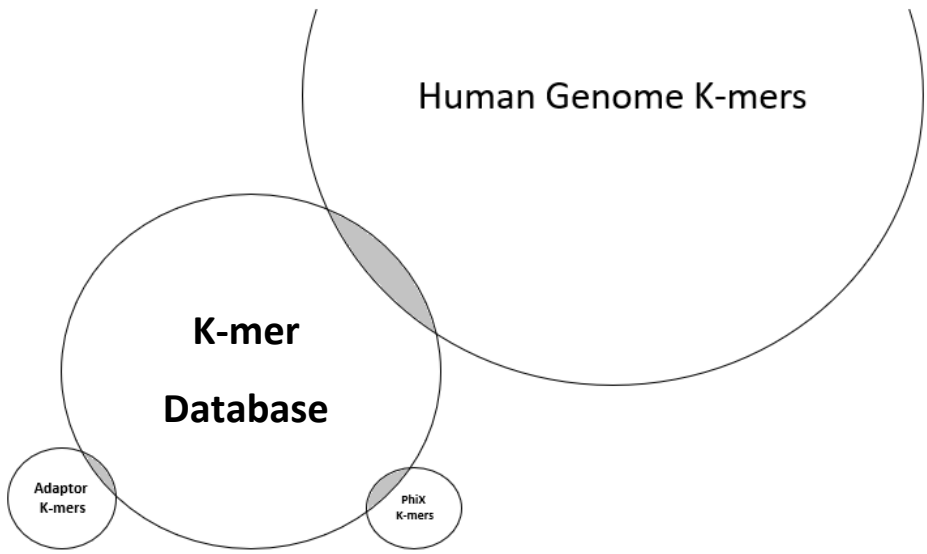


Figure 5. Overlapping k-mers from possible sources of contamination are removed from the final database.

Chapter 3.

Revealing unrecognized species in the genus Streptococcus

3.1. A brief history of streptococcus in clinical metagenomics

Shotgun metagenomics is of great importance in order to understand the composition of microbial community associated with a sample and the potential impact it may exert on its host. For clinical metagenomics, one of the initial challenges is the accurate identification of a pathogen of interest and ability to single out that pathogen within a complex community of microorganisms. However, in absence of an accurate identification of those microorganisms, any kind of conclusion or diagnosis based on misidentification may lead to erroneous conclusions, especially when comparing distinct groups of individuals. When comparing a shotgun metagenomic sample against a reference genome sequence database, the classification itself is dependent on the contents of the database.

Focusing on the genus *Streptococcus*, we built four synthetic metagenomic samples and demonstrated that shotgun taxonomic profiling using the bacterial core genes as the reference database performed better in both taxonomic profiling and relative abundance prediction than that based on the marker gene reference database included in MetaPhlAn2. Also, by classifying sputum samples of patients suffering from chronic obstructive pulmonary disease, we showed that adding genomes of genomospecies to a reference database offers higher taxonomic resolution for taxonomic profiling. Finally, we show how our genomospecies database is able to identify correctly a clinical stool sample from a patient with a

streptococcal infection, proving that genomospecies provide better taxonomic coverage for metagenomic analyses.

Taxonomy classification and quantification of each bacterial species within a shotgun metagenomic sample is a primary goal of most microbiome analyses, which still can be a complicated task. Sequencing of a given metagenomic sample generates a large number of sequence reads that are then fed to a bioinformatic process involving searches of each read against the reference sequence database. The general, albeit important, assumption is that reference-based databases currently available for shotgun metagenomics contain the reference genome sequences that we are interested in. When those references are absent, most classifiers (software that can be used to taxonomically profile a shotgun metagenomic sample), based on sequence identity, will match the reads from the metagenomic sample to the closest reference available in the database. The user, not knowing precisely which bacterial species are present in the sample, may assume that the taxonomic identification is accurate. Adding more references to a database can be a complex task as well; adding new genome sequences to a reference database not only will increase the database size, but also will increase the number of comparisons that the classifier has to perform between the references and each read within the sample. Alternatives to the use of full genomic sequences, MetaPhlAn2 [11] proposes the use of marker genes, which are gene sequences that only occur once within a specific taxon, in this case at the species level. However, this approach assumes that we have a reference genome of every species within a genus, if a new species is sequenced, all marker genes must be

recalculated. To solve this limitation, we propose the use of UBCG (Up-to-date Bacterial Core Gene) sequences [12] as reference for shotgun metagenomics. UBCG sequences are core genes that are defined as single-copy and homologous sequences; they are present in almost all bacterial species. At present, a total of 92 core genes are used for the version 3.0 of the UBCG pipeline, and regardless of the assembly level of a genome or its completeness, we can assume that a reference is complete if all 92 core genes are present in the genome.

To demonstrate the usefulness of these 92 core gene sequences in the use of shotgun metagenomic profiling, we focus on the genus *Streptococcus*, a highly diverse taxon in the phylum *Firmicutes*. The genus comprises a large number of species with a variety of pathogenic potentials to humans and animals including opportunistic pathogens like *Streptococcus oralis* [13], and harmless or even considered probiotic *Streptococcus thermophilus* [14]. *Streptococcus*-caused sepsis is often associated with *Streptococcus pneumoniae* and *Streptococcus pyogenes* [15]. *Streptococcus mitis* has been associated to several clinical diseases like VGS shock syndrome in cancer patients [16]. *Streptococcus pneumoniae* is commonly found on sputum samples of patients with community-acquired pneumonia, so rapid detection in a clinical setting is of high importance [17].

As of November 2019, a total of 88 validly named species and 114 genomospecies of *Streptococcus* are recorded in the EzBioCloud reference database [18]; a genomospecies is defined as a tentatively novel species that is

supported by genomic evidence, such as Average Nucleotide Identity (ANI) [19]–[21]. Most publicly available databases for shotgun metagenomics do not encompass these genomospecies references since they are not formally described and named.

Here, by extracting all 92 core genes for each species available in the EzBioCloud database, including 114 genomospecies of *Streptococcus*, we built two core gene-based databases one containing only validly named species (KrakenUBCG-VNS) and a second one containing both validly named species and genomospecies (KrakenUBCG). By focusing on the genus *Streptococcus*, using synthetic and real clinical datasets we demonstrate how bacterial core genes can be a better alternative as reference sequence databases for metagenomic taxonomic profiling, and finally we show the impact that a metagenomic taxonomic profile will have when including or excluding genomospecies from two core gene-based databases.

3.2. Results and Discussion

3.2.1. Building a core gene reference database

Bacterial genomic sequences on public databases have a diverse range of genomic statistics, such as reference size, number of contigs, assembly status (complete, chromosome, scaffold, contig), N50 values among others. For some species, a high-quality assembly may not be available. When using these references for metagenomic shotgun profiling, this variation may provide a bias for

higher quality and complete genomes, making abundance quantification unreliable, particularly for genomes that may be considered incomplete. To avoid this, we extracted UBCG sequences [12] (92 core genes) from full genomic references from the EzBioCloud database [18]. Regardless of the assembly status or genome length, extracting these core genes will remove any bias based on genome quality. Figure 6 (a) shows three streptococcal references with variable genome size, number of contigs and N50 value. When extracting the UBCG sequences, all the references end up being represented by the same number of genes (92) and their sequence size is near identical. This shows that regardless of having a complete genome with one contig, or a contig assembly with 500 contigs, UBCG will provide an unbiased representation for any bacterial reference regardless of their assembly status, making detection and abundance estimation more reliable. Figure 6 (b) shows how genomic sequences from a variable length size for the genus streptococcus can be translated into 92 UBCG sequences with a narrower difference in sequence length. After extracting all 92 UBCG sequences from the bacterial references, we built two Kraken [9] based pipelines containing these sequences as references, one with just validly named bacterial species (KrakenUBCG-VNS) and another one containing also genomospecies references (KrakenUBCG). Table 1 shows the difference between these two pipelines, highlighting their core algorithm and the number of bacterial and streptococcal species that their databases represent.

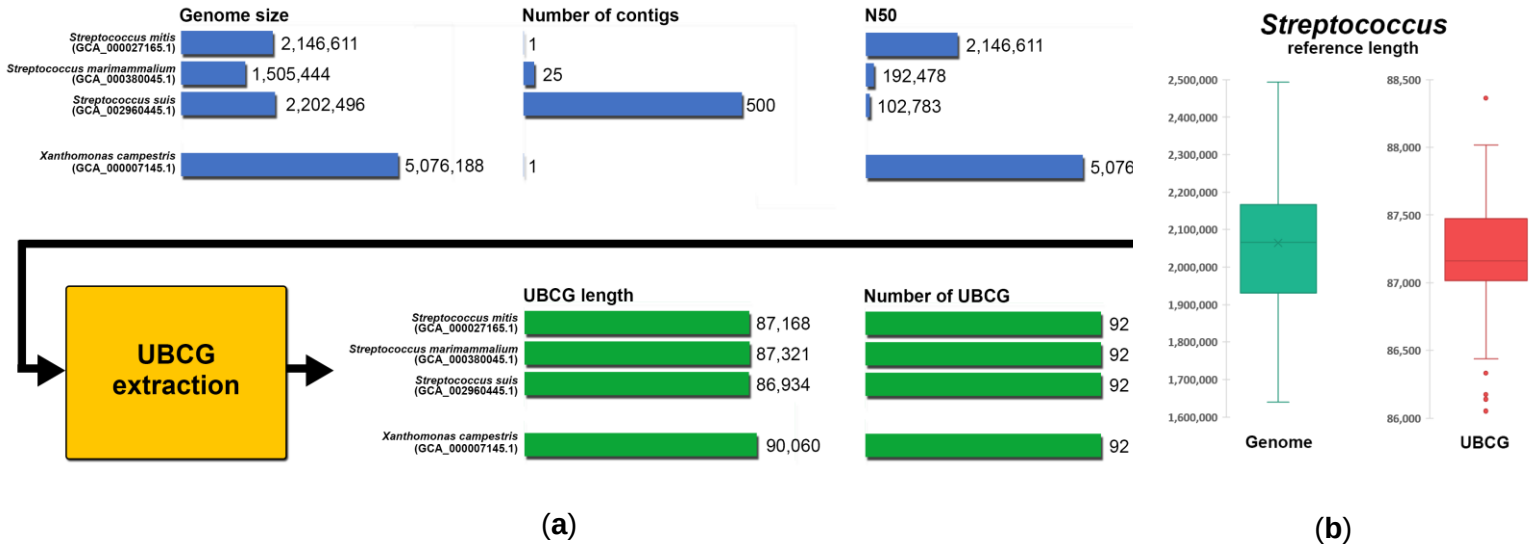


Figure 6. (a) Extracting process of UBCG sequences from genomic sequences with variable length, number of contigs and N50 values, (b) range and the median of the reference size for all the genomes and core genes for the species and genomospecies contained in the genus Streptococcus.

Table 1. Comparison of taxonomic profiling pipelines used in this study.

Pipeline	**Algorithm**	**Database**	**Gene type included in the database**	**Number of bacterial species**	**Number of *streptococcus* species**
KrakenUBCG	Exact match k-mer	EzBioCloud	Bacterial Core Genes (92 UBCG per species)	14,272 species (9,145 validly named+ 5,127 genomospecies)	201 species (88 validly named+ 113 genomospecies)
KrakenUBCG-VNS	Exact match k-mer	EzBioCloud	Bacterial Core Genes (92 UBCG per species)	9,145 validly named species	88 validly named species
MetaPhlAn2	Burrows-Wheeler	NCBI	Marker Genes (variable number)	3,310 species (2,250 validly named+ 1,060 unidentified species)	72 species (53 validly named+ 19 unnamed genomospecies)

3.2.2. Evaluation of Pipelines using Synthetic Metagenomes

To compare the classification accuracies by bacterial core genes as reference sequences, we created four different synthetic metagenomic datasets containing only species from the genus *Streptococcus* using the InSilicoSeq pipeline [22] and proceeded to compare the results with MetaPhlAn2. Table 2 contains the description and accession number of the genomes used on these datasets along with their true abundances. Figure 7 shows the predicted abundances for all four datasets using the KrakenUBCG pipeline and MetaPhlAn2, respectively, along with the true abundance of the synthetic sets. In order to compare some of MetaPhlAn2's predictions to our pipeline KrakenUBCG, we included an alternative truth for some sets, since MetaPhlAn2 clusters several *Streptococcus* species into a single taxon group. Figure 8 shows the log-modulus difference of abundance prediction made by MetaPhlAn2 and the KrakenUBCG pipeline.

The first synthetic set consists of seven different species of public assembled genomes from the NCBI database, which were included in both pipelines. MetaPhlAn2's prediction was compared against a grouped truth, where all the genomes that belong to *S. mitis* and *S. oralis* are clustered into a single group (*S. mitis*, *S. oralis* and *S. pneumoniae;* referred as mitis group hereafter). Both MetaPhlAn2 and KrakenUBCG were able to identify correctly the presence of all species of *Streptococcus* contained in the sample, however MetaPhlAn2's marker gene database cannot distinguish between the species that are contained

in the mitis group. Relative abundance predictions were less consistent for MetaPhlAn2, as it seems to be unable to predict correct abundances for the mitis group. While the expected abundance for *S. mitis* and *S. oralis* was expected at 30% and 20% respectively, MetaPhlAn2 only assigned a total abundance of 21.2% (mitis group), much lower than the expected 50%. KrakenUBCG, on the other hand, accurately predicted 29.9% and 19.3% relative abundance, respectively. Lower abundance prediction by MetaPhlAn2 for the mitis group indicates its overprediction of abundance for the remaining species, most notably *Streptococcus entericus* with a predicted abundance of 26.3% instead of an expected abundance of 15%, while KrakenUBCG accurately predicted its abundance at 15.8%.

Synthetic dataset 2 consists of nine species of *Streptococcus* that were contained in both databases while the genome assemblies representing each species were different. Both pipelines predicted correctly the presence of all the species contained in the sample, and while the KrakenUBCG pipeline contains different reference assemblies for all nine species, it was able to detect the presence of all nine species correctly. KrakenUBCG's abundance predictions were more accurate than MetaPhlAn2's, for the 20% expected abundance for *Streptococcus downei,* KrakenUBCG and MetaPhlAn2 predicted 22.2% and 13.19%, respectively, making KrakenUBCG overpredicting the abundance by 2.2% while MetaPhlAn2 underpredicted the abundance by 6.81%. *Streptococcus pyogenes* had an expected abundance of 15% with a predicted abundance of 13% and 18.64% for KrakenUBCG and MetaPhlAn2, respectively. For the remaining

species, KrakenUBCG's worst prediction was for *Streptococcus pneumoniae*, with a prediction of 6.72% against an expected abundance of 7.5%, while MetaPhlAn2's worst prediction was *Streptococcus agalactiae*, with a predicted abundance of 6.5% against an expected abundance of 5%. Overall, both pipelines performed better than expected, however KrakenUBCG performed favorably over MetaPhlAn's abundance prediction even though it did not contain the assemblies contained in the metagenomic sample.

The third set contains the assemblies of nine different genomes from the genus *Streptococcus*. Those genomes are annotated as genomospecies in the KrakenUBCG pipeline. MetaPhlAn2 contains the same assemblies, however some of them are annotated as the same species under the NCBI database, so we expect MetaPhlAn2's prediction to follow that annotation. Those annotations are as follows: genomospecies KQ969067_s is annotated at the NCBI database as *Streptococcus cristatus*; JPFV_s, LBMT_s and NCVM_s as *Streptococcus mitis*; JUNW_s, KQ970764_s and NCUR_s as *Streptococcus oralis;* and finally, CZEF_s and RSDO_s as *Streptococcus suis*. MetaPhlAn2 is also unable to distinguish between *Streptococcus oralis* and *Streptococcus mitis*, so while the KrakenUBCG prediction is expected to predict nine different genomospecies (KQ969067_s, JPFV_s, LBMT_s, NCVM_s, JUNW_s, KQ970764_s, NCUR_s, CZEF_s and RSDO_s), MetaPhlAn2 is only expected to predict three distinct species (*Streptococcus suis, Streptococcus cristatus* and mitis group), and while it predicted the presence of those three taxa correctly, it also predicted the presence of two other *Streptococcus* species (false positives). KrakenUBCG's

prediction, however, predicted correctly the presence of all nine genomospecies present in the sample. Abundance predictions for MetaPhlAn2 particularly for the mitis group were low, with an abundance prediction of 40.93% against an expected abundance of 66%. For *Streptococcus suis* MetaPhlAn2 predicted 30.21% against an expected 14% abundance. KrakenUBCG's worst abundance prediction was for the genomospecies KQ970764_s with 11.5% against an expected abundance of 13%.

Finally, set 4 consists of species present on the KrakenUBCG pipeline but absent on the MetaPhlAn2 database. MetaPhlAn2 predicted a single species (*Streptococcus infantis*) which is not present in the sample, and while it was expected that MetaPhlAn2 would not be able to predict correctly any of the species due to their absence in their database, it was not unable to predict the presence of more than one species. KrakenUBCG predicted all four species present, as for abundance predictions, each species had an expected abundance of 25% each, and KrakenUBCG predicted 26.4%, 24.2%, 25.3%, 23.9% abundance for *Streptococcus timonensis, Streptococcus respiraculi, Streptococcus plurextorum,* and *Streptococcus pluranimalium*, respectively.

Using four synthetic metagenomic samples containing full genome reads from *Streptococcus*, we demonstrated that using core genes as references not only predict accurate species composition, but also showed little difference when comparing predicted relative abundance versus the absolute abundance. We also showed that MetaPhlAn2 suffers from lower accuracy in predicting relative

abundance, probably because its marker gene references are uneven in coverage, especially for the mitis group.

Table 2. References and truth for the synthetic metagenome samples.

Set	NCBI Accession	Taxon name	EzBioCloud name	Genome size	Truth
1	GCA_000164675.2	Streptococcus parasanguinis ATCC 15912	Streptococcus parasanguinis	2153652	15.00%
	GCA_000257765.1	Streptococcus anginosus subsp. whileyi CCUG 39159	Streptococcus anginosus subsp. whileyi	2294730	10.00%
	GCA_000187465.1	Streptococcus infantis ATCC 700779	Streptococcus infantis	1905984	5.00%
	GCA_000380005.1	Streptococcus didelphis DSM 15616	Streptococcus didelphis	1877438	5.00%
	GCA_000380025.1	Streptococcus entericus DSM 14446	Streptococcus entericus	2036468	15.00%
	GCA_000027165.1	Streptococcus mitis B6	FN568063_s	2146611	30.00%
	GCA_000185265.1	Streptococcus oralis ATCC 49296	GL622184_s	2068336	20.00%
2	GCA_000014485.1	Streptococcus thermophilus LMD-9	Streptococcus thermophilus	1864178	10.00%
	GCA_000007425.1	Streptococcus pyogenes MGAS315	Streptococcus pyogenes	1900521	15.00%
	GCA_000019025.1	Streptococcus pneumoniae Taiwan19F-14	Streptococcus pneumoniae	2112148	7.50%
	GCA_000007465.2	Streptococcus mutans UA159	Streptococcus mutans	2032925	7.50%
	GCA_000380045.1	Streptococcus marimammalium DSM 18627	Streptococcus marimammalium	1505444	10.00%
	GCA_000180055.1	Streptococcus downei F0415	Streptococcus downei	2239421	20.00%
	GCA_000463425.1	Streptococcus constellatus subsp. pharyngis C1050	Streptococcus constellatus subsp. pharyngis	1991156	5.00%
	GCA_000007265.1	Streptococcus agalactiae 2603V/R	Streptococcus agalactiae	2160267	5.00%
	GCA_000314795.2	Streptococcus sp. F0442	KB373315_s	2231248	20.00%
3	GCA_001578775.1	Streptococcus cristatus DD08	KQ969067_s	2206539	20.00%
	GCA_000722685.1	Streptococcus mitis SK667	JPFV_s	2136987	10.00%
	GCA_002005545.1	Streptococcus mitis 321A	LBMT_s	2110680	13.00%
	GCA_002096935.1	Streptococcus mitis B_5756_13	NCVM_s	1896604	5.00%
	GCA_001075675.1	Streptococcus oralis 918_SORA	JUNW_s	1884524	20.00%
	GCA_001579175.1	Streptococcus oralis DD24	KQ970764_s	2129793	13.00%
	GCA_002096595.1	Streptococcus oralis subsp. oralis OD_311844-09	NCUR_s	1951174	5.00%
	GCA_900012395.1	Streptococcus suis 9401240	CZEF_s	2174179	10.00%
	GCA_003934335.1	Streptococcus suis PP422	RSDO_s	2075657	4.00%
4	GCA_900095845.1	Streptococcus timonensis Marseille-P2915	Streptococcus timonensis	1925331	25.00%
	GCA_003595525.1	Streptococcus respiraculi HTS25	Streptococcus respiraculi	2067971	25.00%
	GCA_000423745.1	Streptococcus plurextorum DSM 22810	Streptococcus plurextorum	2103464	25.00%
	GCA_002953735.1	Streptococcus pluranimalium TH11417	Streptococcus pluranimalium	2065522	25.00%

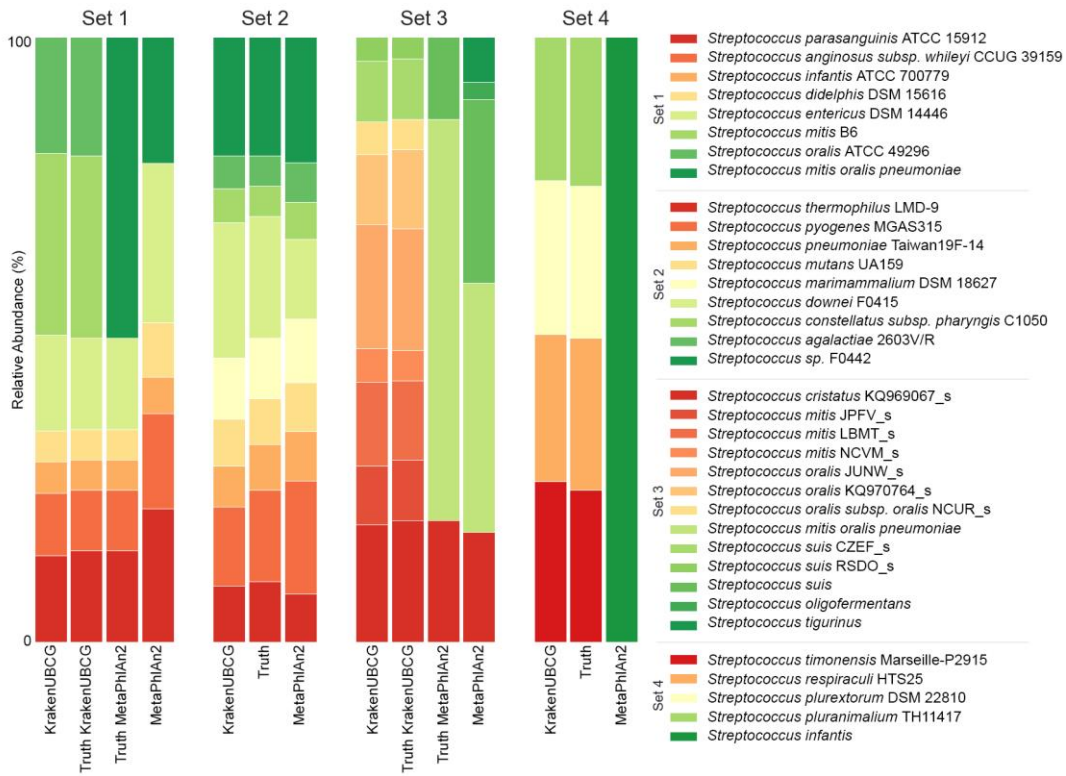


Figure 7. (a) Taxonomic prediction made by KrakenUBCG and MetaPhlAn2 for 4 synthetic metagenome sets containing species from *Streptococcus*; (b) Log-modulus difference between the predicted abundance and the expected abundance (truth).

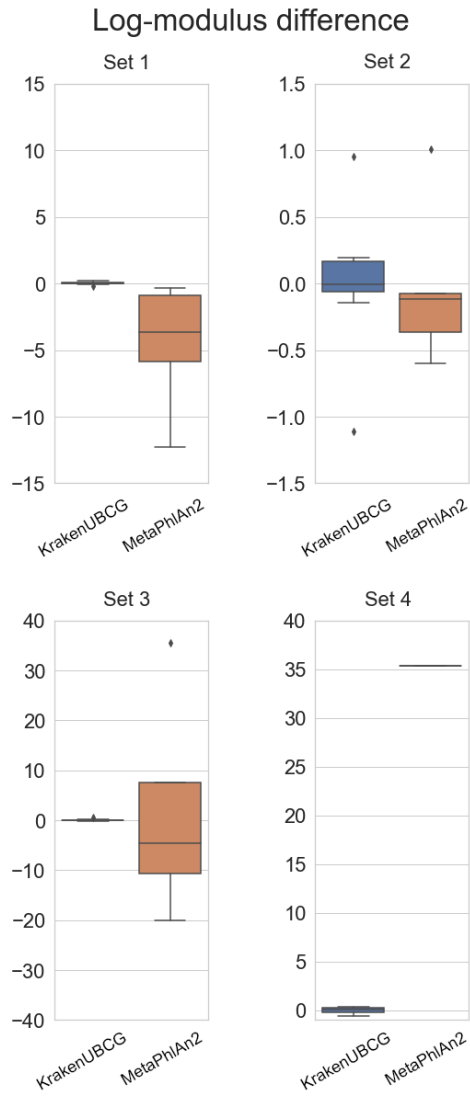


Figure 8. Log-modulus difference between the predicted abundance and the expected abundance (truth).

3.2.3. Chronic obstructive pulmonary disease samples

Chronic obstructive pulmonary disease (COPD) is an obstructive lung disease where the affected individual suffers from long-term breathing problems and airflow. Cameron et al. studied the association between the microbiome and COPD by comparing the metagenomic profiles of two different groups, patients with COPD and 'healthy' smoking controls [23], and showed significant changes in abundance of bacterial species, particularly in the genus *Streptococcus*. They also found the presence of *Streptococcus pneumoniae* in all the samples. Using KrakenUBCG with 201 different *Streptococcus* reference genomes and MetaPhlAn2's database, we profiled all the samples and compared the prediction of *Streptococcus* at genus and species level.

Figure 9 shows the abundance prediction at genus level by both KrakenUBCG and MetaPhlAn2. While the presence of other genera is variable among the databases, they are outside the scope of this study and won't be discussed in detail, however it is important to mention that the number of classified reads for other genera will also have a direct impact on the predicted abundance of *Streptococcus*. As an example, this can be observed on sample copd04, MetaPhlAn2 was unable to detect the presence of any bacteria, while KrakenUBCG detected the presence of three distinct genera.

Figure 10 shows heatmaps for predictions made by KrakenUBCG and MetaPhlAn2 along with their respective ANI dendrogram. Only samples that contained any *Streptococcus* in either of the predictions were included.

References from the MetaPhlAn2 database were matched with the KrakenUBCG pipeline and aligned on the heatmap accordingly. Predictions with a minimum of 0.5% abundance are shown, also only the species and genomospecies that were contained in at least one of the samples were included.

One of the major differences between both pipelines is the mitis group on MetaPhlAn2, which is covered by 21 different species and genomospecies at KrakenUBCG, showing clear differences as expected. On KrakenUBCG, only 3 samples showed the presence of *S. pneumoniae*, and another 3 showed presence of *S. mitis*, but none of the samples showed the presence of both at the same time; genomospecies CP016207_s, KV802702_s, KB373321_s, JPFY_s, JPFT_s, NCVM_s, CP012646_s, JUQO_s, JPFU_s and JUUO_s were predicted only once among the samples, while the remaining genomospecies were found in two or more of the samples. While MetaPhlAn2 found zero presence of bacteria in the mitis group for sample copd02, copd06, scon05 and scon10, KrakenUBCG found the presence of ASZZ_s and AFUU_s for copd02, and JPFV_s and JVWC_s for copd06. For samples scon05 and scon10, KrakenUBCG did not found the presence of any species or genomospecies closer to *S. pneumoniae* or *S. mitis* matching MetaPhlAn2's prediction. Looking at some samples individually we can observe distinct predictions, for example, MetaPhlAn2 predicted only one species of *Streptococcus* (*S. salivarius*) with a 17.5% abundance, while KrakenUBCG found three distinct species (ALIF_s, JYGT_s and *S. intermedius*) with a total abundance of 20.7%. If we analyze each sample individually, we can observe more differences than coincidences. These differences are the result of the lack

of references on MetaPhlAn2's database, and the argument can be made that only formally named species should be added to a reference database. However if MetaPhlAn2's authors decide to add genomospecies references to their database, they would require to recalculate all gene markers to those genus affected with newly added genomospecies, and even if somehow, these recalculations can be done within reasonable computational time, we can already observe that these gene markers cannot differentiate between three already recognized species (*S. mitis, S. oralis* and *S. pneumoniae*), so adding any more references may result in more larger groups of species, just like the mitis group.

Figure 11 and Figure 12 shows these predictions separated between control and COPD samples between EzBioCloud and MetaPhlAn2 respectively. While the purpose of this study is not to find any associations between COPD and healthy controls, we want to show that any kind of conclusion driven by distinct databases may lead to different findings. MetaPhlAn2's prediction did not show the presence of any species that it was not included on the healthy controls, however it did show a significant drop of abundance of *S. salivarius*, while the healthy controls showed the presence of a small abundance of other species. EzBioCloud however showed the presence of GL698454_s, while ALIF_s (same reference used on MetaPhlAn as *S. salivarius*) showed a nearly identical presence between both groups. A higher abundance of JUPI_s was also predicted for COPD samples, along as well a higher abundance on *S. peroris*. Control samples however, showed the presence of a higher diversity of Streptococcal species, similar to MetaPhlAn2's prediction.

Figure 13 shows the taxonomic biomarkers found by LEfSe [24] by KrakenUBCG and MetaPhlAn2. While the study was not aimed at finding associations between COPD and healthy controls, our analyses demonstrated that any kind of conclusion driven by the use of distinct databases may lead to different findings. Both databases found a great diversity of streptococcal species found in the control samples that were absent in the COPD samples. KrakenUBCG found 24 streptococcal biomarkers (8 validly named species and 16 genomospecies) contained in the control metagenomic samples while MetaPhlAn2 detected 6 streptococcal biomarkers, demonstrating the impact that a high species and genomospecies coverage would have on the biomarker discovery process.

Figure 14 shows four biomarkers found by both pipelines while the remaining biomarkers were exclusive to each pipeline. These common biomarker predictions done by both pipelines show that regardless of the core algorithm or database, when a species is predicted accurately, the process for taxonomic biomarker discovery can reliable. However, it also shows how the absence of streptococcal species on MetaPhlAn2's database would miss several biomarkers that were detected by KrakenUBCG (Figure 13). Absence of species may also incur on false taxonomic predictions, particularly when a reference species has no genomically close species. For example, MetaPhlAn2 detected *Streptococcus infantis* as a biomarker, however this may be explained by Figure 10, where it can be seen that MetaPhlAn2's database lacks reference species genomically close to its genome, while KrakenUBCG detected the presence of *Streptococcus infantis*

only in two control samples while the rest of the prediction was assigned to other genomically close genomospecies. This example showcases how an outdated database might impact metagenomic profiling and biomarker discovery, since the lack of genomic diversity around the reference of *Streptococcus infantis* will potentially incur in taxonomic false positives and result in a false biomarker prediction.

