Characterising the source of errors for metagenomic taxonomic classification

Alba Crespi i Boixader

Doctor of Philosophy
University of Edinburgh
2021

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

Alba Crespi i Boixader

Abstract

Characterising microbial communities enables a better understanding of their complexity and the contribution to the environment. Metagenomics has been a rapidly expanding field since the revolution of next generation sequencing began, and it has a wide range of application including for medicine, agriculture, forensics, archaeology and even domestic use [Sarkar et al., 2021, Holman et al., 2017, Khodakova et al., 2014, Santiago-Rodriguez et al., 2017, Vilanova et al., 2015]. Sequencing amplicon data, such as 16S rRNA, is now commonly used to characterise the microbiome in a variety of biological samples. However, their correct taxonomic identification still remains a challenge, and often short reads are identified, correctly or not, at several ranks of the taxonomic tree other than species or subspecies level.

Every metagenomic study is designed for specific needs, and it is often complicated to find a suitable bioinformatics pipeline and reference database. There is currently a lack of systematic benchmarking of in-house methods for metagenomics. The work presented in this thesis aims to establish an approach for the *in silico* validation of 16S rRNA metagenomic data. A method to generate realistic *in silico* metagenome data that resembles project-specific sequencing data is presented, including a new process to generate synthetic negative controls for amplicon data, which can be employed regularly to assess the appropriateness and optimisation of methods for specific metagenomic projects. To aid the benchmarking process, new metrics have been defined based on a measure of taxonomic distance.

A k-mer based method with the lowest common ancestor approach was selected to investigate a range of factors that influence meta-taxonomic classification success. It includes the comparison of database quality filtered at various levels, and as well as a comparison of different taxonomic annotation methodologies. The experimental findings reveal the importance of having highly curated taxonomic annotations of the genetic sequences in the database, and that a missing fraction of the tree of life can lead to misclassification of any related or unrelated organisms. In some cases, it is shown that longer reads can help to improve assignment, with mutations and sequencing errors having a relatively low negative impact.

The marker gene 16S rRNA has well-defined conserved and variable regions, which help to distinguish species. Therefore, these regions were studied and also recalculated using information theory, to investigate which parts of the sequence are discriminative for metagenomic taxonomic identification. In addition, linguistics methods, Term Frequency — Inverse Document Frequency (TF-IDF) coupled with multinomial naive Bayes, is shown to provide understanding of genetic signatures and is applied to generate a new method to classify taxonomically metagenomics short reads.

Biological samples were taken from cattle respiratory tract, DNA was extracted and sequenced to provide metagenomic data. Two sets of experiments were carried out, *(i)* to compare sampling and extraction methods and *(ii)* to characterise the microbial community observed in young cattle in the different lung lobes and nose. The data reveal that the composition of the microbial community observed is highly dependent on the sampling method.

Lay summary

Microorganisms or microbes are present everywhere in the environment and in living things. Identifying them in a particular sample involves the use of a technique known as 'metagenomics', which is based on genetic fingerprints (DNA fragments) specific for each microbe. Metagenomics has been applied to a wide variety of areas, for example to discover which microbes are responsible for a particular disease, determine which microbes are beneficial for growing crops, investigate ancient microbes present in mummies or even to know what is living in your coffee machine. A common way of identifying them in bacteria involves the analysis of the DNA sequence to look for a specific marker gene. However, correctly identifying and quantifying all the microorganisms present in a particular sample is very challenging.

Each metagenomics project is unique, and often it is not a simple task to choose which set of tools is the most appropriate to analyse the data generated. Also, in general, these tools are not adequately tested to determine their suitability for the specific purpose at hand. In this thesis, a method has been developed to simulate 'real-life' metagenomic data to improve microorganism identification. In addition, new ways of measuring the accuracy of the identification has been proposed.

The project has also tried to determine why microbes are sometimes incorrectly identified. To identify them, data was compared with other genetic information currently held in public databases. The quality of this data including the associated taxonomic information (specific identification of the organism and its scientific classification) was found to be a key factor in the correct identification. Furthermore, it was determined that longer DNA fragments resulted in greater accuracy. In contrast, sometimes microorganisms do not have a previously known DNA sequence in the reference database, thus making identification more difficult. Investigations revealed that having as many sequences as possible belonging to a wide range of organisms can also improve the identification of already known organisms.

Genes can contain regions that are identical or almost identical between microorganisms, while others are much more different. These latter ones, depending on the methods used, can be further investigated to improve microorganism identification, making use of DNA fragments of length k (k-mers) found in different proportions, known as genetic signatures, in each microorganism. Thinking of a gene as a sentence and a k-mer as a word enables the use of machine-learning methods from linguistics to create a new approach to identify microorganisms.

Finally, data from cattle lung was collected and analysed. Different methodologies were explored to improve the quality and quantity of microbe DNA for analysis to determine which microorganisms were present and in which compartments of the lung they were found, to enable future exploration of the causes of disease in cattle with pneumonia.

Acknowledgements

I would like to thank both of my supervisors, Ian and David, for all their support, encouragement, and guidance through my PhD journey. Our discussions were really inspiring and exciting, and kept me motivated at moments when I was struggling. Ian is always quick to make connections that help clarify a topic and to raise interesting questions, and is an excellent communicator. He was a truly inspiring supervisor. David's extensive expertise across microbiology applied to animal health helped me widen the horizon of my own knowledge. His perseverance, talent, and brilliant scientific mind were fundamental to the experimental results.

Thank you to all those who have collaborated on projects connected to my thesis, without whom this work would not have been possible. Rob Finn and Alex Mitchell, from EBI (Mgnify) helped me to push me to explore my limits, and raised many ideas for investigation. The Moredun team, especially Mara, Kevin, Morag, Mark and Chris were all very helpful in sharing their exceptional skills and time. I really appreciated the support of my BioSS colleagues throughout.

The support network from colleagues based in the Informatics Forum was invaluable, including endless discussions over lunch in MF1. Also, I would like to thank my friends and family for their infinite interest and support. In particular, I would like to thank Moray, for his continuous encouragement, support, interesting scientific discussions and infinite patience, and Laura, for all her vital support and help during all my years of study, without whom my world wouldn't be the same.

Contents

List of Figures

List of Tables

Abbreviations

AICc Akaike information criterion with a correction for small sample sizes.

ANI average nucleotide identity.

ASVs Amplicon sequence variants.

BAL Bronchoalveolar lavage.

bp base pair.

BRD Bovine Respiratory Disease.

BRDC Bovine Respiratory Disease Complex.

CM Covariance Model.

DA Differential Abundance.

GeFSATI GEnetic Fragments Score Aided Taxonomic Identification.

GTDB Genome Taxonomy DataBase.

HMM Hidden Markov Model.

INSDC International Nucleotide Sequence Database Collaboration.

JC69 Jukes and Cantor, 1969.

kb kilo bases.

KDE Kernel Density Estimation.

KLD Kullbac-Leiber Divergence.

LCA Lowest Common Ancestor.

LEfSe Linear discriminant analysis effect size.

LSU Large subunit ribosomal ribonucleic acid.

MGC My Goldstandard Community.

ML Maximum Likelihood.

NGS Next-Generation Sequencing.

NJ Neighbor Joining algorithm.

OTU Operational Taxonomic Unit.

PBD Phylogenetic Branch Distance.

PCA Principal Components Analysis.

PM post-mortem.

PND Phylogenetic Node Distance.

RTL root-to-leave path.

SSU Small subunit ribosomal ribonucleic acid.

taxID Taxonomic ID.

TDD Taxonomic Distance Divergence.

TF-IDF Term-Frequency Inverse Document Frequency.

TND Taxonomic Node Distance.

TNDSR Taxonomic Node Distance Standard Ranks.

WGS Whole-genome shotgun.

Chapter 1

Introduction

1.1 Motivation

The revolution in Next-Generation Sequencing (NGS) technologies has enabled a step-change in the way that sequence data is collected and used in Biology, including in metagenomics, the sequencing of mixed source nucleic acid samples. These studies have profound implications for human, animal and plant health and disease as well as in diverse areas such as forensic science, environmental pollution monitoring and climate modelling [Sarkar et al., 2021, Holman et al., 2017, Piombo et al., 2021, James et al., 2021, Kajale et al., 2021]. Figure 1.1 shows the main application areas for metagenomics.

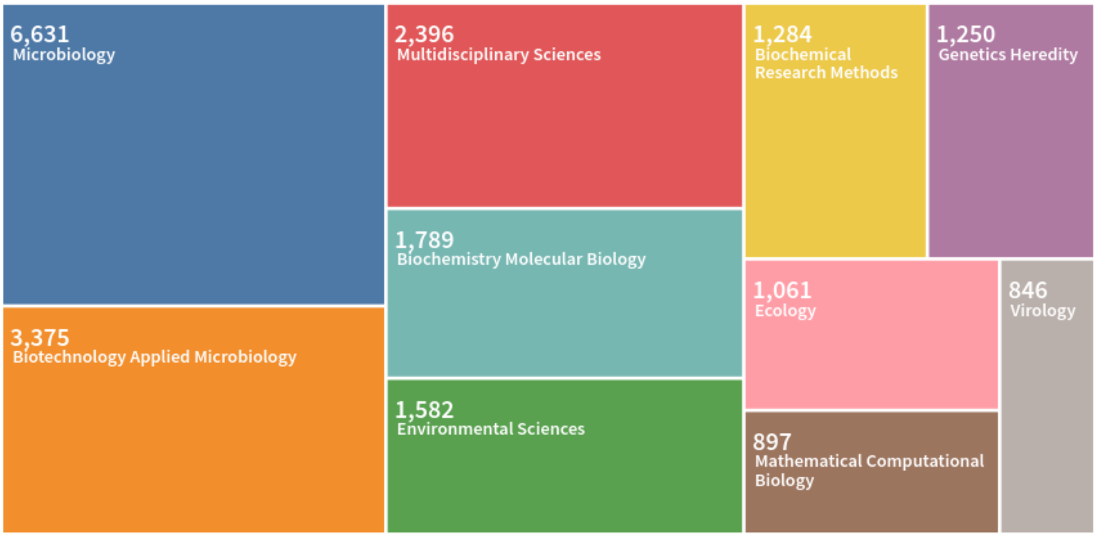


Figure 1.1: Ten top metagenomic applications. Top 10 categories on the web of science. Search terms: metagenomics or metagenomic, 21.366 results on 17/12/2021. They include microbiological related studies, environmental and ecological sciences. Despite viruses being common, their study remains complex, and in here it shows in the last position. Source:web of science

Whole-genome shotgun (WGS) metagenomics can, in principle, sequence all existing organisms present in a sample, as well as functionally characterise them. Targeted or marker gene metagenomics or metagenetics only sequence a highly conserved marker gene or part of it [Breitwieser et al., 2019] to determine the taxonomic composition of a sample. This works well for samples which contain host genomes, is often cheaper than WGS, is widely spread and generally faster, and its biases are well-characterised and understood [Pérez-Cobas et al., 2020, Knight et al., 2018]. The most common marker genes are 16S rRNA for bacteria and archaea and internal transcribed spacer for fungi. This work focuses on the targeted characterisation of bacteria and archaea.

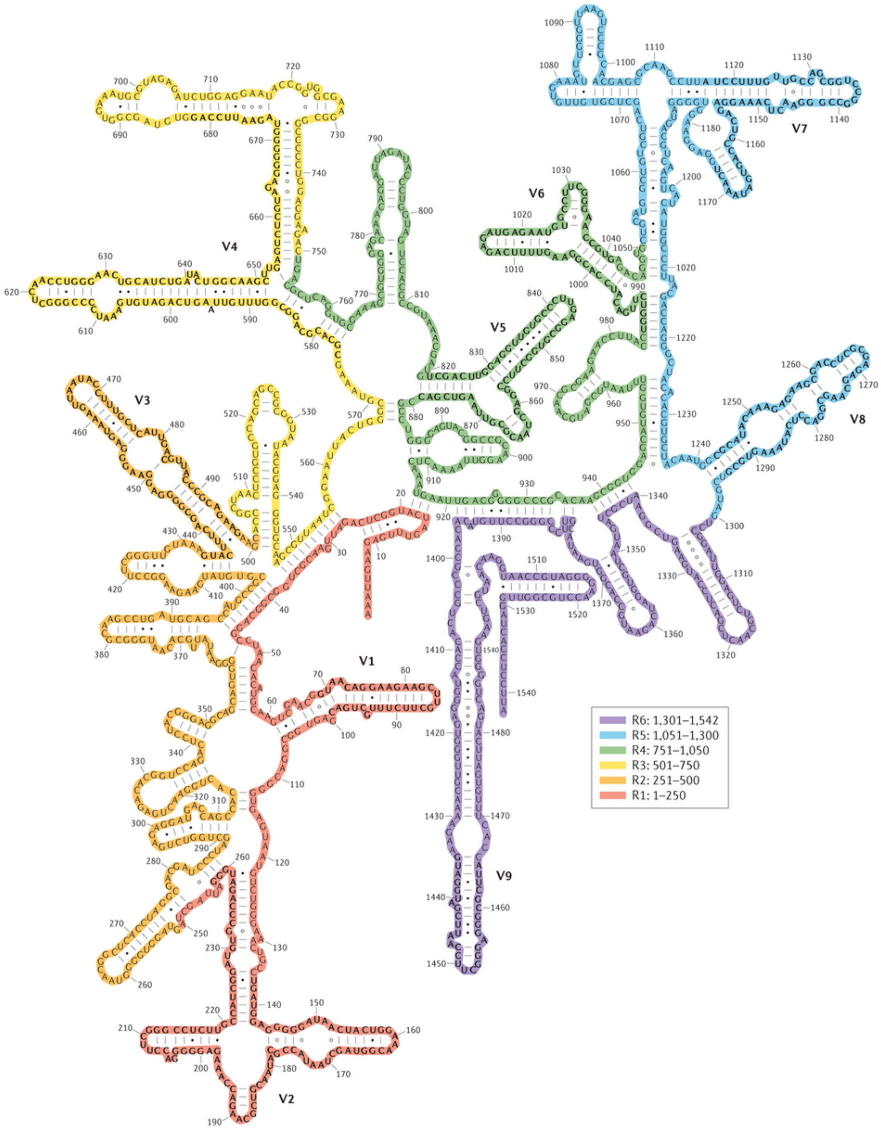
1.2 Marker gene 16S rRNA

The 16S rRNA gene is a housekeeping gene found in Bacteria and Archaea. It is widely used for targeted metagenomic studies and allows multiplexing of samples for sequencing, which reduces costs significantly. It is approximately 1600 base pairs (bp) long, and it is well characterised. It has 9 hyper-variable regions, as shown in figure 1.2. Different variable regions are more suitable to identify certain taxonomic groups, and therefore may be environment-dependant [Yang et al., 2016, Johnson et al., 2019]. In-depth studies of the different variable regions of this gene have revealed that the regions V1-V4 can improve the bacterial biodiversity estimates and accuracy due to its higher divergence [Kim et al., 2011]. However, some contradictory results have been obtained: Yang et al. [Yang et al., 2016] found that the regions V4-V6 are optimal for bacterial phylogenetics and regions V3-V6 are more suitable for extreme environments. Other studies revealed that a full length 16S rRNA gene is necessary for improved taxonomic identification [Yarza et al., 2014, Johnson et al., 2019]. Importantly, it should not be used to identify at the species or strain level [Bharucha et al., 2020, Knight et al., 2018, Pérez-Cobas et al., 2020] because it is not specific enough to confidently distinguish them.

The 16S rRNA gene amplification efficiency can vary depending on the primers used, because affinity varies from sequence to sequence and consequently PCR amplification bias is introduced [Knight et al., 2018, Sunagawa et al., 2013]. Some primers can target most of the bacterial species and in some cases a few Eukaryotic species (close ortholog 18S rRNA with an approximate length of 1800bp) [Kim et al., 2011]. Also, bacterial species often have multiple copies of the 16S rRNA gene, commonly up to 10 but in some cases as many as 17 [Espejo and Plaza, 2018].

1.3 DNA extraction

The challenge of DNA extraction for metagenomics is to obtain all the genetic material (DNA) of all organisms present in a given sample. Some types of samples, e.g.



Nature Reviews | Microbiology

Figure 1.2: 16S rRNA gene. Secondary structure of the 16S rRNA gene sequence belonging to *E. coli*. The conserved and variable regions are marked in bold letters, and the name of the region (V1-9) is printed nearby. The numbers correspond to the position of the ribonucleic base in the sequence which is pointing. The colours of the sequences indicate groups of regions, which are only relevant for the source study. The secondary structure of this gene is complex. The diagram illustrates the interactions of the bases, from which some are located nearby in the sequence, but others are distant. An example of this are the positions 17, 18 and 19 which bonds with the position 915, 916 and 917. Source [Yarza et al., 2014]

clinical, can contain vast amounts of host DNA and also RNA, dominating most of the recovered material. This can be especially problematic for samples where host genetic material is much larger than the average microbial genome. In such cases, additional steps will be required and can consist of depletion of host DNA/RNA, methylation or selective lysis [Afshinnekoo et al., 2017].

Extraction kits can introduce bias to the composition of the microbiome. Also, often, contamination is introduced during the extraction process, known as *kit-ome*. Levels of contamination can vary hugely across different batches and can have huge impacts, especially for samples with a low biomass [Salter et al., 2014].

1.4 Sequencing platforms

Several sequencing platforms are available. The current underlying technologies can be classified by their type of output, either short (second generation) or long read (third generation). Long read sequencing can generate sequences longer than 10 kilo bases (kb). However, despite the continuous improvements, sequences still contain around to 10% of error across each read. In contrast, short-read sequencing is still the most common for metagenomics [Pérez-Cobas et al., 2020]. It requires more complex library preparation, but allows multiplexing which reduces costs and the output is more accurate than for long reads [Hu et al., 2021].

One of the main short read platforms is provided by Illumina. Since Illumina short read sequencing (2×150 bp) became popular over a decade ago, there was a clear need to obtain longer sequences to aid in assembly of genomes, closing gaps. It is particularly useful when there are lots of rearrangements, and also to allow better and more reliable identification. Illumina currently has on the market platforms that allow the sequencing of up to 300 bp reads[1].

Illumina sequencing has popular for a number of years, which allowed the scientific community to develop numerous tools specifically designed to understand its sources of errors in detail. The error rate is estimated to be 0.24 ± 0.06% per base [Pfeiffer et al., 2018] with a higher concentration towards the end of the reads [Tan et al., 2019].

1.5 Bioinformatics analysis steps

Sequencing data contain errors, and it is important to make sure to distinguish them as much as possible from natural variation. Quality control is essential before proceeding to ensure understanding of the nature of the data and any other potential

[1] according to specifications from www.illumina.com, on 02/09/2020

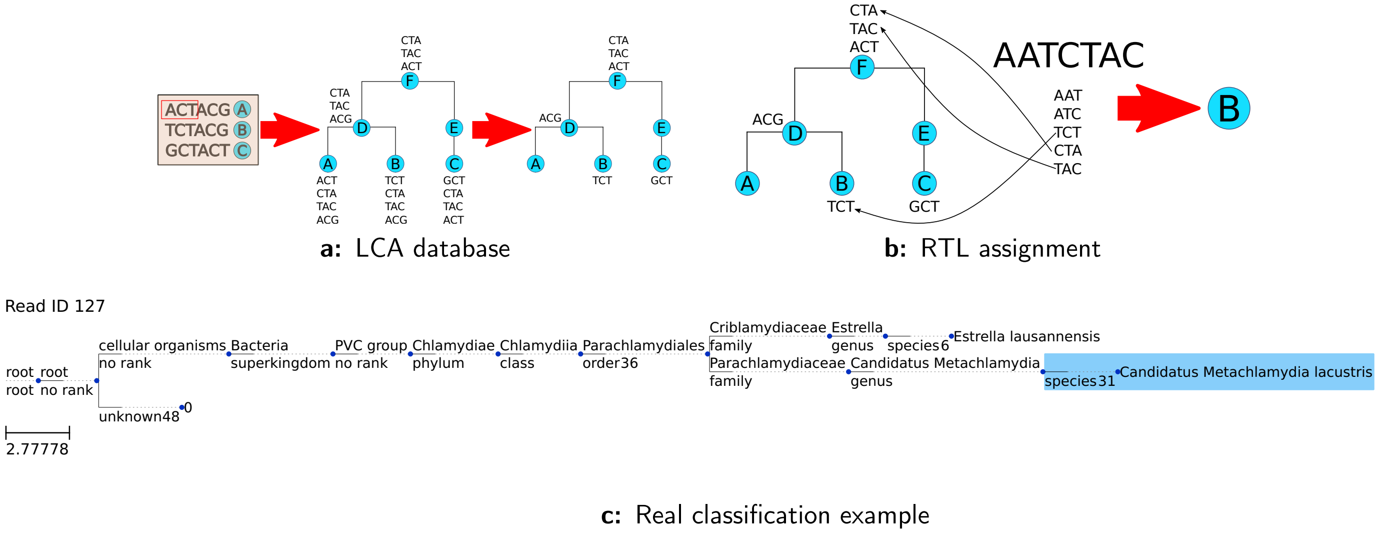
peculiarities. If necessary, reads can be quality trimmed and length filtered. Also, this step remove contamination, adaptors, and host DNA [Bharti and Grimm, 2021, Pérez-Cobas et al., 2020]. Several tools are available for quality control, for example one of the most popular ones is FASTQC [Andrews, 2010], which provides an exhaustive report including per base quality, GC content and over-represented sequences. Quality trimming is also a common practice, especially for Illumina data, and includes Sickle [Joshi NA, 2011], which allows setting a quality threshold and a minimum sequence length to keep.