Here, by comparing classification results using samples from COPD patients with our genomospecies database KrakenUBCG and we found the presence of thirty-three different genomospecies and nineteen species across the samples. On the other hand, MetaPhlAn2 only found the presence of seventeen species of *Streptococcus*, demonstrating the importance of keeping a database up to date. This finding indicates that lack of all species and genomospecies in the database being used and their taxonomies inconsistencies can exert a dramatic impact on the discovery of reliable biomarkers and biodiversity estimates. While MetaPhlAn2's algorithmic predictions may not be incorrect, their use of an outdated database with an incorrect taxonomy annotation may have an impact on their prediction performance, specifically since they rely on that taxonomy structure in order to generate their reference marker genes. Also, the use of marker genes as reference sequences makes the process of updating their database difficult. While generating those marker genes itself requires substantial time and efforts, the process is further complicated by the fact that every time when a new species is added, all the marker genes have to be recalculated. To this end, using UBCG as a reference appears to be more convenient, as once these core

genes have been extracted, if a new species needed to be added to the database, only the core genes for that species have to be calculated.

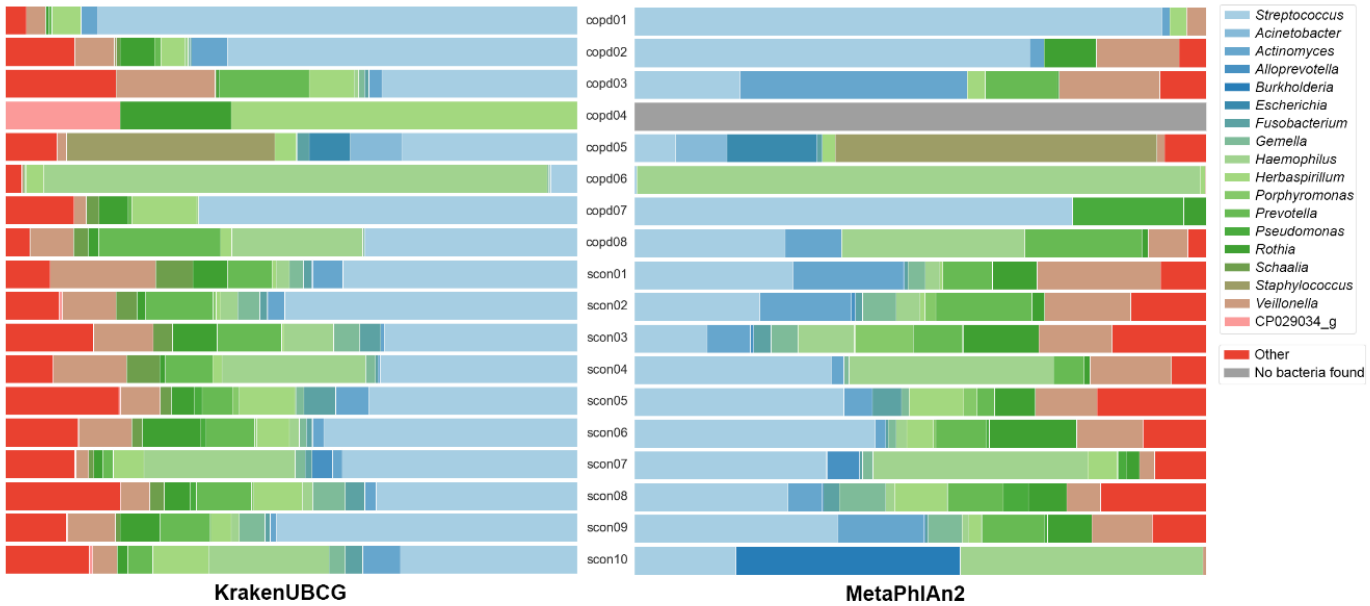


Figure 9. Predictions of taxonomic abundance at genus level between the KrakenUBCG database with genomospecies and the MetaPhlAn2 database.

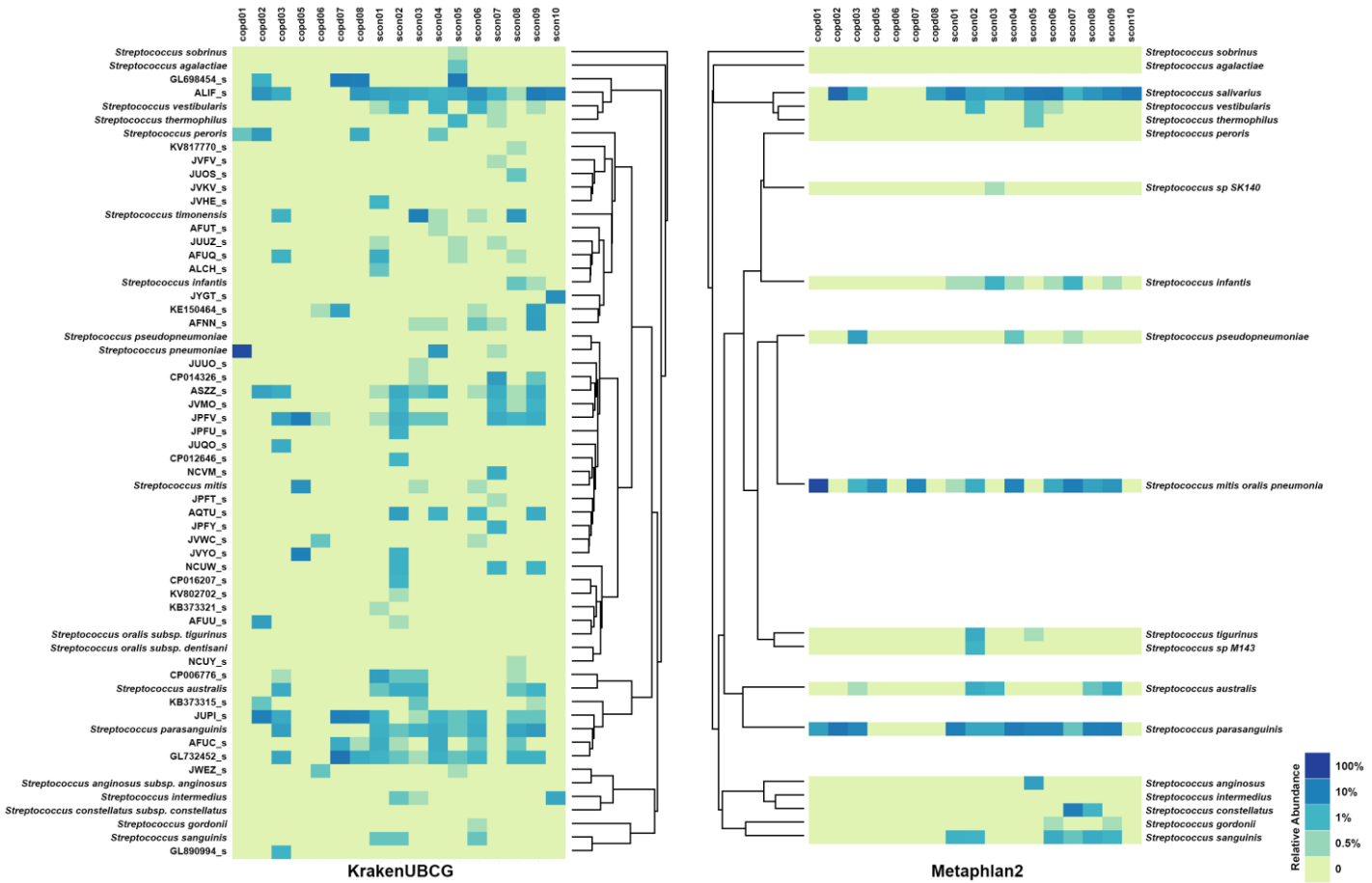


Figure 10. Species predicted for *Streptococcus* using the KrakenUBCG and the MetaPhlAn2 pipeline.

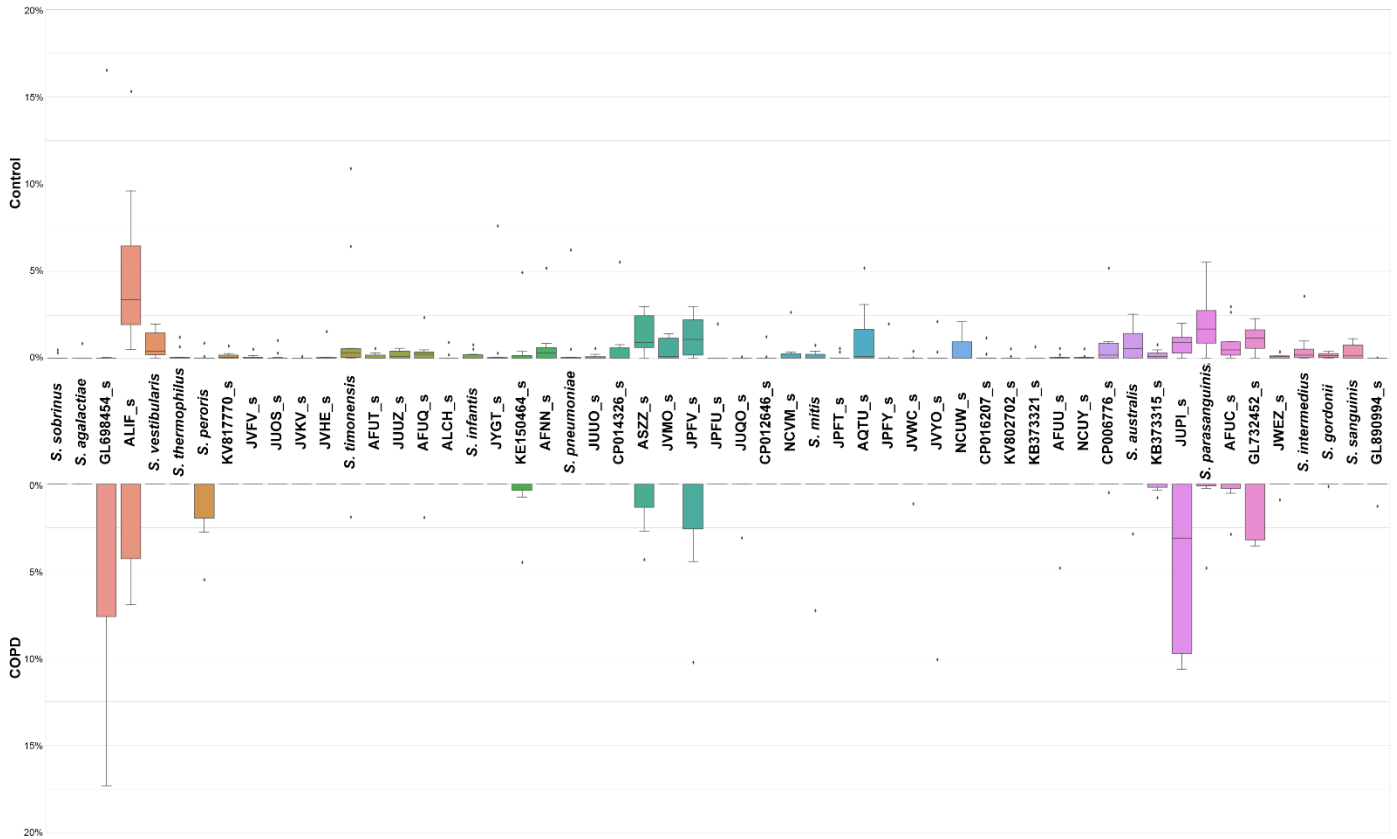


Figure 11. Abundance differences between COPD samples and control samples using the genomospecies database from EzBioCloud.

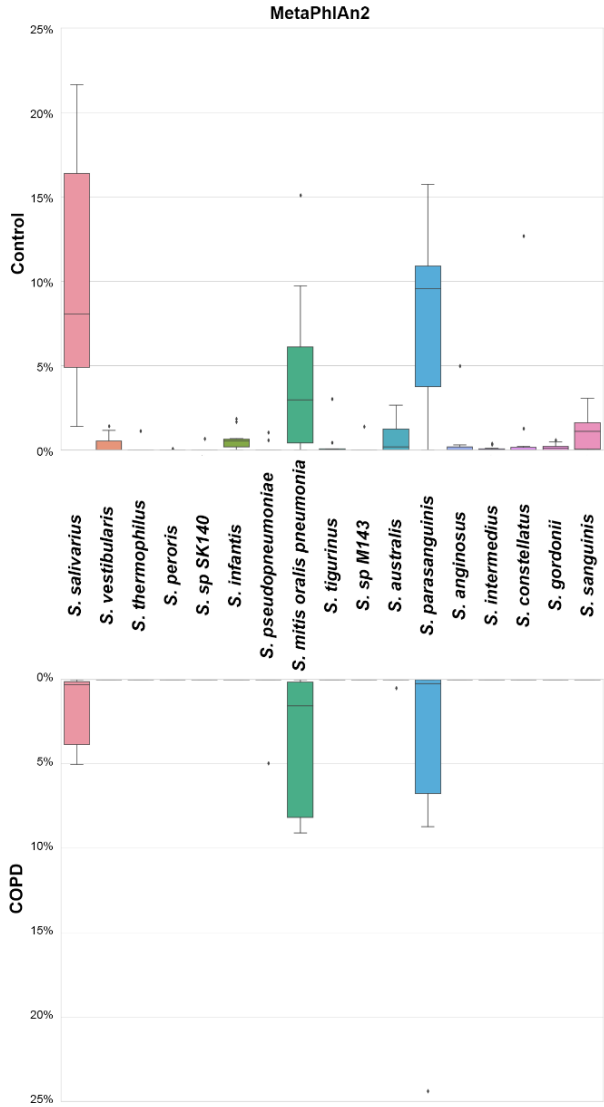


Figure 12. Abundance differences between COPD samples and control samples using MetaPhlAn2.

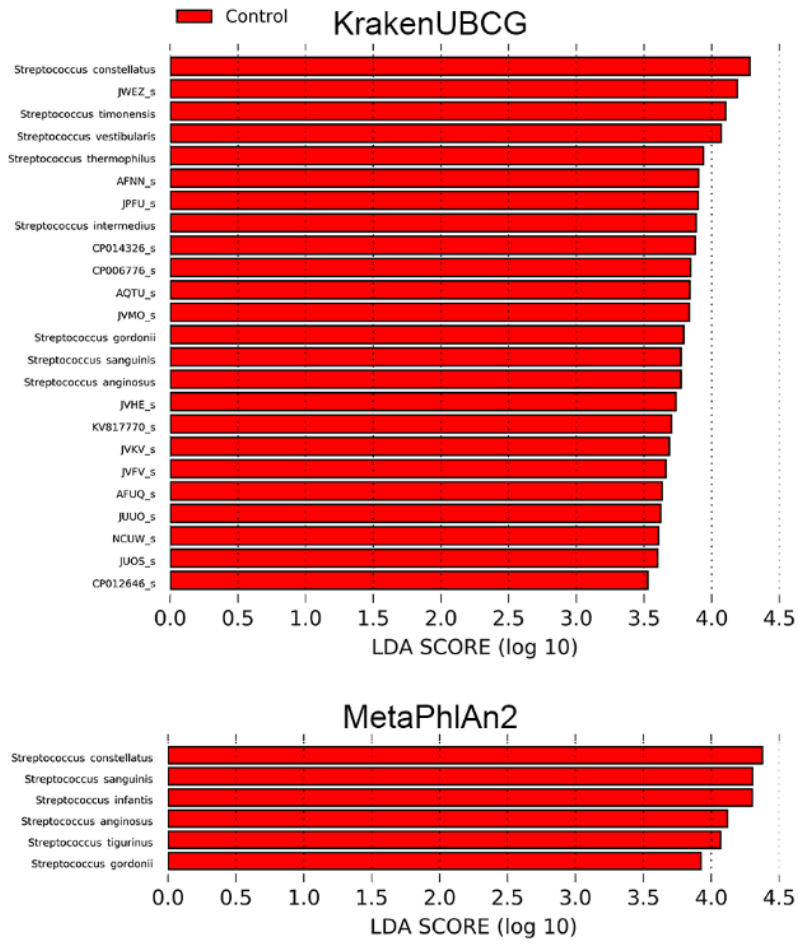


Figure 13. Taxonomic biomarkers found by LEfSe using KrakenUBCG and MetaPhlAn2.

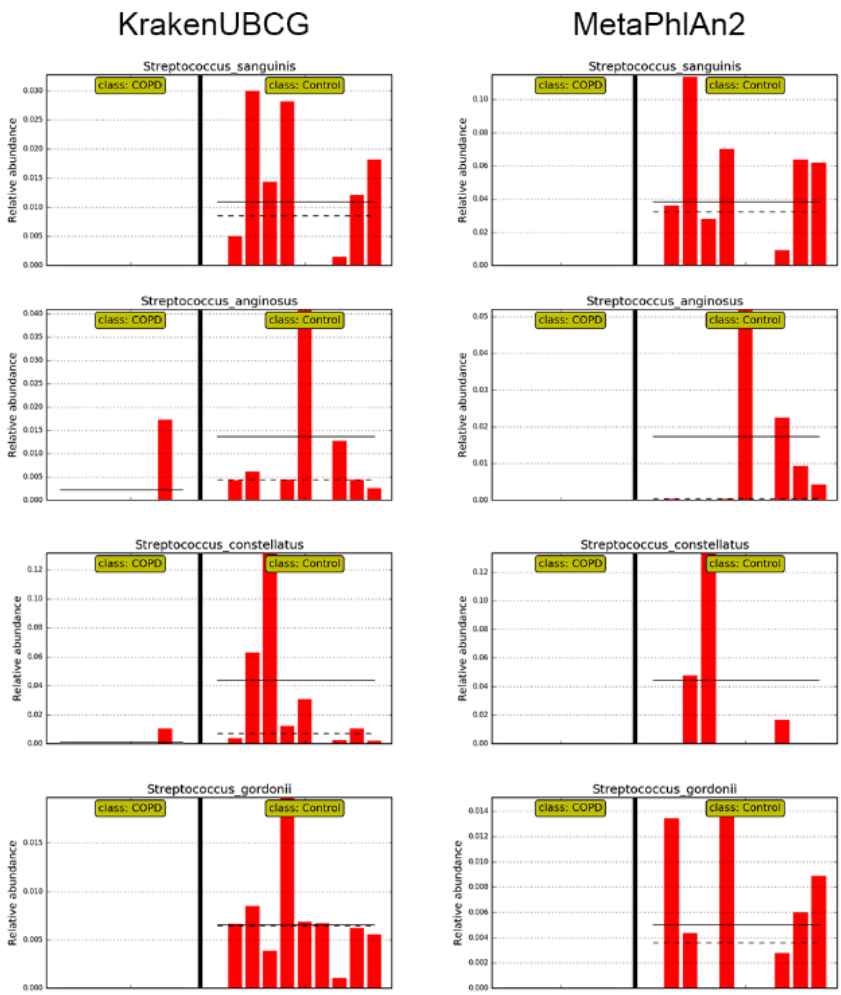


Figure 14. Biomarker features found in common between KrakenUBCG and MetaPhlAn2.

3.2.3. Evaluating the value of genomospecies references in a metagenomic database

To assess the effects of adding genomospecies to the reference database, we classified the same samples with two custom pipelines, one with only valid named *Streptococcus* species (KrakenUBCG-VNS), and a second one with species and genomospecies (KrakenUBCG), the rest of the bacterial references are identical, as described on Table 1.

Additionally, we also ran the classifications using different cutoff values for kraken. These cutoff values represent the k-mer coverage required for a read to be considered classified. A cutoff value of 0 means that only one k-mer is required for the read to be classified, while a cutoff value of 1 requires that the entirety of the read needs to be covered by *N* number of k-mers belonging to the database, where *N* = read size – k + 1, in other words, an exact match.

Figure 15 shows the fold change of classified reads for each sample at each cutoff value. The fold change represents the ratio of number of reads classified by KrakenUBCG against the number of reads classified by KrakenUBCG-VNS. At a cutoff value of 0 (no filter), the fold change of classified reads is closer to 1 (median value of 1.04), meaning that either pipeline (with or without genomospecies) will classify almost the same number of reads at genus level, however at higher cutoff values, the fold change of classified reads will increase greatly when using the genomospecies database KrakenUBCG. Just with a cutoff value of 0.3, the median fold change is 1.227, this means that having

genomospecies references present in the reference database will classify more than 22% additional reads. A more strict cut off value of 0.7 will have a median fold change of 1.935, resulting in 93% additional classified reads when comparing with the database without genomospecies KrakenUBCG-VNS. As expected, including genomospecies references not only increased the number of reads classified for the genus *Streptococcus*, but also by increasing the cutoff value of kraken allowed us to see how closer the references are to the reads in the samples. The argument can be made that, inclusion of additional references to any database would incur in a higher number of classified reads, however, when comparing the fold change for each sample, we can see that this is not always the case. As seen previously on Figure 10, sample copd01 only contains the valid named species *Streptococcus pneumoniae* and *Streptococcus peroris*, and this prediction can be confirmed by the low fold change (Figure 17) between pipelines (with or without genomospecies) and their different cutoff values (median fold change of 1.076). The rest of the samples showed the presence of at least one genomospecies, explaining the median fold difference between classifications (median range between 1.37 and 1.89). Sample copd07 showed the highest median fold change among all samples (median of 1.89), this is explained by Figure 10, showing that this sample contains 5 genomospecies without containing a single validly named species of *Streptococcus*.

These observed fold changes between classifications and distinct cut off values can only be explained if these samples indeed contain one or more genomospecies in their sample. Not including these genomospecies will result in

the loss of classified reads and misidentification at species level. Samples that do not contain any genomospecies showed no significant change between databases and cutoff values. Also, by using the same 92 core genes as our reference for every species, it allows us to discard any potential contamination or bias in our analysis based on genome length, assembly status or incompleteness of the genome. All species and genomospecies are represented equally in our databases (92 core genes per reference, one reference per species), so any additional classified reads can only be explained by the presence or absence of the reference that matches closely with the reads within the samples.

Figure 18 shows the fold difference between the total number of reads classified between the KrakenUBCG and KrakenUBCG-VNS. When the classification thresholds are removed (no filter), both pipelines are able to classify almost the same number of reads, since only one k-mer match is necessary to label a read as classified. When this identity threshold is increased gradually, the number of reads classified with KrakenUBCG-VNS decrease at a higher rate when comparing with KrakenUBCG, concluding that a majority of the reads have higher coverage with most of these genomospecies references. At an identity threshold of 0.8, KrakenUBCG is able to classify more than 2-fold the amount of reads.

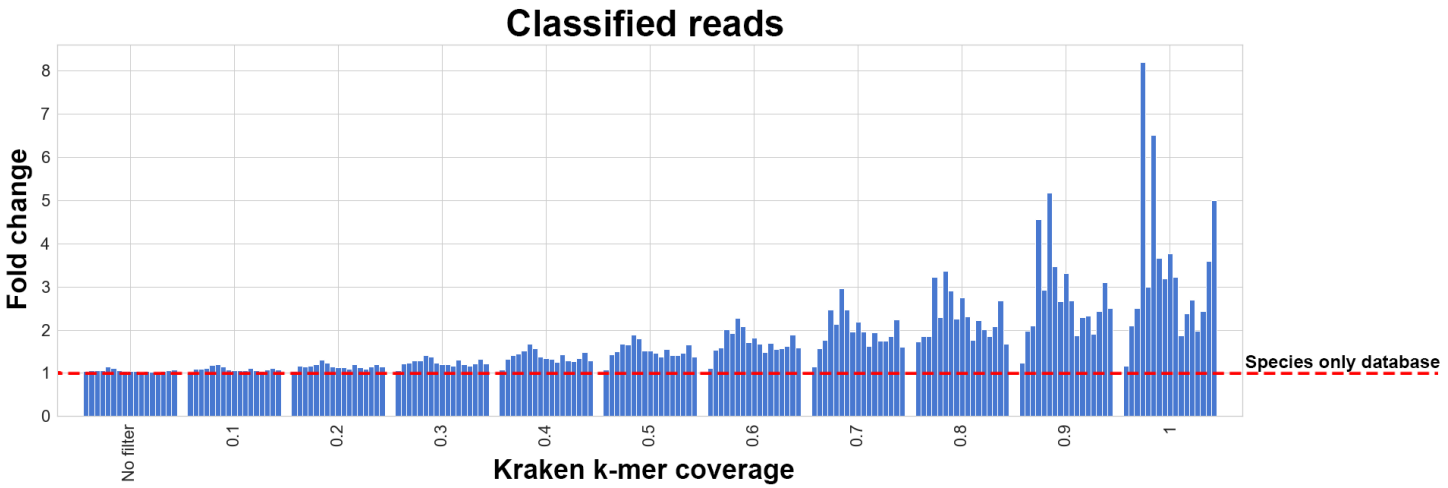


Figure 15. Fold change for k-mer coverage thresholds that the KrakenUBCG pipeline has against KrakenUBCG-VNS, illustrating that with higher thresholds, higher read classification rate is achieved by the genomospecies database.

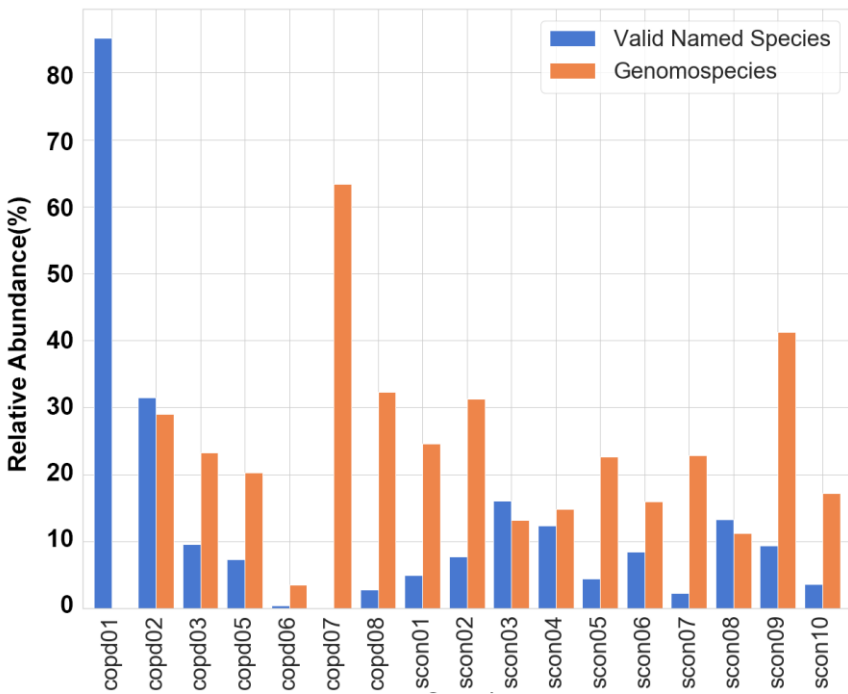


Figure 16. Abundance predicted by KrakenUBCG, separating the abundance assigned to valid species and genomospecies from *Streptococcus* (genus).

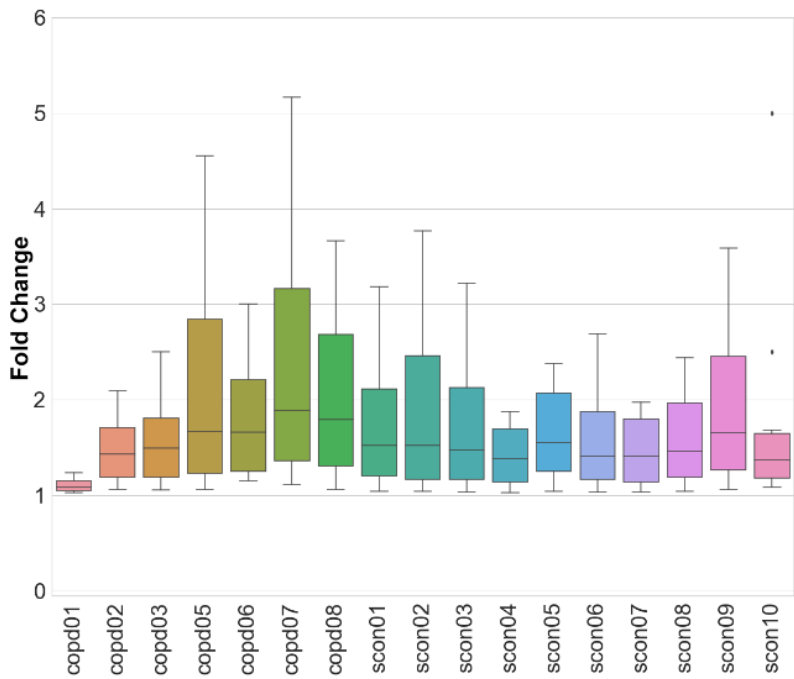


Figure 17. Range and median fold change per sample for all the different k-mer coverage thresholds.

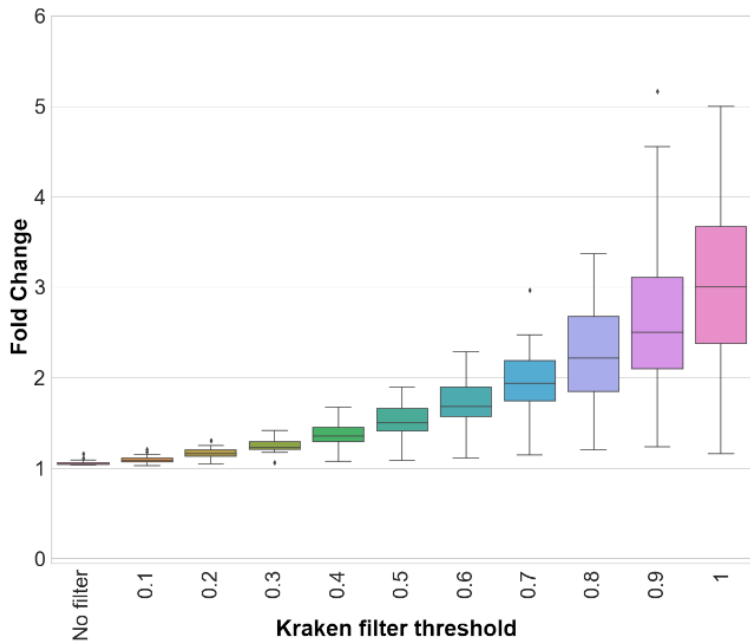


Figure 18. Range and median fold change for several identity thresholds when comparing KrakenUBCG with KrakenUBCG-VNS.

3.2.4. Identifying accurately a Streptococcal infection using clinical data

Tracing the origin of pathogenic bacteria in bloodstream infections can be a challenge. Tamburini et al. [25] propose that the source of infection on some patients is the human gut. They isolated and sequenced pathogenic bacteria present in the bloodstream of patients at Stanford University Hospital and performed shotgun metagenomics on stool samples for those same patients. For this research, we will focus on the patient 22, whereby using his blood isolate and stool sample, the authors were able to match the genome assembled from the bloodstream isolate with the metagenomic assembly generated form the stool sample and classified it as *Streptococcus mitis*.