There are two main types of taxonomic identification methods, the first Operational Taxonomic Unit (OTU) is sequence similarity clustering based, and has been long-established, and the second Exact sequence variants (ESV) also known in some contexts as Amplicon sequence variants (ASVs) are denoising based [Pereira et al., 2020]. OTU methods typically cluster sequences at 97% of identity for identification, which reduces the computation power needed. ASVs taxonomic profilers are designed for feature exact matching to taxonomically identify sequences [Pérez-Cobas et al., 2020] and are much more reproducible [Callahan et al., 2017]. For example, there is no need for rebinning of clusters when datasets are merged [Glassman and Martiny, 2018]. The main difference between OTU and ASVs is the fact the latter methods have the capability, in theory, to distinguish small natural variations from technical errors[Joos et al., 2020]. While some claimed that it is better to avoid OTU based methods [Callahan et al., 2017, Knight et al., 2018, Bharti and Grimm, 2021, Pereira et al., 2020], their performance clearly depends hugely on the context in which they are applied. Recent studies demonstrate that OTU and ASVs methods are comparable in some contexts[Glassman and Martiny, 2018], while in others OTU methods can discriminate better the lower taxonomic ranks [Joos et al., 2020]. Nevertheless, a more in depth research is needed to disentangle the applicability of each type.

To obtain accurate results, it is important to understand each classifier well. For example, the abundance estimation can vary significantly, depending on whether sequence or taxonomic abundance is reported [Sun et al., 2021], and this can have huge effects for the downstream analysis [Jeske and Gallert, 2022].

1.5.1 Kraken taxonomic profiler

Kraken [Wood and Salzberg, 2014], and its newer version Kraken2[Wood et al., 2019], is a composition or denoising based method for metagenomic taxonomic identification, which splits genetic sequences in fragments of size k or k-mers. To create the database, it divides sequences into overlapping fragments of length k (k-mer. Default k=35). These fragments can be unique or in common with several other species, and therefore each k-mer is added to the lowest common ancestor (LCA)[2], an example on how this works is shown in figure 1.3. The classification phase for each read is done by counting the matches of each k-mer against the LCA database and adding them up for each taxon (the scoring system), and then each individual sequence is assigned to the highest rated root-to-leave path (RTL). An example of a real Kraken classification can be found in figure 1.3*c*, and more examples of Kraken classification are found in appendix figure B.1.



a: LCA database

b: RTL assignment

c: Real classification example

Figure 1.3: Kraken taxonomic identification diagram. (a) Kraken creates the classification database by assigning fragments of length k to the lowest common ancestor. (b) For each k-mer of the read, the number of hits matching each node are added together. The final assignment is to the highest weighted root-to-leaf path. In this case, *F* has a weight of 2 and *B* of 1. The highest weighted RTL is *B* (2+1). (c) This is a real example of a Kraken output. Highlighted in blue, the origin of the read. On top of the branch there is the taxon name, underneath each node there is the corresponding taxonomic rank followed by the number of k-mers hitting this taxon. "0" (unknown) are k-mers not present in the reference database. The read ID127 has 3 potential RTL: *unknown* with a total weight of 48, *Estrella lausannensis* with a RTL weight of 42 (36+6) and *Candidatus* Metachlamydia lacustris with an RTL weight of 67 (36+31), which is where it has been assigned.

K-mer based approaches, like Kraken, rely on exact k-mers match. The size of the k-mer is fundamental: too short might not be specific enough, while too long is more likely to be affected by sequencing errors or natural variability of the species [Breitwieser et al., 2019] given genomes tends to accumulate mutations.

Kraken2 is one of the fastest and most efficient taxonomic profilers, and is one of the most accurate for classifying strains not present in the reference database at the species level [Meyer et al., 2022]. Moreover, it assigns each read (or pair of reads) to a taxon while showing the number of hits in each taxon, which allows better understanding of its classification process.

[2] A lowest common ancestor method consist of finding the first common ancestor between 2 or more leaves in a tree.

1.6 Reference databases

Reference databases are key for taxonomic identification of metagenomic reads. There are two main components to this; the genomic sequence content and the associated taxonomic annotation. The genomic databases can be generic, but more often are specialised, for instance containing a set of genes like rRNA, or marker genes such as the 16/18S rRNA, full or nearly completed genomes.

1.6.1 Taxonomy reference databases

There are numerous public databases available for taxonomic classification. Most of them are still based on the Linnaeus taxonomy methods, published in 1753-58 [Miralles et al., 2020], based on physical and other observable traits at the time. Nowadays, evolutionary and phylogenetic theories are widely used but fail to be integrated systematically and reliably within the taxonomic classification systems. This is partly due to the fact that it is hard to compare genomes since the content from more complex organisms, like mammals, to more simple organisms, like bacteria or viruses, differ greatly in size and content. Recent advances allowed the scientific community to propose several approaches to solve the problem.

Long-established reference organisms do not reflect the variability amongst species, and new taxon names have scarcely increased in recent decades officially, however, many are being proposed and used by specialised database curators. This is due to the lack of metadata and specimens associated with the available data, which normally aids the classification [Miralles et al., 2020].

BacDive

An example of a database for bacterial and archaea metadata is BacDive (the Bacterial Diversity Metadatabase) [Söhngen et al., 2014, Reimer et al., 2019]. It was created in 2012 to manually annotate metadata (such as morphology, geographic location, physiology, metabolism, etc.) for culturable prokaryotes. In 2021 [3] it contained 82892 strains belonging to 14350 species. Only about 5% (16,723) of the entries contain an associated 16S rRNA sequence.The strain description relies on the information provided by the submitter. However, without strict control, information from contaminated samples and annotation errors may be included.

Despite being probably the most comprehensive resource of its kind, only one public release a year is available and the FAIR (Findable, Accessible, Interoperable

[3] Search on 04 October 2021 https://bacdive.dsmz.de/dashboard

and Reusable) principles are currently being implemented. Which can make it harder to use in practice.

Partial genetic sequences are frequently the only source available for unculturable or newly discovered microorganisms. Their physical and metabolic state are often extremely hard or impossible to characterise and validate. Phylogenomics is usually applied to compare and place the newly discovered microbiota in a taxonomic tree. However, uncertainty due to the quality of sequencing and lack of supporting phenotypic metadata is not uncommon.

NCBI taxonomy database

The NCBI taxonomy database [Schoch et al., 2020] is one of the most comprehensive resources for taxonomic classification. It contains 2,367,188 taxonomic nodes of which approximately 67.0% are eukaryotic organisms, 22.2% bacteria, 9.4% viral, 0.6% archaeal and the rest, 0.8%, is not specified [4].

The NCBI taxonomy database is manually curated, and the annotation system is based on the guidelines proposed by Linneaus (phenotypic observations such as morphology and physico-chemical properties). However, where possible, 16/18S rRNA phylogenetic reconstruction is used. Monophyly[5] is assumed across all organisms. With some exceptions, lineages generally contain 7 main ranks (superkingdom, phylum, class, order, family, genus, and species) with additional ones where necessary. Often there is little correlation with phylogeny.

Some taxonomic lineages are highly likely to be changed and can be identified by their nomenclature [Schoch et al., 2020]. For example, 'Candidatus' refers to newly proposed prokaryotes while 'sp.' indicates temporary names assigned to prokaryotic taxa until it is given a valid or published name.

Fungi classification has been challenging since this kingdom was first recognized by Linneaus. There are a highly diverse group, and moreover there are some closely related groups of organisms that sometimes can be challenging to discriminate [Naranjo-Ortiz and Gabaldón, 2019]. Currently, NCBI contains 9 phyla out of a potential 16[Schoch et al., 2020].

Several lineages are poorly annotated and not fully hierarchically assigned [Schoch et al., 2020]. Normally, these are identified by the terms 'unclassified' or 'environmental' in their names.

NCBI taxonomy database downloadable FTP files are updated every 24 hours, while the metadata (e.g. names, lineages, etc.) on a weekly basis. There is no

[4] Based on data from 4 October 2021. Source: https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=statisticsperiod=from=to=

[5] Group of organisms descendant from a single ancestor

versioning system, although some changes are easy to track. For example, TaxNodes that are removed or merged are simply stated (without a timeframe).

Genome Taxonomy DataBase (GTDB)

The Genome Taxonomy DataBase (GTDB) is probably the most exhaustive phylogenetic taxonomy for prokaryotes at present. It compares similarities between 120 and 122 single-copy ubiquitous common genes for bacteria and archaea, respectively, presented in two separate trees.

Full genomic sequences are obtained from RefSeq and metagenome-assembled genomes (MAGs) are extracted from the Sequence Read Archive (SRA) metagenomes. After quality checks for completeness, contamination, multiple sequence alignment and de-replication, a taxonomic tree was generated from the concatenated bac120 and annotated by the NCBI taxonomy names.

This approach is able to resolve better potentially polyphyletic[6] grouping compared to 16S rRNA phylogenetic, it does not present PCR bias, and also it solves the issue of copy number variability of the 16S rRNA gene. Although it was first developed for bacteria and later expanded to include archaea [Parks et al., 2020], but still lacks greatly in diversity, e.g Eukaryotes.

1.6.2 Genomic databases

There are numerous resources of generic and specialised genomic databases. Some well-known generic genomic databases belong to the International Nucleotide Sequence Database Collaboration (INSDC) [Arita et al., 2021]. Multiple collections for amplicon data often include their own taxonomies, some of the most well-known are listed in table 1.1. In addition to these, there are multiple reference genome databases, which often are a subset of bigger databases with complementary data from other resources.

Databases generally accumulate data over time, and also guidelines for submission, storage, and checks change [Breitwieser et al., 2019]. Therefore, different versions are likely to generate distinctive results even with the exact same algorithms.

Sequences deposited in available resources have errors. Most of the assembled genomes contain contaminants of all sorts [De Simone et al., 2020] from sequences belonging to other organisms (mainly human, and also other genomes in human sequences) and may include fragments from sequencing controls [R. Marcelino et al., 2020, Breitwieser et al., 2019].

[6] Group of unrelated organisms descendant from multiple ancestors

Table 1.1: Selection of publicly available genomic databases for meta-genome analysis. Description of some of the major and widely used databases that can be used as a reference for metagenomic identification. They are classified by type.

Type	Database	Description	reference
Generic	European Nucleotide Archive (ENA)	Nucleotide sequence archive by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI), INSDC	[Harrison et al., 2021]
	GenBank	nucleotide sequence archive, National Center for Biotechnology Information (NCBI), USA, INSDC	[Sayers et al., 2020]
	DNA DataBank of Japan (DDBJ)	nucleotide sequence database, the National Institute of Genetics, Japan, INSDC	[Kodama et al., 2018]
RNA	Silva	small and large subunit (SSU and LSU) rRNA gene database	[Glöckner et al., 2017]
	Greengenes	16S rRNA gene database chimera-checked.	[DeSantis et al., 2006]
	RDP	Prokaryotic 16S rRNA and fungal 28S rRNA high-quality annotated collection.	[Cole et al., 2014]
	RNA central	non-coding (nc)RNA collection from multiple sources, includes 2D structure.	[Sweeney et al., 2021]
Genomes	RefSeq	Curated high-quality stable reference genomes, transcripts and proteins from INSDC. It contains eukaryotes, prokaryotes, viruses and organlles.	[O'Leary et al., 2016]
	Genomes OnLine Database (GOLD)	Collection of projects which contain both genomic sequencing and metadata.	[Mukherjee et al., 2016]
	RefSeq microbial genomes database	Collection of microbial genomes submitted at the INSD. Reference genomes are manually selected from the most reliable and annotated available.	[Tatusova et al., 2014]
	Integrated Microbial Genomes (IMG)	Partial and complete microbial genomes (archaea, bacteria and virus) including plasmidic sequences.	[Markowitz et al., 2012]
	Ensembl Bacteria	non-redundant prokaryotic genomes according to UniProt criteria.	[Howe et al., 2021]

1.6.3 Silva database

The Silva database [Glöckner et al., 2017] is a collection of Small subunit ribosomal ribonucleic acid (SSU) and Large subunit ribosomal ribonucleic acid (LSU) rRNA sequences. For each release, sequences annotated as relevant are retrieved from EMBL-EBI/ENA releases. rRNA genes are predicted with an HMM model (RNAmmer) [Lagesen et al., 2007]. Then sequences are aligned, quality checked and manually curated, and include partial and environmental sequences. These are generally classified as uncultured, unknown, environmental, and metagenomic.

The Silva accession identifier consists of the original ENA archive ID followed by a dot, then the position of the first nucleotide of the 16/18S rRNA gene in the original sequence, followed by another dot and the position of the terminal nucleotide.

LSU and SSU sequences are archived separately. The database consists of *Parc*, whole data; *Ref*, with archaeal sequences of minimum length of 900 bp and bacterial and eukarotya of 1200 bp whose alignment score is 50 or higher and identity above 70%. Finally, the *Ref NR* is a subset of the Ref which contains the longest representative of each cluster after clustering at 99% identity with UCLUST [Edgar, 2010].

The SILVA SSU version 138.1 contains 9,469,124 aligned sequences in SSU Parc, 2,224,740 in Ref and 510,508 in Ref NR. Table 1.2 show the number of sequences for each superkingdom in each subset. In this thesis the SSU Parc will be referred to as *Parc*, SSU Ref as *Ref* and SSU Ref NR as *NR99*. Silva also has a *truncated* version of all them (referred to as *Trunc*), which consist of the same sequences truncated to the predicted beginning and end of the gene.

Table 1.2: Silva database version 138.1 content. Number of sequences of the SSU Silva database version 138.1 and taxonomic content according to silva's classification.

Subset	Bacteria	Archaea	Eukaryota	Total
SSU Parc	8,475,540	347,020	646,567	9,469,124
SSU Ref	1,983,022	69,198	172,520	2,224,740
SSU Ref NR	431,329	20,389	58,790	510,508

Silva has its own taxonomy classification system, adapted from the Bergey's outlines [Yilmaz et al., 2014, Glöckner et al., 2017]. Specialised resources are used for nomenclature, as well as databases such as the NCBI taxonomy and GTDB (see on page 9). New taxa are added to the SILVA guide tree based on authors' description. A corresponding NCBI taxon ID is provided for each sequence.

Silva database curates their taxonomies up to the genus level. All the names from species and subspecies ranks are the original provided by the submitter of the sequence. This can lead to inaccurate or inconsistent lineages as there is no consistency checking between genus and species names.

1.6.4 Database limitations for metagenomic taxonomic identification

The content of the database needs to reflect as well as possible the microbial community to avoid biased results. In fact, using the wrong reference data may be counterproductive. For instance, [R. Marcelino et al., 2020] generated a mock sample of only fungal species. When the reference database only contained amphibian genomes, the vast majority of the sequences were wrongly assigned to some taxon.

Contamination and annotation in the reference databases impact negatively taxonomic classification. Viral fragments are commonly found in bacterial genomes, and more frequently human and sequencing artefacts in many other organisms. It is estimated that over 2,000,000 sequences in GenBank are contaminated (0.54%, only considering kingdom level contamination to a different lineage) [Steinegger and Salzberg, 2020]. Also, sometimes sequences are labelled to the wrong species. There are currently mechanisms to detect these cases, but some escape the controls.

Some of the main factors that might impact taxonomic identification for metagenomic data are:

- **Incomplete genes and genomes**

  Many species genomes are still not complete. Although some individual genes have been successfully assembled (for instance the housekeeping marker gene 16S rRNA), others have been less successful and only have partial genomic sequences.

- **Contamination**

  One of the major issues for many genome databases is contamination of the reference genomes [Steinegger and Salzberg, 2020]. There are two main types of sources for these problems: computational and experimental. The first ones are commonly caused by homology-based algorithms and the propagation of past erroneous annotations [Bagheri et al., 2020]. The second type is frequently caused by reagents during the genomic extraction and sequencing. The most prevalent are *E. coli* and PhiX174, a phage used as Illumina sequencing control, as well as human genetic fragments and many others [Salter et al., 2014, Breitwieser et al., 2019]. Some laboratories perform contamination checks and remove these artefacts from the sequence data. However, some might remain undetected and therefore incorporated into assembled genomes.

- **Incorrect annotations**

  Some genomic sequences available in public databases are assigned to the wrong species [Breitwieser et al., 2019]. Existing mechanisms allows the detection of most of them, and when such issues are detected in GenBank, the submitters are requested to correct them. However, there is a lack of control as the sequence is owned by the submitter.

- **Number of representatives in the tree of life and databases**

  There is a huge variability in the number of species and subspecies described in public resources. The genomic sequences of a significant number of organisms are not yet know [Pearman et al., 2020], and it is hard to estimate the organismal missing fraction, especially for unculturable organisms.

  For instance, *E. coli* contains 3,372 subspecies in the taxonomy database[7], 5,004,146 entries in INSDC (GenBank)[8] and 4,341,804 in RefSeq, whereas the species *Chlamydia abortus* contains 3 subspecies in the taxonomy database[9], 9,815 entries in INSDC[10] and 1,571 in RefSeq. An example of a less known organism is *Fabrella tsugae*, which is a fungus (a needle cast pathogen of *Larix spp.*) belonging to the phylum of *Ascomycota*. It has a single entry in the taxonomy database[11] and 3 in the INSDC database[12]; its first partial sequence was published in the year 2000, and it is not present in RefSeq.

  To address this, there are innovative solutions such as used in GTDB or Pangenomes project[13].

- **Highly conserved genomic regions**

  Many wrongly identified sequences are due to highly conserved genomic fragments across a wide range of organisms. Reference databases that include an extensive range of organisms have the potential to mitigate the number of false-positives and have more chances to detect any contamination [Breitwieser et al., 2019, R. Marcelino et al., 2020].

[7] Search on https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi with the term "*Escherichia coli*" [05/08/2021]

[8] Search on https://www.ncbi.nlm.nih.gov/nuccore/ with the terms (((*Escherichia coli*) AND "*Escherichia coli*"[Organism]) AND *E coli*) AND *E.Coli* [05/08/2021]

[9] Search on https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi with the term "*Chlamydia abortus*" [05/08/2021]

[10] https://www.ncbi.nlm.nih.gov/nuccore/ search term "*Chlamydia abortus*" [05/08/2021]

[11] Search on https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi with the term "*Fabrella tsugae*" [05/08/2021]

[12] https://www.ncbi.nlm.nih.gov/nuccore/ search term "*Fabrella tsugae*" [05/08/2021]

[13] https://progenomes.embl.de/

1.7 Microbial diversity

Habitats present very diverse microbial communities. Differential abundance, evenness, and richness are measures to quantify unequal representation of the community where there are a few dominant species and many that are rare or of low-abundance. Other parameters that describe the community are *(i)* $alpha$-diversity which measures the local diversity of the community such as richness (includes Chao1, for estimating the diversity of true species; phylogenetic metrics (like Faith's phylogenetic diversity); and evenness (e.g. Shannon index), which makes them less sensitive to the number of sequences per samples, *(ii)* $gamma$-diversity measures the total regional diversity that includes many communities and *(iii)* $beta$-diversity which is a metric that measures how different the samples are in an area (linking alpha and gamma), which can be qualitative (Bray-Curtis, Canberra, weighted UniFrac) or qualitative (binary-Jaccard, unweighted UniFrac), but also there are others like Aitchison distance (for compositional data)[Knight et al., 2018, Luz Calle, 2019, Escobar-Zepeda et al., 2015].

Rarefaction curves estimate the maximum number of species or OTU or ASVs observed in the sample and compares samples with different sizes, which is a normalisation method that consists of subsampling reads such that all samples have the same number [Luz Calle, 2019, Escobar-Zepeda et al., 2015]. However, this exclusion process can create biases and high false positives rates[Nearing et al., 2022]

Non-parametric estimators are used to measure heterogeneity: *(i)* Simpson's index (D) is the probability of assigning two independent sequences taken randomly to the same species, and *(ii)* Shannon-Weaver index H' is an entropy measure which increases with the number of species.

1.8 Abundance comparison

Differential Abundance (DA) analysis of microbial communities tries to identify which taxa are significantly different (in abundance terms) between groups of samples belonging to two or more environments. Many tools are commonly used, some of them adopted from differential expression methods for other types of omics data, such as DESeq and EdeR (RNA-seq), which assume negative binomial distribution, while others are specifically designed, like the popular tool Linear discriminant analysis effect size (LEfSe)[Segata et al., 2011].

Given metagenomic samples present large variation in sequencing depth, the richness can vary drastically. Some tools address this problem by applying rarefaction prior to DA testing (e.g. LEfSe). Another issue is whether or not to filter out rare taxa, which in some contexts and for some statistical methods, this might make sense. Some methods raised issues with the statistical distributions, like the ones designed for RNA-seq data [Nearing et al., 2022], because metagenomic samples in general violate the principle of stable abundance for most species [Weiss et al., 2017]. The data resulting from aggregating metagenomic samples taxa is often highly dimensional and sparse, in other words, it is compositional[Nearing et al., 2022, Pérez-Cobas et al., 2020, Luz Calle, 2019, Knight et al., 2018]. Compositional data analysis employ a range of techniques to use ratios of read counts within a sample [Nearing et al., 2022].

When several of the previous methods were evaluated with a range of environments, it revealed caveats in all them, which need to be taken into account when choosing a tool. For example, some can have high levels of FDR in some cases, and some compositional methods detected fewer significant taxa (although highly correlated with most of the methods at the top 20 hits). Overall, the effect of filtering rare taxa remains unclear, and it is not evident so far, whether any type of methods necessarily outperforms others. [Nearing et al., 2022].

1.9 Benchmarking

Benchmarking metagenomics specifically designed tools is important to establish their performance. Studies evaluating some of the available methods for metagenomics [Lindgreen et al., 2016, Mavromatis et al., 2007, McIntyre et al., 2017] are mostly focused on testing and comparing a limited selection of tools. Their results are not directly comparable, as they use different mock communities [Capella-Gutierrez et al., 2017] and numerous types of metrics. There are two essential components to assess performance and accuracy: gold standards representing ground truth and meaningful and well-defined metrics.

CAMI (Critical Assessment of Metagenome Interpretation) is a community-driven assessment to establish gold standards for metagenomics [Sczyrba et al., 2017]. The second challenge [Meyer et al., 2021] included three types of mock communities (Marine, high strain diversity and a plant-associated including fungi and host material). Their results show that all the tools in the study classify better at higher taxonomic ranks. Kraken2 [Wood et al., 2019] demonstrated the best performance for known marine species and new strains, and is also the fastest and most efficient taxonomic profiler.

LEMMI (A Live Evaluation of Computational Methods for Metagenome Investigation) [Seppey et al., 2020] is a containter based repository for independent pipeline comparison. To be able to compare results, where possible a subset of RefSeq database from 2018 is employed from organisms that have both DNA and protein sequences in the reference. The datasets analysed consist of the CAMI1 data and two *in silico* sets (low and medium diversity) specifically designed for the challenge. Methods are ranked with a performance score, which is a harmonic mean of several metrics applied.

All of them provide hints on potential methods to choose for metagenomic analysis. But it is hard for generic synthetic data to reproduce the variability found in nature [Weber et al., 2019] or account for specific project pitfalls.

In this thesis, the focus is to identify factors that influence metagenomic taxonomic classification.

1.9.1 Mock communities

Mock microbial communities, employed as ground truth by benchmarkers, can be generated *in vitro* or *in silico*. The work presented here, uses the latter one because it is cheaper and easier to manipulate. The College of American Pathologists recommends validating the metagenomic classification pipeline with simulated environments from previously analysed samples [Schlaberg et al., 2017].

In silico data is created from already known genetic sequences used as templates. They have to be independent of previous training datasets, reliable, well-curated, robust [Gardner et al., 2019] and imitating real data. Trustworthy experimental design is crucial to avoid over-fitting, over-optimistic or biased results [Weber et al., 2019], especially for clinical data [Bharucha et al., 2020].

Importantly, the simulated data must include positive and negative controls. In this thesis, positive control refers to genetic sequences found in the reference database and is expected to be identified when performing a meta-taxonomic classification. In contrast, negative control refers to sequences that are not present in the reference database. They might be related to previously unknown organisms or might be completely novel.

There are several publications that describe metagenomic profiling and include negative controls in their studies, the strategies followed include: *(i) randomised sequences* that include the generation of sequences from scratch, shuffling parts of real sequences and introduction of noise [Lindgreen et al., 2016]; *(ii) 'unexpected' sequences* that would not normally be in the reference database [McIntyre et al., 2017], for example protein coding genes when profiling only for the gene 16S rRNA or another example is in Velsko 2018 [Velsko et al., 2018] who used non-oral bacteria for the analysis of dental plaque; and *(iii) identity based* use sequences with low identity match against the reference database, e.g. alignment query <60% and the average nucleotide identity <95% to the best match in the reference database [Liang et al., 2020].

For the positive controls, optimal mock communities should be generated from sequences that are not present in the database [Gardner et al., 2019]. In order to ensure that, several exclusion strategies can be applied:

- *Clade exclusion* approach is where the templates used for the simulation are removed from the classification database at different taxonomic ranks, for example species, genus, or family [Peabody et al., 2015].
- *Random noise* strategy introduces random mutations at each sequence template, for example change the nucleotide in 2% of the positions [Almeida et al., 2018].
- *Simulated evolution* generates new sequences at different evolution rates from the reference [Lindgreen et al., 2016]. The idea is to emulate the natural behaviour. Species genomes accumulate mutations at different rates depending on the genetic region. For example, the 16S rRNA gene consists of well-defined conserved regions with very low mutation rate and variable regions where mutations are much more prevalent.
- *New sequences*: recently deposited genomes have fewer chances to have been incorporated into reference databases [Sczyrba et al., 2017].
- *Random sequences*: this strategy consists of a completely random generation of genomic fragments [Soverini et al., 2019]. The main drawback is that in most cases the simulated sequences will not resemble anything found in nature.
- *Damage simulation*: especially useful for archaeological type of studies, it is the simulation of DNA damage accumulated through time [Velsko et al., 2018].

Despite all, there is still a lack of consensus on what constitutes gold standard data for meta-taxonomics benchmarking: from the number of samples and sequences each set should include to how to simulate realistic microbial communities [Mangul et al., 2019].

Grinder metagenomic simulator

Grinder [Angly et al., 2012] is a versatile metagenomic sequence simulator implemented in Perl. It is capable of generating data for both amplicon, e.g. 16S rRNA data, and shotgun sequencing data. For amplicon data, it contains an additional step to identify the PCR primers. Primers are commonly used in targeted type of studies that flank specific genetic regions based on their conservation, for amplification prior to sequencing.

The abundance profile for the simulated community can be user-specified or based on α and β-diversities.

Reads are selected from a bag of template sequences. For amplicon data they are selected from the start of the specified region, or from a random start position for when simulating shotgun sequence data

It is capable of simulating data for many types of sequencing platforms, including Illumina paired-end.

Finally, Grinder is capable of introducing, if desired, errors for indels, substitutions and homopolymers.

1.10 linguistics methods applied to genomics

In the same way novels contain written stories, genomes can be thought of as a book where chromosomes are chapters and genes are paragraphs. And similarly, each gene contains *phrases* or *words* that are key for the function. Therefore, machine learning approaches from natural language processing can be applied to retrieve information which will identify genomic features of interest. Topic modelling methods are powerful to predict relationships between documents, or organism sequences in this case, and have the potential to identify taxa.

Linguistics methods have been successfully applied in genomics. For example, Liu et al [Liu et al., 2012] created a naive Bayes classification for fungal LSU taxonomic classification. First k-mer size and regions were optimised with Shanon entropy. Two regions were selected for classification. A k-mer size of 8 genus level was applied to a naive Bayes approach to taxonomically classify short reads. Their accuracy levels range from 60 to 80%. It can also be used to identify similar metagenomic samples. In this case, Sener et al [Şener et al., 2018] combined a Term-Frequency Inverse Document Frequency (TF-IDF) approach with Latent Semantic Analysis (LSA) for a range of k-mer size between 2 and 13, for the forward and reverse complement strands.

1.11 The bovine respiratory disease

Bovine Respiratory Disease (BRD) is one of the major causes of morbidity (70-80%) and mortality (40-50% in the US) in dairy calves and feedlot cattle worldwide. It is estimated that the disease causes at least US$1 billion of lost revenue annually in the US alone [Ng et al., 2015, Tizioto et al., 2015] and more than US $3 billion worldwide [DeDonder and Apley, 2015] (including prevention, treatment, and loss of productivity).

BRD is caused by a number of factors, principally multiple environmental stress factors, susceptibility of the host and infectious microbiological agents. Stress resulting from transportation and fasting (especially for long hours), weaning, mixing cattle from different sources and introductory diet have been shown to be extensively linked with a higher incidence of the disease. Climate has also been linked to BRD peaks where commonly outbreaks occur in autumn and winter, although rapid and sudden changes in temperature can also have a big effect as well [Cusack et al., 2003].

It is believed that stress in combination with a primary viral infection in the upper respiratory tract debilitates the host immune system (with mild clinical signs). This is thought to lead to a secondary bacterial infection of the lower respiratory track, ultimately resulting in bronchopneumonia [Grissett et al., 2015, Tizioto et al., 2015]. The viruses and bacteria involved in BRD are known as the Bovine Respiratory Disease Complex (BRDC). These include viruses such as bovine respiratory syncytial virus (BRSV), bovine herpes virus type 1 (infectious bovine rhinotracheitis, IBR), bovine parainfluenza 3 virus (PI3V), and bovine viral diarrhoea virus (BVDV), and the most common bacteria are *Arcanobacterium pyogenes*, *Mycoplasma bovis*, *Pasteurella multocida*, *Histophilus somni* and *Mannheimia haemolytica*. The three latter bacterial pathogens are frequently found in the upper respiratory tract of healthy cattle and, following damage to the respiratory epithelium resulting from viral infection, these opportunistic bacteria colonize the lungs [Gershwin et al., 2015, Holman et al., 2015]. *M. bovis* and *M. haemolytica* are often found in synergy.

BRD symptoms can be quite severe and therefore debilitate the animal. The most common clinical signs are fever, loss of weight, depression, respiratory signs (rapid shallow breathing, coughing—early mild versus advanced acute), adventitious lung sounds, nasal and eye discharges, salivation, diarrhoea and decreased milk production [Ng et al., 2015]. Despite efforts to improve our understanding of BRD, the vast majority of the pathogens which cause individual outbreaks are unknown. However, the disease can be diagnosed by observation of typical clinical signs and by assessment of tracheal washes, blood samples, nasopharyngeal swabs and post-mortem material. In the feedlot, about 7% of the cattle require treatment [Cusack et al., 2003]. Antibiotics, non-steroidal anti-inflammatory and other therapeutic drugs are commonly administrated in the treatment of this disease.

Many studies have been published to characterise the potential microbiota involved in BRD through high throughput sequencing [Holman et al., 2017, Zeineldin et al., 2017a, Johnston et al., 2017, Hause et al., 2015, Davids et al., 2016, Holman et al., 2015]. Most of them are on either samples from swab or BAL and only a few on tissue. They are normally focused on bacteria, commonly targeted to the 16S rRNA gene. Only a minority are trying to detect viruses.

1.12 Objectives

- Develop a method and define suitable metrics for benchmarking real-like metagenomics data. This tool should enable the evaluation and optimisation of any chosen analysis pipeline.
- Identify the main source of errors of the metagenomics taxonomic identification of 16S rRNA sequencing data.
- Implement a strategy to improve the taxonomic classification.
- Analysis of real sequencing data: characterise the microbiome of calf lung lobes.

All the work presented here is for 16S rRNA metagenomics data, and mostly synthetic data. However, some of the methods can also be applied to a more general context. Also, it is based on k-mer methods for taxonomic classification.

With the rapid growing field of metagenomics, it is often challenging to find an adequate method to suit specific research projects. Moreover, there is a need to perform systematic in-house validation of metagenomic bioinformatic methods that fits the available resources. However, this is often challenging due to the lack of software. This work presents a novel methodology to simulate in silico meta-genome data based on real sequencing data.

Organisms are taxonomically organised in hierarchical lineages. And metagenomic short reads can be labelled to any of them. However, precision and recall metrics are based on true or false positives and negatives and cannot capture how far in the taxonomic tree a read had been assigned to. For example, a specific read can be correctly labelled at the phylum level, which is not very specific rank, and quite far from species.Therefore, we define new metrics specifically designed metagenomic taxonomic identification. These are based on distances between two nodes in a tree.

There are multiple factors impacting metagenomics taxonomic classification: from sequencing errors to the reference database. Each reference database contains two essential components: one are sequences representing organisms, and the other is the taxonomic lineages associated with them. However, databases are incomplete, and any metagenomic methods needs to be able to classify as accurately as possible novel sequences.

All genes have regions that are more essential than others, therefore they contain areas much more conserved throughout evolution. Once informative genetic areas have been identified, they can be used to improve future taxonomic identification.

In the same way some words are more relevant to discover the topic of text, some k-mers, or genetic sequence fragment of size "k", are key to determine the species of a given metagenomic read. A novel strategy based on linguistic methods for taxonomic identification is developed.

Finally, calf lungs were sampled, DNA extracted and sequenced for metagenomics 16S rRNA locus. The microbiota for healthy animals of different lung lobes is characterised, compared, and biomarkers are determined. Several sampling methods were tested to establish the best for microbial community recovery.

Chapter 2

Methods

2.1 General programming

All scripts were implemented in python, unless otherwise stated. Seaborn, matplotlib and venn libraries were used to generate images in python. Also, the python libraries in scikit-learn were used to generate Principal Components Analysis (PCA) plots. For the images generated in R, the ggplot2 and ggtree were used. Interactive hierachical pie-charts were created with Krona tools [Ondov et al., 2011]. When necessary, shell scripting was also used, for example to run pipelines.

2.2 Reference databases

2.2.1 NCBI taxonomy database

The NCBI taxonomy database and its metadata were downloaded on [18/11/2020].

2.2.2 Silva database

Fasta files were downloaded from Silva database archive version 138.1 for SSU Parc, ParcTrunc, RefTrunc, refNR99, refNR99Trunc and NR99. Sequences were mapped to their corresponding NCBI Taxonomic ID (taxID) and updated old, merged and deleted nodes by the files provided by NCBI taxonomy.

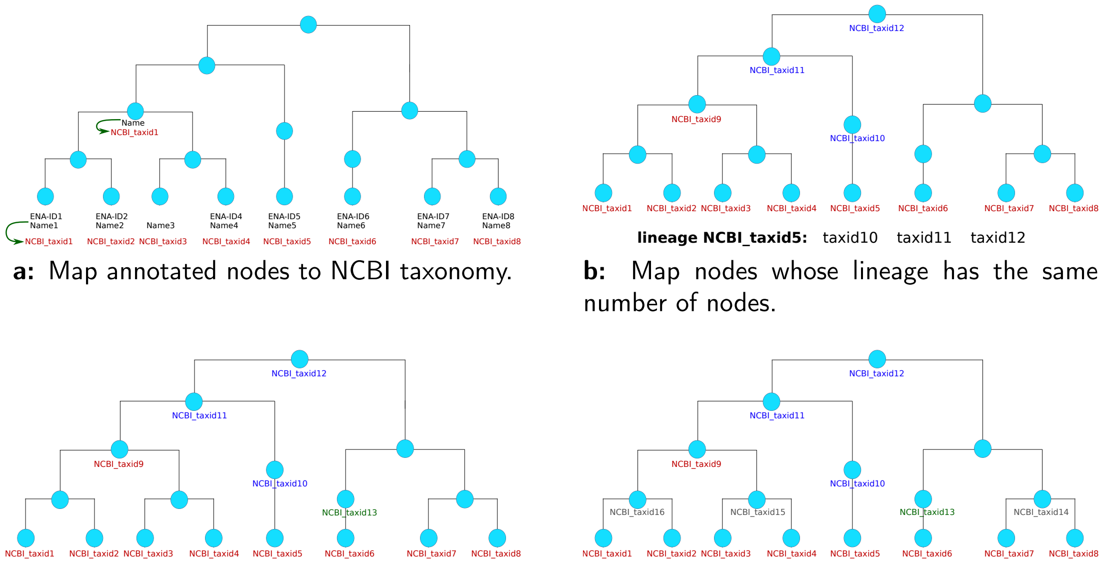
ParcClean and ParcTruncClean were generated by selecting sequences that contain labels of at least the 7 main ranks (species, genus, family, order, class, phylum, superkingdom) in the NCBI taxonomy. The total number of sequences was reduced from over 6 million to just over 1.2 million in both cases.

All the subsets were reverse transcribed and formatted for Kraken2 input.

2.2.3 Mapping NCBI taxonomy database to GTDB

Several steps have been implemented to annotate the NCBI taxonomy to the phylogenetic Genome Taxonomy DataBase (GTDB) tree version 95, as shown in figure 2.1.

First, leaves were directly mapped into Taxonomic ID (taxID) as well as other nodes with annotation. Then, branches which have the same number of nodes on both trees were annotated. Parent nodes were annotated when all the children agree. For those cases where there was disagreement, a strategy of finding the Lowest Common Ancestor (LCA) was applied. All these steps were applied from root to leave and vice versa. Finally, unnamed children of nodes with taxID and unassigned children taxa were labelled.



a: Map annotated nodes to NCBI taxonomy.

b: Map nodes whose lineage has the same number of nodes.

c: Map parent's nodes where all children agree.

d: Map lowest common ancestors.

Figure 2.1: Mapping NCBI taxonomic IDs to GTDB phylogenetic tree. This diagram illustrates the process followed to map NCBI names and IDs into the GTDB. GTDB follows phylogenetic principles based on selected genes (Bacterial and Archaeal) whereas the NCBI was originally based on Linneaus principles, later expanded and in some cases phylogenetic information is included to resolve conflicts where necessary. The mapping process is complex due to the non-matching nature of the two taxonomic trees, specially for those branches with different number of nodes. Starting from the taxonomic node names already mapped by GTDB, first, the leaves were mapped to the NCBI taxonomic ID. Next, those branches with the same number of nodes in both trees were annotated. Finally, a lowest common ancestor approach was applied.

2.3 Mock communities

2.3.1 Chlamydia dataset

A taxonomic tree was plotted of all the Chlamydiae phylum (NCBI taxID 205528) descendants present in the Silva database [Quast et al., 2013] version 128 (RefNR99), see figure B.2 in the appendix.

Sixteen tips were manually selected according to the following criteria: *(i)* belonging to 2 distinct orders; and *(ii)* selecting unevenly across families and genus, such that some have a much higher number of representatives than others.

An abundance profile was created by a gamma variate function (alpha = 1 and beta = 2). This is to emulate the behaviour of microbial communities, where normally 1 or 2 species were disproportionately much more abundant compared to the others. Then the abundance percentage was calculated (see table 2.1).

Table 2.1: Selected Chlamydia species. Chlamydial species selected and their corresponding sequence ID from the Silva database. The percentage was used as the abundance profile, which was randomly generated with gammavariate function (alpha = 1 and beta = 2). This was the input Grinder sequence simulator.

Organism	Sequence ID	percentage
Candidatus Rhabdochlamydia crassificans	AY928092.1.1495	16.7200145863
Chlamydia trachomatis RC-L2(s)/46	CP002672.850541.852092	13.84881612
Candidatus Amphibiichlamydia ranarum	JN402380.1.1473	10.4517322605
Candidatus Rhabdochlamydia porcellionis	AY223862.1.1366	8.13402043305
Chlamydia pecorum PV3056/3	CP004033.960840.962380	7.91822031861
Candidatus Fritschea eriococci	AY140911.1.1515	6.90637230667
Candidatus Amphibiichlamydia salamandrae	JN392919.1.1477	5.63533898192
Simkania negevensis Z	FR872582.370012.371557	4.74672826415
Chlamydia psittaci 84/55	CP003790.1022893.1024430	4.60687103842
Candidatus Metachlamydia lacustris	GQ221847.1.1355	4.301364458
Chlamydia abortus	U76710.1.1548	3.89957847974
Candidatus Fritschea bemisiae	AY140910.4332.5873	3.86787407858
Neochlamydia hartmannellae	AF177275.1.1529	2.9721176161
Criblamydia sequanensis CRIB-18	CCEJ010000017.3833.5371	2.46327063495
Estrella lausannensis	CWGJ01000021.98.1637	1.81694372235
Chlamydia suis MD56	AYKJ01000039.25.1560	1.71073670068

A total of 2000 paired-end reads were generated with grinder [Angly et al., 2012]. Their average length was 150 with insert size of 250. Illumina-like errors were simulated by the recommended 4th degree polynomial derived model and a mutation indel ratio of 80-20 (1 indel for every 4 mutations). The phred scores were recorded as 75 if no error, or 35 otherwise. The simulated reads were then separated into a forward and reverse files.

Phylogenetic tree of the Chlamydia dataset

Sixteen selected species representatives sequences (16S rRNA gene) were aligned with infernal [Nawrocki and Eddy, 2013] with the Covariance Model (CM) of Small subunit ribosomal ribonucleic acid (SSU) rRNA bacteria (entry RF00177) from the RFam database [Kalvari et al., 2018].

A Maximum Likelihood (ML) phylogenetic tree was generated with the r-cran package *phangorn* [Schliep, 2011] following the default method for DNA. The distance matrix was calculated with the function *dist.ml* and model Jukes and Cantor, 1969 (JC69) which was the simplest substitution model, and it assumes equal base and equal mutation rates. An initial tree was estimated with *bionj* function, an improved version of the Neighbor Joining algorithm (NJ). To optimise the tree, the best model that fits the data was chosen according to Akaike information criterion with a correction for small sample sizes (AICc).

Network classification

Directed network was created with igraph for R for a graphical visualisation of where the reads were taxonomically identified compared to reality.

Nodes of the network corresponded to nodes of the taxonomic tree of the 16 chlamydial species, where leaves were coloured in blue and the rest of the taxa were in orange and the unclassified sequences were presented in red. The self-loops were removed from the network for clarity. The size of each leaf node was proportional to the number of correctly identified sequences.

Edges were directed. The edge starts where reads were originally from and point towards where they have been assigned. The width of the edges was proportional to the number of events connecting the two nodes. The number in the middle of the edges was the taxonomic distance or number of taxa between the two nodes in the taxonomic tree. Grey arrows symbolise nodes within the same taxonomic lineage compared to the red edges, which were different.

2.3.2 Genetic regions test data

Metagenomic reads were simulated from the abundance detected in a published sample analysis, accession MGYA00140743 on MGnify. Sequences to be used as templates were selected from the Silva Parc database version 132. The abundance profile often contains higher taxonomic ranks other than species or subspecies. For the simulation, the number of sequences templates to be used was limited to 2 for genus, 3 for family, 4 for order, 5 for class, and 6 for phylum.

Sequences targeting the full length of the 16S rRNA were generated with Grinder, consisting of 48930 Illumina paired-end-like reads, 150nt length.

The final community contains 262 different species from 349 sequence templates. The composition can be observed in figure 2.2.

Subsequently, quality scores from the original samples were imputed, adjusting lengths if necessary (adding N or removing bases). The quality scores at each individual position were then used to generate mutations to the simulated sequence according to the probability associated with each phred score.