We analyzed the streptococcal bacterial isolate from this patient (Sample id: SRR7407865) using the TrueBac ID system [26] which utilizes the exact same species and genomospecies references from the EzBioCloud database included in our KrakenUBCG pipeline.

Table 3 shows the results of TrueBac ID when comparing the isolate with the EzBioCloud database. Our analysis indicates that the closest reference to the isolate is actually streptococcal genomospecies JPFV_s (not *Streptococcus mitis)* with an ANI identity value of 94.61% with a coverage of 86.9%. Since the ANI identity value is below the species threshold (95%) [27], we will assume that this isolate belongs to a tentative new genomospecies and not belong to *Streptococcus mitis* .

To further illustrates the utility of broader species coverage and need to update the reference databases frequently, we built a third metagenomic reference database (KrakenUBCG+ SRR7407865) with the addition of this tentative new genomospecies by assembling the genome and extracting the 92 core genes using the UBCG pipeline and simply adding them to the KrakenUBCG database. We then perform a metagenome classification analysis using all three databases.

Figure 19 shows the classification results for all the available pipelines. Relative abundance at genus level for *Streptococcus* yield an increase of 0.07% after classifying with KrakenUBCG, while it only showed an increase of 0.002% when adding the bloodstream isolate from the sample (KrakenUBCG+ SRR7407865). However, at species level, the classifications for all three databases showed big differences (Figure 20). For the classification with only valid named species (KrakenUBCG-VNS), the highest, most abundant species of *Streptococcus* is *Streptococcus mitis*, however by using KrakenUBCG, the most abundant species is JPFV_s. By observing those two different classifications using the ANI tree shown on Figure 21, we can observe that they are both correct, with the absence of references from streptococcal genomospecies the classification should go to the closest reference, in this case, *Streptococcus mitis*. This is also consistent with the initial classification of the isolate in the original paper as *Streptococcus mitis* [25].

Finally, the third classification using KrakenUBCG+ SRR7407865, showed that by comparing the stool sample with an updated database that contains a bloodstream isolate will successfully assign the top streptococcal species as this tentative novel genomospecies.

While for this case we had access to the bacterial blood isolate, the majority of the time, the user will have to rely on precompiled databases, and in this specific case, if we were to use our genomospecies database, the classification for this streptococcal species would be assigned to the closest reference available in the database, in this case the genomospecies JPFV_s. We also show that in the absence of any genomospecies in the reference database, this streptococcal strain would be classified as *Streptococcus mitis*. This demonstrates the importance of updating reference databases frequently, even if those references have not been published or named formally.

Table 3. TrueBac ID analysis of the Streptococcus isolate from the bloodstream of patient 22.

No.	Hit taxon	ANI (%)	ANI coverage (%)	16S (%)	recA (%)	rplC (%)
1	JPFV_s	94.61	86.9	99.93	94.86	99.36
2	JVMO_s	94.34	85.8	99.59	95.72	98.88
3	JPFU_s	93.06	74.9	99.59	94.55	99.20
4	JUUO_s	93.44	78.5	99.39	94.60	99.68
5	JYGP_s	93.40	75.3	99.18	94.77	99.20

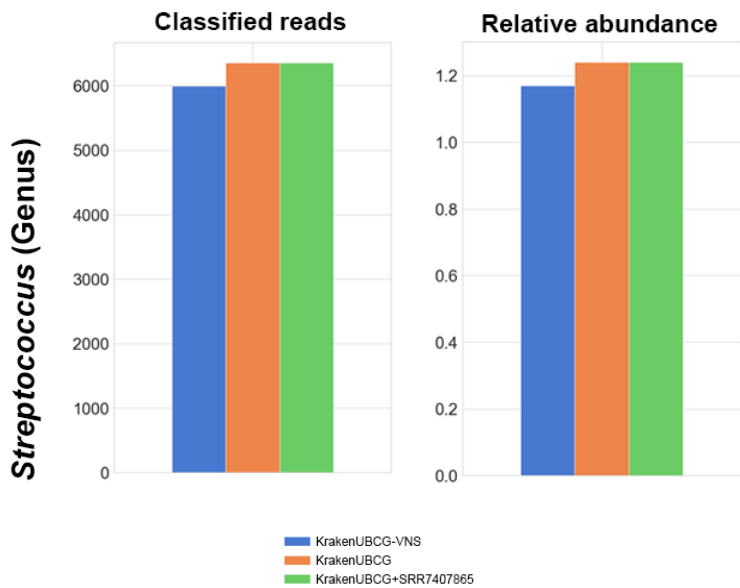


Figure 19. Number of classified reads and relative abundance at genus level for the stool sample of patient 22.

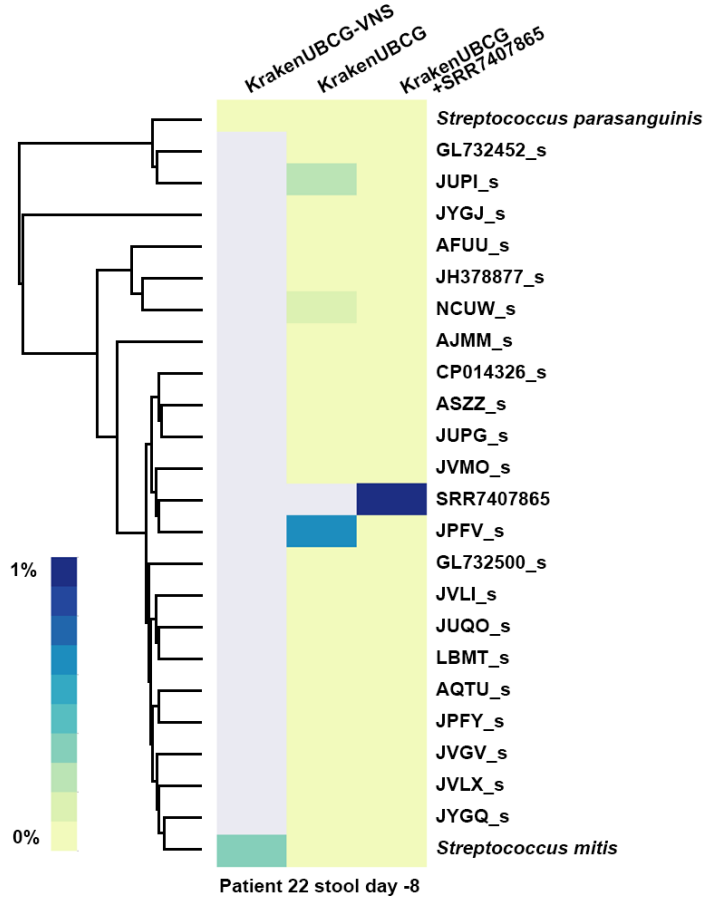


Figure 20. Taxonomy classification of *Streptococcus* for the three pipelines containing different references.

3.2.5. Effects of different ANI thresholds on the classification of genomospecies

The ANI threshold for genomospecies identification proposed by Chun et al. [27] is 95%, and this threshold is applied at the EzBioCloud database. As the authors mention in their article, these are minimal standards, which means that in some cases, a different ANI threshold could be applied in order to set species boundaries for a specific taxon. Similarly, the Genome Taxonomic Database (GTDB) [28] also proposes a 95% ANI threshold for these novel species using FastANI [29], although using a different nomenclature for these genomospecies, supporting this ANI threshold as a consensus in the prokaryote taxonomic community. In the previous example shown on Figure 20, the blood isolate from patient 22 was recognized as a potential new species because the ANI value between the isolate and the closest genomospecies JPFV_s is 94.24%, which is below the proposed species boundary. However, if a different ANI threshold was set, for example 94%, this blood isolate would be considered as a part of the genomospecies JPFV. With this in mind, we decided to change the ANI threshold for genomospecies for the case of patient 22 in order to see the effects that a different ANI threshold would have on the classification of streptococcal bacteria present in the stool sample. Figure 21 shows the ANI dendrogram marked with the different ANI thresholds used on this analysis.

For this experiment, we built three additional databases, each one containing a different number of streptococcal genomospecies depending on the

ANI threshold set for each database. When setting a new ANI threshold, only one reference would be used as a reference for that genomospecies. For example, at a ANI threshold of 94%, the genomospecies JUQO_s and LBMT_s would be considered as a single genomospecies, so in order to select which reference would be used to represent this new genomospecies, we selected the genome with the highest N50 value. After building these three additional databases, we profiled the stool sample of patient 22 and generated heatmaps with their respective ANI dendrogram. With a 94% ANI cutoff (Figure 22), the blood isolate from patient 22 is now considered as the JPFV_s genomospecies, at the same time, JUPI_s now is considered a strain of GL732452_s; despite these changes, the main streptococcal species found on the sample is JPFV_s, which is the genomospecies that now represents the blood isolate from the patient. A small abundance is also detected for the genomospecies NCUW_s. Lowering the ANI threshold at 93% (Figure 23), decreases the number of genomospecies present in this tree, with CP014326_s now representing a vast majority of genomospecies from the 95% ANI tree, including ASZZ_s, JUPG_s, JVMO_s, JPFV_s and the blood isolate SRR7407865. In this case CP014326_s is classified as the main streptococcal species with NCUW_s also being found present with a low abundance. Lastly, with an ANI threshold of 92% (Figure 24) only six species remain, this time, the genomospecies CP014326_s maintains its predominant presence of streptococcal abundance, with a slight abundance increment, and this could be mainly due to the absence of NCUW_s, now being represented by JH378877_s.

In the previous examples, we demonstrated that regardless the ANI threshold set for genomospecies, the detection of a specific species can be achieved if a representative species close to the one of interest is present in the database. With an ANI threshold of 92%, genomospecies CP014326_s represents our blood isolate SRR7407865, and with an ANI distance between these two of 93.591%, it can be seen that with a small sacrifice of abundance detection, the presence of either genomospecies can be achieved regardless of the ANI threshold set by the user.

Here, we showed that even with different ANI thresholds, the classification of the streptococcal strain was consistent with the location of the blood isolate within the original ANI dendrogram. However, a loss of abundance was seen mainly because of the higher ANI distance between the species present in the stool sample and the reference used to represent the genomospecies on each of the examples. While the ANI threshold for setting boundaries for new species can be debatable, when using a computational method that is based on sequence similarity, we believe that a fixed sequence similarity ANI threshold should be implemented, regardless of the taxonomic rank we set to given a taxon node (species, genomospecies or even strain).

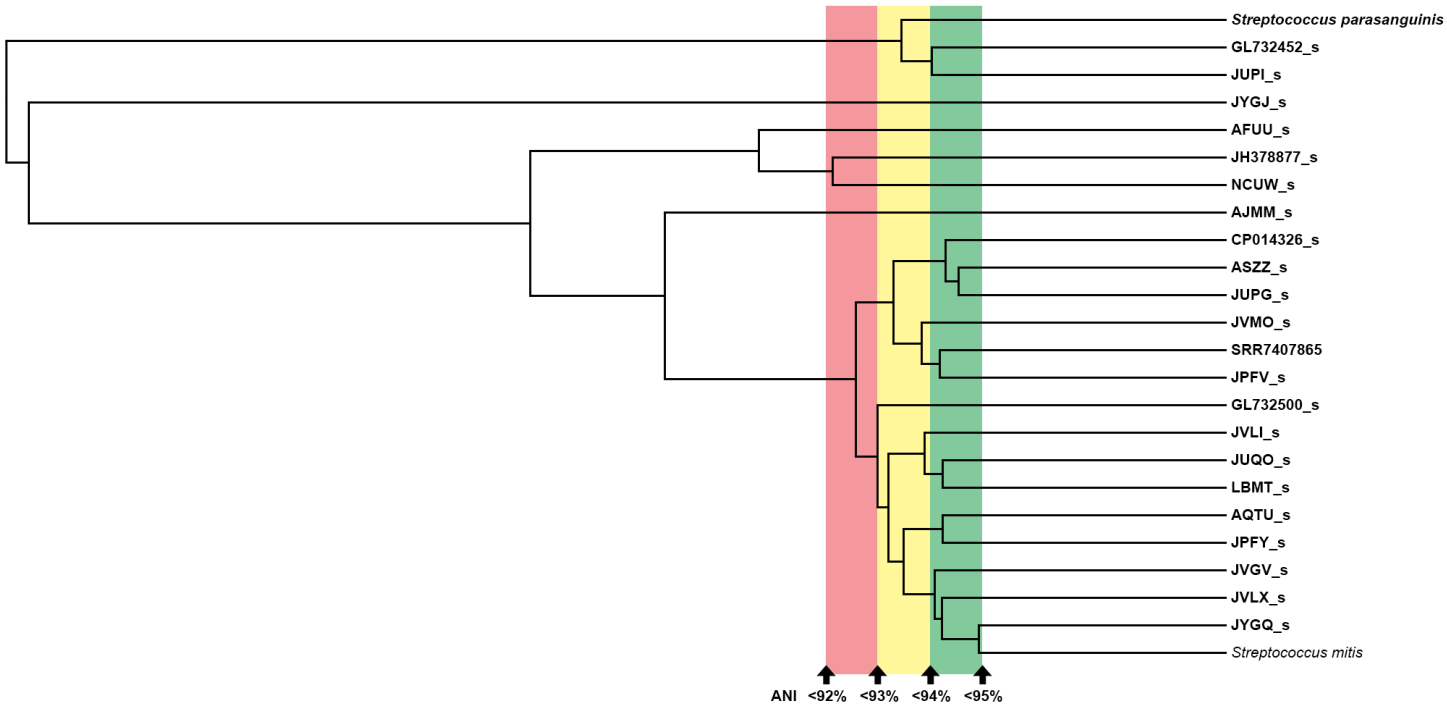


Figure 21. ANI dendrogram with different ANI thresholds highlighted for the sample of patient 22, showing distinct possible genomospecies for this streptococcal subtree.

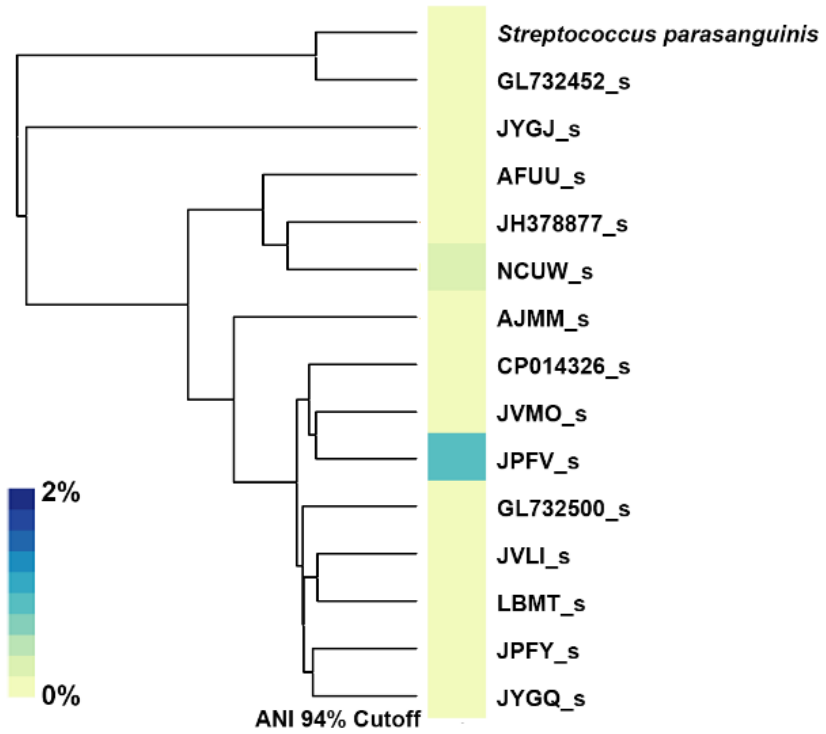


Figure 22. Classification of streptococcal species using 94% ANI threshold for genomospecies boundaries, showing the changes of species detection depending on the ANI threshold used.

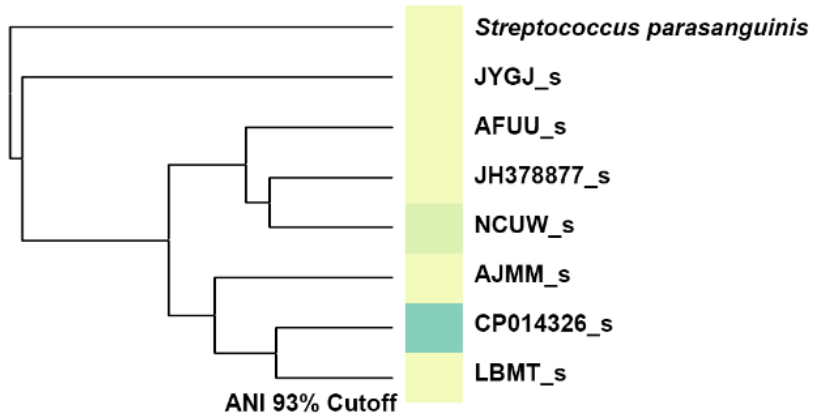


Figure 23. Classification of streptococcal species using 93% ANI threshold for genomospecies boundaries, showing the changes of species detection depending on the ANI threshold used.

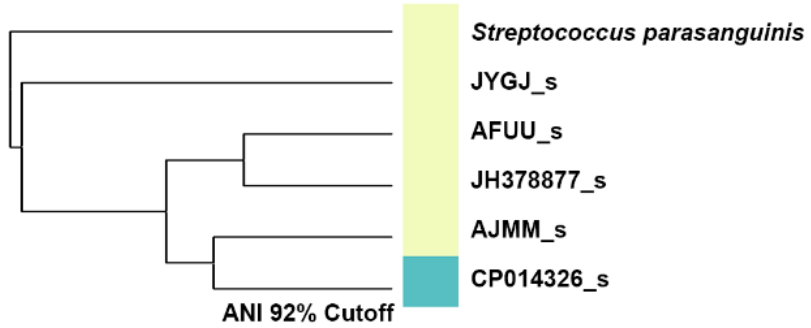


Figure 24. Classification of streptococcal species using 92% ANI threshold for genomospecies boundaries, showing the changes of species detection depending on the ANI threshold used.

3.3. Materials and Methods

3.3.1. Selecting the reference genomes

We used the EzBioCloud database [18] as a source for our reference genomes and selected one reference per species with a total of 9,145 bacterial species, including 201 species (88 validly named species and 113 genomospecies) in the genus *Streptococcus*. Supplementary Table S1 shows the list of references used, with their respective EzBioCloud and NCBI accession numbers. Supplementary Figure S1 shows the ANI dendrogram of all the streptococcal references used in this study. For section 2.4, we merged taxonomic nodes depending on the new ANI threshold set by selecting only one reference with the highest N50 value for their assembly; N50 is defined as the length of the shortest contig that accumulatively show 50% or more of the genome size [27].

The genome sequence of a blood isolate from patient 22 (NCBI accession SRR7407865) was incorporated to the database for section 2.3. The raw data was downloaded from NCBI and assembled using the pipeline SPAdes[30] with the default parameters.

3.3.2. Average nucleotide identity and hierarchical clustering

OrthoANIu[31] was used to calculate ANI. Hierarchical clustering was carried out from the ANI matrix by applying the UPGMA(unweighted pair group method with arithmetic mean) algorithm using the R library phangorn [32].

3.3.3. Synthetic and Real metagenomic samples

The four simulated datasets used to compare our pipeline with MetaPhlAn2 were generated with InSilicoSeq [22] using the MiSeq model provided. These datasets are available at https://bitbucket.org/streptosynth/metagenome/downloads/.

Supplementary Table S2 shows the genomes used for these sets. The chronic obstructive pulmonary disease metagenomic samples[23] were downloaded from https://www.ncbi.nlm.nih.gov/bioproject/PRJEB9034/, while the metagenome and isolate samples of the patient 22[25] was from https://www.ncbi.nlm.nih.gov/bioproject/PRJNA477326/.

3.3.4. Extracting the core genes

The 92 genes that are defined as Up-to-date Bacterial Core Gene (UBCG) were extracted for all species from EzBioCloud databases, including those belonging to the genus *Streptococcus* as described earlier [12]. Figure 25 shows the distribution and location of all 92 UBCG of the *Streptococcus suis* reference genome (NCBI accession GCA_900143575.1). All 92 UBCGs were also extracted from the genome assembly of a blood isolate (NCBI SRA accession SRR7407865).

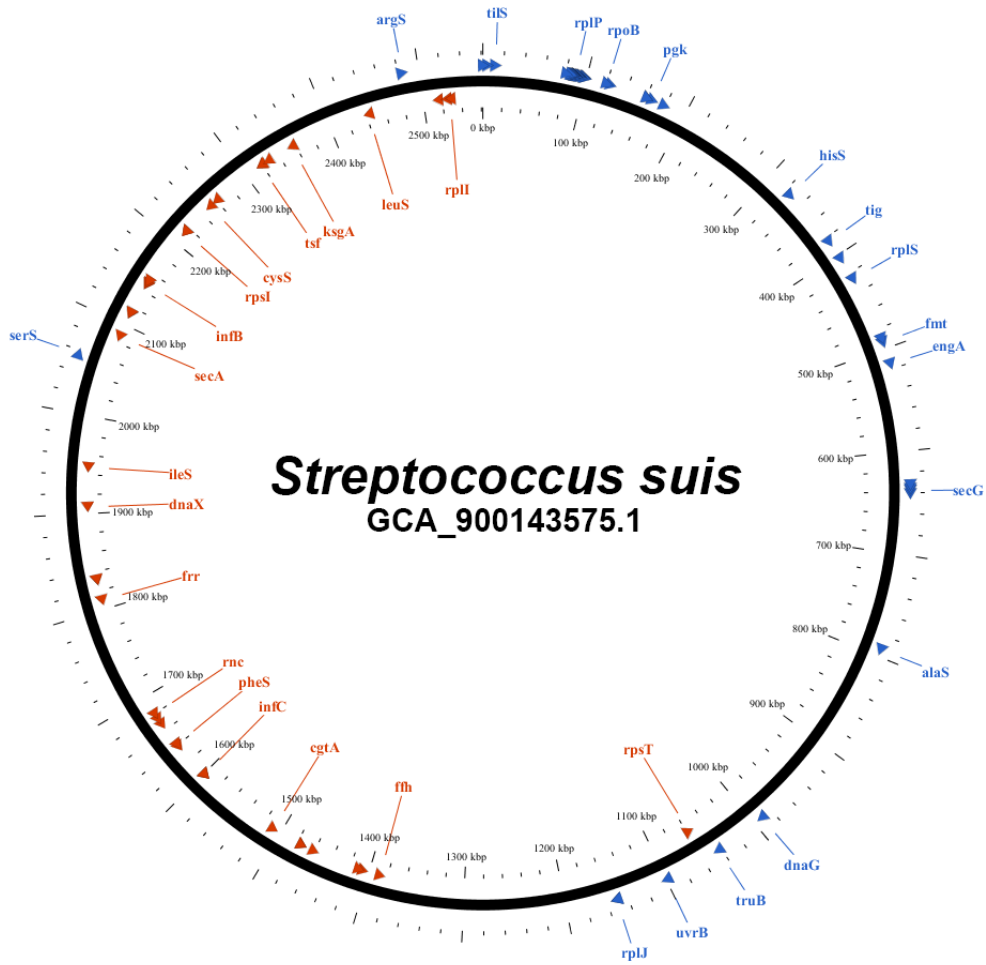


Figure 25. Location of the 92 core genes on the reference for *Streptococcus suis*.

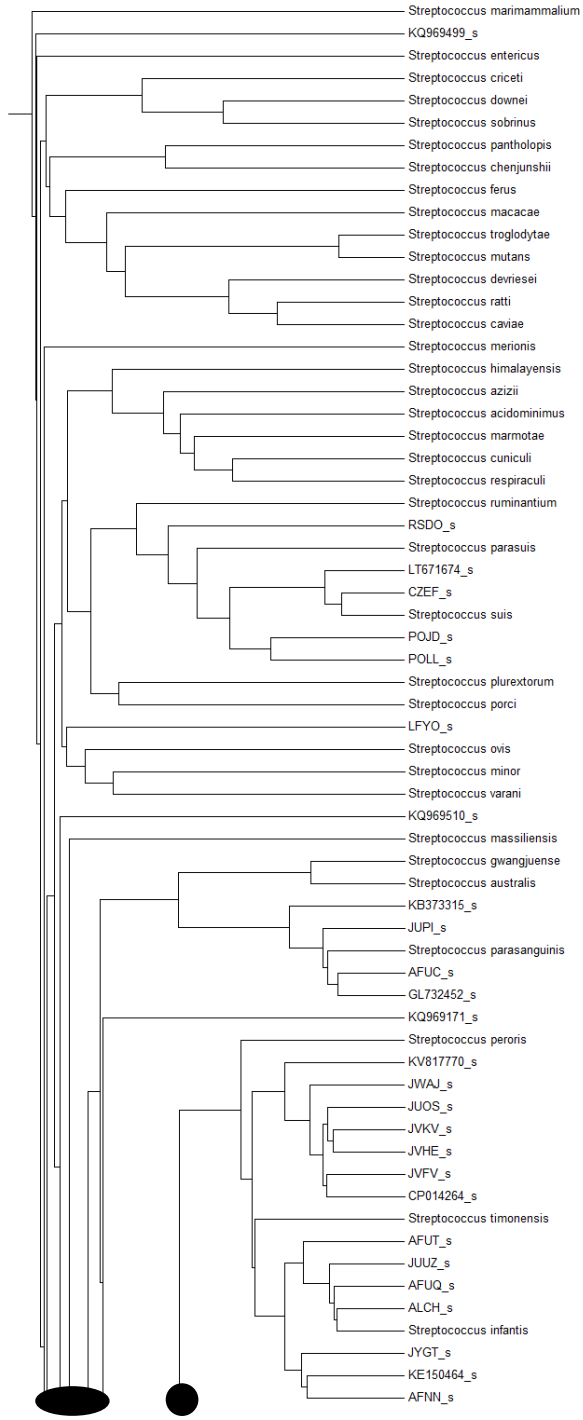


Figure 26. ANI Dendogram for the genus *Streptococcus.*

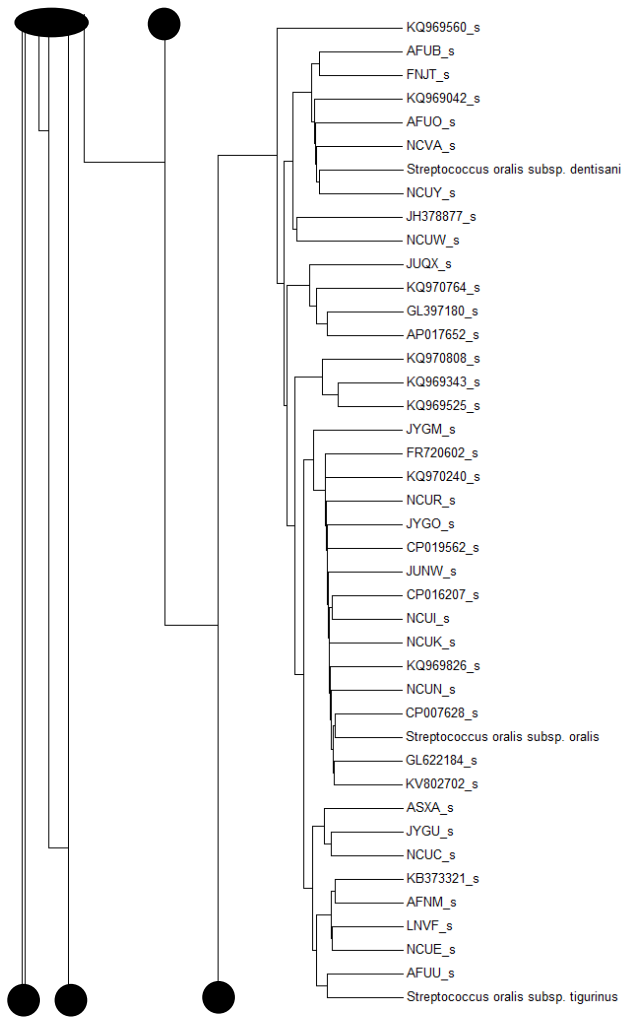


Figure 26. Continuation.

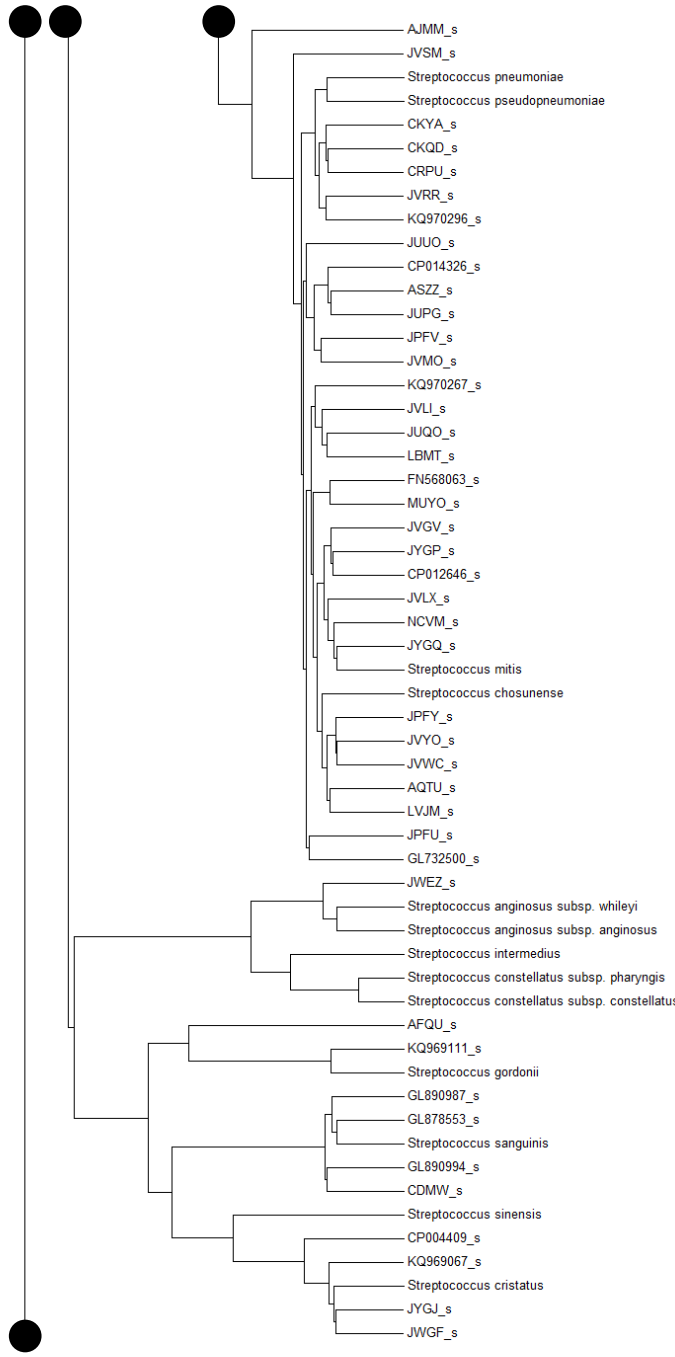


Figure 26. Continuation.

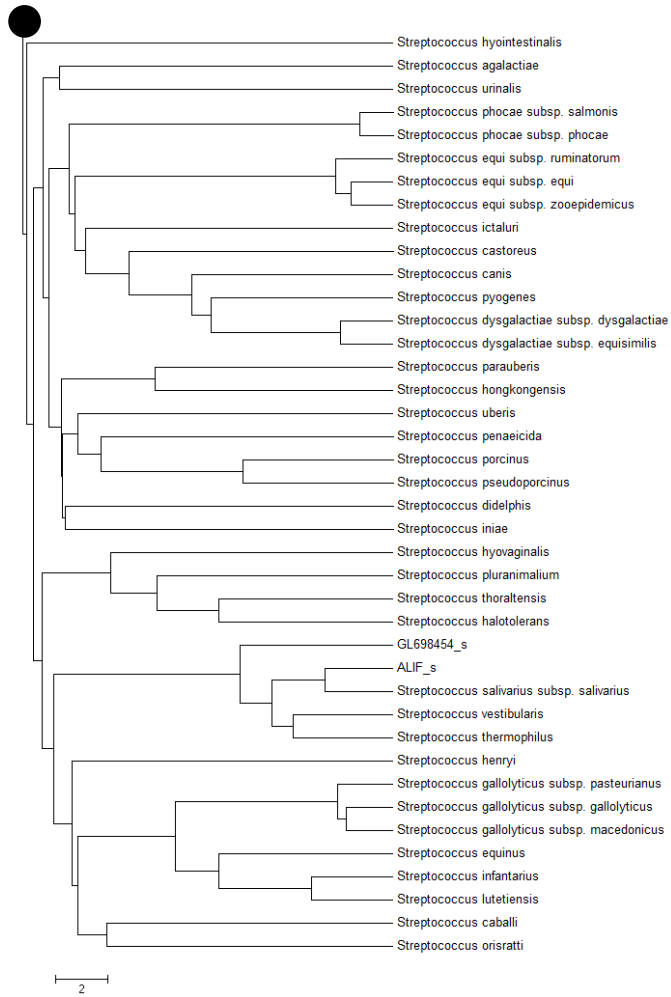


Figure 26. Continuation.

3.3.5. Taxonomic profiling

The publicly available software Kraken was used to build the UBCG k-mer database. Kraken is a k-mer based metagenomic taxonomic classifier that utilizes a modified HashMap created by Jellyfish [33] in order to classify sequences with higher speeds and accuracy. First, using EzBioCloud's taxonomy, we created a taxonomic structure compatible with Kraken and then we converted all the UBCG references to fasta format. Once the references were added, we compiled the database with a k-mer size of 26 bp. A normalization database using Bracken [10] was also generated for different read lengths (100, 200, 250). Bracken distributes reads classified at higher taxonomic levels to a species level classification.

In this study, a total of 5 databases using UBCGs were build, one containing just validly named species (KrakenUBCG-VNS), a second one containing the same valid named species with the addition of genomospecies references (KrakenUBCG) defined by the ANI threshold of 95% proposed by Chun et.al [27]. Lastly, three databases were build using different ANI thresholds (94%, 93% and 92%). For each of these databases, we maintained the 'one reference per taxon' approach, and the method of selecting the representative genome for these databases can be found on section 4.1.

To classify the metagenomes using kraken, all samples were classified with a kraken filter threshold of 0.1 with the exception of section 2.2.1 where different thresholds were used to demonstrate closeness of the reference with the metagenomic reads. Kraken's threshold per read is calculated as the fraction C/Q,

where C is the number of k-mers mapped to the lowest common ancestor values in the clade rooted at the label, and Q is the number of k-mers in the sequence that lack an ambiguous nucleotide.

After the classification by Kraken and the normalization step by Bracken, all metagenomic abundances predicted with our UBCG database were normalized using the total length of all the UBCG belonging to a single species.

To compare our results, we also predicted the abundances at species level using MetaPhlAn2 [11] with the v20 version of database and default parameters. In order to compare fairly our predictions with MetaPhlAn2, we matched their reference markers using the NCBI accession number provided in their database and EzBioCloud's NCBI's accession numbers and matched them when present in their database as seen in Figure 4. Abundance accuracy for the synthetic datasets was measured using the log-modulus difference between the truth and predicted abundances of both pipelines [3]. The log-modulus was calculated as:

$$y' = \text{sign}(y)*\log10(1 + |y|) \quad (1)$$

to preserve the sign of the difference between estimated and expected abundance, y.

3.3.6. Biomarker discovery

Biomarker discovery shown on Figure 5 was calculated using LEfSe [24] which utilizes a linear discriminant analysis (LDA) effect size method to support

high-dimensional class comparisons. LEfSe's biomarkers were found using a P-value <0.05 (Kruskal–Wallis test) and an LDA score (log 10) > 2.0.

3.4. Conclusions

In this study, we demonstrated that UBCG sequences can be used as references for metagenomic classification, showing that they are easy to extract from genome sequences and accurate when predicting relative abundance. However, UBCG sequences can only represent bacterial organisms, so in order to represent other type of organisms, a different set of core genes needs to be used. We also demonstrated that inclusion of the genomospecies in the reference databases significantly improve the classification accuracy of bacterial species within a metagenomic sample. Furthermore, our study implicated that, regardless of the ANI threshold used for species/genomospecies boundaries, if the genomic reference sequence doesn't change, only the name for the given species will be different. Finally, we showed that while public available pipelines and databases are easily accessible, for accurate and reliable taxonomic classification, an updated database with proper taxonomic and genomic curation must be used.

Chapter 4.

A large-scale shotgun metagenomic analysis on *Bacteroides*

4.1. Introduction

In recent times, the increasing recognition of gut microbiota and its role in host metabolism and immunity has promoted an unprecedented interest in developing gut microbiota-related diagnostic and therapeutic targets for many diseases. Next-generation sequencing techniques and multiomics approaches have dramatically expanded our knowledge of the microbial world. The genus Bacteroides includes some of the predominant gut bacteria in humans and is known to have an important role in maintaining a healthy gut ecosystem.7 Individuals classified as enterotype [34], which is characterized by low levels of Bacteroides, have a higher incidence of symptomatic atherosclerosis. Moreover, Bacteroides abundance was found to be decreased in patients with atherosclerotic ischemic stroke and transient ischemic attack.[35].

Also, bacterial species present in the genus Bacteroides are significant clinical pathogens and are found in most anaerobic infections. However, they maintain a complex and generally beneficial relationship with the host when retained in the gut, but when they escape this environment they can cause significant pathology. Genomic and proteomic analyses have vastly added to our understanding of the manner in which Bacteroides adapt to, and thrive in, the human gut. Bacteroides fragilis, which accounts for only 0.5% of the human colonic flora, is the most commonly isolated anaerobic pathogen due, in part, to its potent virulence factors. Species of the genus Bacteroides have the most

antibiotic resistance mechanisms and the highest resistance rates of all anaerobic pathogens.

Bacteroidetes is one of the major lineages of bacteria and arose early during the evolutionary process [36]. Bacteroides species are anaerobic, bile-resistant, non-spore-forming, gram-negative rods. Bacteroides may be passed from mother to child during vaginal birth and thus become part of the human flora in the earliest stages of life [37]. The bacteria maintain a complex and generally beneficial relationship with the host when retained in the gut, and their role as commensals has been extensively reviewed [38]. Bacteroides have also been associated with the development of IBD. Bacteroides expresses polysaccharide A, which can induce regulatory T cell growth and cytokine expression that are protective against colitis [39]. However in recent studies, the correlation between geographical distribution and the presence of specific gut bacteria have been proposed [40][41][42].

In this study, we aim to massively profile thousands of gut microbiome samples and catalog them based on geographic distribution. Are specific species of Bacteroides solely correlated to disease or can it be that the presence of some species can be correlated to proximity? Additionally, is the animal gut microbiome also host for Bacteroides, if so, do they have anything in common with humans?

4.2. Bacteroides on the human gut

4.2.1. Collecting the samples

Public repositories were searched and a total of 2,719 human gut metagenomic samples were collected. Table 5 describes the accession numbers divided by country and their respective study title. These samples were downloaded and stored, comprising a total of 45 terabytes of raw fastq data. Table 6 shows the number of samples downloaded per continent. The number of samples from South America was low and for the most part they were excluded from this analysis.

4.2.2. Methods

4.2.2.1. Reference Genomes

A bacterial reference database was built as described on Chapter 2. References used for the genus Bacteroides can be seen on Table 4. EzBioCloud's [18] taxonomy was implemented and a policy of one genome per bacterial species was implemented as described on Chapter 3. A total of 92 Up-to-date Bacterial Core Genes (UBCG) [12] sequences were extracted for each representative genomic reference from the database. Figure 27 displays the ANI dendrogram tree calculated using the OrthoANIu [43] algorithm. Hierarchical clustering was carried out from the ANI matrix by applying the UPGMA (unweighted pair group method with arithmetic mean) algorithm using the R library phangorn [32]. Figure 28

shows the unrooted maximum likelihood tree using UBCG sequences inferred using the GTR + CAT model using the RAxML pipeline [44].

4.2.2.2. Metagenome profiling

Using the k-mer approach described on Chapter 2 and 3 (KrakenUBCG), all samples were profiled in a sequential manner. All 2719 human samples were profiled in 14.07 days of computational time using a server with 30 threads. Figure 29 shows a Histogram demonstrating how much computational time was required for al the samples, being most samples taking around 5 minutes of computational time required for processing. On average, KrakenUBCG classifies 12.7 million reads per minute on a 30 CPU server (Figure 30) while MetaPhlAn2 classifies 604 thousand reads per minute. This makes KrakenUBCG 21-Fold faster than MetaPhlan2, potentially taking 295 days of processing time for MetaPhlAn2 to profile all 2,719 samples (Figure 31). Samples with less than 1% of any *Bacteroides* species were excluded. Abundance percentage was normalized using the total length of the UBCG sequences for every *Bacteroides* species. Zero values were ignored when calculating the median abundance for any given species.

Table 4. Genomes used for this study.

Accession Number	BioProject	Taxon name
GCA_000011065.1	PRJNA62913	Bacteroides thetaiotaomicron
GCA_000012825.1	PRJNA58253	Bacteroides vulgatus
GCA_000186225.1	PRJNA62135	Bacteroides helcogenes
GCA_000190575.1	PRJNA63269	Bacteroides salanitronis
GCA_000212915.1	PRJNA66921	Bacteroides coprosuis
GCA_000381365.1	PRJNA201685	Bacteroides salyersiae
GCA_000382445.1	PRJNA201686	Bacteroides massiliensis
GCA_000613465.1	PRJNA224116	Bacteroides nordii
GCA_000613745.1	PRJNA224116	Bacteroides propionicifaciens
GCA_000613805.1	PRJNA224116	Bacteroides paurosaccharolyticus
GCA_000297695.1	PRJNA181634	Bacteroides JH815484_s
GCA_000315485.1	PRJNA182882	Bacteroides oleiciplenus
GCA_000428105.1	PRJNA224116	Bacteroides pyogenes
GCA_000428125.1	PRJNA224116	Bacteroides graminisolvens
GCA_000210075.1	PRJNA39177	Bacteroides xylanisolvens
GCA_000154205.1	PRJNA54547	Bacteroides uniformis
GCA_000154525.1	PRJNA54825	Bacteroides stercoris
GCA_000154845.1	PRJNA54879	Bacteroides coprocola
GCA_000172175.1	PRJNA54881	Bacteroides intestinalis
GCA_000156195.1	PRJNA54985	Bacteroides finegoldii
GCA_000187895.1	PRJNA54991	Bacteroides plebeius
GCA_000156075.1	PRJNA54993	Bacteroides dorei
GCA_000158035.1	PRJNA55279	Bacteroides cellulosilyticus
GCA_000157915.1	PRJNA55301	Bacteroides coprophilus
GCA_000513195.1	PRJNA224116	Bacteroides timonensis
GCA_000195635.1	PRJNA66157	Bacteroides fluxus
GCA_000374365.1	PRJNA199285	Bacteroides gallinarum
GCA_000374585.1	PRJNA199296	Bacteroides barnesiae
GCA_000226135.2	PRJNA86875	Bacteroides faecis
GCA_000517545.1	PRJNA224116	Bacteroides reticulotermitis
GCA_000499785.1	PRJNA224116	Bacteroides neonati
GCA_000614125.1	PRJDB600	Bacteroides rodentium
GCA_001314995.1	PRJNA289334	Bacteroides ovatus
GCA_900129065.1	PRJEB18171	Bacteroides faecichinchillae
GCA_900142015.1	PRJEB18311	Bacteroides stercorirosoris
GCA_900128905.1	PRJEB18170	Bacteroides luti

Table 4 continuation.

Accession Number	BioProject	Taxon name
GCA_900129655.1	PRJEB18217	Bacteroides clarus
GCA_900128455.1	PRJEB18046	Bacteroides mediterraneensis
GCA_900128495.1	PRJEB18049	Bacteroides ilei
GCA_900130135.1	PRJEB18269	Bacteroides togonis
GCA_900104585.1	PRJEB16348	Bacteroides ihuae
GCA_900108345.1	PRJEB16738	Bacteroides ndongoniae
GCA_900130125.1	PRJEB18268	Bacteroides congonensis
GCA_900155865.1	PRJEB18812	Bacteroides bouchesdurhonensis
GCA_001688725.2	PRJNA317592	Bacteroides caecimuris
GCA_002160055.1	PRJNA377666	Bacteroides DQ456084_s
GCA_002222615.2	PRJNA393727	Bacteroides caccae
GCA_002632415.1	PRJNA397629	Bacteroides NQMG_s
GCA_900241005.1	PRJEB22714	Bacteroides cutis
GCA_002998435.1	PRJNA282954	Bacteroides zoogleoformans
GCA_900291465.1	PRJEB24949	Bacteroides LT985808_s
GCA_003096855.1	PRJNA439857	Bacteroides galacturonicus
GCA_003437535.1	PRJNA482748	Bacteroides QSQT_s
GCA_003472565.1	PRJNA482748	Bacteroides QRME_s
GCA_003474285.1	PRJNA482748	Bacteroides QROH_s
GCA_003479375.1	PRJNA482748	Bacteroides QUHA_s
GCA_003865075.1	PRJDB7416	Bacteroides faecalis
GCA_007341375.1	PRJNA553582	Bacteroides koreensis
GCA_007341395.1	PRJNA553353	Bacteroides kribbi
GCA_010206385.1	PRJNA551571	Bacteroides acidifaciens
GCA_900540105.1	PRJEB26432	Bacteroides caecicola
GCA_900544075.1	PRJEB26432	Bacteroides caecigallinarum
GCA_000155815.1	PRJNA54989	Bacteroides eggerthii
GCA_000025985.1	PRJNA57639	Bacteroides fragilis
GCA_000432695.1	PRJNA221957	Bacteroides gallinaceum
GCA_004342845.1	PRJNA519314	Bacteroides heparinolyticus
GCA_000614185.1	PRJDB603	Bacteroides sartorii
GCA_000511775.1	PRJNA224116	Bacteroides tectus

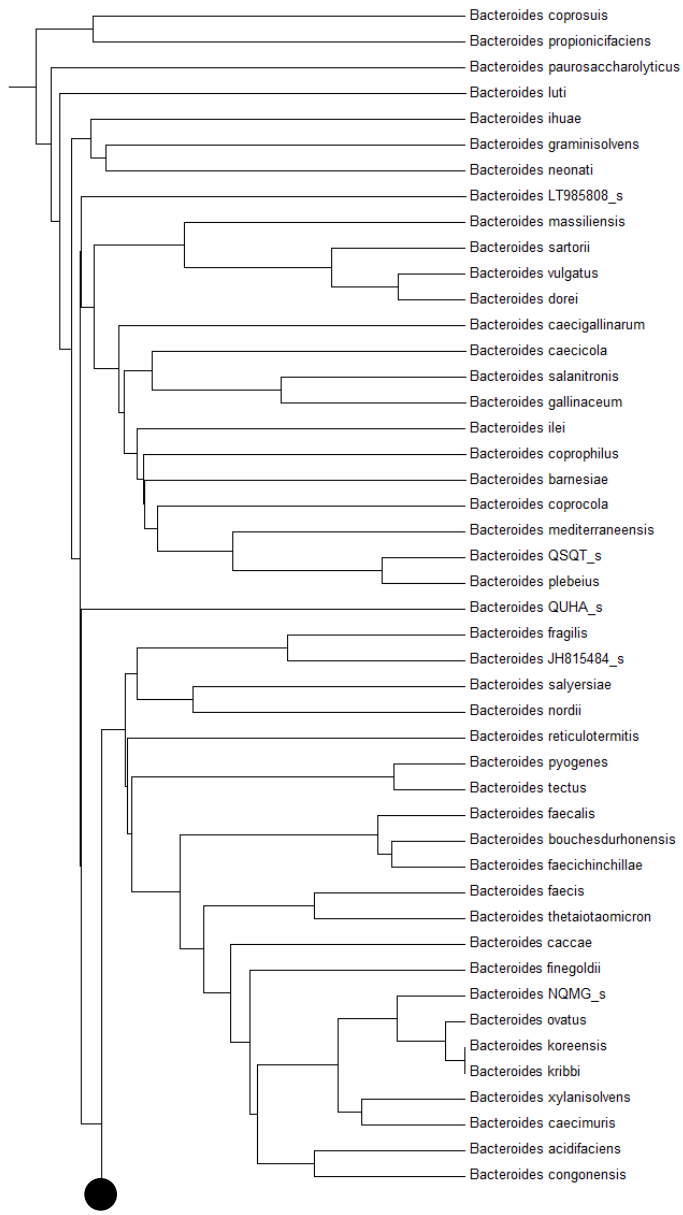


Figure 27. ANI dendrogram for the genus *Bacteroides*.

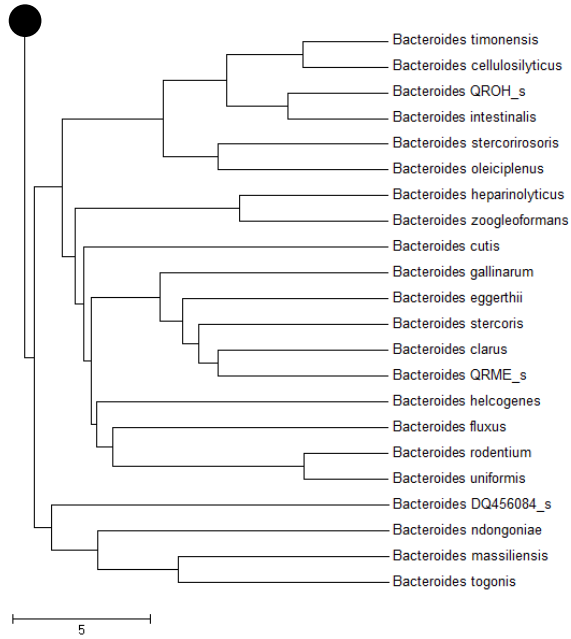


Figure 27. Continuation.

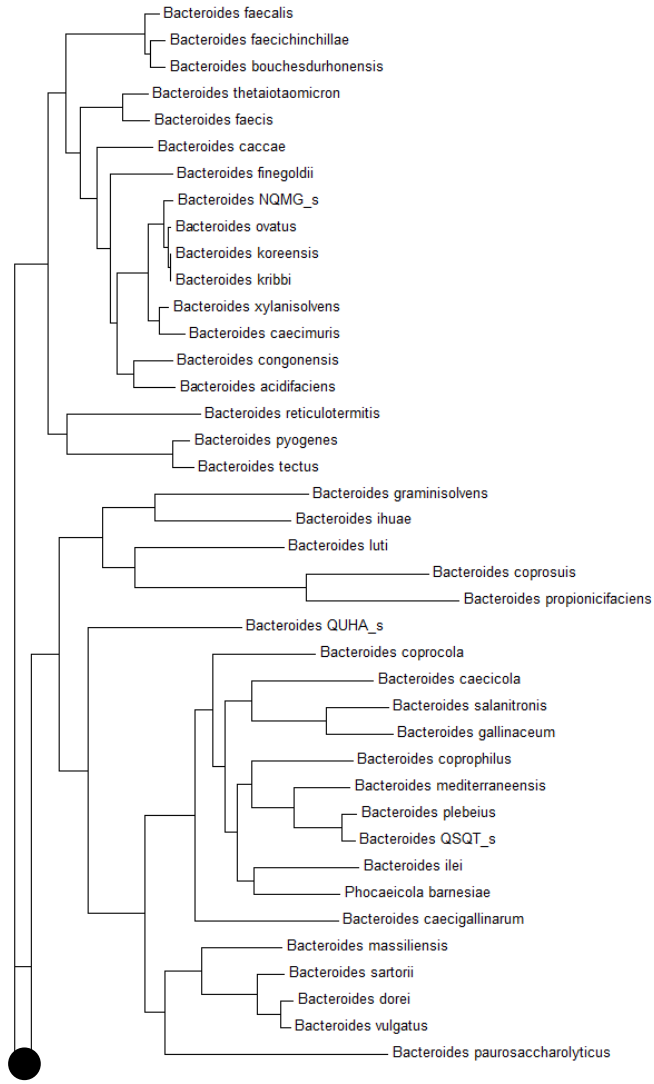


Figure 28. UBCG phylogenetic tree of the genus *Bacteroides.*

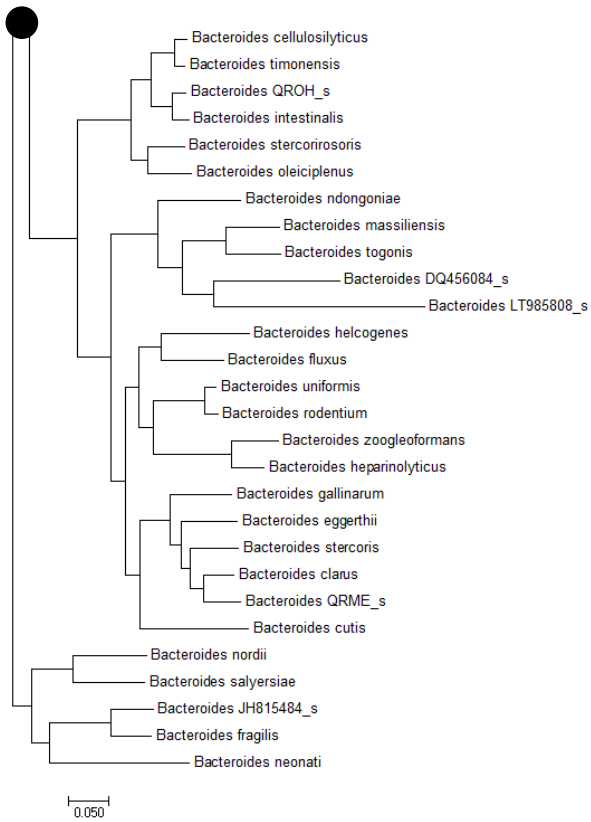


Figure 28. Continuation.

Table 5. Table describing the human metagenomic samples used on this study.

Country	# of subjects	Title	Accessions
Austria	63	Gut microbiome development along the colorectal adenoma-carcinoma sequence	ERP008729
Bangladesh	34	Gut microbial succession follows acute secretory diarrhea in humans.	PRJEB9150
Canada	24	The initial state of the human gut microbiome determines its reshaping by antibiotics	PRJEB8094
China	114	Alterations of the human gut microbiome in liver cirrhosis	ERP005860
China	53	Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer	PRJEB10878
China	41	Gut microbiota dysbiosis contributes to the development of hypertension	PRJEB13870
China	31	The Gut Microbiome Signatures Discriminate Healthy From Pulmonary Tuberculosis Patients	SRP118759
China	71	Breast cancer in postmenopausal women is associated with an altered gut metagenome	PRJNA453965
Denmark	177	Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes	ERP002061
Denmark	292	Richness of human gut microbiome correlates with metabolic markers	ERP003612
Ethiopia	24	The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations	PRJNA529124, PRJNA529400, PRJNA504891, PRJEB31971
France	61	Potential of fecal microbiota for early-stage detection of colorectal cancer	ERP005534, ERA000116, ERP003612
Germany	5	Potential of fecal microbiota for early-stage detection of colorectal cancer	ERP005534, ERA000116, ERP003612
Germany	7	Temporal and technical variability of human gut metagenomes	ERP009422
Germany	28	Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients.	ERP019674
Ghana	23	The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations	PRJNA529124, PRJNA529400, PRJNA504891,PRJEB31971
India	88	The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches	PRJNA397112, SRR059347, ERP000108, SRR341581
Israel	20	Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses	PRJEB17643
Italy	11	Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota	SRP056480, mgp8810
Italy	5	Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling.	PRJNA339914
Italy	101	Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets	SRP126540, SRP083099
Japan	106	The gut microbiome of healthy Japanese and its microbial and functional uniqueness	PRJDB3601

Table 5 Continuation.

Country	# of subjects	Title	Accessions
Korea	20	Stability of gut enterotypes in Korean monozygotic twins and their association with biomarkers and diet	ERP002391
Korea	17	The effects of sequencing platforms on phylogenetic resolution in 16 S rRNA gene profiling of human feces	PRJEB17896
Luxembourg	11	Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes	PRJNA289586
Madagascar	111	Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.	PRJNA485056, PRJNA504891
Mongolia	110	Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis	SRP080787, ERP00586014, SRA04564614, ERA0001162
Netherlands	393	Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity.	PRJNA319574
USA (Oklahoma)	18	Gut microbiome diversity among Cheyenne and Arapaho individuals from western Oklahoma	PRJNA268964, PRJNA299502
Peru	9	Subsistence strategies in traditional societies distinguish gut microbiomes	PRJNA268964
Peru	7	Subsistence strategies in traditional societies distinguish gut microbiomes	PRJNA268964
Russia	95	Human gut microbiota community structures in urban and rural populations in Russia	PRJNA176385
Spain	24	Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes	ERP002061
Sweden	39	Gut metagenome in European women with normal, impaired and diabetic glucose control	ERP002469
Sweden	100	Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life	ERP005989
Switzerland	11	Association of the Intestinal Microbiome with the Development of Neovascular Age-Related Macular Degeneration	PRJEB13835
Switzerland	30	Retinal artery occlusion is associated with compositional and functional shifts in the gut microbiome and altered trimethylamine-N-oxide levels	PRJEB24557
Tanzania	36	The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations	PRJNA529124, PRJNA529400, PRJNA504891, PRJEB31971
Tanzania	22	Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota	SRP056480, mgp8810
Thailand	15	US Immigration Westernizes the Human Gut Microbiome	PRJEB28687
USA	19	Subsistence strategies in traditional societies distinguish gut microbiomes	PRJNA268964
USA	37	Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans	PRJEB29065
USA	213	Strains, functions and dynamics in the expanded Human Microbiome Project	PRJNA48479, PRJNA275349
Venezuela	3	The microbiome of uncontacted Amerindians	SRP049631

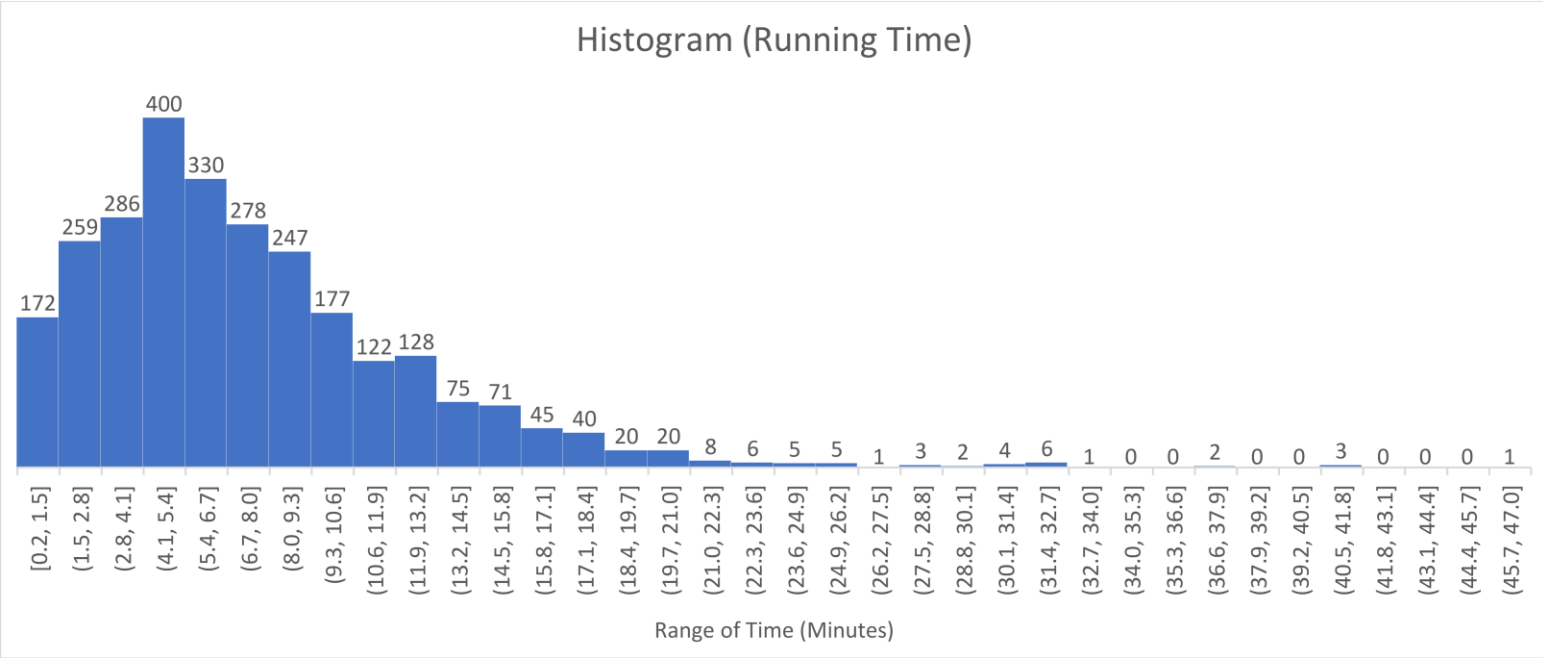


Figure 29. Histogram showing the range of running time that it took for all human samples to be profiled.

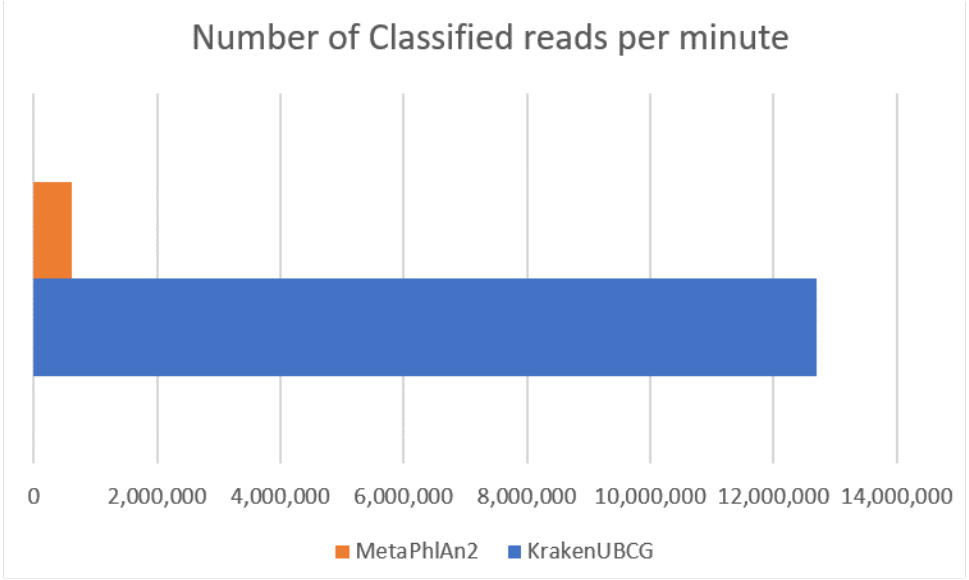


Figure 30. Number of reads that can be classified per minute at 30 threads.