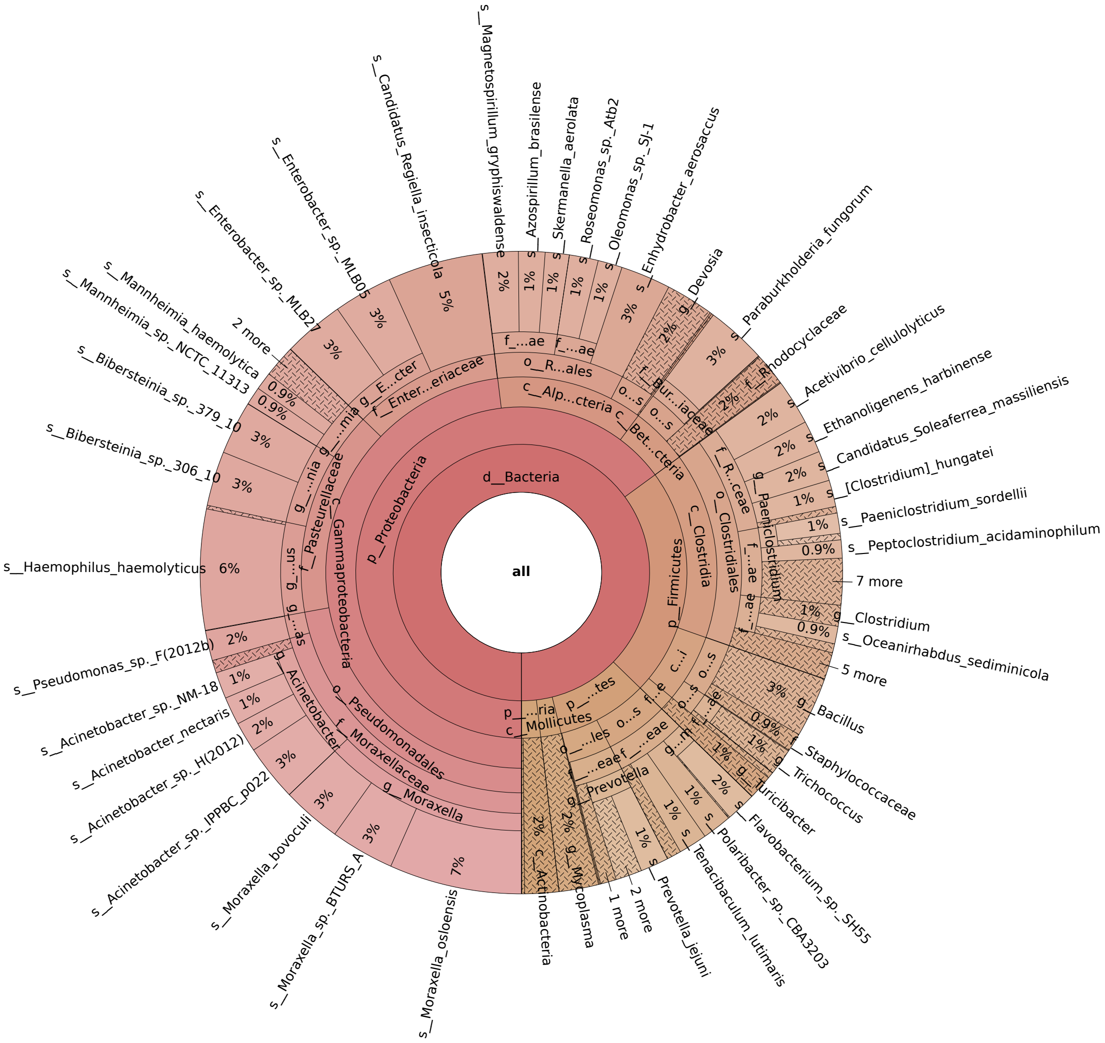


Figure 2.2: Abundance profile of the genetic regions test data. Hierarchical representation of the abundance profile of the simulated data, which contains 262 species and, 48930 reads. All the reads are bacterial. The main dominant phylum is Proteocteria followd by Firmicutes.

2.3.3 Selection of publicly available environmental samples

Samples analysed by MGnify [Mitchell et al., 2020], which were publicly available and include raw sequencing reads, were selected from nine diverse biome projects with the following criteria: *(i)* amplicon samples sequenced with the Illumina platform that *(ii)* have a standard normal quality, *(iii)* different environments to cover a wide range of clans, *(iv)* at least 60% of the sequences were taxonomically identified, *(v)* reasonable amount of number of reads (e.g. discard extremely low number of reads), *(vi)* include single and paired end samples and *(vii)* a range of average read length across samples. Table 2.2 contains a description of them.

Table 2.2: Samples for the simulation of the mock communities. Description and main characteristics of the amplicon analysis from MGnify selected to emulate different environments. The SSU OTU tsv file was used as abundance template. The fastq files were used to obtain the number of sequences, their length and quality. The phred scores were used as templates to create mutations on the simulated sequence according to probability associated. In italics were the names the samples were referred to. The number of reads and read length refers to the unmerged reads (if paired). The proportion of unassigned reads is the output of the pipelined applied to the data.

Sample	MGnify ID	Pipeline version	Instrument model	Library layout	Environment	Number of reads	Read length	Unassigned reads	Description
soil	MGYA00038984	2.0	Illumina MiSeq	single	soil	24,937	151	26.15%	Soil community for preservation Metagenome.
fish	MGYA00169868	4.1	Illumina MiSeq	single	fish	65,850	238	0.0%	Alginate oligosaccharide-induced shift in the intestinal microbiota of salmon.
human	MGYA00147801	4.1	Illumina MiSeq	paired	human	25,089	408	0.0%	Oral and nasal microbiome in Parkinson's disease.
ice	MGYA00175862	4.1	Illumina MiSeq	paired	deep marine sediment	104,163	236	0.0%	Microbial diversity of the central Arctic Ocean.
gut	MGYA00247748	4.1	Illumina MiSeq	paired	Mouse gut microbiome	32,219	250	0.0%	Prior dietary practices and connections to a human gut microbial metacommunity alter responses to diet interventions.
plant	MGYA00087222	3.0	Illumina MiSeq	paired	plant	30,977	241	5.97%	Effects of crown gall disease on natural microbiota of Vitis vinifera.
sludge	MGYA00094787	3.0	Illumina MiSeq	paired	sludge	45,887	313	40.77%	Evaluation of the impact of DNA extraction and primer choice on the observed microbial community in activated sludge using 16S rRNA amplicon sequencing.
faeces	MGYA00109029	3.0	Illumina HISeq 2500	paired	human faeces	112,432	236	12.88%	Genetic determinants of the gut microbiome in the TwinsUK cohort.
reactor	MGYA00246008	4.1	Illumina MiSeq	paired	archaeal community	32,248	393	0.0%	Characterization of archaeal and bacterial communities in UASB reactors fed pig manure supernatant, operated at high ammonia concentrations.

2.3.4 Generation of the Test Dastaset

The SSU Operational Taxonomic Unit (OTU) tsv file from MGnify was the input for the abundance profile for each sample. This abundance was checked and updated with the NCBI taxonomy database (29 January 2019). Taxa matching species/subspecies levels were selected. In the case of taxa higher up on the tree of life, sequences belonging to the specific branch were picked randomly. Thirteen sets of varying complexity were created.

Positive controls

Sequences from the Silva database [Quast et al., 2013] (reference version 132) that were also included in the GTDB database [Parks et al., 2018] were chosen as templates.The lineage of each one of these had been mapped and updated to the NCBI taxonomy database on 29 January 2019.

Negative controls

Synthetic 16S rRNA-like sequences were generated. An HMM profile was created from the alignment of unidentified and unclassified 16S rRNA sequences present in the RNA central database [Sweeney et al., 2019] (on 31/01/2019) after duplicate removal (CD-hit [Fu et al., 2012], global sequence identity flag set to 1, version 4.6). Finally, 100 sequences were generated with the function hmmemit from HMMER [Wheeler and Eddy, 2013].

Mock community

Single or paired-end mode was used according to the input dataset. The sequence length profile of each set was the same as the original experimental fastq file(s). Similarly, the quality phred scores were imputed from the original sequenced data. This ensured that the lengths of the sequences and quality profiles were preserved to provide a more realistic dataset. This first set did not contain any mutations of any type. Then random mutations was introduced at 1% , 2% and 3% as noise. Next, sequencing error was introduced according to the probability associated with each phred score per base. For the final sets, the random noise was reverted, which makes a total of 8 single or paired-end files per sample. 3 replicas from each of the 13 different levels of complexity were created.

A total of 312 synthetic sets of data were simulated with Grinder [Angly et al., 2012].

2.4 Meta-taxonomics classification

Kraken [Wood and Salzberg, 2014] methodology was applied for taxonomic assignment. Both Kraken1 and Kraken2 were used for the analysis, which is indicated in each case.

First, a database of LCA k-mers had to be created. It required the NCBI taxonomy database format and a formatted kraken fasta file, which consisted of adding *kraken:taxid|NNNNN* after each sequence ID (NNNNN indicates the corresponding taxID).

The *Chlamydia* data was analysed with Kraken [Wood and Salzberg, 2014] version 1.0 with the paired-end mode. A customised database was built consisting of the 16 original sequences with the corresponding fraction of the NCBI taxonomy database.

The test data was analysed with Kraken1 with the reference database Silva Ref version 132.1 after quality trimming with sickle (quality = 30 and minimun length=100 base pair (bp)). NCBI taxonomy was mapped to Silva with the metadata provided by Silva. taxID were updated with the NCBI taxonomy database, available on 29/01/2019. It consisted of merging taxIDs and deleting sequences whose taxID no longer existed.

All the datasets for chapter 5 were taxonomically labelled with Kraken2 classification with the Silva database subsets NR99, NR99Trunc, Ref, RefTrunc, Parc, ParcTrunc, ParcClean and ParcTrunClean mapped to NCBI taxonomy and NR99 with Silva taxonomy.

2.5 Metrics

In table 2.3 there is a description of the metrics calculated. They were implemented in python3.

Table 2.3: Metrics. The metrics were implemented with python3 for each class. TP - true positives, TN - true negatives, FP - false positives, FN - false negatives.

Metric	**Formula**	**Definition**
Precision	$\frac{TP}{TP+FP}$	Ratio of correctly predicted positive observations to the total predicted positive observations.
Recall or sensitivity	$\frac{TP}{TP+FN}$	Ratio of correctly predicted positive observations to all the observations in the actual taxon.
F1 Score	$2 * \frac{Precision}{Precision+Recall}$	Weighted average of precision and recall.
Specificity	$\frac{TN}{TN+FP}$	Ability to determine the true negatives correctly
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Ability to differentiate the positive and negative cases correctly
False Positive Rate	$\frac{FP}{FP+TN}$	Proportion of false positives within the entire set of tests.

2.6 Term-Frequency Inverse Document Frequency

In chapter 4 the Term-Frequency Inverse Document Frequency (TF-IDF) approach was implemented with the functions provided by the Scikit-learn library. Full details can be found at the revelant sections.

2.7 Lung data

The taxonomic data used for the experiment to evaluate the DNA extraction methods was input into the MEGAN Community Edition (version 6.5.2, built 25 Aug 2016) to analyse and generate plots.

The taxonomic assignment data from the experiment, for characterising the microbiome of the respiratory tract, was plotted with python3 standard libraries described previously.

The α-diversity was calculated and plotted with the function provided by scikit-bio with the abundance results directly from Kraken.

The Lasso method was used to select phyla by sampling method and lung location that enhance the prediction accuracy. All the nasal samples were removed because of their small sample size. Those samples with less than 1000 counts were discarded . Those phyla with no counts in any of the grouped samples were discarded. Next, the relative abundance by sample was calculated. This step is necessary to solve the problem of uneven sequencing depths and eliminates technical variation.

The Lasso method was implemented with R by Magdalena Navarro. Once the data was aggreagated by phyla, the matrix was randomly separated into training (80%) and testing (20%) per class for each group. The first step for the lasso method was to get the optimal value for lambda (penalisation or regularisation parameter) for a multinomial and alpha 1 per group type (sampling method, lung location, animal). The Lasso regression was trained with the best lambda (determined at the previous step).

DA was performed with LEfSe version 1.0 was applied to determine differentially abundant taxonomic groups for the lung data. The method was applied after normalising the data.

Chapter 3

Benchmarking taxonomic profiling

3.1 Motivation

Numerous methodologies have been implemented for assigning taxonomic information to metagenomic sequencing data. It is often challenging to choose a suitable analysis strategy that fits the data [Mangul et al., 2019, Weber et al., 2019]. Often popular methods based on literature are used, which might not necessarily be the best performing or adequate [Capella-Gutierrez et al., 2017].

Sometimes reads are correctly identified as the original species by the binner or taxonomic classifier. Others show similarities to multiple species and therefore might be labelled to a common ancestor [Hornung et al., 2019] (higher taxonomic rank in the tree of life). Occasionally, reads might be misclassified into unrelated taxonomic lineages, as shown in figure 3.1. Other times there is not enough information to confidently assign them. It is important to understand why, when and how often these errors occur.

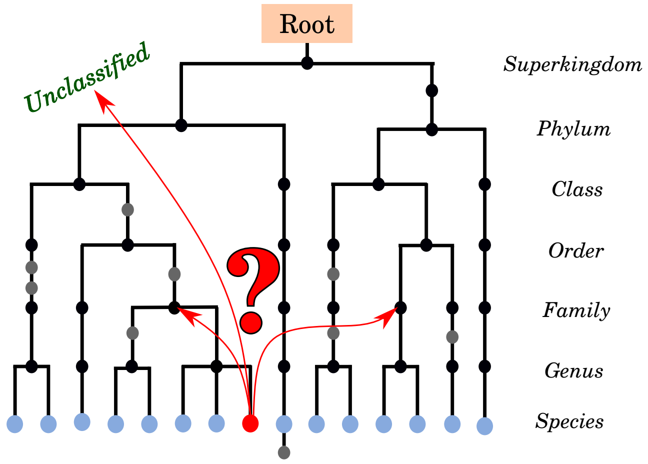


Figure 3.1: Classification in the tree of life. Example of a potential taxonomic tree. Text to the right are the standard taxonomic ranks. Blue nodes are species, black belong to standard taxonomic ranks and grey to other ranks. The red dot represents the original species of a metagenomic sequence. Red arrows are pointing to potential taxonomic nodes where it can be assigned. Frequently sequences are identified to their species of origin. There are cases where they can be labelled to a common ancestor, to some unrelated taxonomic node or unclassified.

3.1.1 Objectives

The main goals of this chapter are to:

- Produce a new method for simulating *In Silico* meta-genetics data based on real experimental sequencing amplicon data.
- Define new meaningful metrics specifically designed for meta-taxonomics quality assessment.
- Evaluate the taxonomic assignment of the marker gene 16S rRNA gene.

3.2 New proposed taxonomic distance based metrics

When profiling metagenomics data, it is essential to understand where misclassification occurs. A new type of metric is proposed based on distance in a given taxonomic or phylogenetic tree. The main concept is to understand how far from the original organism, sequences are classified. Four subtypes are defined below.

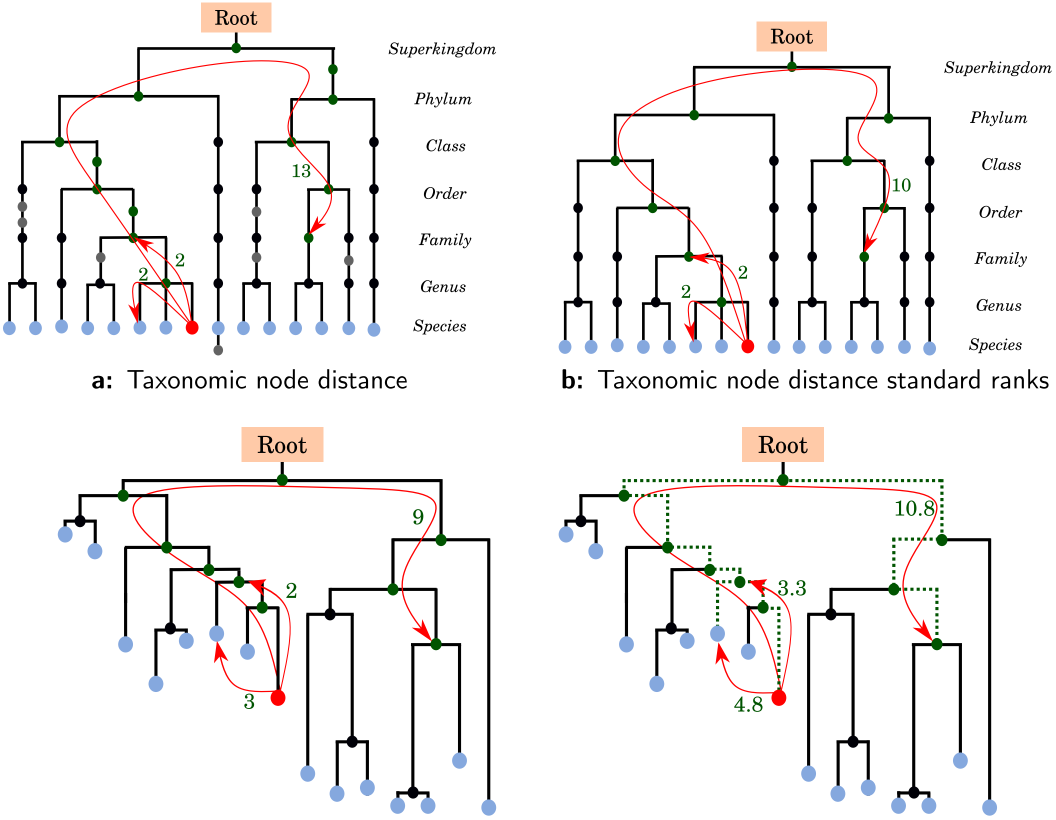
Taxonomic Node Distance (TND) — Number of nodes that differ between the original taxon and the classification result in the taxonomic tree.

Taxonomic Node Distance Standard Ranks (TNDSR) — Number of standard ranks taxa that differ between the origin of the sequence and the classification result in the taxonomic tree. The standard ranks being: superkingdom, phylum, class, order, family, genus, species and the root node of the taxonomic database.

Phylogenetic Node Distance (PND) — Number of nodes between the original and the assigned node in the phylogenetic tree.

Phylogenetic Branch Distance (PBD) — The accumulated branch distance between the original and the assigned node in the phylogenetic tree.

Figure 3.2 represents a hypothetical classification of 3 different reads belonging to the same species. When a read is correctly identified, the distance is 0. Distances increase with classification occurring in further away nodes. The distance assigned to non-classified reads will be the maximum number of nodes or maximum branch distance between the two most distant leaves in the representative tree.



a: Taxonomic node distance

b: Taxonomic node distance standard ranks

c: Phylogenetic node distance

d: Phylogenetic branch distance

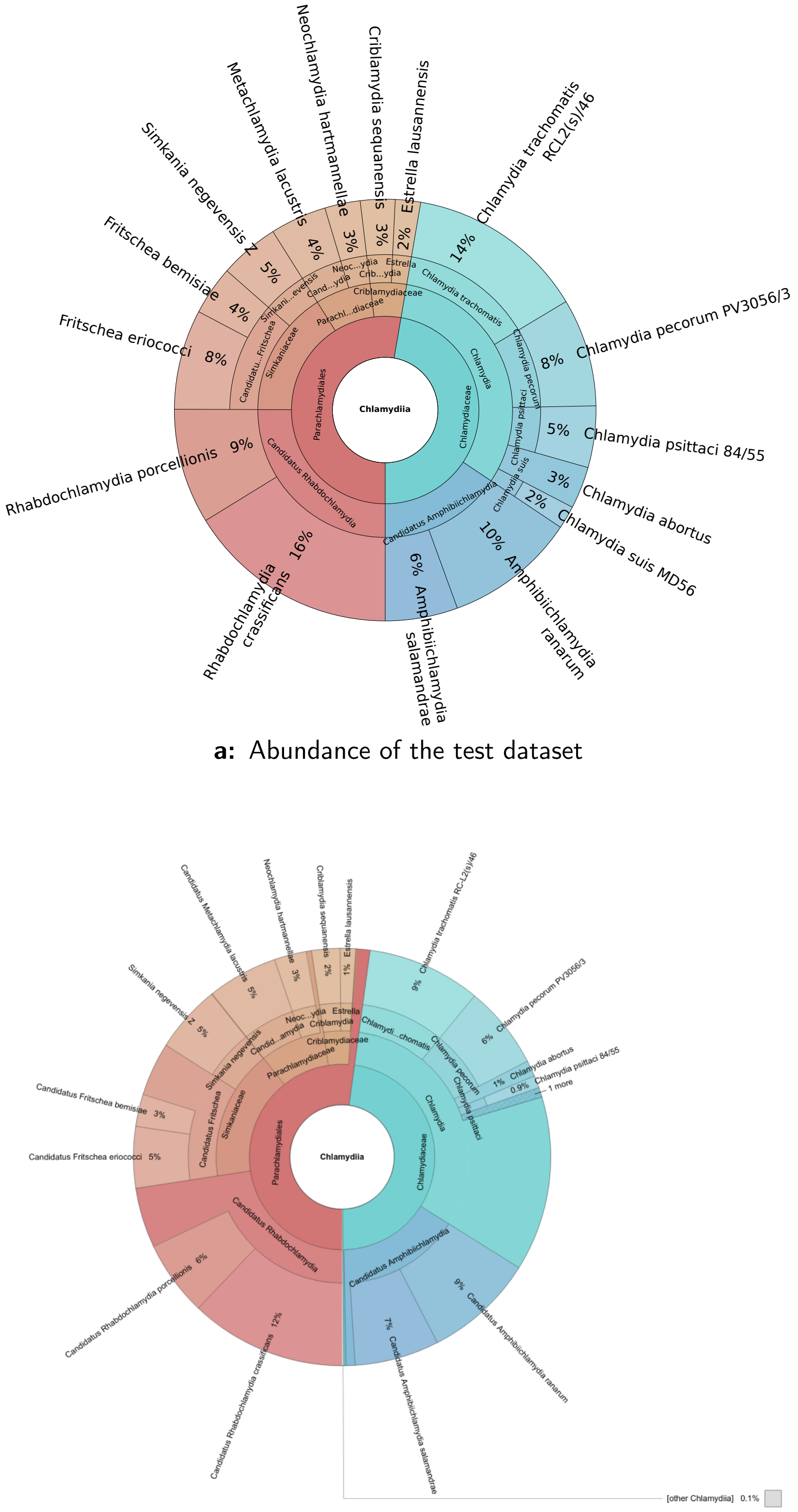
Figure 3.2: Taxonomic and phylogenetic distance based metrics. *(a)* TND is the number of nodes between the real sequence taxon and the assignment by the taxonomic profiler in the taxonomic tree. *(b)* TNDSR is the number of nodes belonging to standard ranks between the real sequence taxon and the assignment by the taxonomic profiler in the taxonomic tree.*(c)* PND is the number of nodes between where the real sequence belongs and where in the phylogentic tree it has been classified. *(d)* Similarly, PBD is the accumulated branch length between the origin and the labelled result. Taxonomic and phylogenetic trees often do not agree on the number of nodes and in the exact clustering of the sequences. Blue dots are species in the trees. The red dot respresent the origin of sequenced result and the arrows are pointing towards where it could have been classified. Green text next to the arrows, are the corresponding distances.

3.3 Characterisation of the source of taxonomic classification errors at small scale

The small mock community, referred to as the *Chlamydia* set (defined in chapter 2 section 2.3.1), contains 16 simulated *Chlamydia* species (16S rRNA sequences). This set was analysed with Kraken2 (taxonomic classifier) and the reference database contains the same organisms employed for the simulation.

It is expected that Kraken correctly identifies most of the simulated data, and that sequences with a higher number of errors are classified at higher taxonomic ranks in the same lineage. However, that is not always the case according to the results.

Kraken is not able to fully match the sequences at their origin with this small



a: Abundance of the test dataset

b: Result abundance

Figure 3.3: Real and result abundance profile. *(a)* is the composition of the test simulated reads from the Chlamydia clan. *(b)* is the abundance profile from analysing the data with kraken. The profiles look relatively similar. The vast majority of sequences are correctly classified. However, some sequences are labelled at higher taxonomic ranks, indicating that the profiler, even in this small scale test (16 species in the reference database and the exact same 16 species are used to create a simple mock community) is not capable to identify 100% of the short reads at the species level.

data set. None of the reads have been labelled as unidentified. Figure 3.3 shows the abundance profile of the original dataset and the resulting abundance from the analysis. At the rank phylum the proportion is almost identical, but differences are increasingly higher at lower taxonomic ranks.

Most of the sequences, 74,80% (1496 out of 2000) are correctly labelled to the original taxon. All the *Estrella lausannensis* reads are correctly identified at the sequence level. 24,55% (491) of reads are classified 1 to 3 nodes away in the taxonomic tree, and as little as 13 (0.65%) are assigned to 4 or more nodes away from their origin. A few reads (9), all belonging to *Chlamydia psittaci 84/55*, are assigned to the order of Criblamydiaceae, which is not part of the same lineage. In the appendix, figure B.3, is a visual representation of how far in the taxonomic tree are taxonomically identified for reads not assigned to their real orgin.

Accuracy metrics are calculated at the different rank levels (species, genus, family, order). Taxon identification is more accurate at higher taxonomic ranks, as previously described [Sczyrba et al., 2017], shown in figure 3.4.

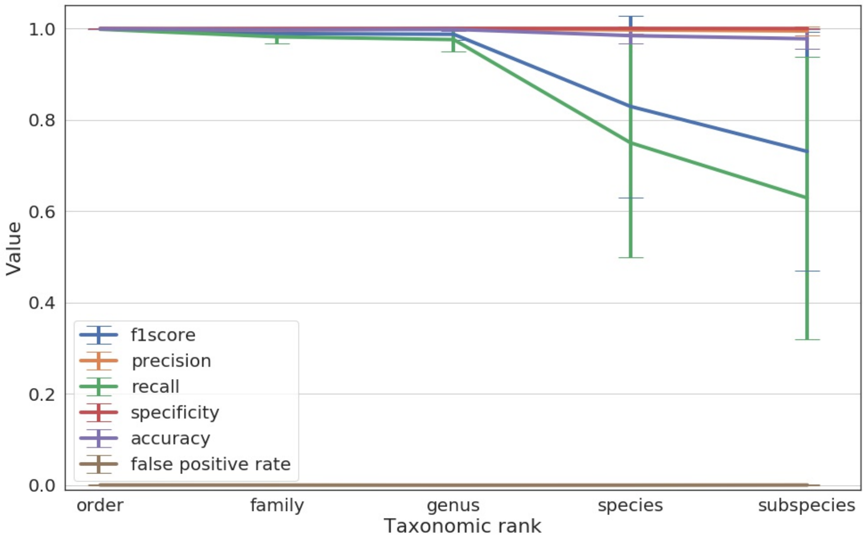


Figure 3.4: Test data metrics by rank. Metrics (f1score, precision, recall, specificity, accuracy and false positive rate) are averaged by rank. The error bars are the standard deviation. Higher taxonomic ranks (order, family, genus) present the best results for each metric. The recall and f1score, drop significantly at the species and subspecies level, indicating that the classification is less accurate at his lower taxonomic ranks..

Figure 3.5 shows the metrics at the sequence level clustered by the taxonomic tree. Results at the sequence level suggest that clans with more representatives, e.g. genus *Chlamydia* with 5 sequences, have lower levels of recall and lower f1 scores as well. Approximately a third of the short reads belonging to this clade were classified at the genus level (*Chlamydia*, 271 out of 730).

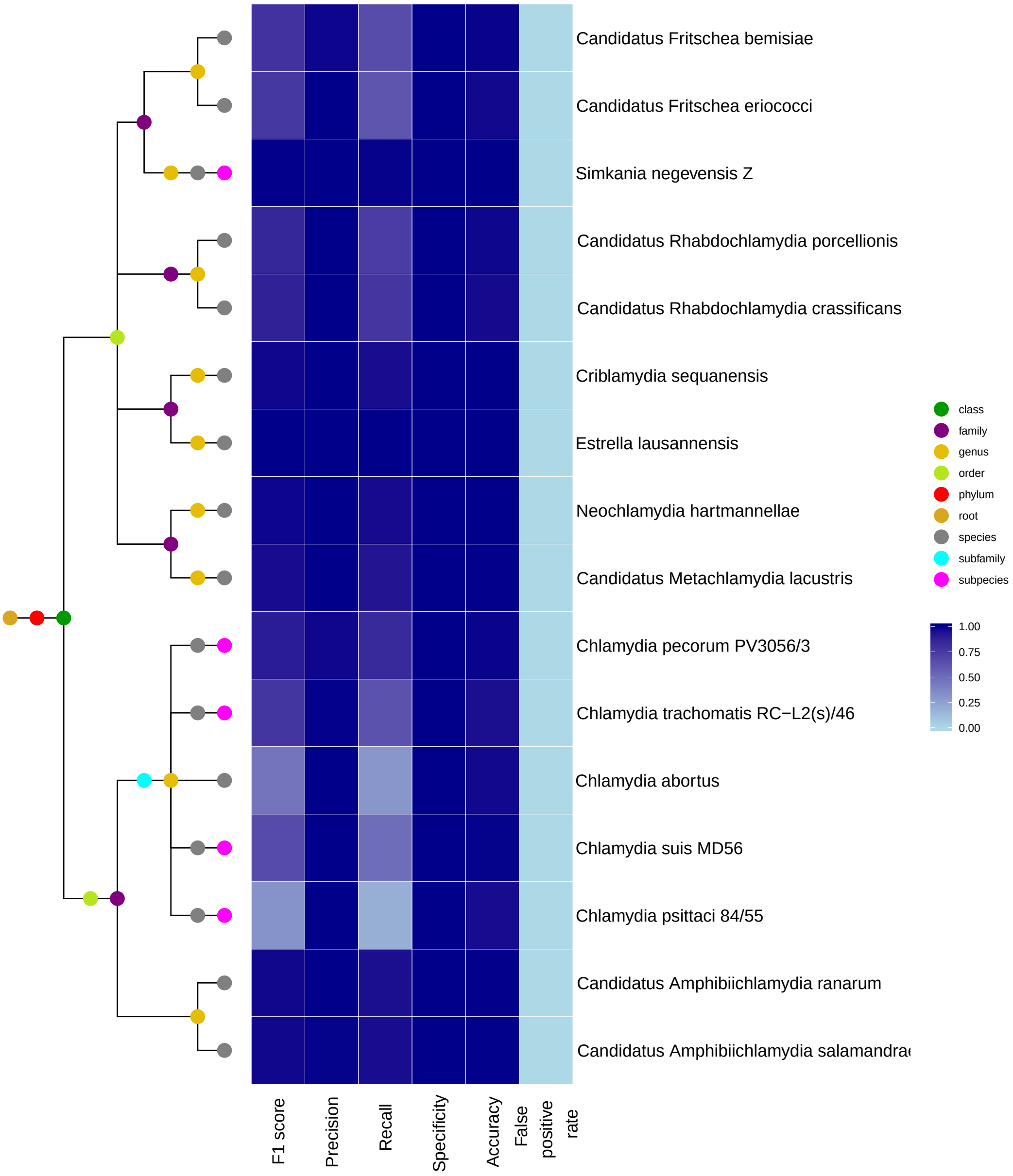


Figure 3.5: Test dataset metrics at sequence level. On the left, is the taxonomic tree of the 16 selected species, where nodes are coloured according to the taxonomic rank. In the middle, is a heatmap of the metrics aggregated by species/subspecies accordingly. On the right are the names of the selected species. The genus Chlamydia, which contain 5 species/subspecies, contains the lowest values of recall and F1 score, which also correspond to the lineage with more representatives in the database.

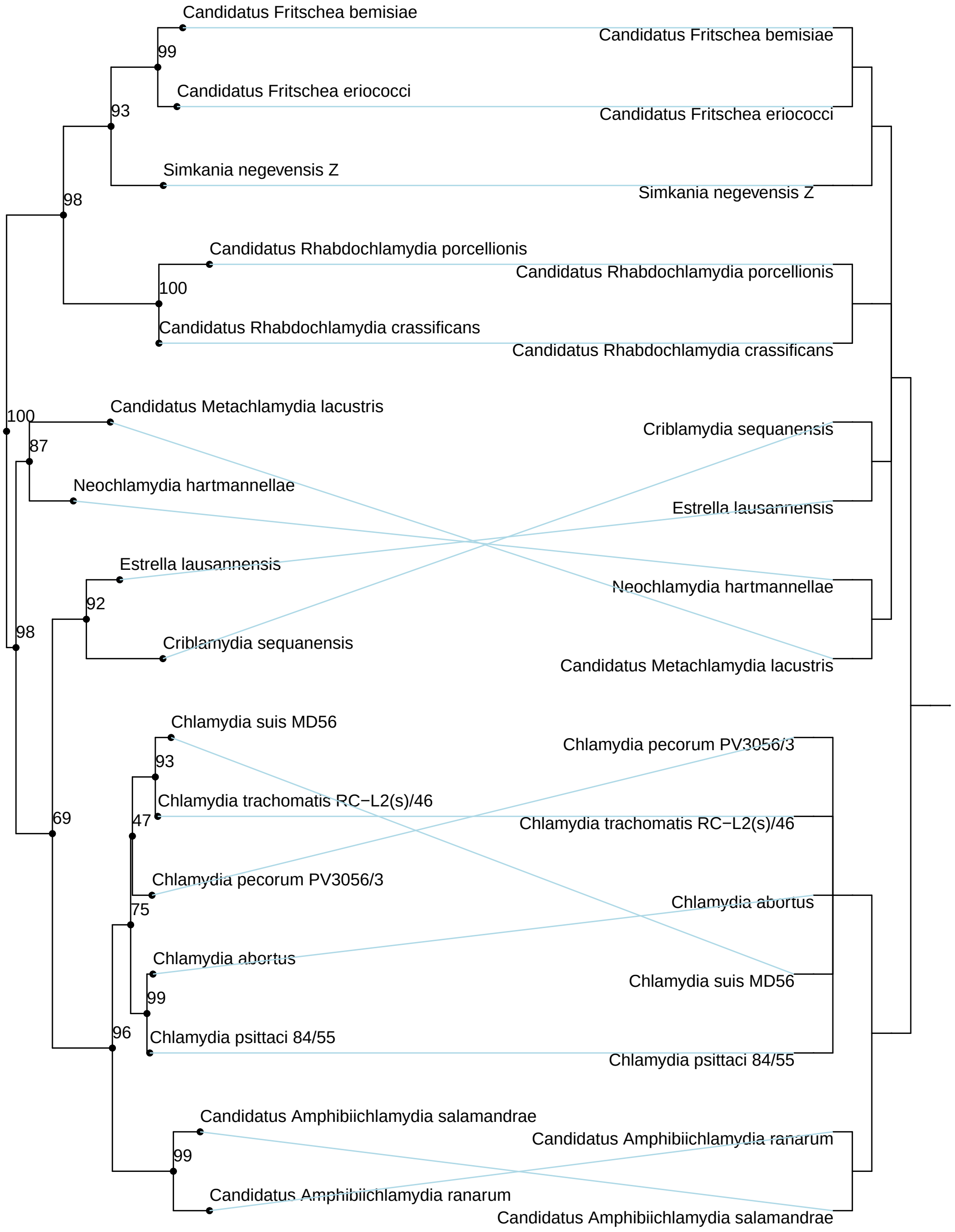

Candidatus Fritschea bemisiae
99
Candidatus Fritschea eriococci
93
Simkania negevensis Z
98
Candidatus Rhabdochlamydia porcellionis
100
Candidatus Rhabdochlamydia crassificans
100
Candidatus Metachlamydia lacustris
87
Neochlamydia hartmannellae
Estrella lausannensis
92
Criblamydia sequanensis
98
Chlamydia suis MD56
93
Chlamydia trachomatis RC−L2(s)/46
47
69
Chlamydia pecorum PV3056/3
75
Chlamydia abortus
99
96
Chlamydia psittaci 84/55
Candidatus Amphibiichlamydia salamandrae
99
Candidatus Amphibiichlamydia ranarum
Candidatus Fritschea bemisiae
Candidatus Fritschea eriococci
Simkania negevensis Z
Candidatus Rhabdochlamydia porcellionis
Candidatus Rhabdochlamydia crassificans
Criblamydia sequanensis
Estrella lausannensis
Neochlamydia hartmannellae
Candidatus Metachlamydia lacustris
Chlamydia pecorum PV3056/3
Chlamydia trachomatis RC−L2(s)/46
Chlamydia abortus
Chlamydia suis MD56
Chlamydia psittaci 84/55
Candidatus Amphibiichlamydia ranarum
Candidatus Amphibiichlamydia salamandrae

Figure 3.6: Phylogenetic and taxonomic tree of 16 species belonging to the Chlamydiae phylum. To the left is the maximum likelihood phylogenetic tree of the 16 species present in the test dataset and database (loglikelihood -2502.929). Each node has its corresponding bootstrap value. The length of the branch corresponds to the phylogentic distance. To the right is the taxonomic tree, the length of the branches are only representing the structure. Light blue lines join the corresponding tips. The 5 sequences belonging to the genus Chlamydia cluster much closer together than any other clan. The taxonomic order Parachlamydiales (top 9) is more diverse than the order Chlamydiales (bottom 7) according to the phylogeny. Some sequences in the phylogenetic tree group distinctively from their taxonomy (middle 4).

In fact, the sequences belonging to the genus *Chlamydia* are phylogenetically much closer than any other clan in the tree, as shown in figure 3.6. This suggests that in such cases, Kraken is not able to assign reads confidently at lower taxonomic levels. At the same time, this genus is phylogenetically much more distant from the rest. Some wrongly labelled reads of this clan clustered to evolutionary close taxa (*Candidatus* Amphibiichlamydia salamandrae and *Criblamydiaceae*, which contains the species *Criblamydia sequanensis* and *Estrella lausannensis*).

While the taxonomic order Chlamydiales (the bottom 7 species in figure 3.6) cluster together in the phylogenetic tree, the order Parachlamydiales is much more diverse (top 9 species). This clade shows some disagreements between the trees. This fact highlights the importance of the lineage for lowest common ancestor methods, such as Kraken, for taxonomic identification. Sequences may be labelled differently depending on the clustering.

Mutations, insertions, deletions and sequencing errors can have a great impact on metataxonomics accuracy. There is only one sequence without errors, which belongs to *Chlamydia suis MD56*, and it is classified correctly at the species level. Thirteen others do not have mutations but have between 1 and 5 insertions and or deletions and are correctly classified at their original taxon. About three quarters of the reads without any insertion or deletion are accurately labelled (458 out of 602) and 9 are labelled 3 ranks or further apart.

Table 3.1 contains a summary of the number of sequences assigned to each taxon and the mean number of events (mutations, insertions and deletions) for each group. Although the majority of correctly identified reads have a lower average of events per read, reads classified further have a slightly higher average number of events.

Table 3.1: Chlamydia dataset taxonomic classification summary.

Origin	**Assignation**	**Occurrences**	**Assigned rank**	**TND**	**events mean***
Candidatus Amphibiichlamydia ranarum	*Candidatus* Amphibiichlamydia	9	genus	1	5.22 ±1.39
	Candidatus Amphibiichlamydia ranarum	171	species	0	5.00 ±1.66
Candidatus Amphibiichlamydia salamandrae	*Candidatus* Amphibiichlamydia	5	genus	1	6.00 ±1.22
	Candidatus Amphibiichlamydia salamandrae	130	species	0	5.12 ±1.54
	Chlamydiaceae	1	family	2	5.00 ±0.00
Candidatus Fritschea bemisiae	*Candidatus* Fritschea	25	genus	1	5.24 ±1.48
	Candidatus Fritschea bemisiae	49	species	0	4.98 ±1.56

Origin	Assignation	Occurrences	Assigned rank	TND	events mean*
Candidatus Fritschea eriococci	*Candidatus* Fritschea	58	genus	1	5.38 ±1.48
	Candidatus Fritschea eriococci	95	species	0	5.47 ±1.58
	Parachlamydiales	1	order	3	4.00 ±0.00
Candidatus Metachlamydia lacustris	*Candidatus Metachlamydia lacustris*	109	species	0	5.40 ±1.64
	Parachlamydiaceae	7	family	2	5.86 ±1.21
	Parachlamydiales	2	order	3	6.50 ±0.71
Candidatus Rhabdochlamydia crassificans	*Candidatus* Fritschea bemisiae	1	species	7	6.00 ±0.00
	Candidatus Rhabdochlamydia	59	*genus*	1	5.13 ±1.55
	Candidatus Rhabdochlamydia crassificans	242	species	0	4.88 ±1.58
	Chlamydiia	2	class	4	6.50 ±0.71
	Parachlamydiales	8	order	3	5.13 ±1.55
Candidatus Rhabdochlamydia porcellionis	*Candidatus* Rhabdochlamydia	34	genus	1	5.44 ±1.74
	Candidatus Rhabdochlamydia porcellionis	117	species	0	5.20 ±1.60
	Parachlamydiales	7	order	3	5.57 ±0.98
Chlamydia abortus	*Chlamydia*	55	genus	1	5.57 ±0.98
	Chlamydia abortus	25	species	0	4.92 ±1.58
	Chlamydiaceae	1	family	3	5.00 ±0.00
	Chlamydia trachomatis RC-L2(s)/46	1	subspecies	4	4.00 ±0.00
Chlamydia pecorum PV3056/3	*Chlamydia*	24	genus	2	5.04 ±1.78
	Chlamydiaceae	1	family	4	7.00 ±0.00
	Chlamydia pecorum PV3056/3	126	subspecies	0	4.77 ±1.66
	Chlamydiia	1	class	6	4.00 ±0.00
Chlamydia psittaci 84/55	*Chlamydia*	79	genus	2	4.90 ±1.64
	Chlamydia pecorum PV3056/3	1	subspecies	5	7.00 ±0.00
	Chlamydia psittaci 84/55	19	subspecies	0	5.16 ±1.21
	Criblamydiaceae	1	family	9	6.00 ±0.00
Chlamydia suis MD56	*Chlamydia*	15	genus	2	5.20 ±1.61
	Chlamydia suis MD56	15	subspecies	0	4.53 ±2.45
Chlamydia trachomatis RC-L2(s)/46	*Candidatus* Amphibiichlamydia salamandrae	1	species	7	7.00 ±0.00
	Chlamydia	98	genus	2	6.52 ±1.75
	Chlamydia pecorum PV3056/3	2	subspecies	5	5.00 ±0.00

Origin	Assignation	Occurrences	Assigned rank	TND	events mean*
	Chlamydia trachomatis RC-L2(s)/46	175	subspecies	0	5.90 ±1.58
Criblamydia sequanensis	*Criblamydia sequanensis*	44	species	0	5.45 ±1.61
	Parachlamydiales	2	order	3	6.50 ±2.12
Estrella lausannensis	*Estrella lausannensis*	24	species	0	5.13 ±1.70
Neochlamydia hartmannellae	*Neochlamydia hartmannellae*	50	species	0	4.80 ±1.67
	Parachlamydiaceae	1	family	2	6.00 ±0.00
	Simkaniaceae	1	family	5	8.00 ±0.00
Simkania negevensis Z	*Parachlamydiales*	1	order	4	6.00 ±0.00
	Simkania negevensis Z	105	subspecies	0	5.22 ±1.56

**Mean number of events (mutations, insertions and deletions) introduced in each sequence and the corresponding standard deviation.* This table contains a count of sequences classified at the different taxa. Most of the sequences are classified to the original taxon or same clade (in green). There are very few sequences labelled to a far away node (taxonomic distance >=4). The lowest average of events (in red) does not always correspond to the perfect identification.