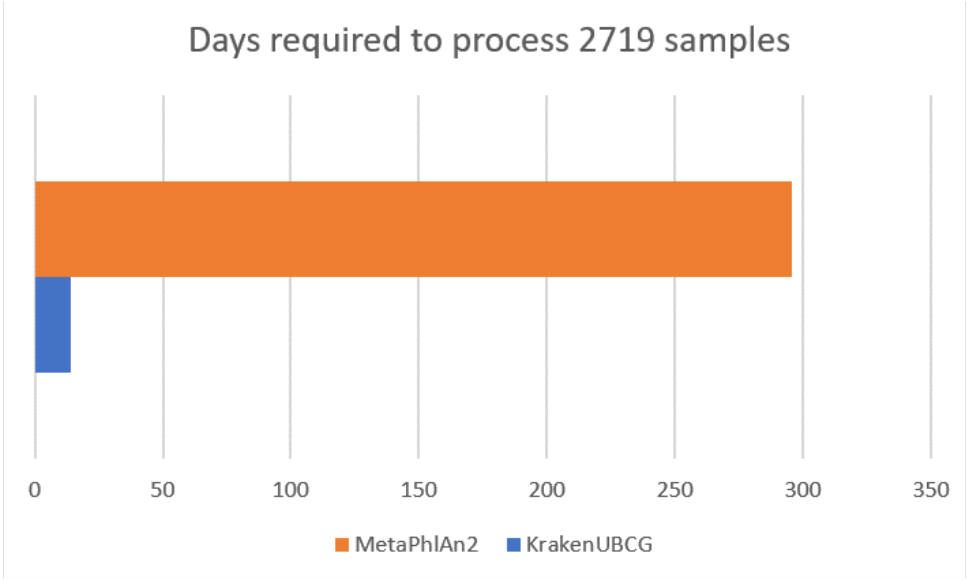


Figure 31. Number of days required to profile all 2719 samples.

Table 6. Number of metagenomic samples per country.

Continent Name	Number of Samples
Asia	720
South America	19
North America	311
Africa	216
Asia Europe	95
Europe	1358

4.2.3. Results

2719 human gut samples were profiled and searched for Bacteroides. Samples with less than 1% of total abundance of Bacteroides were removed from the analysis. A total of 2371 profiles were left for the analysis. From the 69 different species and genomospecies contained on the genus *Bacteroides*, 47 were found present in the human gut samples.

Bacteroides vulgatus was present in 1,598 of the samples (Figure 32), making it the most present *Bacteroides* species in the human gut samples profiled in this study. *Bacteroides uniformis* was the second most present in the samples, with 1,547 samples, followed by *Bacteroides galacturonicus* with 1,263 samples containing this species. Finally, *Bacteroides dorei* was the 4th most populous species with 730 samples containing this bacteria. If we exclude all *Bacteroides* with low abundance (Figure 33), only 725 samples contain more than 5% abundance of *Bacteroides vulgatus,* making a decrease of more than half of the samples, however, it remains to be the most present *Bacteroides* in the metagenome samples. *Bacteroides uniformis* also decreased in a similar way, but also remaining at second place with 609 samples. The biggest decrease was *Bacteroides galacturonicus,* with only 192 samples containing 5% abundance of this bacteria; this is a 6-fold decrease. Finally, *Bacteroides dorei* similarly decreased by 3-fold, with only 233 samples containing more than 5% abundance.

Figure 34 shows a Venn diagram, illustrating how many samples contain one or all of these top 4 *Bacteroides* species. If we consider all samples with 1%

abundance or more, 215 samples contain all 4 species; 393 samples contain all except *Bacteroides dorei;* 172 samples contain all except *Bacteroides galacturonicus*. The most common set of *Bacteroides* pair was *Bacteroides vulgat*us and *Bacteroides galacturonicus* with 166 samples; on the other hand, only 19 samples *Bacteroides dorei* and *Bacteroides vulgatus*; similarly, only 25 samples contained *Bacteroides dorei* and *Bacteroides galacturonicus*.

Figure 35 shows a similar Venn diagram, but this time only considering the samples with more than 5% abundance. As shown, only one sample contains all four *Bacteroides* species with more than 5%. In fact, at this high abundance threshold, it can be seen that the highest number of samples are for those than only contain one species. This could mean that once a gut contains a high abundant *Bacteroides* species, it is difficult for a second one to co-exist, although pairings can still be seen, with *Bacteroides dorei* and *Bacteroides galacturonicus* being the most abundant.

Analyzing abundances per country as shown on Figure 36 can yield interesting findings. The Hadza (Tanzania) samples only contained one species of Bacteroides (*Bacteroides galacturonicus*) at a median abundance of 1.2%. Similarly, samples from rural tribes contained little diversity of Bacteroides species. Samples from Peru only contained a median abundance of 2.2% of *Bacteroides galacturonicus; Y*anomami Amerindian samples contained 2 different species, *Bacteroides intestinalis* and *Bacteroides sartorii* at 3.4% and 2.8% median abundance, respectively. Matses (also from Peru) showed no presence of any

species of *Bacteroides*. These findings demonstrate how the common belief that the presence of Bacteroides is a consequence of industrialism, since all the other samples contain high diversity and median abundance of *Bacteroides*.

As seen on Figure 36, *Bacteroides togonis* is highly abundant in European samples. With the exception of China, there is no presence of this species in North America, Asia and Africa. *Bacteroides coprocola* can be found in most countries except for African samples (Ethiopia, Madagascar, Tanzania) and countries in center Asia (Mongolia, Bangladesh).

Respecting its name, *Bacteroides mediterraneensis* can only be seen in samples from countries close to the Mediterranean (Russia, France, Netherlands, Switzerland, Denmark, Sweden).

While abundance of a specific species can be significant, geographicaly just the presence could be meaningful. Figure 38 shows the number of simultaneous Bacteroides species in a given individual per location. It can be observed that individuals from developed countries have the highest number of simultaneous *Bacteroides* species in their gut. Ethiopia, Ghana, Madagascar, Peru, Hadza tribe, Hmong tribe and the Yanomami all had a median presence of one *Bacteroides* species. China had a median of 6 *Bacteroides* species, the highest of any Asian country, followed by South Korea with 4. The USA and Canada had a median of 6 *Bacteroides* per sample.

If we observe the presence (Figure 39) of a specific species from *Bacteroides* and quantify per country how many samples (ratio) contained this

given species, interesting findings can be seen. Figure 40 shows a world map figure, displaying the percentage of samples from a specific country that contains the presence of *Bacteroides vulgatus*. In this case, this species is greatly present worldwide, with several countries having a 100% of samples with this species (China, France, Germany, India, Japan, Netherlands, South Korea and the USA). Only 1% of the Madagascar samples contained this species and 27% of the Hmong samples, these two being the geographical groups with less samples with this species. In the same way, Figure 41 shows the ratio of samples that contain *Bacteroides stercoris*. There highest countries with this species are China and the USA with 54% and 52% of the samples. *Bacteroides uniformis* is also highly present worldwide (Figure 42) with the exception of samples from Africa; only 3% of the samples from Madagascar had the presence of this species, 30% of Tanzanian samples and 18% from the Hmong group.

Bacteroides dorei (Figure 43) one of the most abundant *Bacteroides* species worldwide, is also highly present in most locations, although there is no country with a high percentage of presence, with Japan being the highest with a 69% of presence. *Bacteroides galacturonicus* on the other hand, is dominant on African samples (Figure 44); Ghana, Madagascar, Tanzania and Hmong show a 100% presence of this bacteria. This is the only *Bacteroides* species that is dominant on African samples. Other countries with a 100% presence are Russia and Mongolia. *Bacteroides caccae* (Figure 45) has a 53% presence on Indian samples, 47% on USA samples and 45% on South Korean samples.

All remaining countries showed a low abundance between 10-35% with the exception of Madagascar, showing zero presence of this bacteria. North American samples (USA and Canada) show the highest presence of *Bacteroides xylanisolvens* (Figure 46) in their samples (57% and 45% respectively). European samples have relatively low presence of this species, with their presence being less than 18%.

Bacteroides ovatus (Figure 47) is also another species that is only highly present on North American samples (USA 47% and Canada 45%), while the rest of the world shows lower presence. Countries with zero *Bacteroides ovatus* are Israel and Tanzania; only 1% presence in Madagascar and 2% presence in Italy and Russia. Bacteroides fragilis (Figure 48) was found on a 100% of the samples from Bangladesh, this could be explained by the study where the original researchers analyzed samples with diarrhea (Table 5).

Originally identified as *Bacteroides plebeius*, *Bacteroides* QSQT_s was submitted to NCBI in August 2018 from Guangdong China. ANI evidence shows that this assembly has less than 95% ANI identity with *Bacteroides plebeius* and it was reclassified as a genomospecies by the EzBioCloud database (Figure 49). We discovered that 90% of human samples from South Korea (Figure 50) had the presence of this genomospecies, followed by India (56%) and China (46%). The fact that the genome assembly originated in Asia could explain why samples from countries geographically close contain the presence of this species.

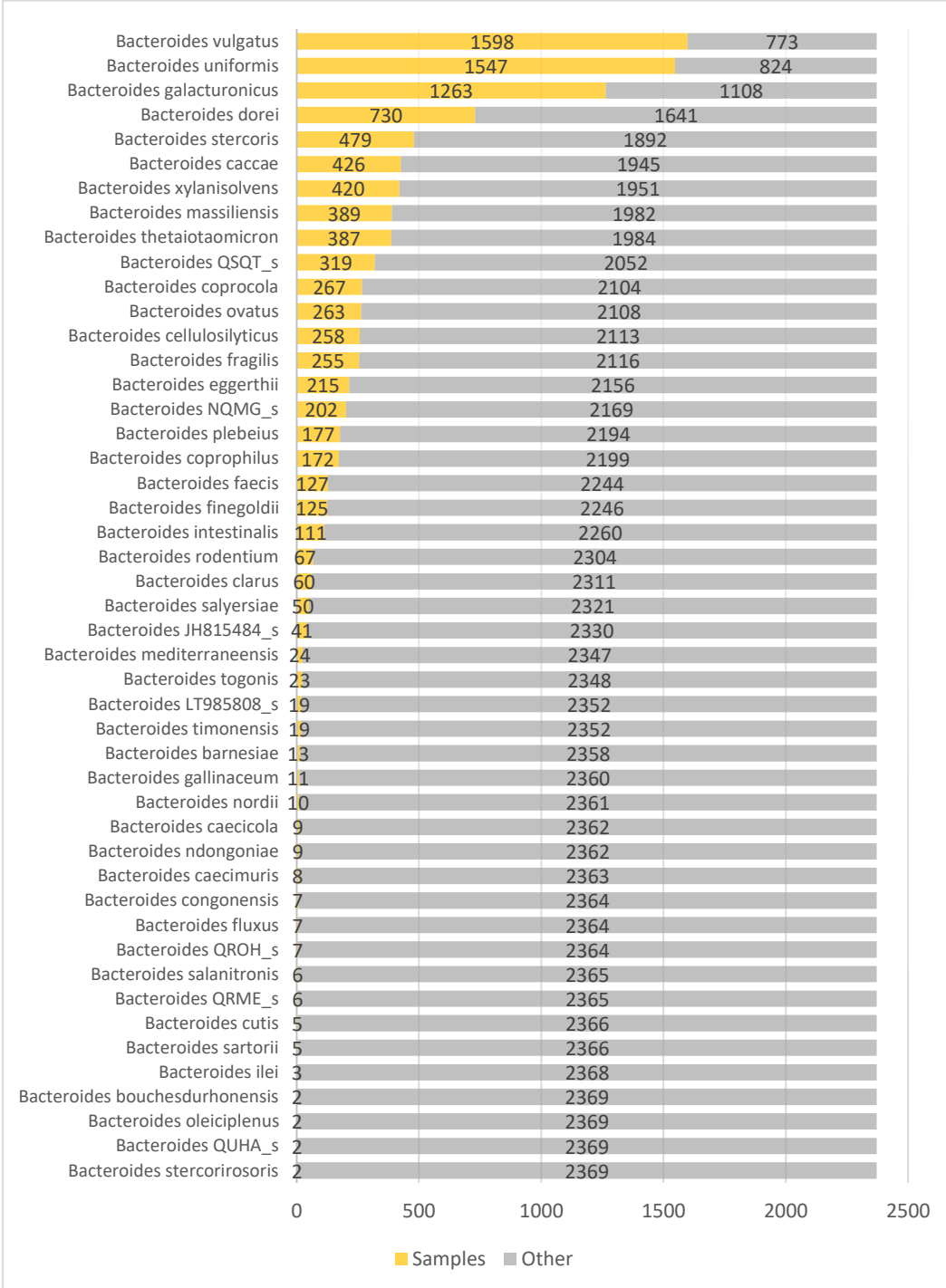


Figure 32. Metagenome samples containing <1% Bacteroides abundance and the presence of a specific species.

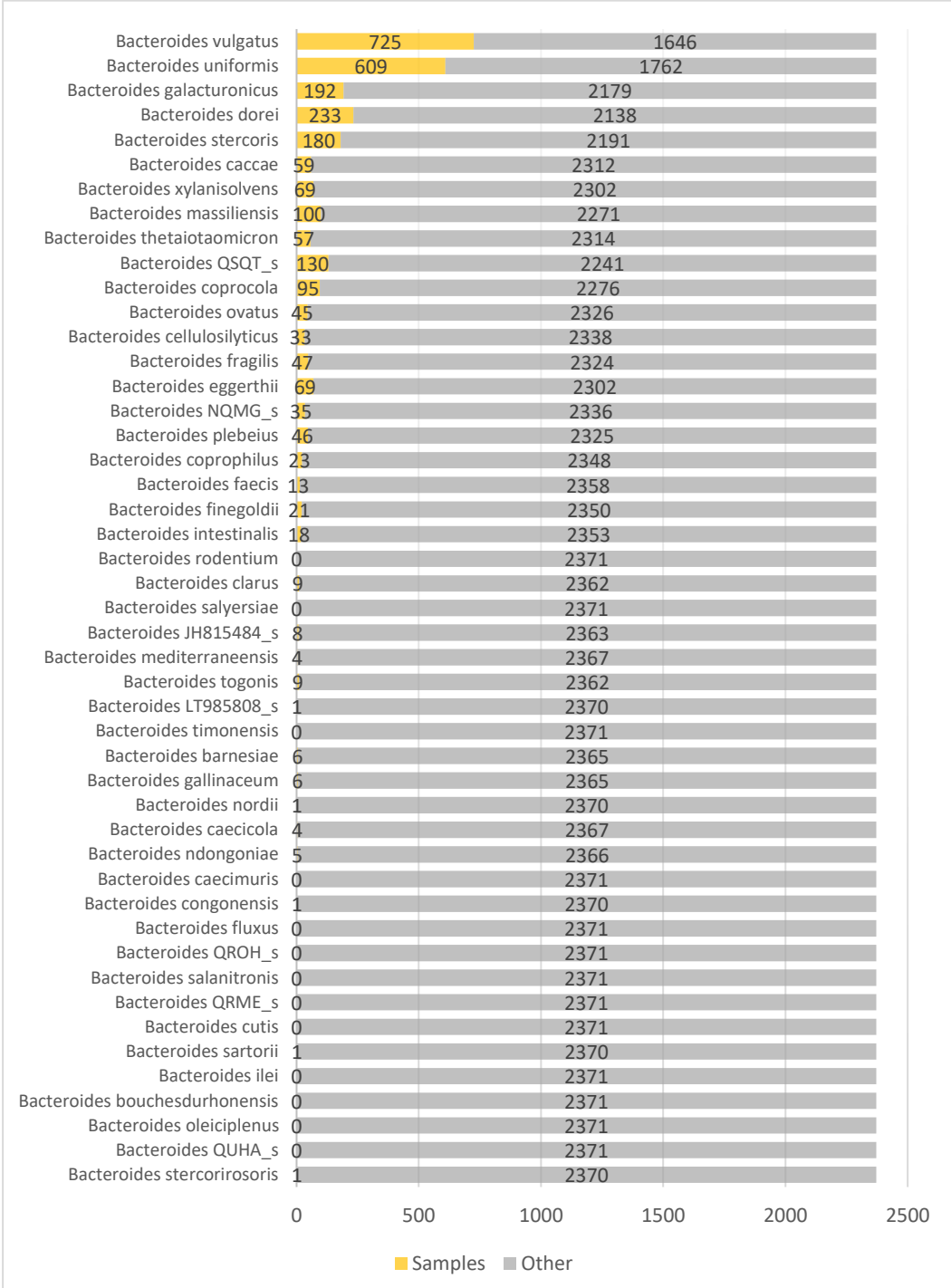


Figure 33. Metagenome samples containing <5% Bacteroides abundance and the presence of a specific species.

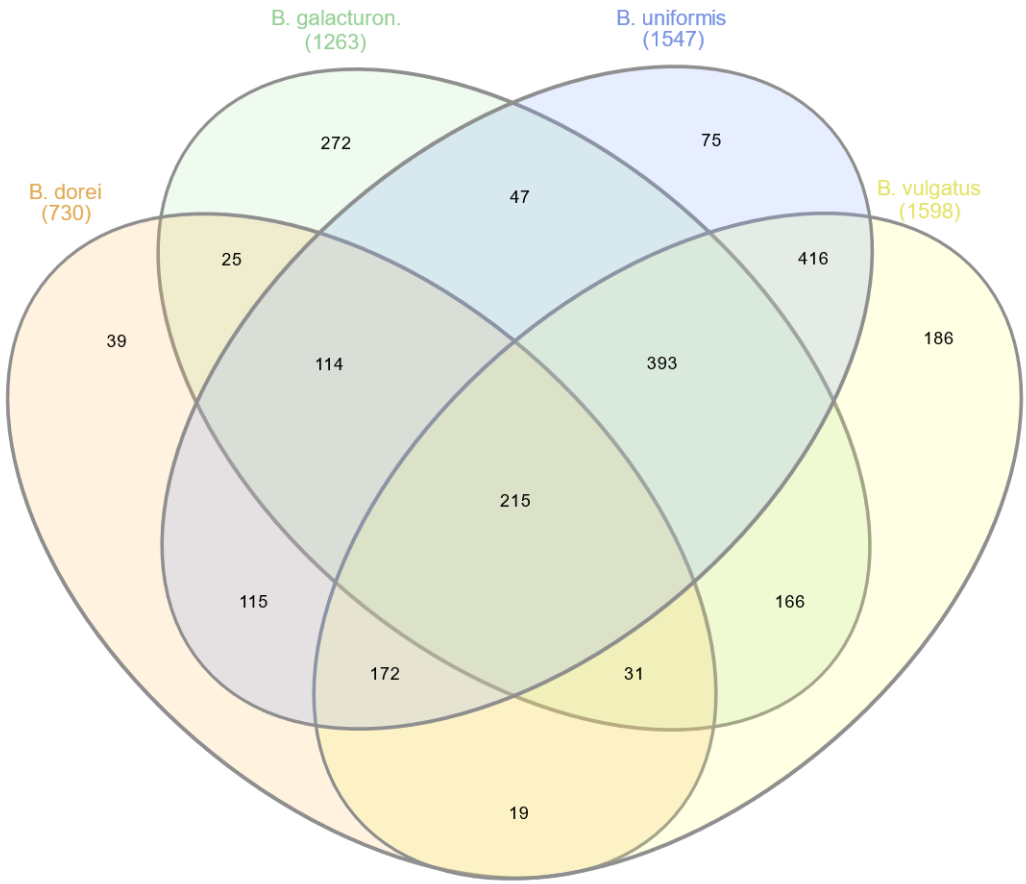


Figure 34. Venn diagram showing the number of metagenome samples with 4 most abundant Bacteroides species present with <1% abundance.

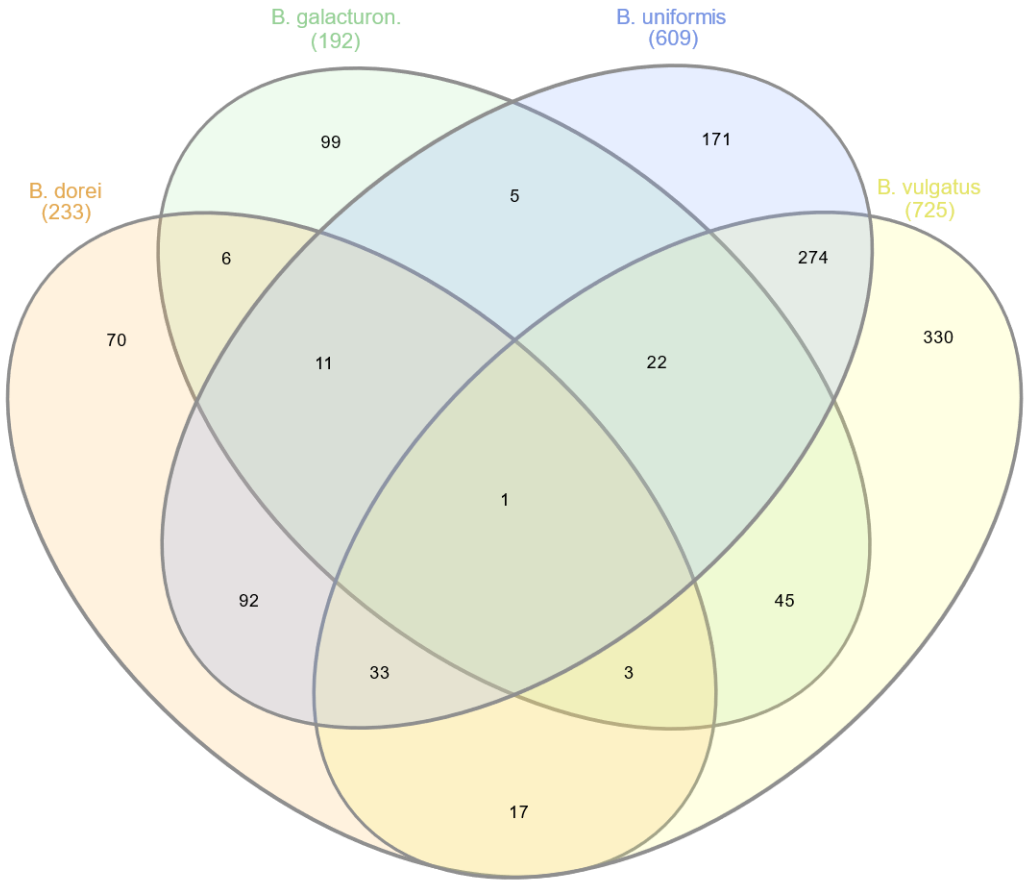


Figure 35. Venn diagram showing the number of metagenome samples with 4 most abundant Bacteroides species present with <5% abundance.

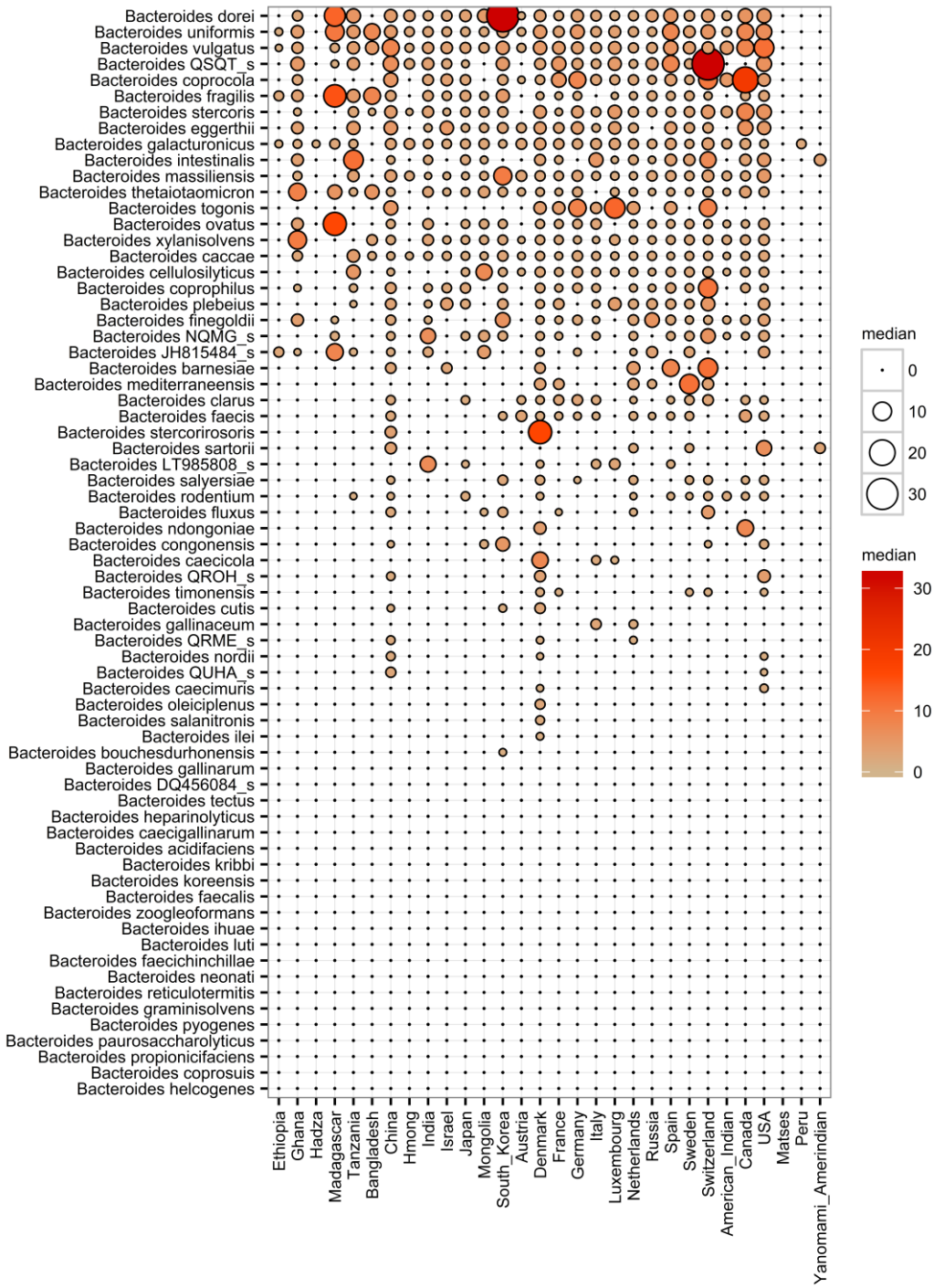


Figure 36. Abundance of Bacteroides per species for each country analyzed.

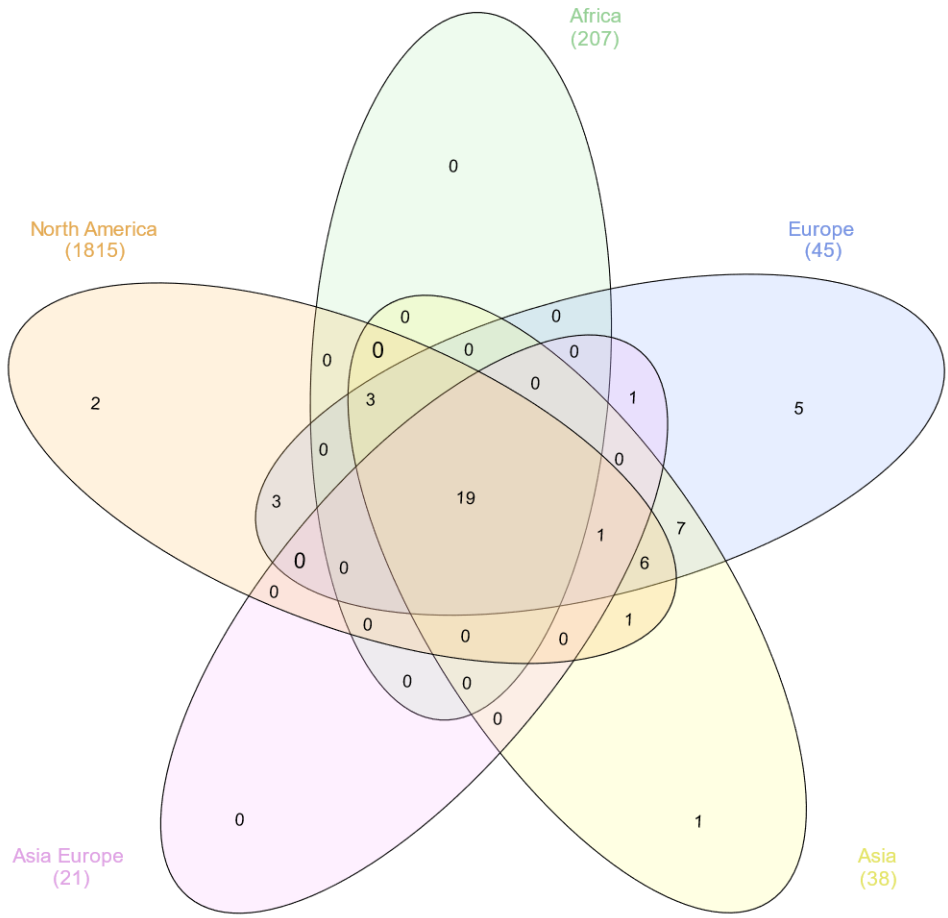


Figure 37. Venn diagram showing the number of *Bacteroides* species overlapping (presence) per continent.

Table 7. *Bacteroides* species present in two or more continents.