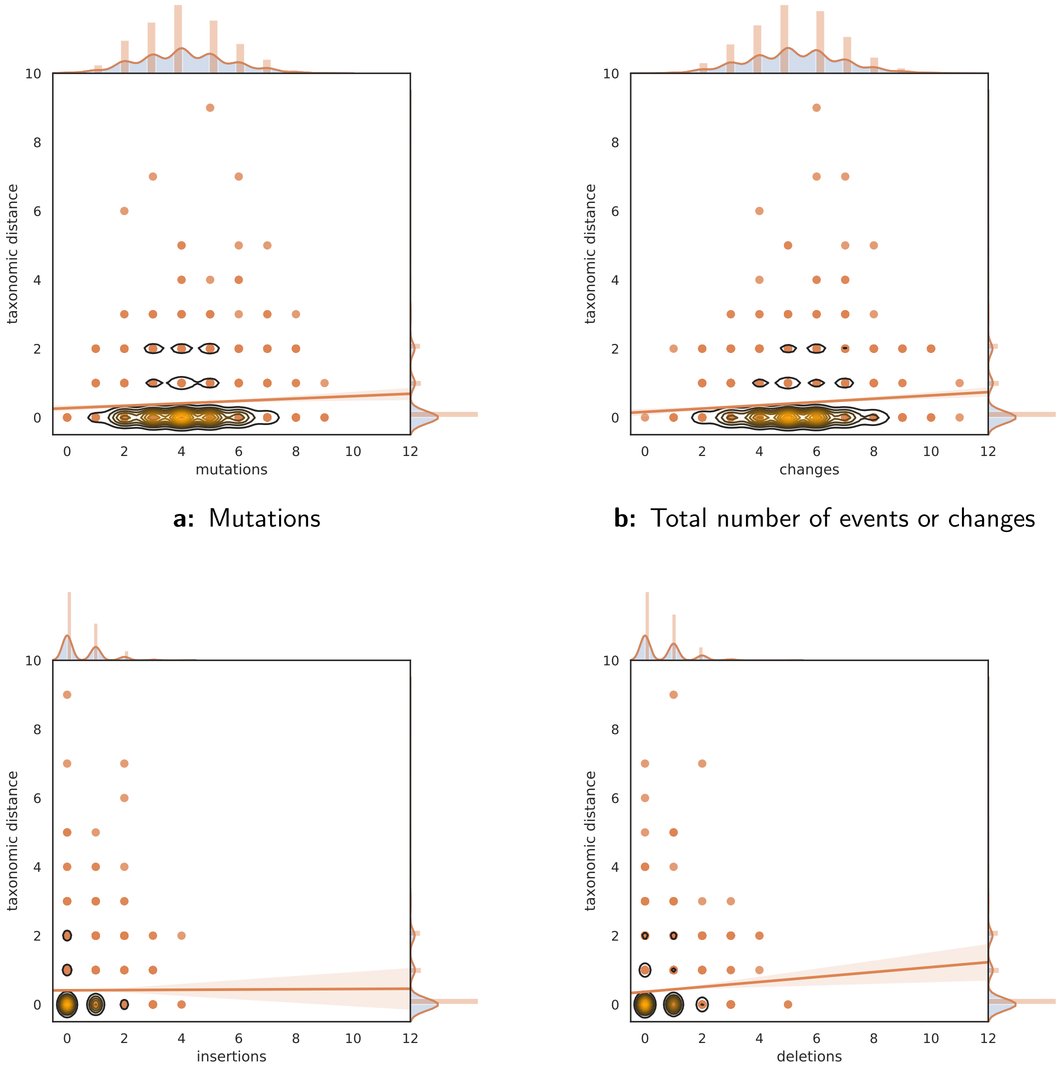
In fact, looking closer to the correlation between events in sequences and taxonomic distance, we can see a weak increasing effect as shown in figure 3.7.

3.4 Distance metrics comparison

The *Test data* (described in chapter 2 section 2.3.4) is designed to compare the newly proposed distance based metrics.

The taxonomic and phylogenetic distances (TND, TNDSR, PBD and PND) slightly increase with the number of mutations. Figure B.4 (see appendix) shows their distribution for each type of experiment. PBD and PND mean of the experiments without noise is higher than the experiments with 1% noise.

The number of accumulated mutations in a sequence, as demonstrated in figure 3.8, gently increase all the taxonomic distance metrics (TND, TNDSR, PBD and PND). This might be indicative of the robustness of Kraken against mutations.



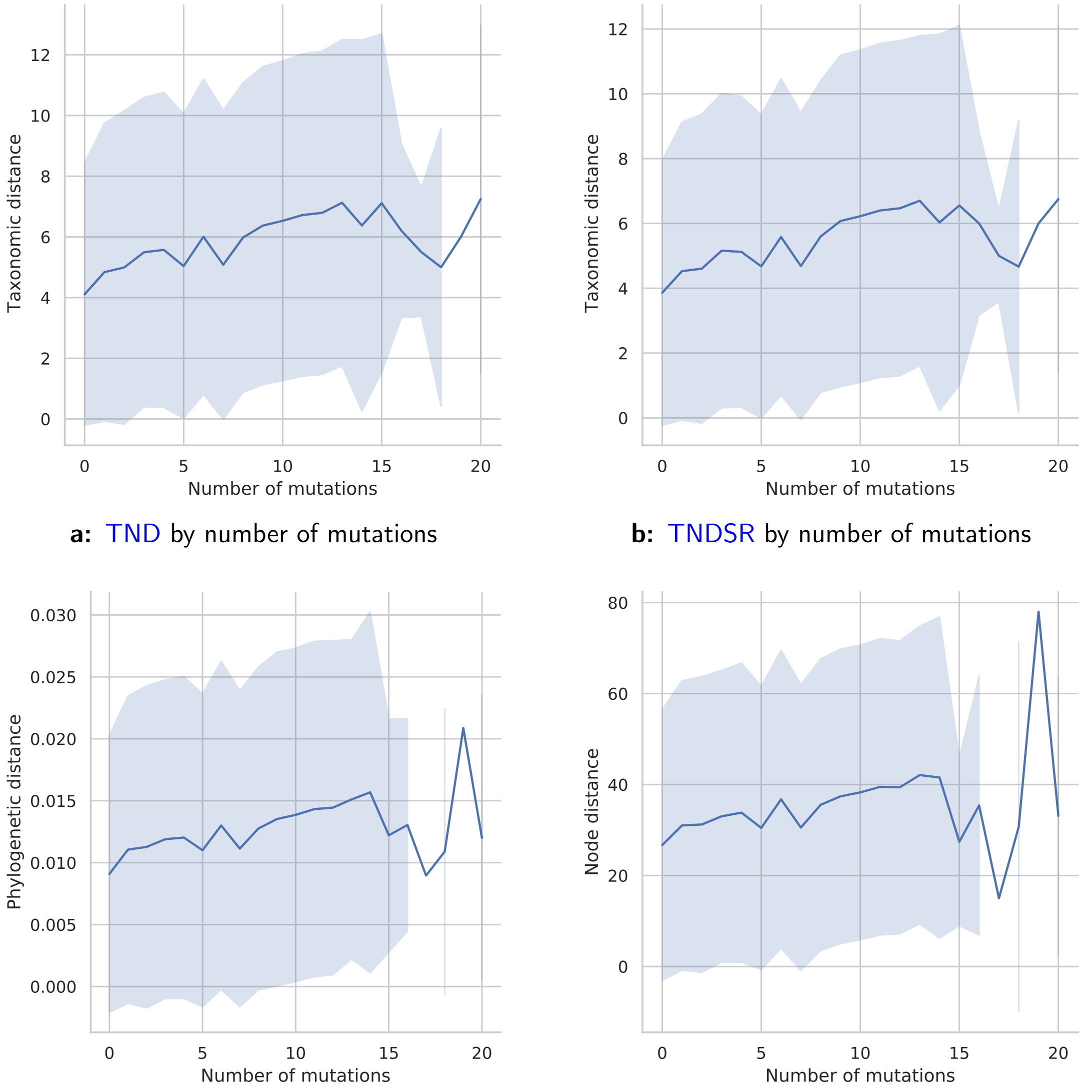
a: Mutations

b: Total number of events or changes

c: Insertions

d: Deletions

Figure 3.7: Distribution and correlation of the number of events with the taxonomic distance. Plots a-d depicts the distribution of events (mutations, insertions, deletions) and the total of them by sequence -changes- by the TND where sequences are classified. The taxonomic distance is the number of taxonomic nodes away from where sequences have been classified. In general the more changes a sequence presents (mutations, insertions and deletions) the higher the TND. However, this seems to be quite a small effect. The number of insertions do not seem to have much impact on classification whereas deletions seem to have a bigger effect. Most sequences that have deletions also have a high number of mutations, therefore it is hard to discriminate the effect.



a: TND by number of mutations

b: TNDSR by number of mutations

c: PBD by number of mutations

d: PND by number of mutations

Figure 3.8: Distance metrics by number of mutations. Average values of the newly proposed metrics. Light blue area represents the standard deviation. The number of mutations present in a short read slightly increase the taxonomic, phylogenetic and node distances. A few short reads have a higher number of mutations, which increases the confidence intervals. However, the increment of distance is inferior to the increment of number of mutations.

3.5 Enabling metagenomics taxonomic classification benchmarking for individual needs

The goal of this section is to establish a system that allows the evaluation of the metataxonomic assignment and parameter optimisation for any method of choice. Each type of project is unique: only certain clades are present in a microbiome, and the quality and length of the sequencing is platform and protocol dependant. It is crucial to ensure the bioinformatics pipeline is fit for purpose.

A new method is developed to generate *in silico* microbial communities, named My Goldstandard Community (MGC), that resembles real sequencing metagenomics data.

3.5.1 My Goldstandard Community

MGC consist of negative and positive controls. Sequencing errors and mutations are introduced as described below. This method was specifically designed for 16S rRNA sequencing data.

Negative controls

Negative controls are sequences expected not to be identified by the taxonomic profiler. In this case, amplicon unclassified-like data was created, as figure 3.9 illustrates.

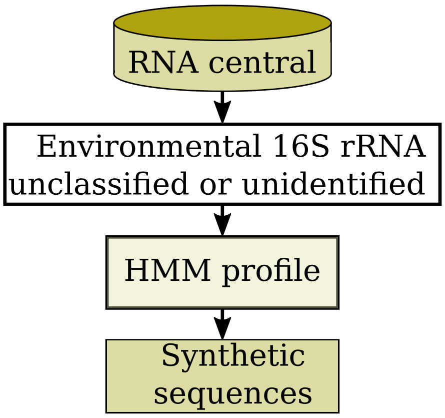


Figure 3.9: Negative control generation. Diagram representing the process for creating the synthetic sequences. First, sequences of the gene 16S rRNA which are labelled as "unclassified" or "unidentified" in the RNA central database are downloaded. The sequences are aligned, after duplicate removal, and an HMM profile is generated. This profile is then used to create synthetic sequences that will be used as *negative controls* for the experiments.

3021 non identified 16S rRNA genetic sequences were selected from environmental samples (on 31/01/2019) from the RNA central database [Sweeney et al., 2019]. The search terms used are shown in the box below:

```
tax_string:"unclassfied" OR tax_string:"unidentified"
AND tax_string:"environmental"
AND 16S* OR small* subunit* AND NOT 18S* NOT large*
AND length:[1300 TO 1900]
AND qc_warning_found:"False" AND rna_type:"rRNA"
```

Duplicates were removed with cd-hit-est [Fu et al., 2012] (global sequence identity flag set to 1) (version 4.6) leaving a total of 2946 entries. They were then aligned with nhmmer [Wheeler and Eddy, 2013] (3.1b1 (May 2013)) with the Bacterial SSU RNA (RF00177) profile from RFam database [Kalvari et al., 2018].

A Hidden Markov Model (HMM) profile was created from the resulting alignment. Finally, a set of 100 synthetic sequences were generated with the function hmmemit from HMMER [Wheeler and Eddy, 2013].

Additionally, these sequences were blasted (blastn) against the reference database SILVA Parc v138.1. Sequences with a match of minimum length of 35 (Kraken2 uses a k=35), percentage identity above 90% and e-value inferior of 0.01 were removed. This left a total of 96 sequences.

Positive controls

Positive controls are sequences expected to be identified by the taxonomic profiler. Their generation is based on previously known and annotated sequences.

Templates for positives controls are selected from a relevant database. For example, Silva database for 16S rRNA data.

Mock community generation

Generating a simulated environmental community requires the NCBI taxonomy database (or another taxonomy database in NCBI's format), metagenomic sequencing data, the corresponding microbiome abundance profile generated by a taxonomic profiler, a mapping file between the positive controls and the taxonomic database, and the file containing the negative controls.

The templates for positive controls are assigned from the existing name or corresponding taxID from the selected database for templates. Those names in the abundance profile which are in non-leaf nodes are randomly assigned a sequence template from a species or subspecies from their corresponding clan. Five percent of unclassified sequences are introduced from the synthetic negative controls.

Sequences were simulated with Grinder with the abundance values arising from analyses of the original sequence data, and with the templates previously selected. The number of sequences to simulate was set to the same number as the fastq file from sequencing, the length was set to the longest read from the fastq file. Three replicates were created.

Next, each sequence was randomly assigned a quality score from the original sequence data and the length was adjusted accordingly. This newly created data did not contain mutations and is copied across to generate the rest of experimental data. From the previously created data another 3 sets were generated by introducing random mutations at 1%, 2% and 3% independently. From each of the previous sets, new mock samples were obtained by introducing sequencing errors according to the probability of a base being erroneous given the corresponding phred score. In total, 8 sets of data were generated, with 3 replicas each, to make MGC as figure 3.10 shows.

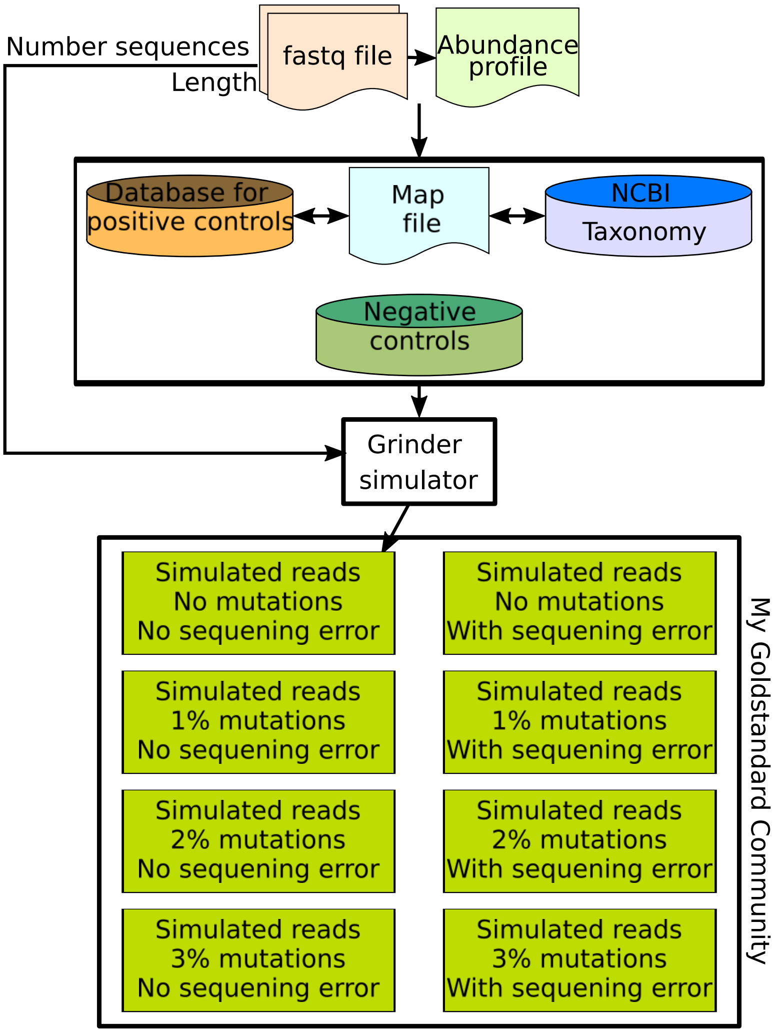


Figure 3.10: My Goldstandard Community diagram. The simulation of realistic datasets. For the process, it is required the sequencing data (fastq file(s)) and the corresponding taxonomic abundance profile. From the abundance profile, reference sequences are selected after mapping to NCBI taxonomy. This constitutes the positive controls of the experiments. The quality is imputed from the original fastq file(s) (although mutations are not yet introduced), and the sequence length adjusted accordingly. MGC is created by using the abundance of the abundance profile and 5% of negative controls. Grinder simulator generate the first profile (of the total of 8), which does not have any background mutations nor sequencing errors. From this one, the rest of MGC files are generated, consisting of the introduction of 1%, 2% and 3% of random mutations (but not sequencing errors). Finally, from the previous 4 files, sequencing errors are introduced according to quality of each base.

3.5.2 Results

Simulation of communities

To mimic microbial communities and at the same time cover a wide taxonomic spectrum, 9 samples were chosen from a variety of environments (described in chapter 2 section 2.3.3 on page 26): from fish to plant and from gut to water, including archaeal. Only the Illumina sequencing platform was taken into consideration, in both single or paired-end mode. A range of sequencing depths and average read lengths were included, as both may influence classification. The abundance profile was downloaded from the analysis results available in MGnify [Mitchell et al., 2020] (previously known as EBI Metagenomics).

Table 3.2: Summary of the number of sequences simulated per sample. Number of sequences of each type of control included per sample and replica.

Sample	Replica	Number of sequences	Negative controls	Positive controls	% positive controls	% negative controls
faeces	1	32,219	1,470	30,749	95.44	4.56
	2	32,219	1,513	30,706	95.30	4.69
	3	32,219	1,539	30,680	95.22	4.78
fish	1	65,850	3,362	62,488	94.89	5.11
	2	65,850	3,466	62,384	94.74	5.26
	3	65,850	3,436	62,414	94.78	5.22
gut	1	112,432	5,059	107,373	95.50	4.50
	2	112,432	5,178	107,254	95.39	4.61
	3	112,432	4,973	107,459	95.58	4.42
human	1	25,089	1,262	23,827	94.97	5.03
	2	25,089	1,341	23,748	94.66	5.34
	3	25,089	1,282	23,807	94.89	5.11
ice	1	104,163	4,571	99,592	95.61	4.39
	2	104,163	4,645	99,518	95.54	4.46
	3	104,163	4,716	99,447	95.47	4.53
plant	1	30,977	1,607	29,370	94.81	5.19
	2	30,977	1,567	29,410	94.94	5.06
	3	30,977	1,561	29,416	94.96	5.04
reactor	1	32,248	1,489	30,759	95.38	4.62
	2	32,248	1,522	30,726	95.28	4.72
	3	32,248	1,532	30,716	95.25	4.75
sludge	1	45,887	2,140	43,747	95.34	4.66
	2	45,887	2,129	43,758	95.36	4.64
	3	45,887	2,219	43,668	95.16	4.84
soil	1	24,937	1,195	23,742	95.21	4.79
	2	24,937	1,192	23,745	95.22	4.78
	3	24,937	1,167	23,770	95.32	4.68

The database Silva Parc version 138.1 was the input for the positive control sequences. The previously described method was applied to simulate realistic mock communities. Table 3.2 contains a brief description of the content of the newly simulated environmental samples.

Metataxonomic identification

The vast majority of sequences were identified as belonging to the taxonomic tree. In fact, only about 4.68% of the simulated data was not classified by Kraken2 as shown in table 3.3, some are from the positive controls and the majority from the negative.

Table 3.3: Number of classified sequences by control type. Almost all the positives controls were classified, and the majority of negative controls were unclassified.

	Classified	Unclassified
Negative	4,867	532,197
Positive	10,834,134	50

The majority of sequences were labelled to leaves of the tree of life (species or below), as shown in figure 3.11, and a considerable proportion of reads classified at the superkingdom rank.

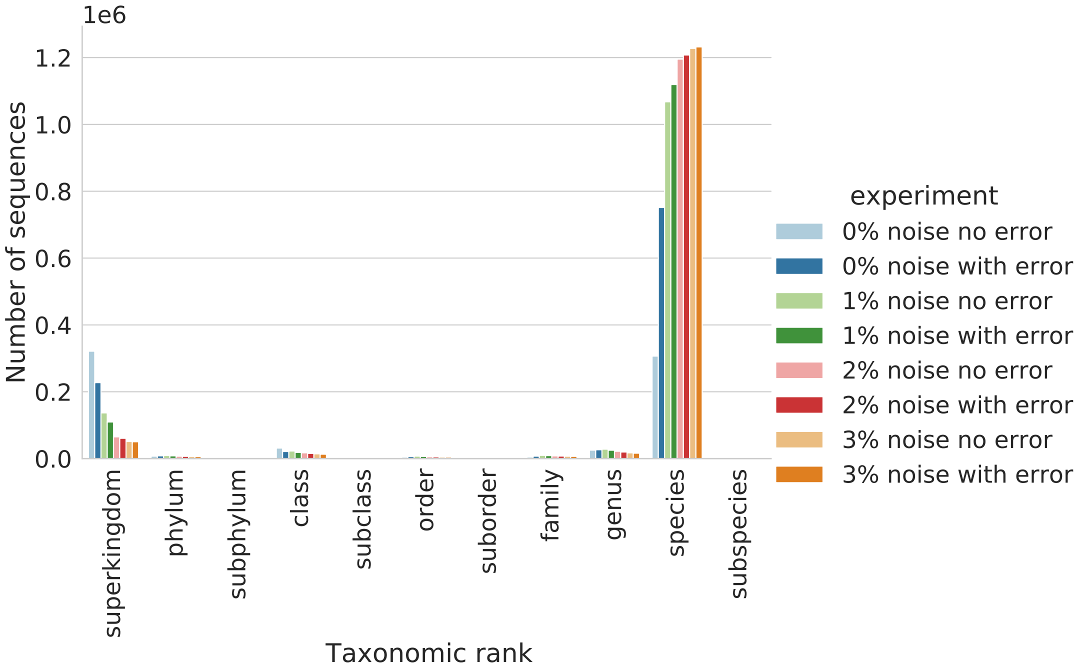


Figure 3.11: Number of sequences classified per taxonomic rank. Most of the sequences were identified at species level, interestingly, the higher proportion found in experiments with more error. A significant fraction could only be assigned to superkingdom. Contrary to previously, those experiments with less errors are more abundant here. This could be indicative of Kraken being less reliable at lower taxonomic ranks, especially with higher number of mutations compared to the reference.