Continents	Total	Species
Africa, Asia, Asia Europe, Europe, North America	19	*Bacteroides plebeius*
		Bacteroides uniformis
		Bacteroides fragilis
		Bacteroides cellulosilyticus
		Bacteroides JH815484_s
		Bacteroides QSQT_s
		Bacteroides finegoldii
		Bacteroides stercoris
		Bacteroides intestinalis
		Bacteroides coprocola
		Bacteroides vulgatus
		Bacteroides galacturonicus
		Bacteroides coprophilus
		Bacteroides ovatus
		Bacteroides xylanisolvens
		Bacteroides caccae
		Bacteroides dorei
		Bacteroides thetaiotaomicron
		Bacteroides massiliensis
Africa, Asia, Europe, North America	3	*Bacteroides NQMG_s*
		Bacteroides eggerthii
		Bacteroides rodentium
Asia, Asia Europe, Europe, North America	1	*Bacteroides faecis*
Asia, Europe, North America	6	*Bacteroides congonensis*
		Bacteroides sartorii
		Bacteroides salyersiae
		Bacteroides QROH_s
		Bacteroides nordii
		Bacteroides clarus
Asia, North America	1	*Bacteroides QUHA_s*
Europe, North America	3	*Bacteroides ndongoniae*
		Bacteroides timonensis
		Bacteroides caecimuris

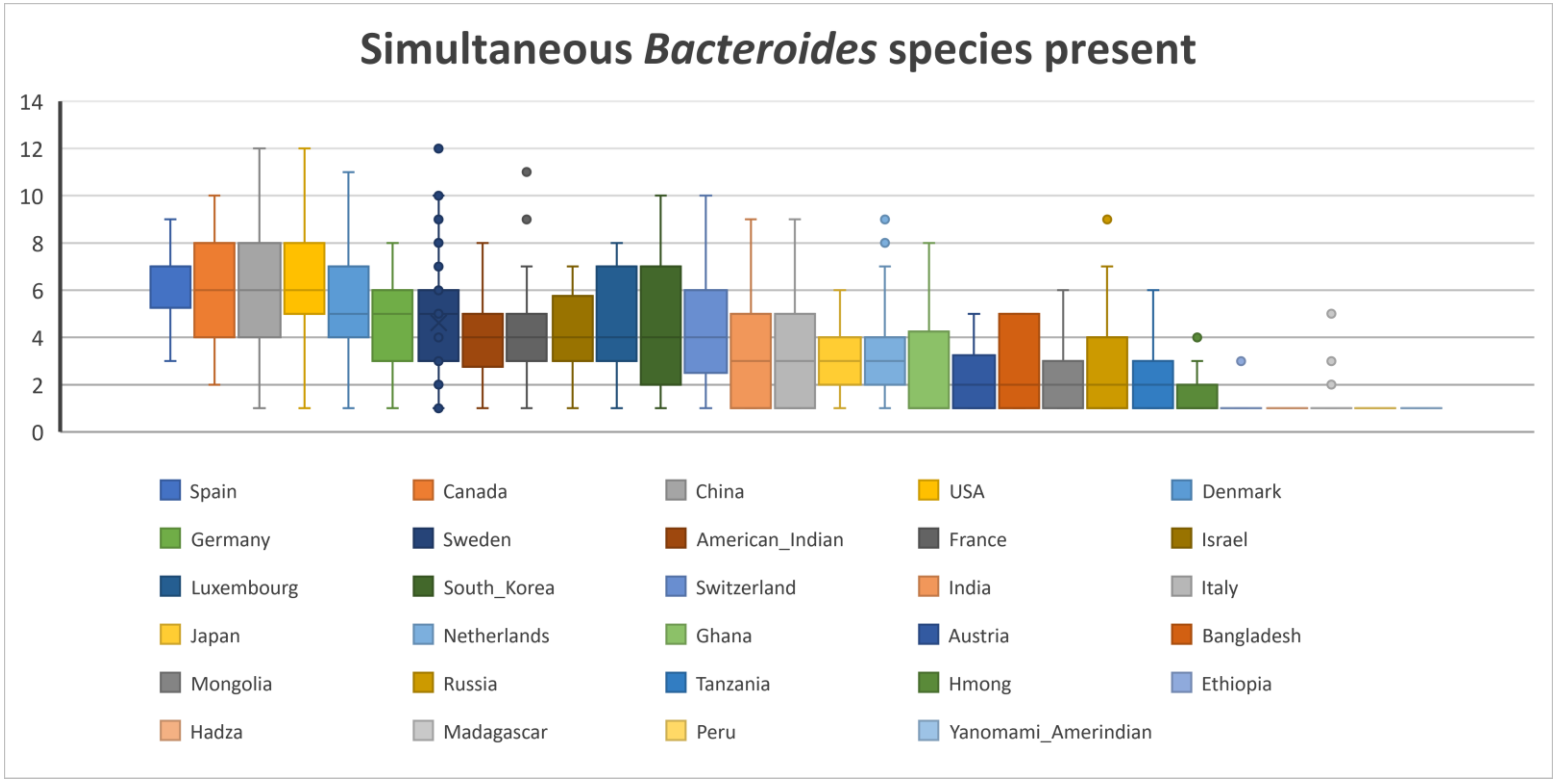


Figure 38. Number of *Bacteroides* species present simultaneously on the human gut.

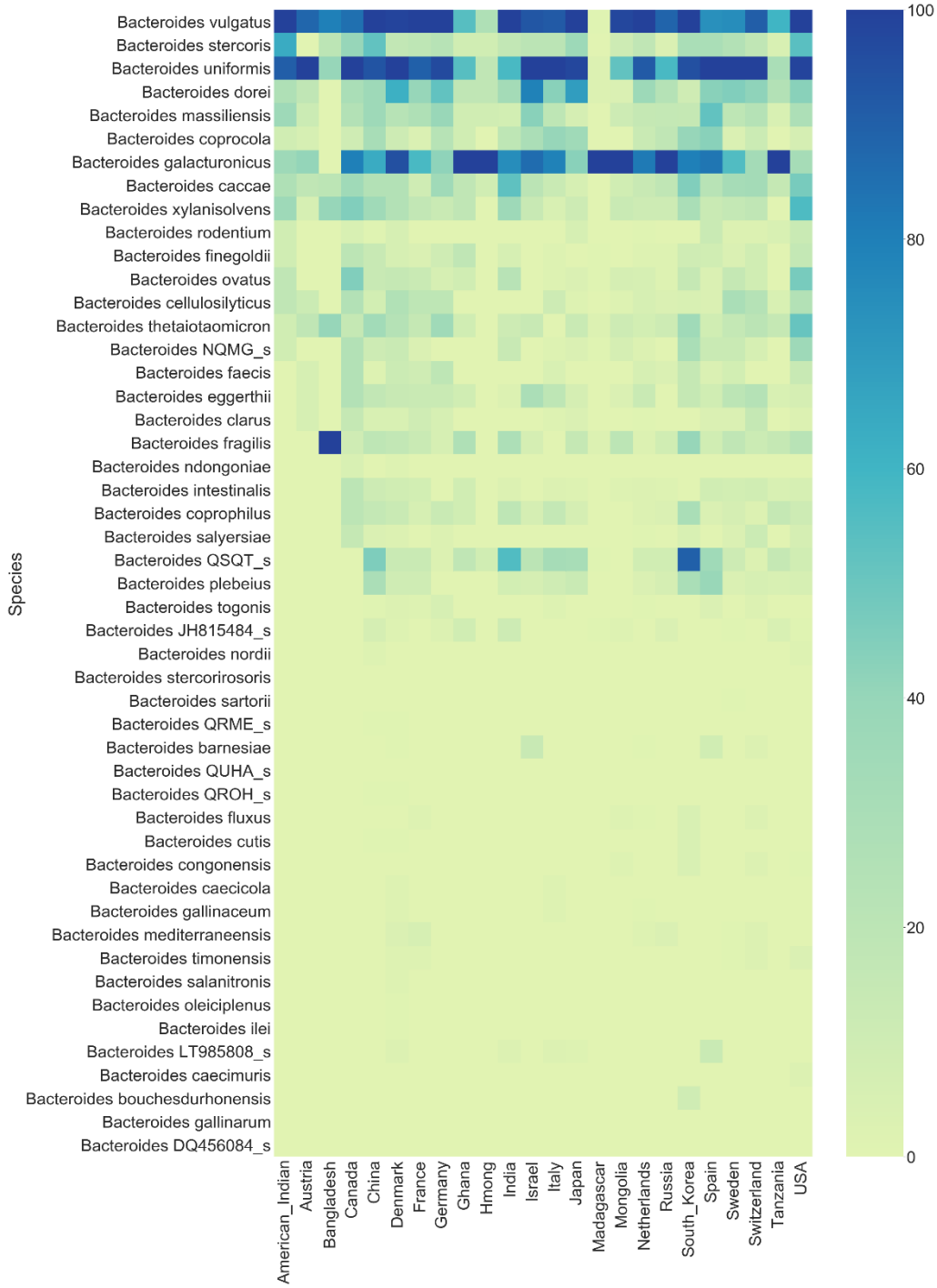


Figure 39. Heatmap showing the percentage of samples for each country that contain any given *bacteroides* species.

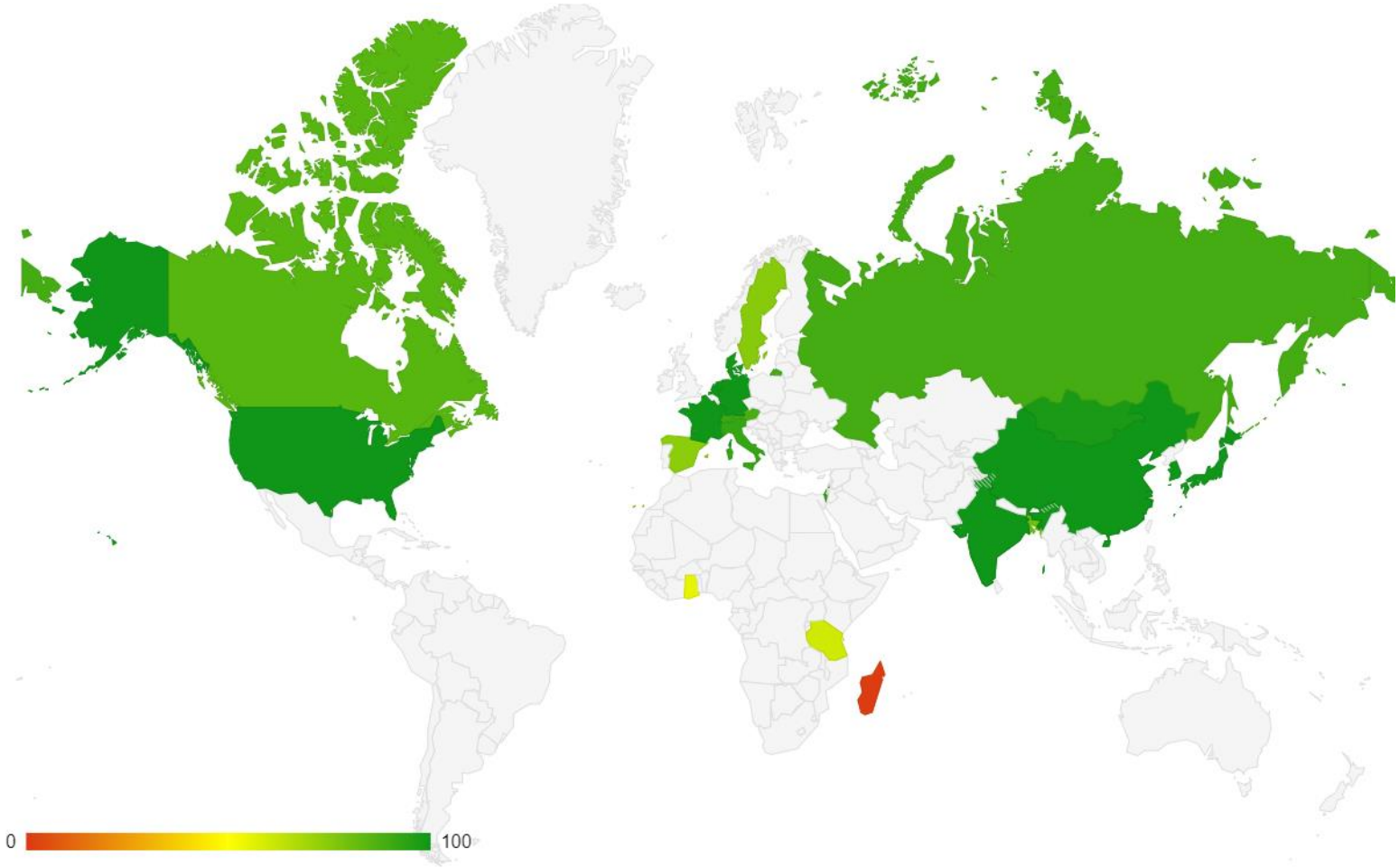


Figure 40. World map showing the percentage of presence of *Bacteroides vulgatus* on the metagenomic samples.

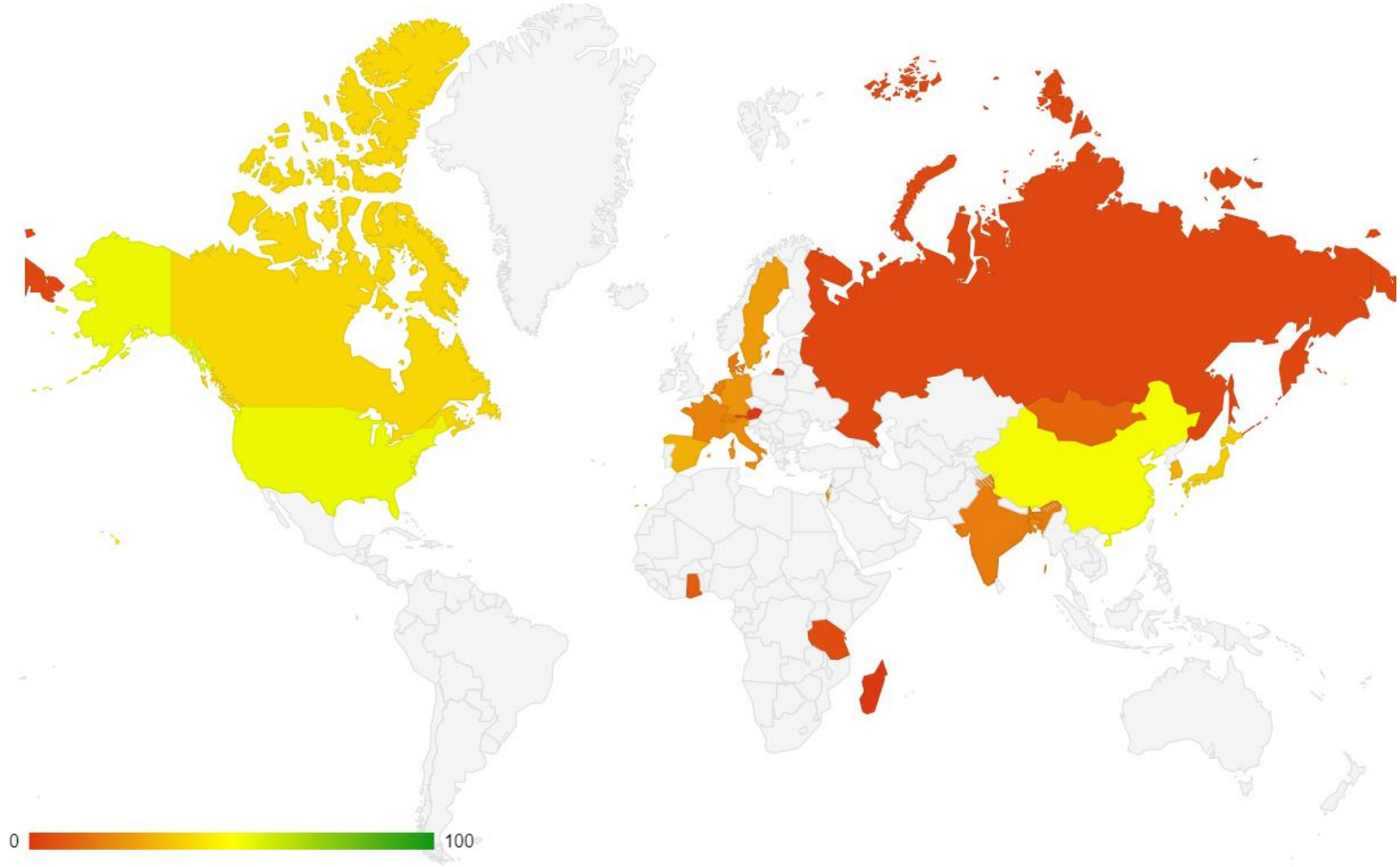


Figure 41. World map showing the percentage of presence of *Bacteroides stercoris* on the metagenomic samples.

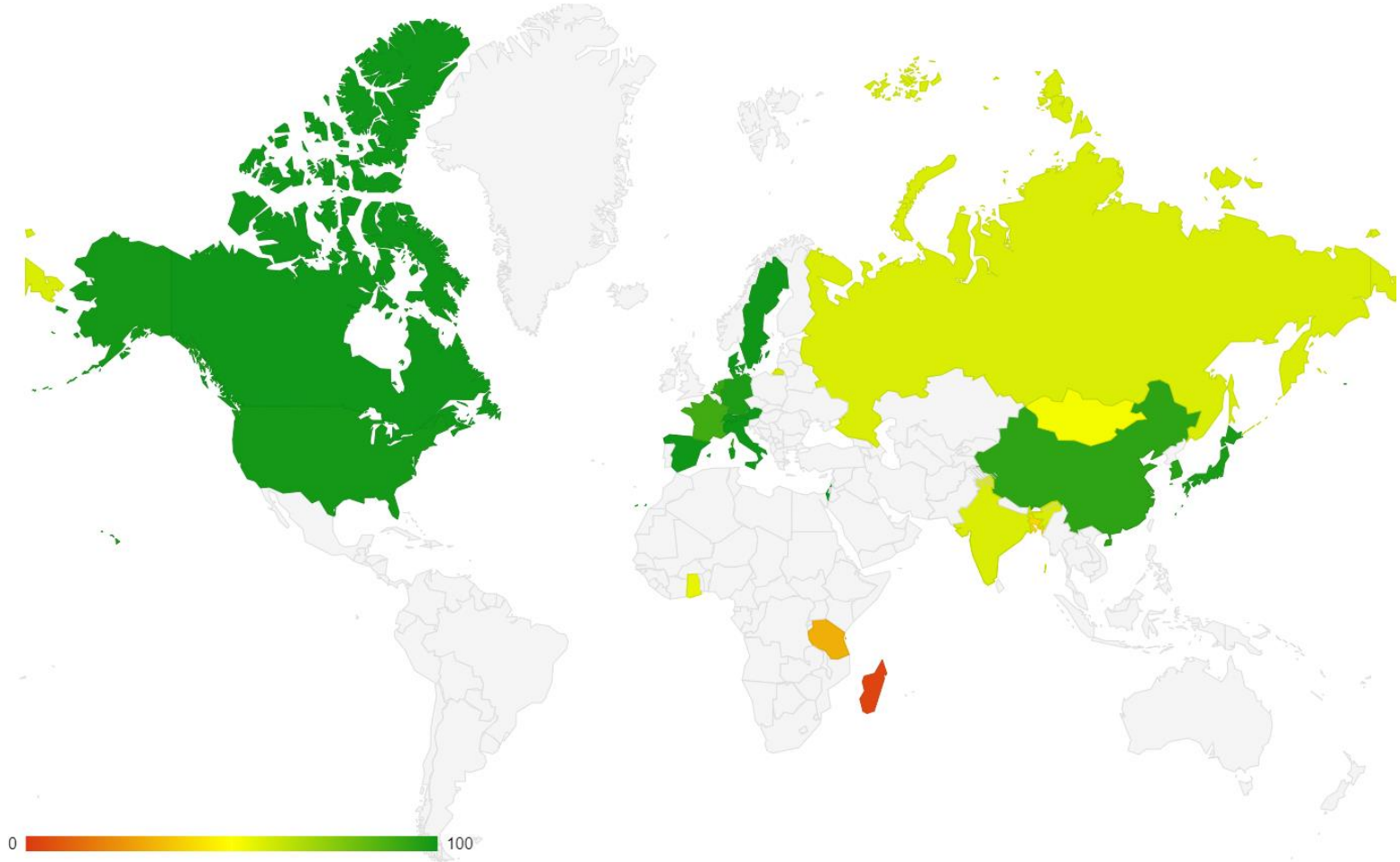


Figure 42. World map showing the percentage of presence of *Bacteroides uniformis* on the metagenomic samples.

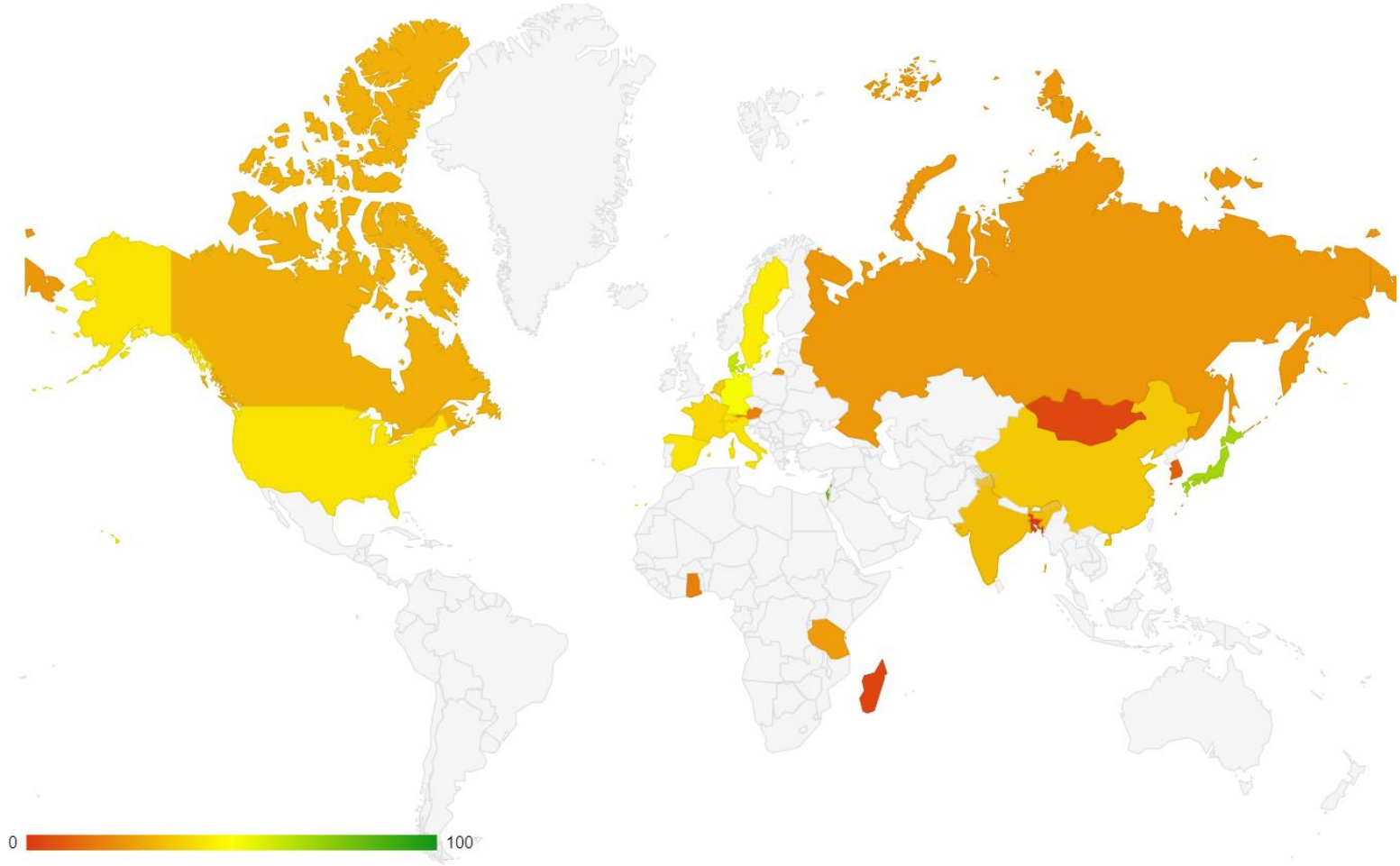


Figure 43. World map showing the percentage of presence of *Bacteroides dorei* on the metagenomic samples.

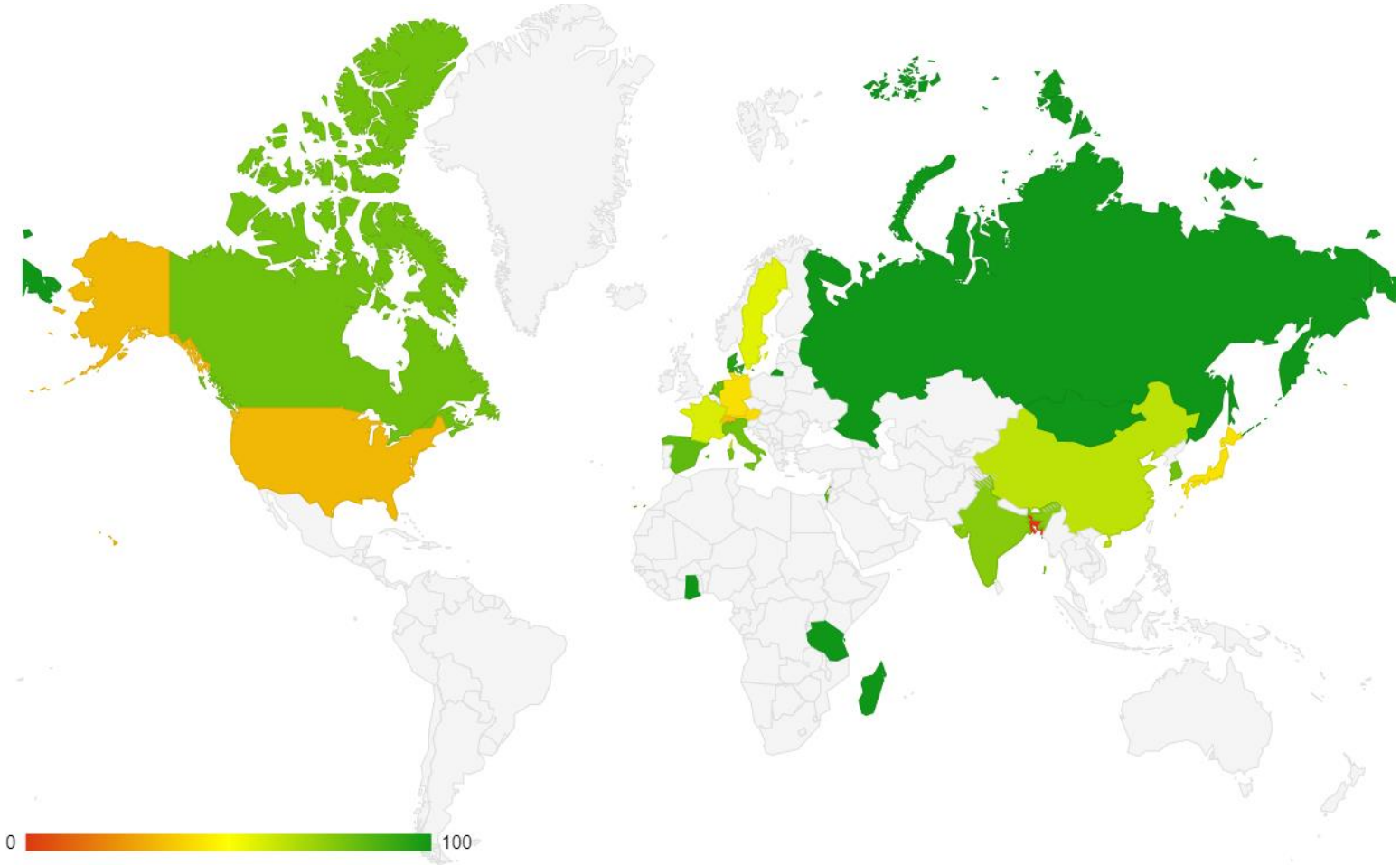


Figure 44. World map showing the percentage of presence of *Bacteroides galacturonicus* on the metagenomic samples.

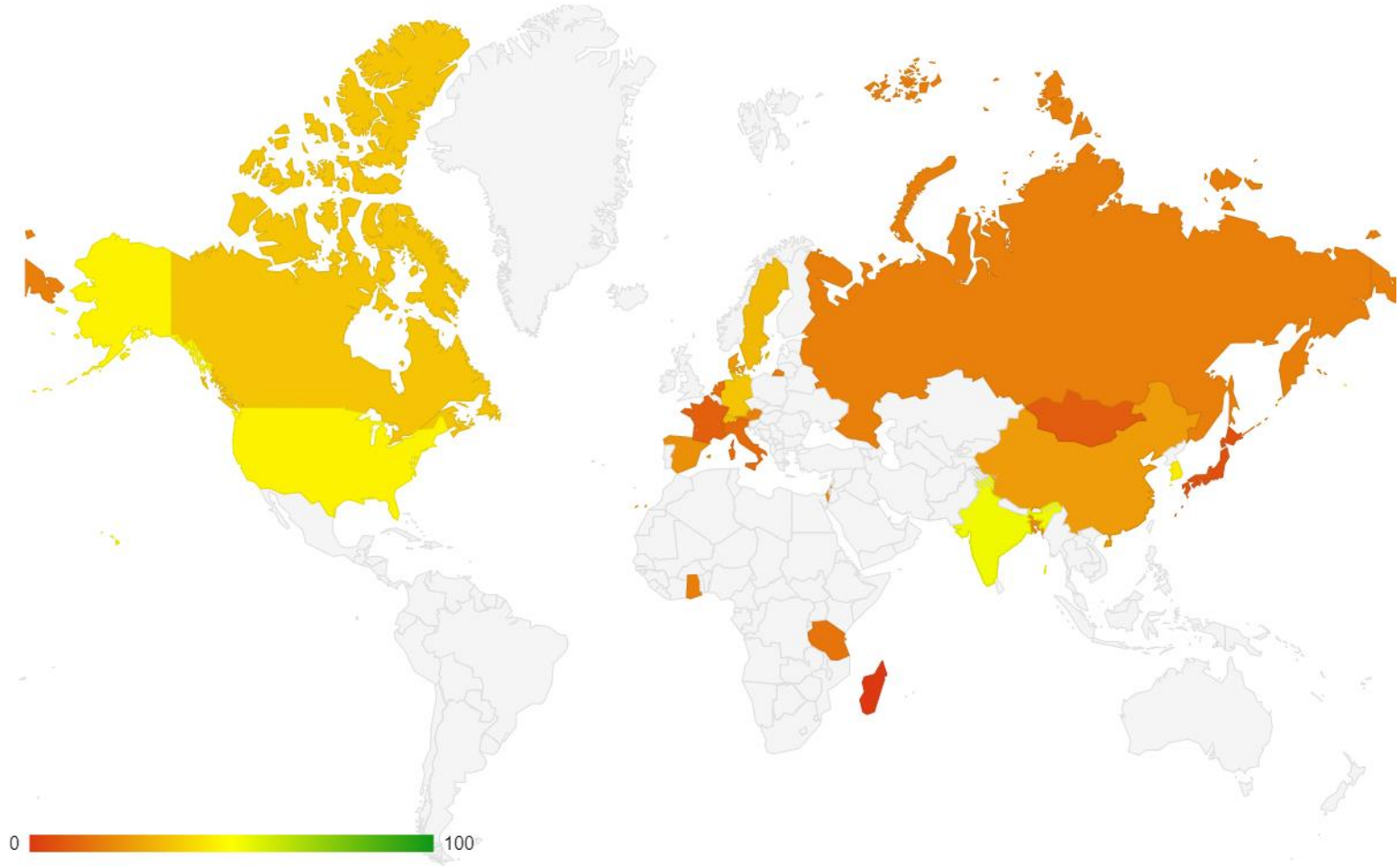


Figure 45. World map showing the percentage of presence of *Bacteroides caccae* on the metagenomic samples.