Figure 3.12 shows the absolute abundance by superkingdom in each sets of samples.

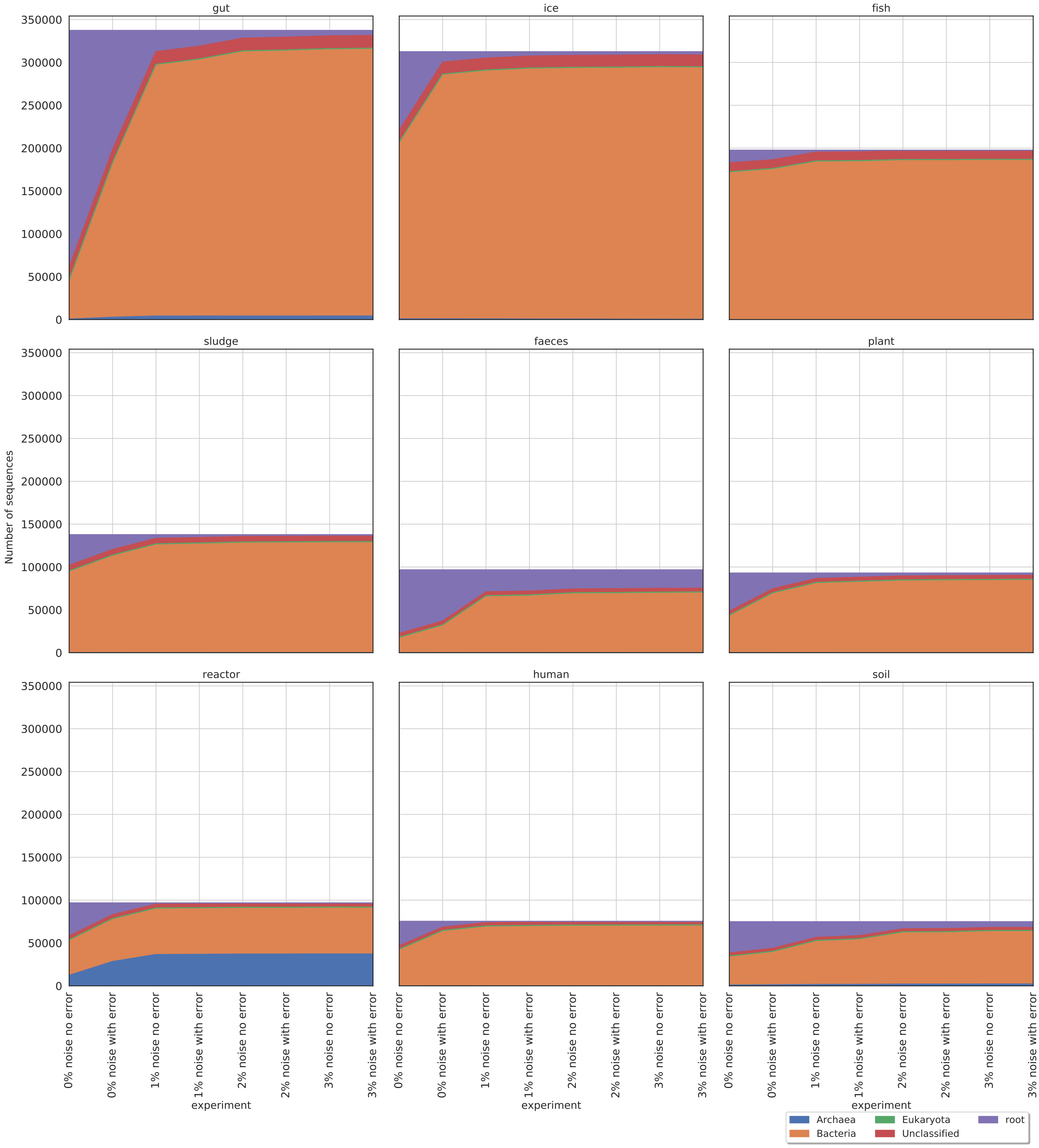


Figure 3.12: Superkingdom abundance by environment and experiment. Absolute abundance at the superkingdom level by environment (sorted from the highest to the lowest number of sequenced reads). Reads classified at "root" is when Kraken2 find kmers matching in the reference database, but cannot assign them to any branch due to too many conflicts. Those samples with more errors have more sequences identified as bacteria.

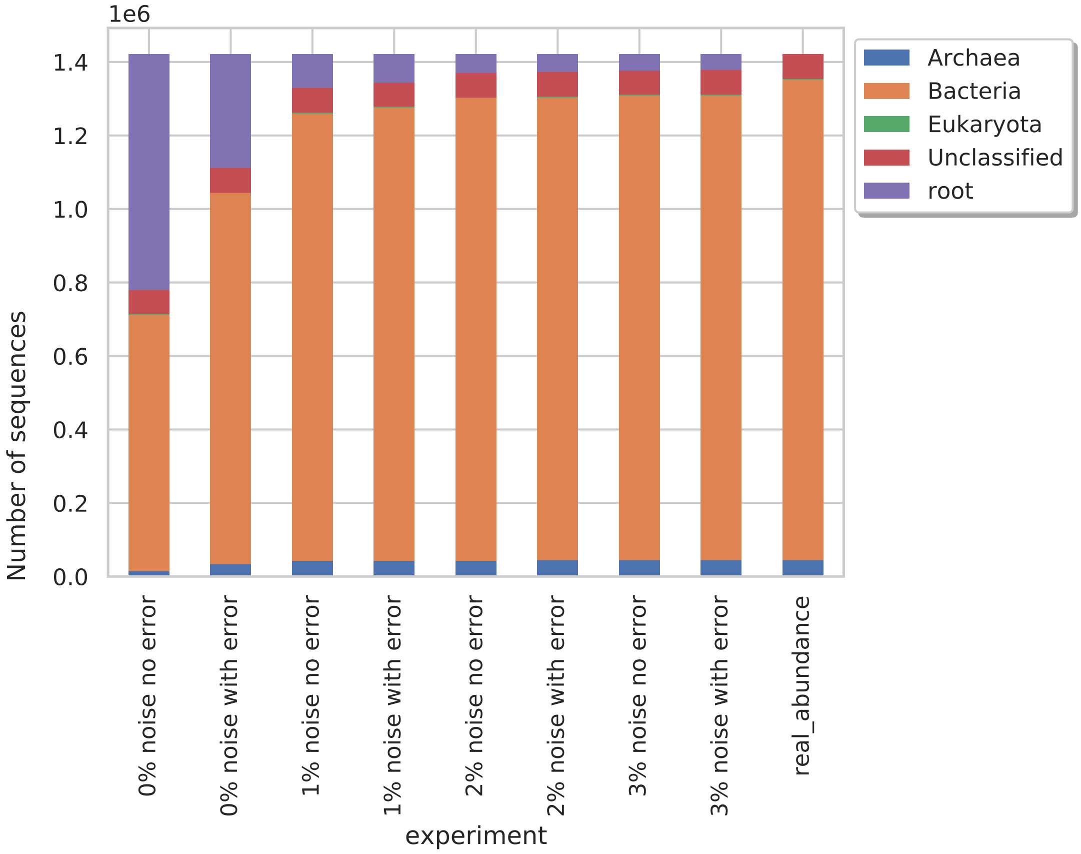


Figure 3.13: Superkingdom abundance by experiment. This figure shows the classification at the Superkingdom rank of all the simulated mock communities by experiment. There is an overall low number of unclassified reads. Reads with higher proportion of mutations have lower proportion of 'root' labelled sequences. The fraction of unclassified reads increase slightly with higher number of errors/mutations in the simulated data.

Kraken sometimes identifies sequences as belonging somewhere in the tree, and are labelled as 'root'. Interestingly, the number of sequences assigned to 'root' decreases with higher number of mutations, as observed in figure 3.13.

Kraken does perform better at higher taxonomic ranks, as observed in figure 3.14, according to the common metrics. However, the performance is highly dependent on the microbial composition present on each sample (figure B.5 in appendix).

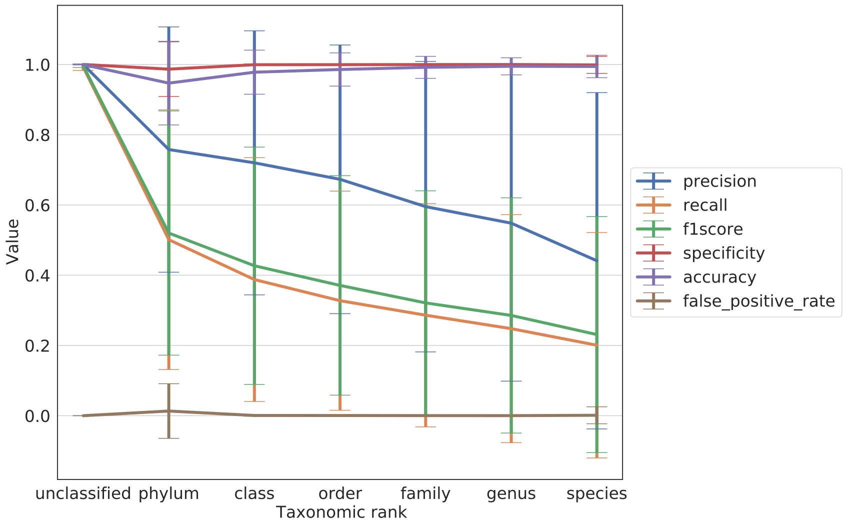


Figure 3.14: Accuracy metrics by rank. The metrics presented here show that identification at higher taxonomic ranks is more reliable, although there is a high standard deviation in all cases. The precision, recall and F1 score are the metrics with more significant drops at lower taxonomic ranks. The unclassified sequences include all the negative controls and have the best results for all the metrics.

The taxonomic distance distribution has a slight tendency to increase with experiments that contain more mutations, as shown in figure 3.15. The negative controls are sometimes identified in the taxonomic tree.

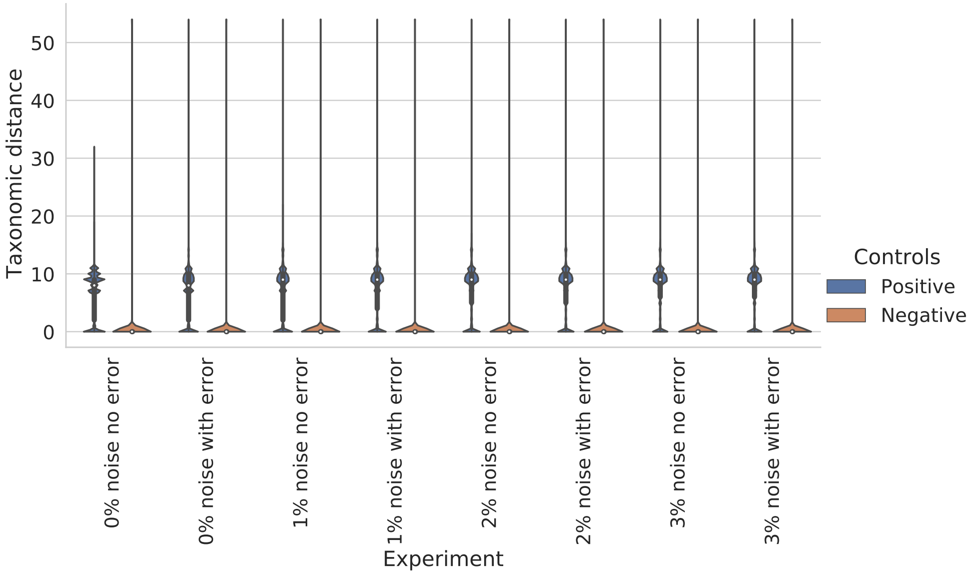


Figure 3.15: Violin plot of the Taxonomic Node Distance by type of control and experiment. This figure illustrates the TND by positive and negative controls (reads which contain or not sequences representing them in the reference database) by experiment. TND distribution is quite similar amongst the different experiments with most of the positives controls being 10 or under, and almost all negative controls at 0 with expections. Values of TND at 0 or close mean that reads have been classified where expected, whereas higher values mean they have been classified in further away nodes compared to the original species.

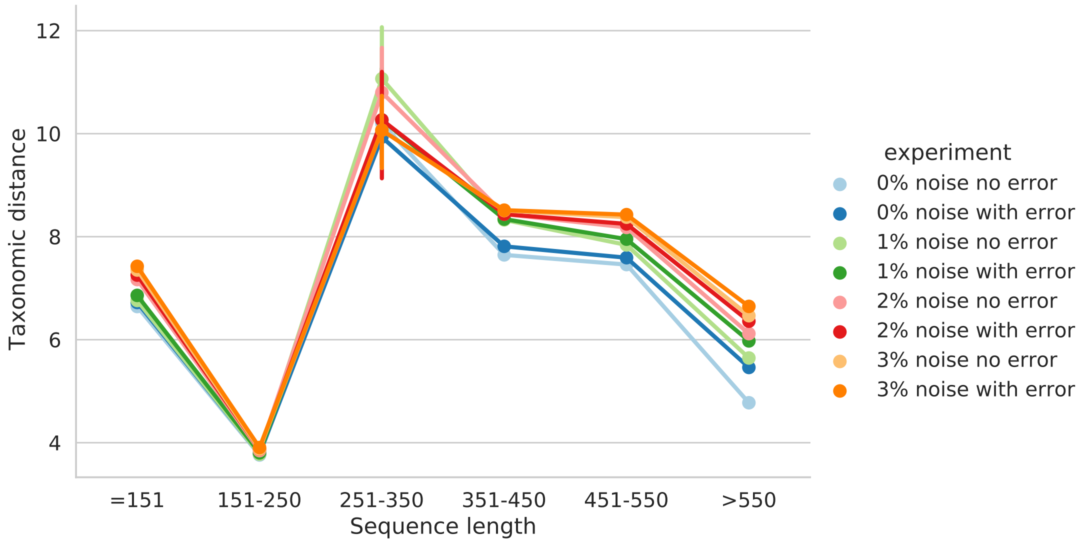


Figure 3.16: Taxonomic Node Distance by length. Short reads have grouped by length. The TND is generally lower for the shorter sequences, and also not much difference is observed by the different experiments, except for the sequences over 550 bp. However, the taxonomies represented in each group is different as the nature of MGC. The two shorter groups belong to soil and fish samples.

TND of the total length of the simulated data is higher for the middle range 251-550 base pair (bp), as figure 3.16 shows. The lowest taxonomic distance corresponds to the second-shortest group. These sequences are single end and belong to 2 of the environments. Also, TND does not differ much between experiments, except for the longest group of sequences.

Depending on the microbial community, in this case each sample, TND can vary greatly. Figure 3.17 shows that TND can vary up to 5 on average from sample to sample. Moreover, the number of errors introduced also differs significantly.

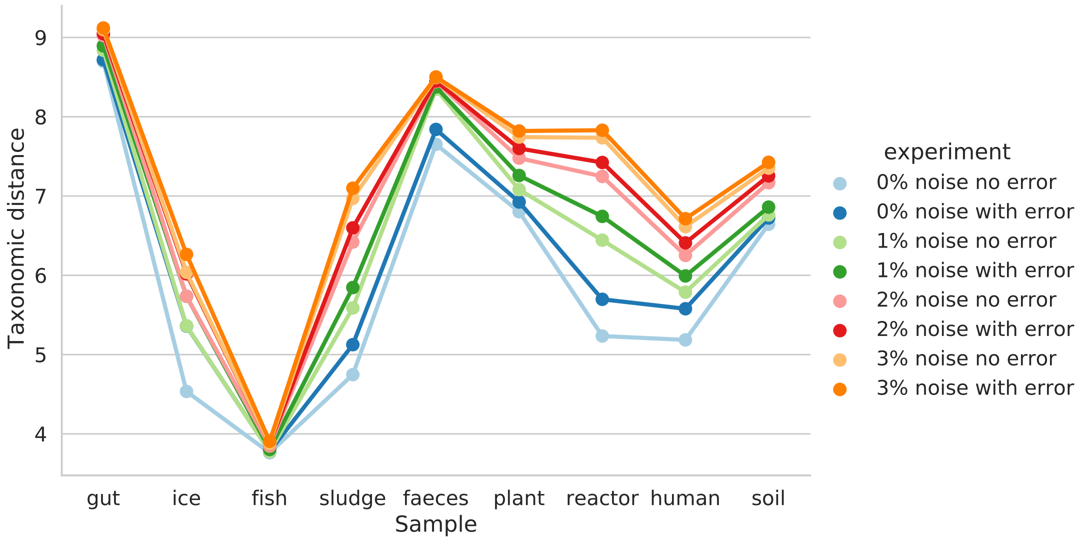


Figure 3.17: Taxonomic Node Distance by sample and experiment. Samples are in descending order by number of reads. The taxonomic distance differs vastly depending on the organisms present on the sample. It could be because of too many or too little representatives in the reference database or due to the fact that some lineages have much larger number of taxonomic nodes than others in the NCBI database, and therefore if reads belonging to very long lineages have a potential risk of presenting higher values of TND even if the accuracy values at specific taxonomic ranks is the same. Also, the effect of the number of mutations in each sample is different. This highlights the importance of checking the fitness of the chosen database for the sample(s) of study, and understand the impact that mutations will have on the chosen overall bioinformatics analysis pipeline.

3.6 Discussion

The *Chlamydia dataset* demonstrates that the classification method, Kraken, is not capable of fully identifying all the sequences with the same organisms employed as reference. Misclassification occurs at all levels, but is more common amongst closer related taxonomic nodes. For benchmarking, appropriate metrics should be used. They should also be meaningful for the purpose, and when necessary new measures should be proposed [Capella-Gutierrez et al., 2017]. For example, precision, recall, f1 score, etc. measure the level of error classification at specified taxonomic ranks, but fail to capture whether sequences are assigned to a close relative or very far away in the tree of life.

The TND measures how far away, in the taxonomic tree, sequences have been assigned. However, the number of ranks can greatly vary from lineage to lineage, and therefore it can introduce a bias. An alternative is to only count the number of standard ranks (*superkingdom, phylum, class, order, family, genus* and *species*), which are common in all branches, named here as TNDSR. However, some clades still contain multiples of some standard ranks in the NCBI taxonomy database.

Independently, Chen et al. [Chen et al., 2019] defined a similar metric based on the concept of taxonomic distance. Their metric is based on the "number of ranks in difference" divided by the "number of unique ranks in two taxa". The measure of the PBD and PND would solve all the previously mentioned issues. But currently there is no single phylogenetic tree that encapsulates all potential organisms, although the GTDB database has a great potential to become the main reference for taxonomic classification.

The *Test data* show comparable profiles (at different scale) for all these measures. However, PBD and PND are quite likely to behave differently in other contexts. For example, when GTDB is mapped with NCBI taxonomy, a number of leaves nodes in GTDB with the exact same species name are discovered. The resulting mapping also contains some nonsensical lineages, e.g. mixing up bacteria and eukaryotes.

Laboratories across the world must ensure the quality of their tests: reliability, reproducibility and consistency [Schlaberg et al., 2017]. However, these principles are not applied systematically for bioinformatic algorithms. A review of the literature reveals contradictions in what are the best methods to use. Nevertheless, benchmarking studies are representative and accurate in specific contexts [Pollock et al., 2018]. Therefore, we propose to validate the bioinformatics methods for each study to ensure the best possible taxonomic identification is achieved.

In this chapter, a novel methodology was proposed to generate *in silico* data that is similar in composition to the one detected by the chosen classifier, and with the same sequencing data profile. In this way, individuals can measure and decide on the fitness of their methods and parameters. Indeed, the data shows the wide disparity in the metrics that can be observed for different types of samples.

Up to 3% of random errors were introduced to the simulated data (MGC). This was based on the fact that many OTU tools and pipelines identify clusters at 97% of similarity [Imelfort et al., 2014, Velsko et al., 2018, Alves et al., 2016] as they consider them to belong to the same species. However, with the increasing number of microbial genomes available, it is proven that organisms belonging to the same species can present much lower or much higher average nucleotide identity (ANI) [Breitwieser et al., 2019]. For instance, ANI values between the genus *Shigella* and *E. coli* species are >97% while some of the members of the *Escherichia* are only 93% identical [Breitwieser et al., 2019, Hornung et al., 2019] at genome level.

Kraken performs better at higher taxonomic ranks, such as superkingdom or phyla. Interestingly, when random noise was introduced in the data set, more short sequences were assigned to lower taxonomic ranks rather than being unclassified.

To the best of my knowledge, I generated a novel methodology to generate synthetic 16S rRNA data to be used as negative controls. This method can be applied to other similar genes.