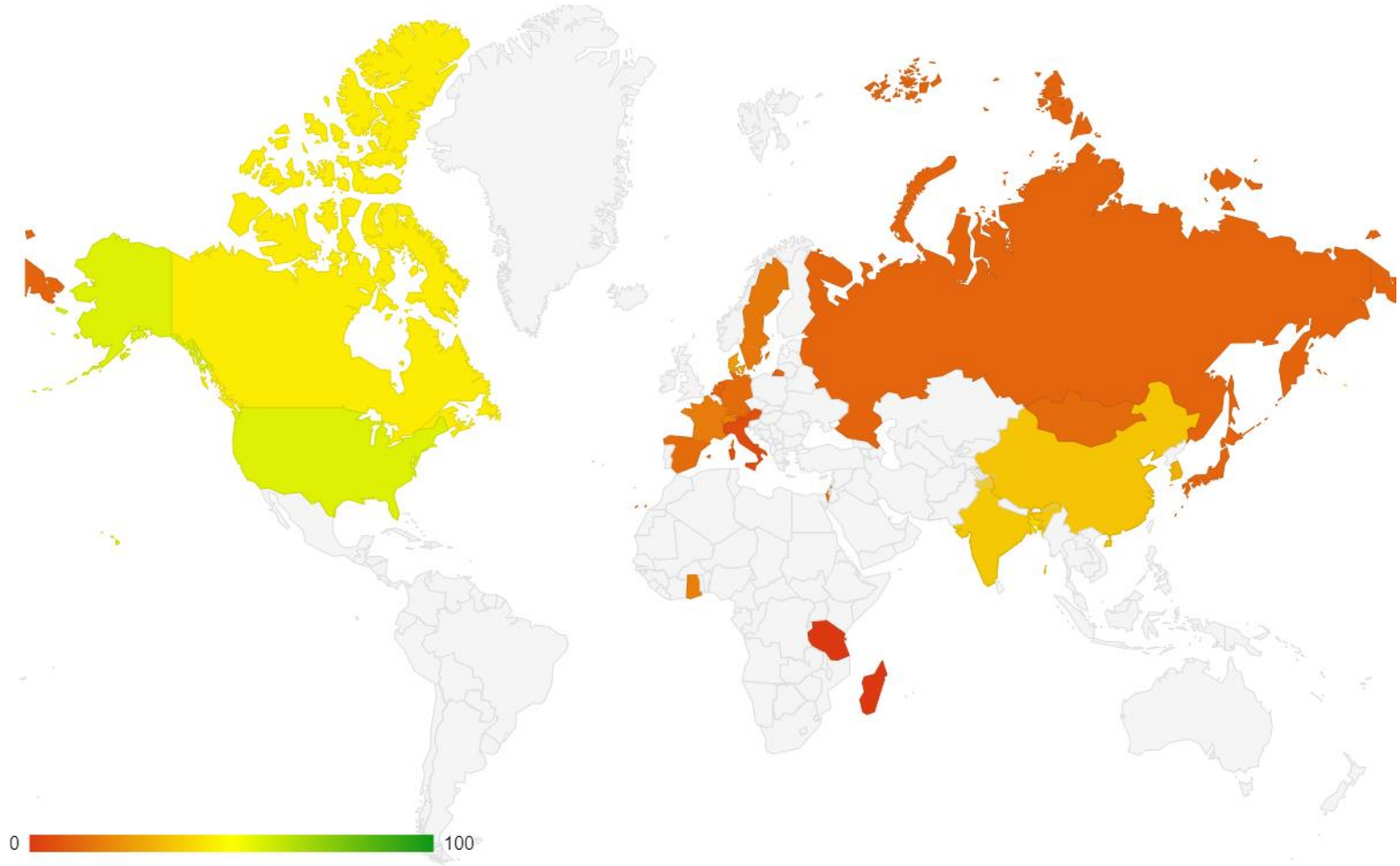


Figure 46. World map showing the percentage of presence of *Bacteroides xylanisolvens* on the metagenomic samples.

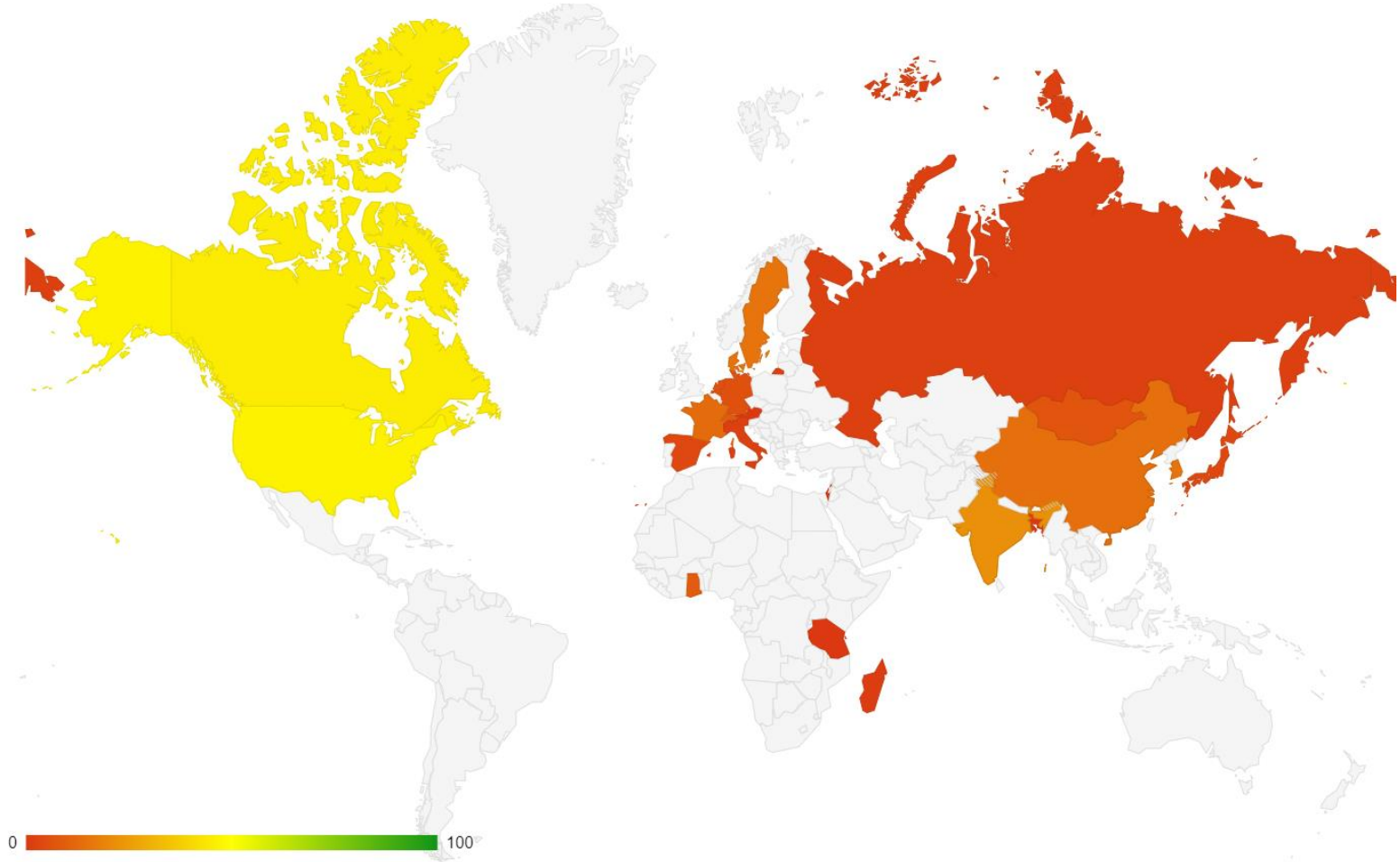


Figure 47. World map showing the percentage of presence of *Bacteroides ovatus* on the metagenomic samples.

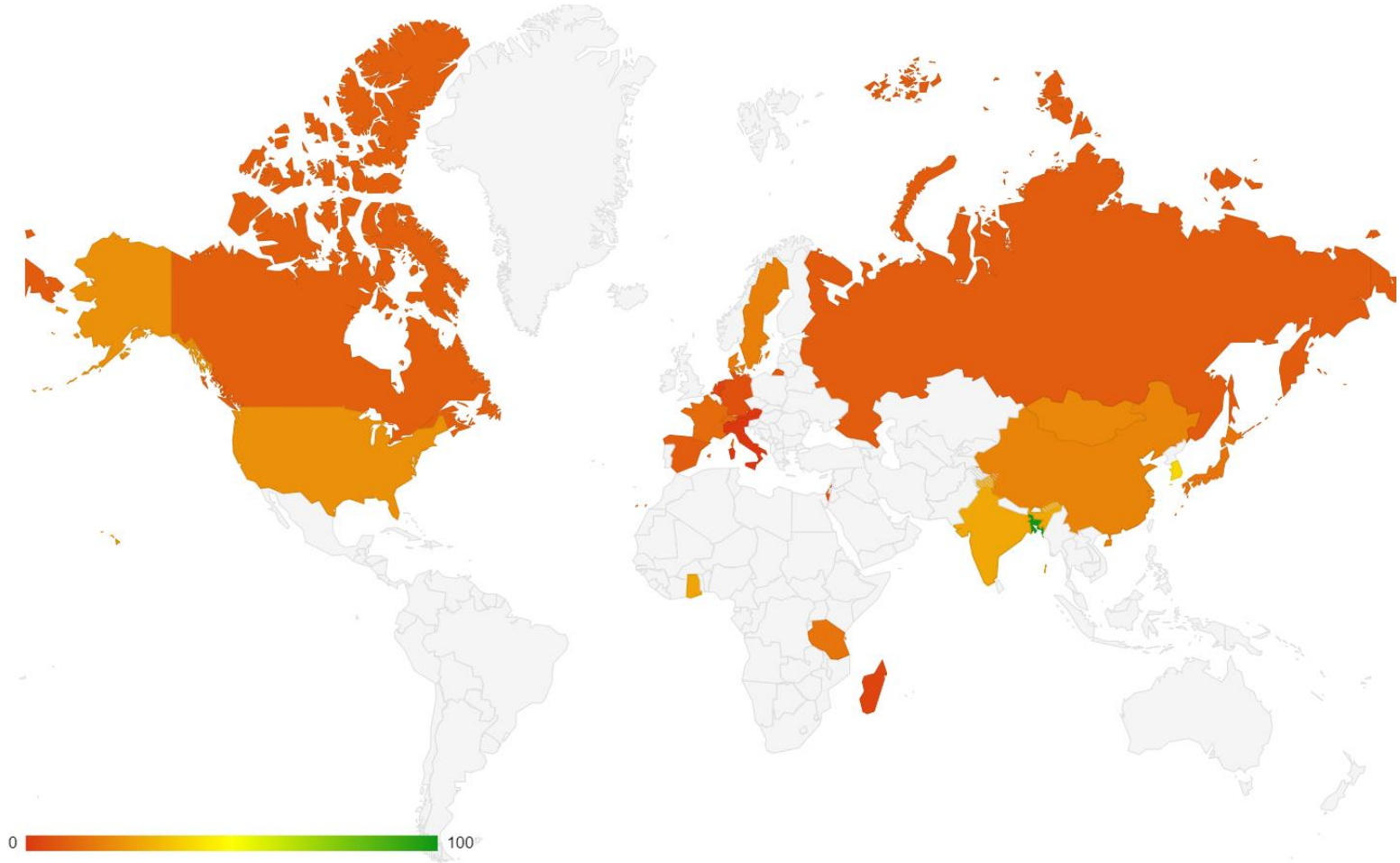


Figure 48. World map showing the percentage of presence of *Bacteroides fragilis* on the metagenomic samples.

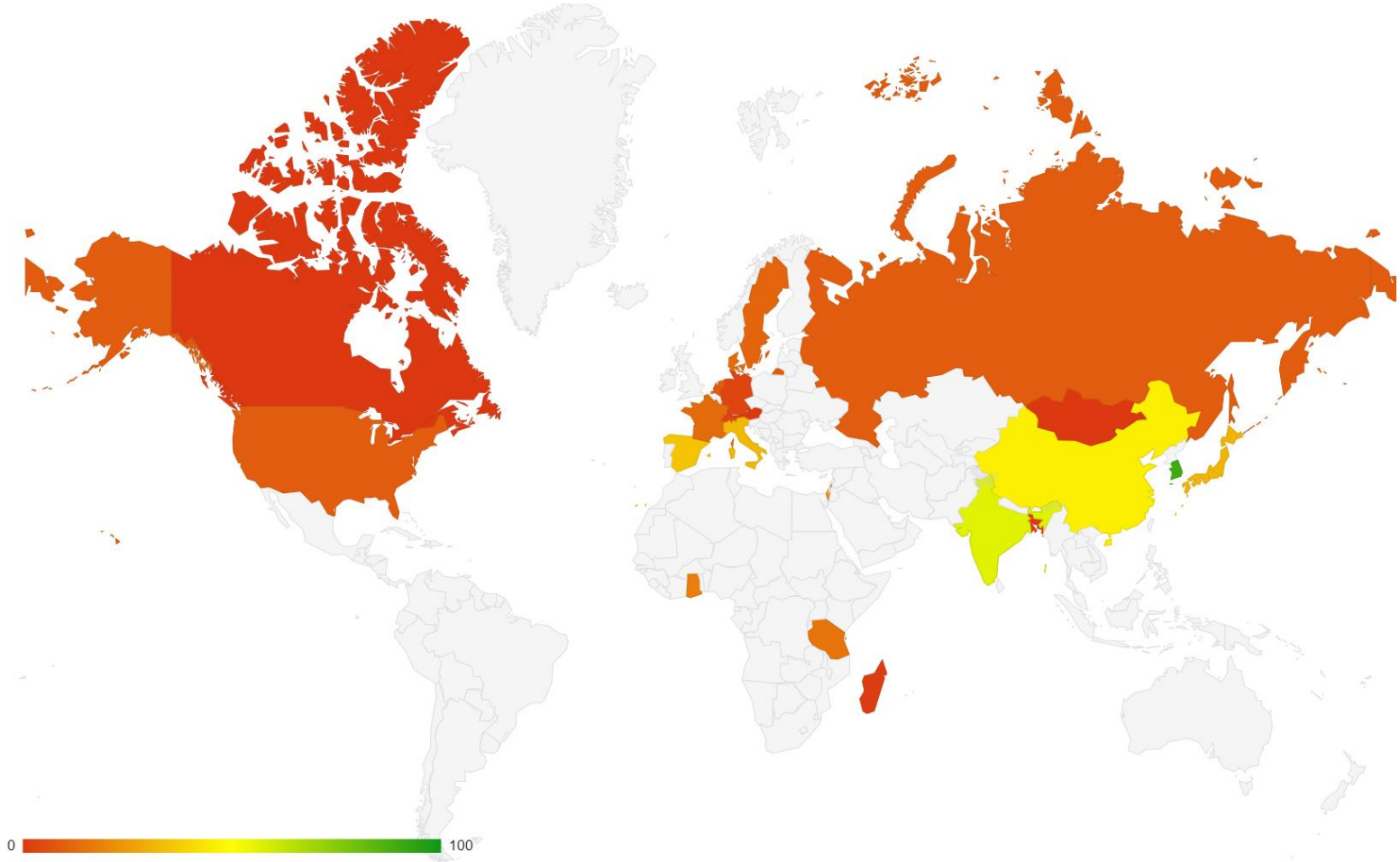
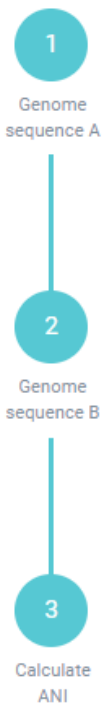


Figure 49. World map showing the percentage of presence of *Bacteroides* QSQT_s on the metagenomic samples.



Fasta QC genomo_GCF_003437535.1_ASM343753v1_genomic.fna

Contigs	Total length (bp)	A	C	G	T	N	GC content (%)
71	3,968,058	1,089,869	853,731	902,033	1,122,147	278	44.25

Fasta QC plebeius_GCF_000187895.1_ASM18789v1_genomic.fna

Contigs	Total length (bp)	A	C	G	T	N	GC content (%)
19	4,421,924	1,232,116	968,209	991,003	1,229,996	600	44.31

Metric	Value
OrthoANIu value (%)	93.96
Genome A length (bp)	3,931,080
Genome B length (bp)	4,413,540
Average aligned length (bp)	1,925,853
Genome A coverage (%)	48.99
Genome B coverage (%)	43.64

Figure 50. OrthoANIu [31] results when comparing *Bacteroides* QSQT_s (1) and *Bacteroides plebeius* (2).

4.3. Bacteroides on Animal Species

4.3.1. Methods

We also downloaded and metagenome profiles 2,095 animal samples as described on Table 8. Using the same database and process from the human samples, we profiled all samples using KrakenUBCG and its database. After analyzing all animal samples, we excluded animal species that contained no *Bacteroides* (Baboon and Gorilla). We considered only those species with at least 1% abundance on any given species for the genus *Bacteroides*. Abundance percentage was normalized using the total length of the UBCG sequences for every *Bacteroides* species. Zero values were ignored when calculating the median abundance for any given species.

4.3.2. Results

Figure 51 displays a bubble map showing the median abundance for those animal species that contained any *Bacteroides*. Human samples separated per continent are also displayed for comparative purposes. It can be seen that animal species contain less diversity of *Bacteroides*. Mouse metagenomic samples showed 19 distinct *Bacteroides* species, with *Bacteroides massilensis* and *Bacteroides vulgatus* having the most median abundance. Species present in Mouse but absent on Human samples are *Bacteroides acidifaciens* and *Bacteroides koreensis*.

Primate samples showed 12 different species, with *Bacteroides fragilis* and *Bacteroides vulgatus* being the most abundant. *Bacteroides koreensis* was also present on primates while being absent in human samples. Chicken samples displayed 10 distinct species, with *Bacteroides clarus and Bacteroides salanitronis* being the most abundant. *Bacteroides pyogenes* while being absent on human samples, was found on chicken metagenomic samples.

House cat samples had 6 different *Bacteroides* species; *Bacteroides coprocola* and *Bacteroides stercoris* were the most abundant. All species from cat samples are also found in human samples. Similarly, dog samples had 10 distinct *Bacteroides* with all of them being present in human samples. Bacteroides propionicifaciens was found in one rumen sample, this is the only organism that contained this species on this study.

If we observe the ratio of presence of *Bacteroides* on the samples (Figure 52) we can observe similarities between house cat and dog samples. The majority of the samples from these two animal species contained *Bacteroides coprocola*, *Bacteroides coprophilus* and *Bacteroides plebeius*; all three species are the most present on these two species in comparison to the human samples. Additionally, 81% of cat samples contains *Bacteroides stercoris*. *Bacteroides vulgatus* was the most present species on mouse samples. For the rest of the animal species, most of the species present were at a lower presence ratio.

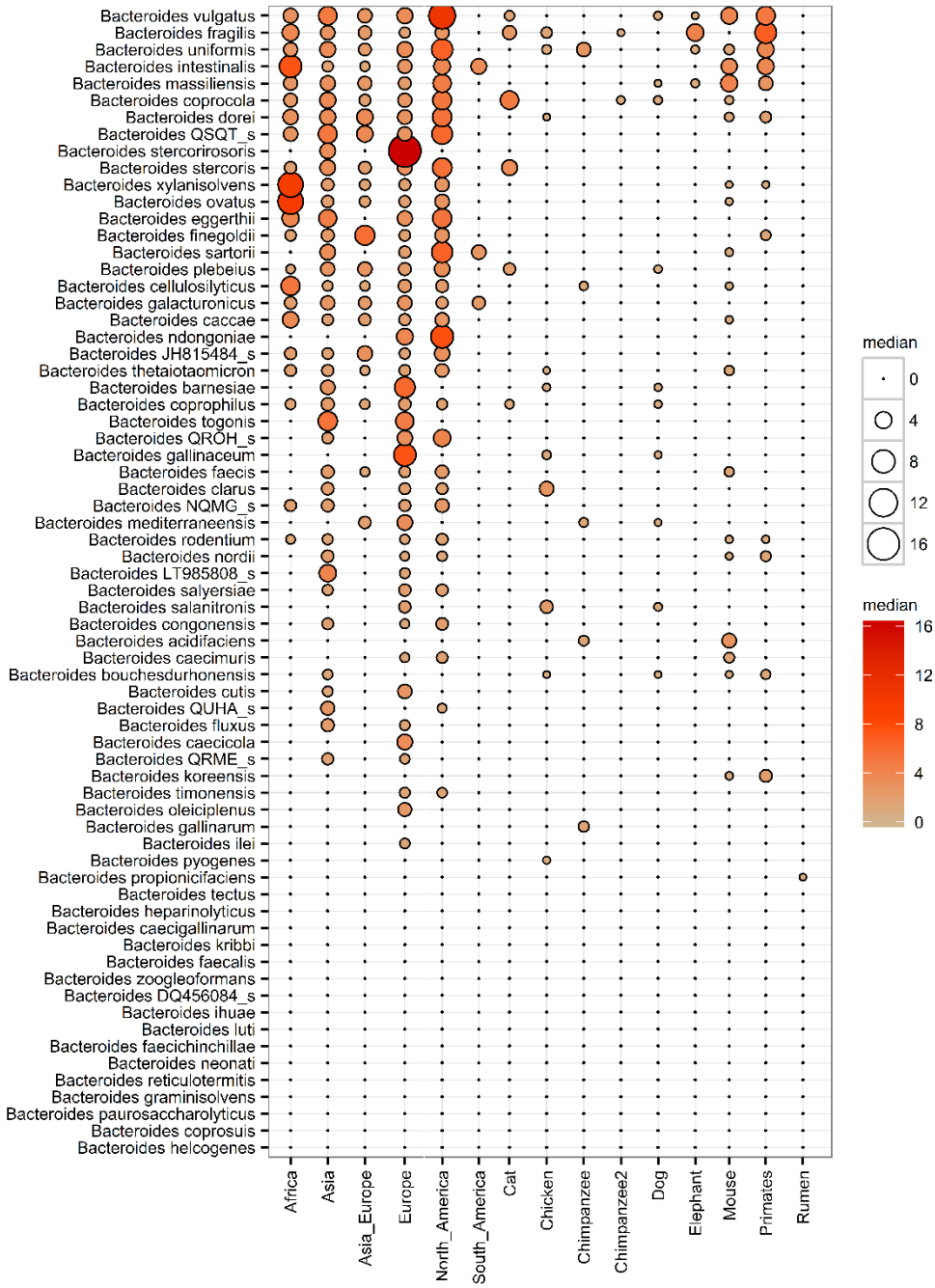


Figure 51. Abundance of *Bacteroides* per species for each continent and animal species.

Figure 52. Heatmap showing the percentage of samples for each continent and animal species that contain any given *bacteroides* species.

Table 8. Description of the 2095 animal metagenomic samples used in this study.

Organism	# Samples	Accession Number	Study
Baboon	48	PRJNA271618	Raw metagenomic sequencing reads collected from samples from 48 adult baboons in two social groups during July - August 2012.
Cat	11	PRJNA49515	Feline cat gastrointestinal metagenome.
Chicken	10	PRJNA338370	The effect of Lactobacillus (L.) plantarum P-8 on the gut microbiome of chickens.
Chicken	502	PRJNA417359	A catalogue of chicken gut metagenome and the microbial responses to antibiotic and plant-derived benzylisoquinoline alkaloids.
Chimpanzee	224	PRJNA505752	Exome sequencing of wild chimpanzee individuals using non-invasively collected fecal samples.
Chimpanzee	72	PRJEB21543	The impact of endogenous content, replicates and pooling on genome capture from fecal samples.
Dog	129	PRJEB20308	Similarity of the dog and human gut microbiomes in gene content and response to diet.
Elephant	4	PRJNA240141	Elephant feces Metagenome.
Gorilla	23	PRJNA419744	Gorilla fecal virome.
Gorilla	19	PRJNA382701	Gorilla Raw sequence reads.
Mouse	380	PRJEB7759	A Catalogue of the Mouse Gut Metagenome.
Primates	249	PRJEB22679	Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes.
Rumen	424	PRJEB23561	A catalog of microbial genes from the bovine rumen reveals the determinants of herbivory.

4.4. Discussion and conclusions

Finding any meaningful results on gut microbiomes has many challenges. In this case, by focusing on Bacteroides we narrowed down our focus on the hundreds of bacterial species present in a metagenome sample. But in order to have any significant findings, thousands of samples are needed. Sampling a community with a few dozens of samples can yield misleading results. While in this study we profiled more than 2,000 samples, we still believe this number is exceptionally low to yield significant results. Additionally, the unevenness on the number of samples per community is also troublesome. Those communities represented by a low number of samples can make the results bias. For the most part, samples from the USA are easy to find on public repositories, while samples from other locations, such as African countries can be difficult to find. Another issue that scientists face; even if the number of samples is no longer an issue, the amount of computational resources required in order to profile those samples can take years with current tools.

In this study, using an efficient pipeline, we profiled thousands of human samples looking for any indication that *Bacteroides* could be correlated to geographical location. The most striking difference between human samples was the low amount of *Bacteroides* on samples from rural areas, particularly African and South American tribes. Other than that, some countries showed certain dominance for a specific *Bacteroides* species, but this could also be explained by the type of study were the samples came from. In case of animal samples, Mice

showed the most diversity of *Bacteroides*, however this can be attributed by the number of bacterial references isolated from this animal. House cat and dog samples showed some correlation between each other, can this be attributed to the similarities of their lifestyle and diet? The rest of the animal samples showed little or no *Bacteroides*, however we cannot discard the possibility that this is due to the lack of bacterial references isolated from these animals in the database. Additionally, should animal research be simplified just by animal species? Or should they receive the same research treatment as human research, for example house dog samples from the USA and house dog samples from South Korea. Unfortunately, in the close future this is not possible until sequencing technologies get cheaper and bioinformatics get faster and easy to use.

This study will be worth revisiting in a few years, with the increase of references in the database, and the inclusion of more metagenomic samples, we are excited to see how if the results change or they maintain their conclusions.

General Conclusion

Shotgun metagenomics is a powerful tool to decipher the contents of a complex microbial community. The initial challenge of the bioinformatic process is the computational time it takes to analyze one sample. The current understanding that sequence homology search is the best way to profile metagenomic samples is flawed. The best methodology is not necessarily the one that gives the best results, but the one that gives the best approximate result within a reasonable amount of time. Exact match k-mer approaches are the best heuristics that combine speed and accuracy; however, they rely on their reference database for the best results.

Here, we propose a well-known method for shotgun metagenomic profiling based on exact match k-mers, optimized to take advantage of RAM memory and an alternative representation of nucleotides in memory. Proving that an algorithm is faster in computer science is not challenging, however, in bioinformatics, we must prove that the method delivers the best results; and to do this, we need a reliable database.

We propose the use of UBCG reference sequences as a complement to our method. Not only these sequences make the database smaller in size making the memory requirements lower, but also, they normalize the references for all bacterial species by having every species represented by the same number of sequences.

Proving that our method was better than current tools was not easy. First, using synthetic data, we showed that not only our method can detect bacterial species accurately, but also quantify them in a precise manner. Then, using clinical

samples, we showed that having an updated database, can yield different conclusions when comparing to a static and outdated database. It is imperative that any type of clinical research based on the presence and abundance of any bacteria is performed by a method that also includes the most updated and complete database available at the time.

Finally, showing the great speed of our method, we profiled more than 2,000 human samples and an additional 2,000 samples from animal species in search for *Bacteroides*. We found that some groups of human samples contain specific species of *Bacteroides*, however we cannot rule out the possibility that these patterns may be correlated to the study and not to the geographic location. We learned that in order to get a more conclusive result, when comparing a specific genus, a great number of samples representing a group are needed. While having a total of 4,000+ samples is a good start, is not enough. Also, when analyzing animal gut samples, a clear bias can be seen. Those animal species that have been studied the most have the most advantage on metagenomic analysis, since they may have some representation in the form of references in the database. At the end, how many isolates from the gut of elephants are contained on public databases? The answer is none.

2020 is the beginning of a new decade, hopefully it will be the decade of shotgun metagenomics. With sequencing technologies getting cheaper, and faster computational heuristics being proposed, new research and bio targets may be

found; all of this in order to diagnose and treat human health conditions around the world in a personalized manner and with lower mortality rates.

References

[1] M. J. Cox, W. O. C. M. Cookson, and M. F. Moffatt, “Sequencing the human microbiome in health and disease,” *Hum. Mol. Genet.*, vol. 22, no. R1, pp. R88–R94, Aug. 2013.

[2] T. J. Sharpton, “An introduction to the analysis of shotgun metagenomic data,” *Front. Plant Sci.*, vol. 5, p. 209, Jun. 2014.

[3] A. B. R. McIntyre *et al.*, “Comprehensive benchmarking and ensemble approaches for metagenomic classifiers,” *Genome Biol.*, vol. 18, no. 1, pp. 1–19, 2017.

[4] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, “Metagenomic microbial community profiling using unique clade- specific marker genes,” vol. 9, no. 8, 2012.

[5] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, “MEGAN analysis of metagenomic data,” *Genome Res.*, vol. 17, no. 3, pp. 377–386, 2007.

[6] T. A. K. Freitas, P. E. Li, M. B. Scholz, and P. S. G. Chain, “Accurate read-based metagenome characterization using a hierarchical suite of unique signatures,” *Nucleic Acids Res.*, vol. 43, no. 10, pp. 1–14, 2015.

[7] A. E. Darling, G. Jospin, E. Lowe, F. A. Matsen, H. M. Bik, and J. A. Eisen, “PhyloSift: Phylogenetic analysis of genomes and metagenomes,” *PeerJ*, vol. 2014, no. 1, pp. 1–28, 2014.

[8] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, “CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers,” *BMC Genomics*, vol. 16, no. 1, pp. 1–13, 2015.

[9] D. E. Wood and S. L. Salzberg, “Kraken : ultrafast metagenomic sequence classification using exact alignments,” *Genome Biol.*, vol. 15, p. R46, 2014.

[10] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg, “Bracken : estimating species abundance in metagenomics data,” *PeerJ Comput. Sci.*, vol. 3, pp. 1–17, 2017.

[11] D. T. Truong *et al.*, “MetaPhlAn2 for enhanced metagenomic taxonomic profiling,” *Nat. Methods*, vol. 12, no. 10, pp. 902–903, 2015.

[12] S. Na, Y. O. Kim, S. Yoon, S. Ha, I. Baek, and J. Chun, “UBCG : Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction §,” *J. Microbiol.*, vol. 56, no. 4, pp. 280–285, 2018.

[13] X. Gao, X. Zhi, H. Li, H. Klenk, and W. Li, “Comparative Genomics of the Bacterial Genus *Streptococcus* Illuminates Evolutionary Implications of Species Groups,” *PLoS One*, vol. 9, no. 6, p. e101229, 2014.

[14] D. M. Linares, T. F. O. Callaghan, P. M. O. Connor, R. P. Ross, and C. Stanton, “*Streptococcus thermophilus* APC151 Strain Is Suitable for the Manufacture of Naturally GABA-Enriched Bioactive Yogurt,” *Front. Microbiol.*, vol. 7, no. November, pp. 1–9, 2016.

[15] W. Krzy, K. K. Pluskwa, A. Jurczak, and D. Ko, “The pathogenicity of the

Streptococcus genus," *Eur J Clin Microbiol Infec Dis*, vol. 32, pp. 1361–1376, 2013.

[16] S. A. Shelburne *et al.*, "*Streptococcus mitis* Strains Causing Severe Clinical Disease in Cancer Patients," *Emerg. Infect. Dis.*, vol. 20, no. 5, pp. 762–771, 2014.

[17] N. Ehara *et al.*, "A novel method for rapid detection of *Streptococcus pneumoniae* antigen in sputum and its application in adult respiratory tract infections," *J. Med. Microbiol*, vol. 57, pp. 820–826, 2008.

[18] S. Yoon *et al.*, "Introducing EzBioCloud : a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies," *Int. J. Syst. Evol. Microbiol.*, vol. 67, pp. 1613–1617, 2019.

[19] P. P. Patil, S. Kumar, S. Midha, V. Gautam, and P. B. Patil, "Taxonogenomics reveal multiple novel genomospecies associated with clinical isolates of *Stenotrophomonas maltophilia*," *Microb. Genomics*, vol. 4, no. 8, p. e000207, 2018.

[20] S. Fischer *et al.*, "Leptospira Genomospecies and Sequence Type Prevalence in Small Mammal Populations in Germany," *Vector-Borne Zoonotic Dis.*, vol. 18, no. 4, pp. 188–199, 2018.

[21] S. J. Salipante *et al.*, "Characterization of a Multidrug-Resistant, Novel *Bacteroides* Genomospecies," *Emerg. Infect. Dis.*, vol. 21, no. 1, pp. 95–98, 2015.

[22] O. Karlsson-lindsjo, J. Hayer, and E. Bongcam-rudloff, “Sequence analysis Simulating Illumina metagenomic data with InSilicoSeq,” *Bioinformatics*, vol. 35, no. July 2018, pp. 521–522, 2019.

[23] S. J. S. Cameron *et al.*, “Metagenomic Sequencing of the Chronic Obstructive Pulmonary Disease Upper Bronchial Tract Microbiome Reveals Functional Changes Associated with Disease Severity,” *PLoS One*, vol. 11, no. 2, pp. 1–16, 2016.

[24] N. Segata *et al.*, “Metagenomic biomarker discovery and explanation,” *Genome Biol.*, vol. 12, no. 6, p. R60, 2011.

[25] F. B. Tamburini, T. M. Andermann, E. Tkachenko, F. Senchyna, N. Banaei, and A. S. Bhatt, “Precision identification of diverse bloodstream pathogens in the gut microbiome,” *Nat. Med.*, vol. 24, pp. 1809–1814, 2018.

[26] S. Ha *et al.*, “Application of the Whole Genome-Based Bacterial Identification System , TrueBac ID , Using Clinical Isolates That Were Not Identified With Three Matrix- Assisted Laser Desorption / Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) Systems,” *An. Lab. Med.*, vol. 39, pp. 530–536, 2019.

[27] J. Chun *et al.*, “Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes,” *Int. J. Syst. Evol. Microbiol.*, vol. Jan, no. 68, pp. 461–466, 2018.

[28] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, “GTDB-Tk:

a toolkit to classify genomes with the Genome Taxonomy Database," *Bioinformatics*, vol. Nov, pp. 1–3, 2019.

[29] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries," *Nat. Commun.*, vol. 9, no. 1, pp. 1–8, 2018.

[30] A. Bankevich *et al.*, "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, 2012.

[31] S. H. Yoon, S. min Ha, J. Lim, S. Kwon, and J. Chun, "A large-scale evaluation of algorithms to calculate average nucleotide identity," *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.*, vol. 110, no. 10, pp. 1281–1286, 2017.

[32] K. Schliep *et al.*, *Package "phangorn": phylogenetic reconstruction and analysis*. 2019.

[33] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, no. 6, pp. 764–770, 2011.

[34] C. M. Lakhani, "*Bacteroides* as a window into the microbiome," *Physiol. Behav.*, vol. 176, no. 3, pp. 139–148, 2019.

[35] J. Yin *et al.*, "Dysbiosis of gut microbiota with reduced trimethylamine-n-oxide level in patients with large-artery atherosclerotic stroke or transient

ischemic attack," *J. Am. Heart Assoc.*, vol. 4, no. 11, pp. 1–12, 2015.

[36] A. A. Salyers, P. Valentine, and V. Hwa, "Genetics of Polysaccharide Utilization Pathways of Colonic *Bacteroides* Species BT - Genetics and Molecular Biology of Anaerobic Bacteria," M. Sebald, Ed. New York, NY: Springer New York, 1993, pp. 505–516.

[37] G. Reid, "When Microbe Meets Human," *Clin. Infect. Dis.*, vol. 39, no. 6, pp. 827–830, Sep. 2004.

[38] J. Xu and J. I. Gordon, "Honor thy symbionts," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 18, pp. 10452–10459, Sep. 2003.

[39] Y. Zhou and F. Zhi, "Lower Level of *Bacteroides* in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis," *Biomed Res. Int.*, vol. 2016, p. 5828959, 2016.

[40] V. Corby-Harris, A. C. Pontaroli, L. J. Shimkets, J. L. Bennetzen, K. E. Habel, and D. E. L. Promislow, "Geographical distribution and diversity of bacteria associated with natural populations of *Drosophila melanogaster*," *Appl. Environ. Microbiol.*, vol. 73, no. 11, pp. 3470–3479, 2007.

[41] B. Senghor, C. Sokhna, R. Ruimy, and J. C. Lagier, "Gut microbiota diversity according to dietary habits and geographical provenance," *Hum. Microbiome J.*, vol. 7–8, no. February, pp. 1–9, 2018.

[42] M. S. Age *et al.*, "Extensive Unexplored Human Microbiome Diversity Resource Extensive Unexplored Human Microbiome Diversity Revealed by

Over 150 , 000 Genomes from Metagenomes Spanning Age , Geography , and Lifestyle," *Cell*, pp. 1–14, 2019.

[43] I. Lee, Y. O. Kim, S. Park, and J. Chun, "OrthoANI : An improved algorithm and software for calculating average nucleotide identity," pp. 1100–1103, 2016.

[44] A. Stamatakis, "RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, 2014.

Appendix I. A list of genomes from the genus *Streptococcus* used on Chapter's 3 analysis.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_000188295.1	Streptococcus vestibularis ATCC 49124	ATCC 49124	Streptococcus vestibularis
GCA_001375655.1	Streptococcus varani	FF10	Streptococcus varani
GCA_000188055.3	Streptococcus urinalis 2285-97	2285-97	Streptococcus urinalis
GCA_900475595.1	Streptococcus uberis	NCTC 3858	Streptococcus uberis
GCA_002355215.1	Streptococcus troglodytae	TKU 31	Streptococcus troglodytae
GCA_900095845.1	Streptococcus timonensis	Marseille-P2915	Streptococcus timonensis
GCA_000380145.1	Streptococcus thoraltensis DSM 12221	DSM 12221	Streptococcus thoraltensis
GCA_900474985.1	Streptococcus thermophilus	NCTC 12958	Streptococcus thermophilus
GCA_900475585.1	Streptococcus suis	NCTC 10234	Streptococcus suis
GCA_900475395.1	Streptococcus sobrinus	NCTC 12279	Streptococcus sobrinus
GCA_000767835.1	Streptococcus sinensis	HKU4	Streptococcus sinensis
GCA_900475505.1	Streptococcus sanguinis	NCTC 7863	Streptococcus sanguinis
GCA_000253335.1	Streptococcus salivarius CCHSS3	JIM8780	Streptococcus salivarius subsp. salivarius
GCA_003609975.1	Streptococcus ruminantium	GUT-187	Streptococcus ruminantium
GCA_003595525.1	Streptococcus respiraculi	HTS25	Streptococcus respiraculi
GCA_000286075.1	Streptococcus ratti FA-1 = DSM 20564	FA-1	Streptococcus ratti
GCA_002055535.1	Streptococcus pyogenes	NCTC 8198	Streptococcus pyogenes
GCA_000188035.3	Streptococcus pseudoporcinus LQ 940-04	LQ 940-04	Streptococcus pseudoporcinus
GCA_002087075.1	Streptococcus pseudopneumoniae ATCC BAA-960	CCUG 49455	Streptococcus pseudopneumoniae
GCA_900475415.1	Streptococcus porcinus	NCTC 10999	Streptococcus porcinus
GCA_000423765.1	Streptococcus porci DSM 23759	DSM 23759	Streptococcus porci
GCA_001457635.1	Streptococcus pneumoniae	NCTC 7465	Streptococcus pneumoniae

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_000423745.1	Streptococcus plurextorum DSM 22810	DSM 22810	Streptococcus plurextorum
GCA_002953735.1	Streptococcus pluranimalium	TH11417	Streptococcus pluranimalium
GCA_000772915.1	Streptococcus phocae C-4	C-4	Streptococcus phocae subsp. salmonis
GCA_001302265.1	Streptococcus phocae	ATCC 51973	Streptococcus phocae subsp. phocae
GCA_000187585.1	Streptococcus peroris ATCC 700780	ATCC 700780	Streptococcus peroris
GCA_002887775.1	Streptococcus sp. CAIM 1838	CAIM 1838	Streptococcus penaeicida
GCA_000187935.2	Streptococcus parauberis NCFD 2020	NCFD 2020	Streptococcus parauberis
GCA_000440555.1	Streptococcus suis 86-5192	86-5192	Streptococcus parasuis
GCA_000164675.2	Streptococcus parasanguinis ATCC 15912	ATCC 15912	Streptococcus parasanguinis
GCA_001642085.1	Streptococcus pantholopis	TA 26	Streptococcus pantholopis
GCA_000380125.1	Streptococcus ovis DSM 16829	DSM 16829	Streptococcus ovis
GCA_000380105.1	Streptococcus orisratti DSM 15617	DSM 15617	Streptococcus orisratti
GCA_002093515.1	Streptococcus oralis subsp. tigurinus	AZ_14	Streptococcus oralis subsp. tigurinus
GCA_000164095.1	Streptococcus oralis ATCC 35037	ATCC 35037	Streptococcus oralis subsp. oralis
GCA_000382825.1	Streptococcus dentisani 7747	CECT 7747	Streptococcus oralis subsp. dentisani
GCA_900475095.1	Streptococcus mutans	NCTC 10449	Streptococcus mutans
GCA_000148585.1	Streptococcus mitis NCTC 12261	NCTC 12261	Streptococcus mitis
GCA_000377005.1	Streptococcus minor DSM 17118	DSM 17118	Streptococcus minor
GCA_900187085.1	Streptococcus merionis	NCTC 13788	Streptococcus merionis
GCA_900459365.1	Streptococcus massiliensis	NCTC 13765	Streptococcus massiliensis
GCA_001623565.1	Streptococcus sp. HTS5	HTS5	Streptococcus marmotae
GCA_000380045.1	Streptococcus marimammalium DSM 18627	DSM 18627	Streptococcus marimammalium
GCA_000187995.3	Streptococcus macacae NCTC 11558	NCTC 11558	Streptococcus macacae

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_900475675.1	Streptococcus lutetiensis	NCTC 13774	Streptococcus lutetiensis
GCA_900475975.1	Streptococcus intermedius	NCTC 11324	Streptococcus intermedius
GCA_001595425.1	Streptococcus iniae	CAIM 527	Streptococcus iniae
GCA_000187465.1	Streptococcus infantis ATCC 700779	ATCC 700779	Streptococcus infantis
GCA_900459445.1	Streptococcus infantarius	NCTC 13760	Streptococcus infantarius
GCA_000188015.3	Streptococcus ictaluri 707-05	707-05	Streptococcus ictaluri
GCA_000420785.1	Streptococcus hyovaginalis DSM 12219	DSM 12219	Streptococcus hyovaginalis
GCA_900459405.1	Streptococcus hyointestinalis	NCTC 12224	Streptococcus hyointestinalis
GCA_000785785.1	Streptococcus uberis	CAIM 1894	Streptococcus hongkongensis
GCA_001708305.1	Streptococcus himalayensis	HTS2	Streptococcus himalayensis
GCA_000376985.1	Streptococcus henryi DSM 19005	DSM 19005	Streptococcus henryi
GCA_001598035.1	Streptococcus sp. HTS9	HTS9	Streptococcus halotolerans
GCA_900475015.1	Streptococcus gordonii	NCTC 7865	Streptococcus gordonii
GCA_900478025.1	Streptococcus pasteurianus	NCTC 13784	Streptococcus gallolyticus subsp. pasteurianus
GCA_900459545.1	Streptococcus gallolyticus	NCTC 13767	Streptococcus gallolyticus subsp. macedonicus
GCA_900475715.1	Streptococcus gallolyticus	NCTC 13773	Streptococcus gallolyticus subsp. gallolyticus
GCA_900475025.1	Streptococcus ferus	NCTC 12278	Streptococcus ferus
GCA_900459295.1	Streptococcus equinus	NCTC 12969	Streptococcus equinus
GCA_900459475.1	Streptococcus equi subsp. zooepidemicus	NCTC 4676	Streptococcus equi subsp. zooepidemicus
GCA_000706805.1	Streptococcus equi subsp. ruminatorum CECT 5772	CECT 5772	Streptococcus equi subsp. ruminatorum
GCA_900156215.1	Streptococcus equi	ATCC 33398	Streptococcus equi subsp. equi
GCA_000380025.1	Streptococcus entericus DSM 14446	DSM 14446	Streptococcus entericus
GCA_900459095.1	Streptococcus dysgalactiae	NCTC 13762	Streptococcus dysgalactiae subsp. equisimilis

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_900459225.1	Streptococcus dysgalactiae subsp. dysgalactiae	NCTC 13731	Streptococcus dysgalactiae subsp. dysgalactiae
GCA_900459175.1	Streptococcus downei MFe28	NCTC 11391	Streptococcus downei
GCA_000380005.1	Streptococcus didelphis DSM 15616	DSM 15616	Streptococcus didelphis
GCA_000423725.1	Streptococcus devriesei DSM 19639	DSM 19639	Streptococcus devriesei
GCA_001921845.1	Streptococcus cuniculi	NED12-00049-6B	Streptococcus cuniculi
GCA_900475445.1	Streptococcus cristatus ATCC 51100	NCTC 12479	Streptococcus cristatus
GCA_000187975.3	Streptococcus criceti HS-6	HS-6	Streptococcus criceti
GCA_000463425.1	Streptococcus constellatus subsp. pharyngis C1050	C1050	Streptococcus constellatus subsp. pharyngis
GCA_900459125.1	Streptococcus constellatus	NCTC 11325	Streptococcus constellatus subsp. constellatus
GCA_003086355.2	Streptococcus sp. Z15	Z15	Streptococcus chenjunshii
GCA_001937065.1	Streptococcus sp. 'caviae'	Cavy grass 6	Streptococcus caviae
GCA_000425025.1	Streptococcus castoreus DSM 17536	DSM 17536	Streptococcus castoreus
GCA_000268305.2	Streptococcus canis FSL Z3-227	FSL Z3-227	Streptococcus canis
GCA_000379985.1	Streptococcus caballi DSM 19004	DSM 19004	Streptococcus caballi
GCA_001984715.1	Streptococcus azizii	Dec-02	Streptococcus azizii
GCA_900476055.1	Streptococcus australis	NCTC 13166	Streptococcus australis
GCA_000257765.1	Streptococcus anginosus subsp. whileyi CCUG 39159	CCUG 39159	Streptococcus anginosus subsp. whileyi
GCA_000463465.1	Streptococcus anginosus C1051	C1051	Streptococcus anginosus subsp. anginosus
GCA_900458965.1	Streptococcus agalactiae	NCTC 8181	Streptococcus agalactiae
GCA_900459045.1	Streptococcus acidominimus	NCTC 12957	Streptococcus acidominimus
GCA_003934335.1	Streptococcus suis	PP422	RSDO_s
GCA_002961305.1	Streptococcus suis	1225	POLL_s
GCA_002960445.1	Streptococcus suis	2219	POJD_s

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_002096935.1	Streptococcus mitis	B_5756_13	NCVM_s
GCA_002096685.1	Streptococcus oralis subsp. dentisani	RH_13585_10	NCVA_s
GCA_002096655.1	Streptococcus oralis subsp. dentisani	RH_70047_11	NCUY_s
GCA_002096335.1	Streptococcus oralis subsp. dentisani	Y_5914_11	NCUW_s
GCA_002096595.1	Streptococcus oralis subsp. oralis	OD_311844-09	NCUR_s
GCA_002096535.1	Streptococcus oralis subsp. oralis	RH_1735_08	NCUN_s
GCA_002096445.1	Streptococcus oralis subsp. oralis	RH_57980_07	NCUK_s
GCA_002096435.1	Streptococcus oralis subsp. oralis	Y_11577_11	NCUI_s
GCA_002096365.1	Streptococcus oralis subsp. tigurinus	B_003802_10	NCUE_s
GCA_002096215.1	Streptococcus oralis subsp. tigurinus	OD_348934_12	NCUC_s
GCA_002014795.1	Streptococcus mitis	CCUG 63687	MUYO_s
GCA_001650315.1	Streptococcus sp. CCUG 49591	CCUG 49591	LVJM_s
GCA_900143575.1	Streptococcus suis	LS9N	LT671674_s
GCA_002093545.1	Streptococcus oralis subsp. tigurinus	AZ_8	LNVF_s
GCA_001182825.2	Streptococcus sp. X13SY08	X13SY08	LFYO_s
GCA_002005545.1	Streptococcus mitis	321A	LBMT_s
GCA_001814775.1	Streptococcus sp. HMSC067H01	HMSC067H01	KV817770_s
GCA_001810825.1	Streptococcus sp. HMSC076C08	HMSC076C08	KV802702_s
GCA_001579625.1	Streptococcus oralis	DD17	KQ970808_s
GCA_001579175.1	Streptococcus oralis	DD24	KQ970764_s
GCA_001579035.1	Streptococcus mitis	DD26	KQ970296_s
GCA_001579045.1	Streptococcus mitis	DD28	KQ970267_s
GCA_001579025.1	Streptococcus oralis	DD27	KQ970240_s

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_001578965.1	Streptococcus oralis	DD20	KQ969826_s
GCA_001578945.1	Streptococcus oralis	DD16	KQ969560_s
GCA_001578935.1	Streptococcus oralis	DD15	KQ969525_s
GCA_001578875.1	Streptococcus sp. DD13	DD13	KQ969510_s
GCA_001578885.1	Streptococcus sp. DD12	DD12	KQ969499_s
GCA_001578855.1	Streptococcus oralis	DD14	KQ969343_s
GCA_001578805.1	Streptococcus sp. DD10	DD10	KQ969171_s
GCA_001578795.1	Streptococcus gordonii	DD07	KQ969111_s
GCA_001578775.1	Streptococcus cristatus	DD08	KQ969067_s
GCA_001578705.1	Streptococcus oralis	DD05	KQ969042_s
GCA_000411475.1	Streptococcus sp. HPH0090	HPH0090	KE150464_s
GCA_000314775.2	Streptococcus sp. F0441	F0441	KB373321_s
GCA_000314795.2	Streptococcus sp. F0442	F0442	KB373315_s
GCA_000960105.1	Streptococcus mitis	UC5873	JYGU_s
GCA_000960085.1	Streptococcus mitis	UC921A	JYGT_s
GCA_000960025.1	Streptococcus mitis	SK137	JYGQ_s
GCA_000960005.1	Streptococcus mitis	OT25	JYGP_s
GCA_000959975.1	Streptococcus mitis	OP51	JYGO_s
GCA_000959945.1	Streptococcus mitis	COL85/1862	JYGM_s
GCA_000959885.1	Streptococcus cristatus	ATCC 49999	JYGJ_s
GCA_001069915.1	Streptococcus cristatus	1015_SOLI	JWGF_s
GCA_001069165.1	Streptococcus anginosus	1043_SSUI	JWEZ_s
GCA_001068775.1	Streptococcus pseudopneumoniae	1172_SPSE	JWAJ_s

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_001068835.1	Streptococcus pseudopneumoniae	1213_SPSE	JVYO_s
GCA_001068945.1	Streptococcus pseudopneumoniae	1271.rep1_SPSE	JVWC_s
GCA_001069865.1	Streptococcus pseudopneumoniae	144_SPSE	JVSM_s
GCA_001070715.1	Streptococcus pseudopneumoniae	163_SPSE	JVRR_s
GCA_001076615.1	Streptococcus pseudopneumoniae	294_SPSE	JVMO_s
GCA_001070815.1	Streptococcus pseudopneumoniae	315_SPSE	JVLX_s
GCA_001072315.1	Streptococcus pseudopneumoniae	330_SPSE	JVLI_s
GCA_001072375.1	Streptococcus pseudopneumoniae	342_SPSE	JVKV_s
GCA_001072925.1	Streptococcus pseudopneumoniae	434_SPSE	JVHE_s
GCA_001076775.1	Streptococcus pseudopneumoniae	445_SPSE	JVGV_s
GCA_001073085.1	Streptococcus pseudopneumoniae	469_SPSE	JVFV_s
GCA_001074155.1	Streptococcus pseudopneumoniae	74_SPSE	JUUZ_s
GCA_001074975.1	Streptococcus pseudopneumoniae	75_SPSE	JUUO_s
GCA_001074565.1	Streptococcus pseudopneumoniae	843_SPSE	JUQX_s
GCA_001074635.1	Streptococcus mitis	850_SMIT	JUQO_s
GCA_001074805.1	Streptococcus parasanguinis	886_SPAR	JUPI_s
GCA_001074825.1	Streptococcus pseudopneumoniae	888_SPSE	JUPG_s
GCA_001075875.1	Streptococcus oralis	900_SORA	JUOS_s
GCA_001075675.1	Streptococcus oralis	918_SORA	JUNW_s
GCA_000722755.1	Streptococcus mitis	SK578	JPFY_s
GCA_000722685.1	Streptococcus mitis	SK667	JPFV_s
GCA_000722695.1	Streptococcus mitis	SK629	JPFU_s
GCA_000722815.1	Streptococcus mitis	SK1126	JPFT_s

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_000235485.1	Streptococcus sp. oral taxon 058 str. F0407	F0407	JH378877_s
GCA_000212855.1	Streptococcus sanguinis SK355	SK355	GL890994_s
GCA_000212815.1	Streptococcus sanguinis SK49	SK49	GL890987_s
GCA_000195025.1	Streptococcus sanguinis SK330	SK330	GL878553_s
GCA_000187745.1	Streptococcus sp. M334	M334	GL732500_s
GCA_000187505.1	Streptococcus parasanguinis ATCC 903	ATCC 903	GL732452_s
GCA_000187445.1	Streptococcus sp. C150	C150	GL698454_s
GCA_000185265.1	Streptococcus oralis ATCC 49296	ATCC 49296	GL622184_s
GCA_000146585.1	Streptococcus mitis ATCC 6249	ATCC 6249	GL397180_s
GCA_000253155.1	Streptococcus oralis Uo5	Uo5	FR720602_s
GCA_900104285.1	Streptococcus sp. NLAE-zl-C503	NLAE-zl-C503	FNJT_s
GCA_000027165.1	Streptococcus mitis B6	B6	FN568063_s
GCA_900012395.1	Streptococcus suis	9401240	CZEF_s
GCA_001096185.1	Streptococcus pneumoniae	SMRU824	CRPU_s
GCA_001983955.1	Streptococcus oralis	S.MIT/ORALIS-351	CP019562_s
GCA_001683375.1	Streptococcus sp. oral taxon 064	W10853	CP016207_s
GCA_001560895.1	Streptococcus mitis	SVGS_061	CP014326_s
GCA_001553685.1	Streptococcus sp. oral taxon 431	F0610 (5-114)	CP014264_s
GCA_001281025.1	Streptococcus mitis	KCOM 1350 (= ChDC B183)	CP012646_s
GCA_000688775.2	Streptococcus sp. VT 162	VT 162	CP007628_s
GCA_000479315.1	Streptococcus sp. I-P16	I-P16	CP006776_s
GCA_000385925.1	Streptococcus oligofermentans AS 1.3089	AS 1.3089	CP004409_s
GCA_001113365.1	Streptococcus pneumoniae	SMRU2014	CKYA_s

Appendix I. Continued.

NCBI Accession	NCBI name	Strain name	EzBioCloud name
GCA_001171885.1	Streptococcus pneumoniae	SMRU946	CKQD_s
GCA_001078705.1	Streptococcus sanguinis	2908	CDMW_s
GCA_000430305.1	Streptococcus mitis 17/34	17/34	ASZZ_s
GCA_000442175.1	Streptococcus tigurinus 2426	2426	ASXA_s
GCA_000385835.1	Streptococcus mitis 13/39	13/39	AQTU_s
GCA_002355895.1	Streptococcus sp. NPS 308	NPS 308	AP017652_s
GCA_000286295.1	Streptococcus salivarius K12	K12	ALIF_s
GCA_000279535.1	Streptococcus mitis SPAR10	SPAR10	ALCH_s
GCA_000259505.1	Streptococcus sp. SK643	SK643	AJMM_s
GCA_000223235.2	Streptococcus oralis SK313	SK313	AFUU_s
GCA_000223255.2	Streptococcus infantis SK970	SK970	AFUT_s
GCA_000223335.2	Streptococcus infantis X	X	AFUQ_s
GCA_000221165.2	Streptococcus mitis bv. 2 str. F0392	F0392	AFUO_s
GCA_000222725.2	Streptococcus parasanguinis SK236	SK236	AFUC_s
GCA_000222705.2	Streptococcus mitis bv. 2 str. SK95	SK95	AFUB_s
GCA_000220065.2	Streptococcus sp. oral taxon 056 str. F0418	F0418	AFQU_s
GCA_000215385.2	Streptococcus infantis SK1076	SK1076	AFNN_s
GCA_000215365.2	Streptococcus oralis SK255	SK255	AFNM_s

국문초록

샷건 메타지노믹스는 미생물과 숙주 또는 환경사이의 미치는 영향을 이해하는데 매우 중요한 역할을 하고 있다. 기술의 발달과 더불어 메타지노믹스를 통한 올바른 미생물 종의 동정과 각 종들의 분포는 마이크로바이옴 연구의 핵심 구성요소가 되었으며, 지난 10 년간 샷건 메타지노믹스 분석을 위한 여러 알고리즘과 데이터베이스들이 개발되어져 왔다. 하지만 서로 다른 기준 데이터 혹은 알고리즘을 사용한 방법들은 서로 다른 분류 정보와 분석 파이프라인으로 인하여 편향된 결과를 나타내기도 하였는데, 이를 보완하고 보다 정확한 분류 동정을 위해 배양이 어려운 표준 균주와 같은 다양한 균주의 유전체 데이터를 포함하는 기준 데이터베이스의 중요성이 대두되고 있다.

샷건 메타지노믹스 분석에서 또 다른 중요한 요소는 분석에 소요되는 시간이라 할 수 있는데 대부분의 생물정보학적 프로그램들은 계산을 수행함에 있어 메모리와 알고리즘 최적화가 되어있지 않아 분석에 상당한 시간이 소요되는 문제점이 있다. 이러한 문제를 해결하기 위해, 본 연구에서는 exact match k-mer classification 과 같은 방법을 사용하여 분석 속도를 향상시켰으며 Up-to-date

Bacterial Core Gene (UBCG)를 기준 데이터베이스로 사용하여 보다 정확한 샷건 메타지노믹 분석을 수행할 수 있게 하였다.

분석의 효율성을 높이기 위해 두개의 기준 UBCG 데이터베이스가 만들어 졌으며 한 개는 박테리아의 분류체계에서 유효한 종명 (Valid names)만을 가지고 있는 데이터베이스와 다른 하나는 유효한 종명과 함께 EzBioCloud 에 있는 genomospecies 를 가지고 생성하였다. 검증을 위해 *Streptococcus* 종을 포함하는 (i) 합성된 메타지놈 샘플과 (ii) 만성 폐쇄성 폐질환(COPD) 환자의 임상 검체 (iii) 혈류 감염 환자의 임상 검체로 이루어진 세개의 데이터 셋을 이용하였으며 기존에 널리 알려진 샷것 파이프라인인 MetaPhlan2 과 본 연구의 파이프라인을 비교 분석하였다.

위 검증 분석에서 UBCG 를 기준 서열로 사용하기에 충분함을 검증하였으며, 빠르고 정확하게 기준 유전체에서 UBCG 서열을 뽑아 샷건 분석에 용이함을 증명하였다. 또한 genomospecies 를 기준 데이터베이스에 추가함으로써, 보다 개선된 분류 정확도를 얻을 수 있음을 제시하였다. 마지막으로 비록 여러 파이프라인과 데이터베이스들이 존재하지만 보다 신뢰할 수 있는 분류결과를 얻기

위해선 기준 데이터베이스의 지속적인 업데이트와 분류 체계의 검증의 중요함을 강조하였다.

이후 본 연구에서 개발된 파이프라인을 이용하여 4,000 개의 샷건 메타지놈 샘플에서 사람에 장내에 가장 많이 발견되는 *Bacteroides* 종에 대한 분석을 수행하였다. 많은 양의 데이터를 분석하여야 하기 때문에 기존에 많이 사용되는 MetaPhlAn2 과 같은 방법은 사용할 수 없었으며 분석 결과 *Bacteroides* 는 도시화된 사람에게 많이 분포하는 반면 아프리카 혹은 남미지역에서 원시적 부족의 삶을 사는 사람에게서는 상대적으로 적게 분포함을 확인할 수 있었다. 또한 각 나라별 인구에서는 우점되는 *Bacteroides* 종이 다름을 확인할 수 있었는데 이는 각 연구의 샘플링 방법 혹은 위치에 따라 설명되어 질 수 있었다. 실험용 쥐의 결과에서는 가장 다양한 *Bacteroides* 를 관찰할 수 있었으며 이는 많은 수의 기준 유전체가 생쥐에게서 나왔기 때문인 것으로 생각된다. 또한 고양이나 강아지 같은 반려동물의 샘플에서도 높은 상관관계를 발견할 수 있었는데 각 동물들의 생활양식과 먹이에 따른 결과인 것으로 보인다.

본 연구를 통해 보다 많은 메타지놈 데이터 분석의 필요성을 강조하고 있으며, 핵심 유전자들을 기준 데이터로 사용하는 방법의 실효성과 성능을 검증하였다. 이러한 핵심 유전자 기반의 기준 데이터베이스는 보다 정확하고 전체 미생물의 풍부도를 예측하는데 중요한 역할을 하는 것을 확인하였고 k-mer 방법을 통해 기존에 존재하던 다른 파이프라인 보다 더욱 빠른 결과를 도출할 수 있었다. 마지막으로 빠르게 기준 데이터베이스를 만들 수 있기 때문에 항상 최신의 데이터를 가지고 분석을 수행할 수 있으며 이는 궁극적으로 본 연구의 파이프라인을 실질적으로 연구나 진단 목적으로 이용하는 연구자들에게 큰 도움이 될 것이다.

주요어: Metagenome, Shotgun, K-mer, Exact match, *Streptococcus*, *Bacteroides*, Core Genes, Sequence classification.