According to the data it can be concluded that: *(i)* Kraken is not always able to recover the original taxon; *(ii)* the number of error or evolutionary events introduced in the simulated data seem to affect very little the assignment by Kraken; *(iii)* sequences that are evolutionary quite close might generate confusion when classifying; *(iv)* assignment is more accurate at higher taxonomic ranks; and *(v)* misclassification seems more likely to happen between phylogenetically closer sequences.

Chapter 4

Discriminatory gene regions for taxonomic assignment

4.1 Motivation

Taxonomic assignment is a major challenge in metagenomics. Often, marker genes are being used, and the most common is the housekeeping 16S rRNA gene. It has nine variable and ten conserved regions. Understanding which genetic regions are more informative, has the potential to aid taxonomic classification.

The classifier Kraken is k-mer based. Genetic sequences contain fragments of size k (k-mers) which are found in varying frequencies within a sequence and across species. Rare k-mers might be more informative for taxonomic classification than the more common ones. Genetic sequences are represented by an alphabet of 4 characters, which can in simple terms be visualised as a document of text on a specific topic. If k-mers are thought of as words and genetic sequences as documents, linguistic methods can be applied to develop a novel approach for metataxonomic classification.

4.1.1 Objectives

The main goals of this chapter were to:

– apply different strategies to improve taxonomic classification based on the information of different genetic regions.
– define database specific regions which contain more useful information for taxonomic classification.
– develop a new classification approach based on linguistic methods, which penalizes highly frequent k-mers for taxonomic classification.

4.2 Genetic regions

The 16S rRNA gene has a complex structure related to its function. To be able to maintain the function, there are key regions, which can be identified when performing evolutionary studies to determine conserved and variable regions across species. The work presented in this section aimed to discover if the regions with more accumulated mutations across bacteria and eukaryotes counterpart present better taxonomic identification compared to the conserved regions, as has been previously described in the literature [Chakravorty et al., 2007].

4.2.1 Selection of reads that contain variable regions

This work was conducted in collaboration with Rob Finn and Alex Mitchell (EMBL-EBI).

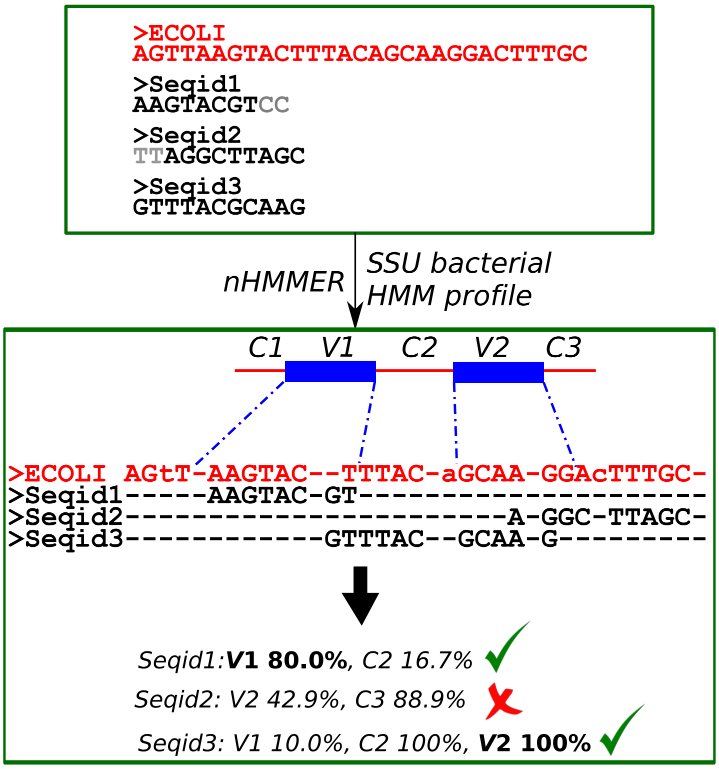

>ECOLI
AGTTAAGTACTTTACAGCAAGGACTTTGC
>Seqid1
AAGTACGTCC
>Seqid2
TTAGGCTTAGC
>Seqid3
GTTTACGCAAG
nHMMER
SSU bacterial HMM profile
C1
V1
C2
V2
C3
>ECOLI AGtT-AAGTAC--TTTAC-aGCAA-GGAcTTTGC-
>Seqid1-----AAGTAC-GT--------------------
>Seqid2---------------------A-GGC-TTAGC-
>Seqid3------------GTTTAC--GCAA-G---------
Seqid1:V1 80.0%, C2 16.7% ✓
Seqid2: V2 42.9%, C3 88.9% ✗
Seqid3: V1 10.0%, C2 100%, V2 100% ✓

Figure 4.1: Diagram of variable regions reads, mapping and selection. This digram illustrates the process of mapping short reads into the 16S rRNA gene regions. An *E. coli* sequence was added to the short reads. Next, they were aligned with nHMMER and the HMM SSU bacterial profile. For each read, the relative position to the overall gene is defined, and the proportion of the variable region calculated. Reads that map to a variable region at the specified threshold were selected for analysis.

Figure 4.1 contains a diagram of the process. Basically, an *E. coli* full-length sequence (Silva accession CP007265.4699050.4700590) was added to the mock community. Then all these were aligned with nHMMER and the HMM profile for bacteria from Rfam (SSU RF00177). The variable region reads, identified from the literature [Yarza et al., 2014], were mapped to the full length *E. coli* sequence to determine their alignment positions. The alignment positions of these variable regions were then used to determine the position of the short read in the 16S rRNA gene and their percentage of coverage to the variable and conserved regions.

The main goal here was to implement a method capable of mapping short sequencing reads to specific regions of the 16S rRNA gene. In the selection phase, only those reads which cover a significant part of the areas of interest, the pre-defined variable genetic regions, were selected for metagenomic taxonomic identification.

This process was evaluated by first testing using the *Chlamydia* dataset (see chapter 2 section 2.3.1) to determine the selection criteria threshold of the mapping reads, in other words, establish how much of each region needs to be covered by the short reads in order to observe a better taxonomic classification. This process of selecting reads that map variable regions was performed prior to the taxonomic identification. Short reads were selected by several thresholds varying from 25% to 100% coverage of the variable regions and minimum length covered up to 50 bp. The taxonomic classification was performed by Kraken with a reference database consisting of 16 chlamydial species.

Figure 4.2 shows the f1 score metric for each of the 16 species in the *Chlamydia* synthetic community, after applying the selection process at different thresholds of coverage of the variable regions, including minimum sequence length (the last variable regions were very short and this was to ensure enough coverage), followed by a taxonomic identification by Kraken. The first striking result observed was that the best performance was obtained for the full data without performing any read selection step. The second overall result was that more stringent threshold values of minimum variable coverage resulted in the worst performance, e.g. coverage of 90 to 100%. The minimum length coverage threshold did not seem to have much of an effect on the performance, therefore this parameter was ignored. However, with this small data set, it was not possible to establish a threshold to determine the minimum read coverage necessary for improving read identification.

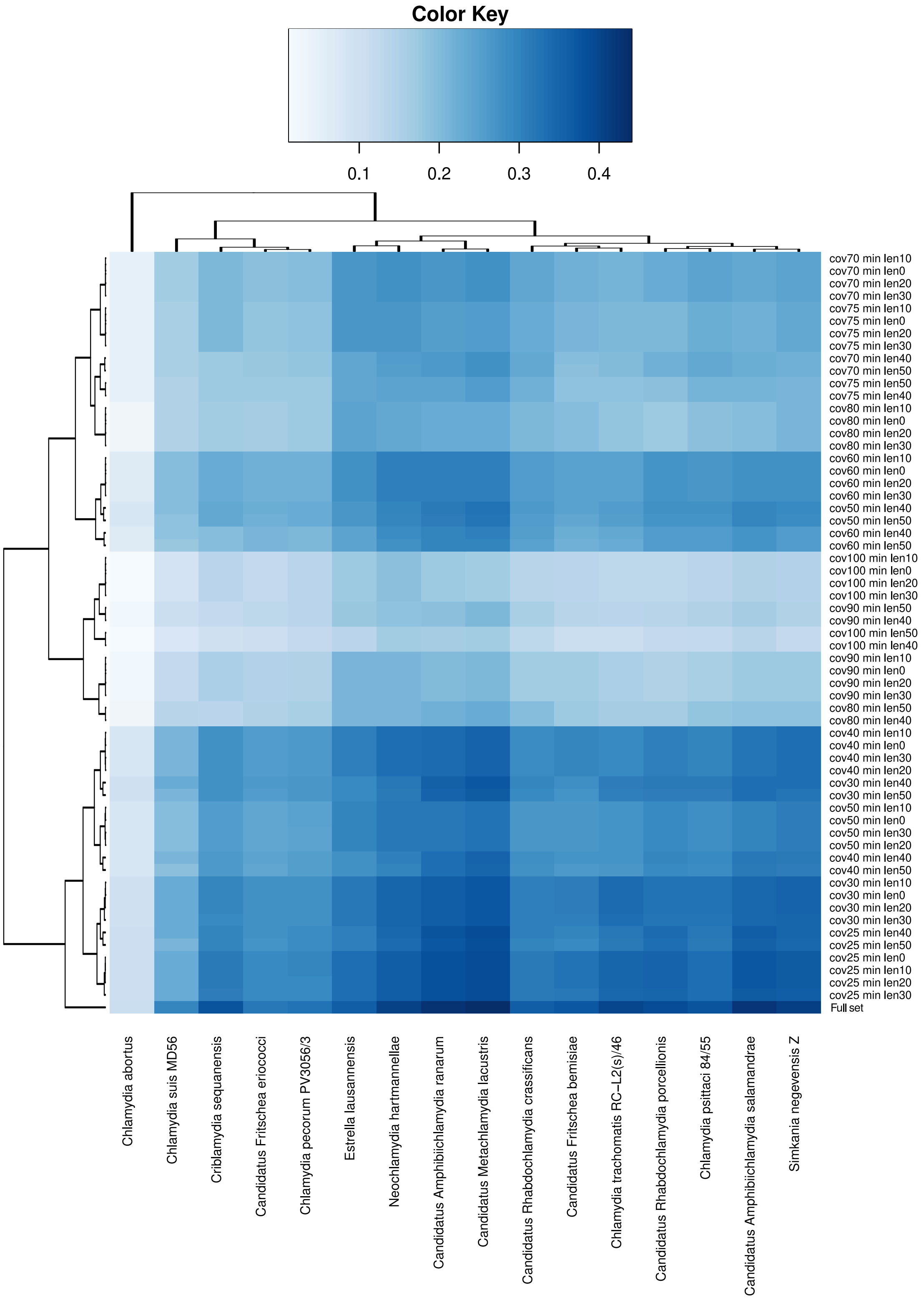
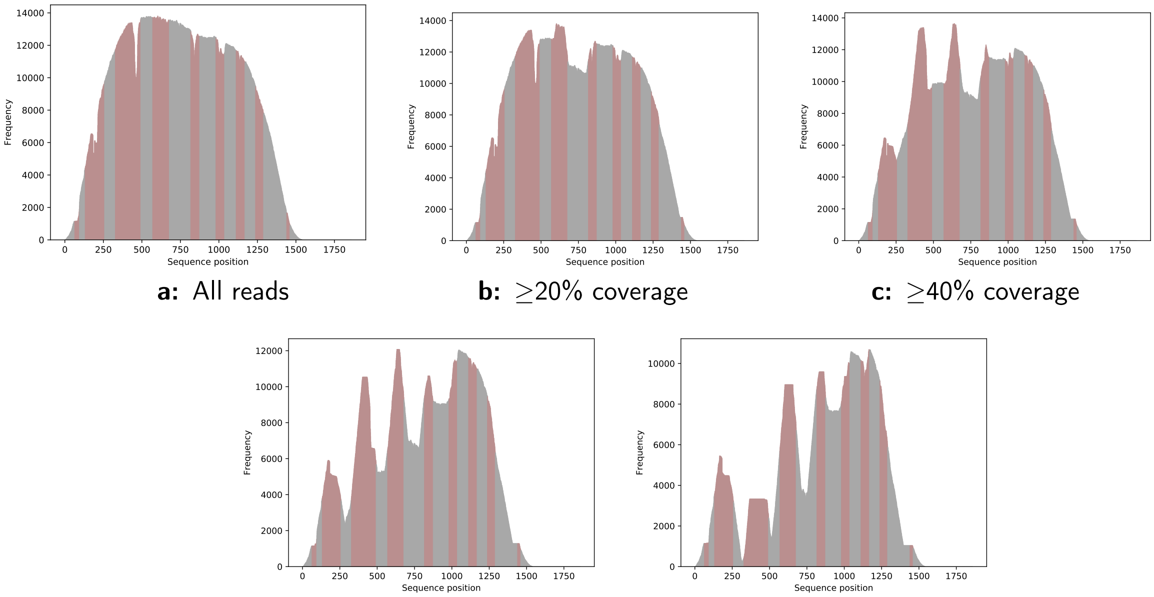


Figure 4.2: *Chlamydia* set f1 Score of selected reads by variable regions' coverage. The dataset consisted of 16 selected Chlamydial species. After the taxonomic assignment with Kraken1 with the same species as reference database, F1 score was calculated for each specie (columns of the heatmap). The rows are the reads that map variable 16S rRNA regions at a combination of thresholds, which include a minimum coverage (indicated by *cov* followed by a number, where the number is the minimum percentage that the short read maps into variable regions, ranging from 25% to 100%), and the minimum number of position that a short reads need to cover a given variable region (e.g. *min len* followed by the number of positions, ranging from 0 to 50bp). High proportion of coverage present low performance, and the minimum length has little effect. However, for this test, the best performance results are for the set without applying any selection of variable regions. This could be due to the nature of Kraken that relies on fragments of size *k*, which at short values of *k* might include repetitions inside and outside variable regions.

The *Chlamydia* data set was a small sample analysed with a tiny reference database. To better understand whether the variable region selection pre-process improve taxonomic classification, a bigger set of data analysed with a more comprehensive database was required. The variable region selection criteria was established to a set of minimum coverage thresholds of 20%, 40%, 60%, and 80% for the dataset *genetic regions test data* described in chapter 2 section 2.3.2, and was analysed with the reference database Silva Parc version 132, to determine if on a more realistic scale the results varied from the previous data set. Figure 4.3 shows the profile of mapping the short reads to the 16S rRNA gene sequence after selecting reads according to variable region coverage.



a: All reads

b: ≥20% coverage

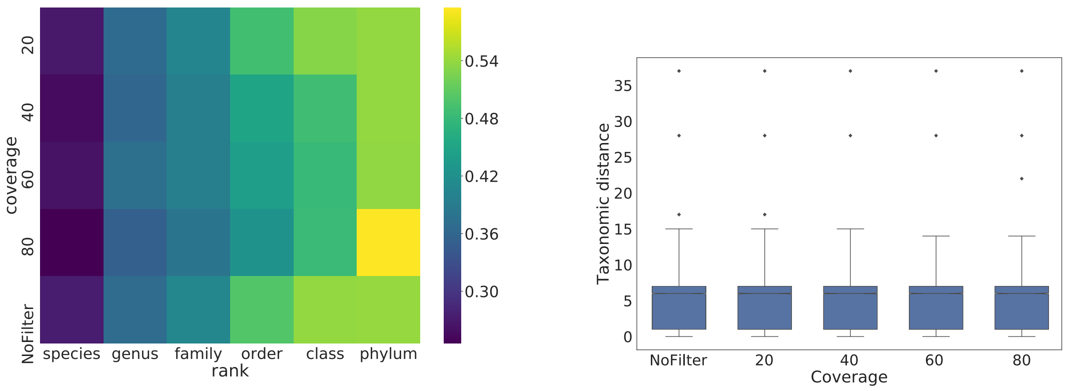
c: ≥40% coverage

d: ≥60% coverage

e: ≥80% coverage

Figure 4.3: 16S rRNA gene sequence coverage of the *genetic regions test data* by selecting reads mapping to the variable regions. The *Chlamydia* dataset is mapped to the 16S rRNA gene sequence. The sub-figures are histograms showing the number of reads that map each individual base. In grey are the conserved regions and in red the variable. The first sub-figure (**a**) is the mapping profile for the whole dataset, and the rest (**b-e**) are the number of reads mapping after selecting those which covers a minimum of any given variable regions (ranging from 20% to 80% correspondingly). Note that in (**e**) a significant proportion of reads covering the V3 region (3rd read area starting from the left) has not been included by the selection process.

The results were evaluated for each taxonomic rank with the f1 score metric, and values, shown in figure 4.4. Similarly, as observed for the *Chlamydia* data set, the f1 score was in general worse for the most stringent coverage threshold and best results were for the full data set. The exception being for the most strict coverage threshold of 80% at the phylum level. This could be due to the relatively low number of phyla present in the set, and by chance the ones with better outcome were kept. The taxonomic distance results did not show any differences between full or more



a: F1 score

b: Taxonomic distance

Figure 4.4: Test data set metrics after selecting reads by variable regions' coverage. This figure shows the *Chlamydia* data set F1 score and TND by after selecting reads mapping variable 16S rRNA regions at different coverage thresholds. While the F1 scores present similar and low results across the coverage thresholds from species to family rank, higher taxonomic ranks show improved value, specially for phylum ath 80% coverage. However, as seen before, at this level of selection (4.3) the lost of reads is significant. On the contrary, the TND remain stable across the different threshold levels.

stringent values of variable region coverage.

Our results suggest that both regions, conserved and variable, contribute equally to the identification of the metagenomic reads.

It is also possible that the variable regions or regions which contain discriminative fragments may vary with the addition of new sequences in the database, and therefore they should perhaps be re-defined. Moreover, different approaches need to be considered on how to use the variable regions to explore how to improve results.

4.2.2 Informative genetic regions

Each genetic database contains different sequences, and taxonomic clan representation might vary. These factors can easily influence evolutionary models which determine conserved and variable gene regions. Here, a methodology was established to define variable and conserved regions for the 16S rRNA gene. The ultimate goal was to understand whether new limits on these types of regions have more discriminatory power for taxonomic classification. Therefore, the next step was to evaluate by masking out the potentially conserved regions from the reference database.

The methodology employed to determine database specific variable regions (see figure 4.5) was as follows: First, only those sequences considered of high quality from the Silva database NR99 (510,503 sequences) were selected, which consisted of selecting those sequences whose taxonomic annotations include the seven main taxonomic ranks (superkingdom, phylum, class, order, family, genus, and species), which left 195,352 sequences, and that do not contain ambiguous bases; Second, all the 172,928 sequences remaining were then structurally aligned with Infernal [Nawrocki and Eddy, 2013] and the Covariance Model (CM)[1] from Rfam bacterial.

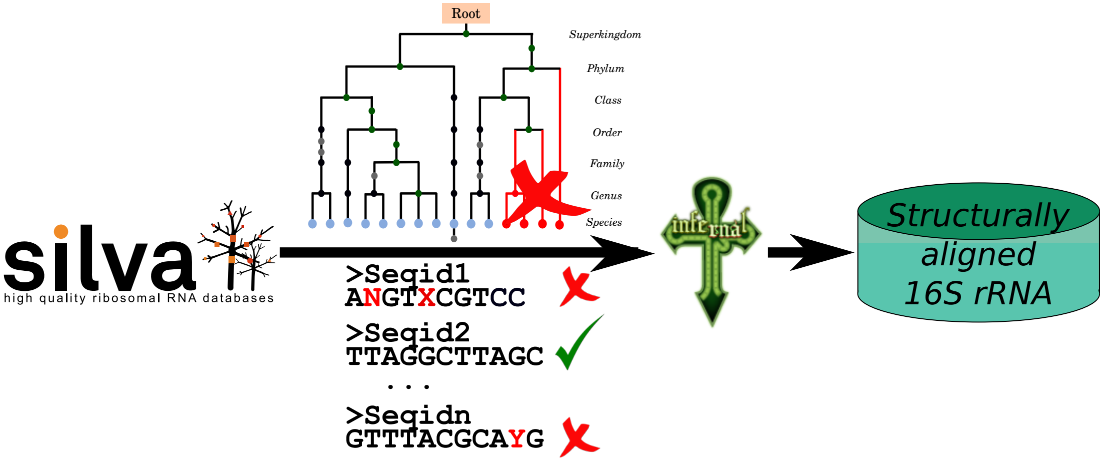


Figure 4.5: 16S rRNA gene structural alignment diagram. The diagram illustrates the process followed to generate structurally alignment of 16S rRNA. First, high quality sequences were selected from the Silva database NR99 (containing the seven main ranks and without ambiguous bases). Then they were aligned with Infernal and the covariance model from Rfam SSU bacterial.

Once the sequences were structurally aligned, the information content for each position of the alignment was measured with the Kullbac-Leiber Divergence (KLD) method, which was used to measure the conservation status of the genetic regions. KLD is the loss of information when the observed distribution is compared to the expected, and was calculated as shown in equation 4.1:

$$KLD_j = \sum_{i=1}^{N} P_{ij} log \left(\frac{P_{ij}}{Q_i} \right) \tag{4.1}$$

$$P_{ij} = \frac{C_{ij}}{\sum_{i=1}^{N} C_{ij}} \tag{4.2}$$

$$Q_i = \frac{C_i}{\sum_{i=1}^{N} C_i} \tag{4.3}$$

KLDj The information content for the j-th column in an alignment.

Pij Relative frequency of a particular letter *i* in the j-th column.

Qi The expected frequency of a letter *i*.

Cij Number of counts of letter *i* int he j-th column.

Ci Total number of counts of letter *i*.

[1] Covariance model definition: "approach to several RNA sequence analysis problems using probabilistic models that flexibly describe the secondary structure and primary sequence consensus of an RNA sequence family". Source [Eddy and Durbin, 1994]

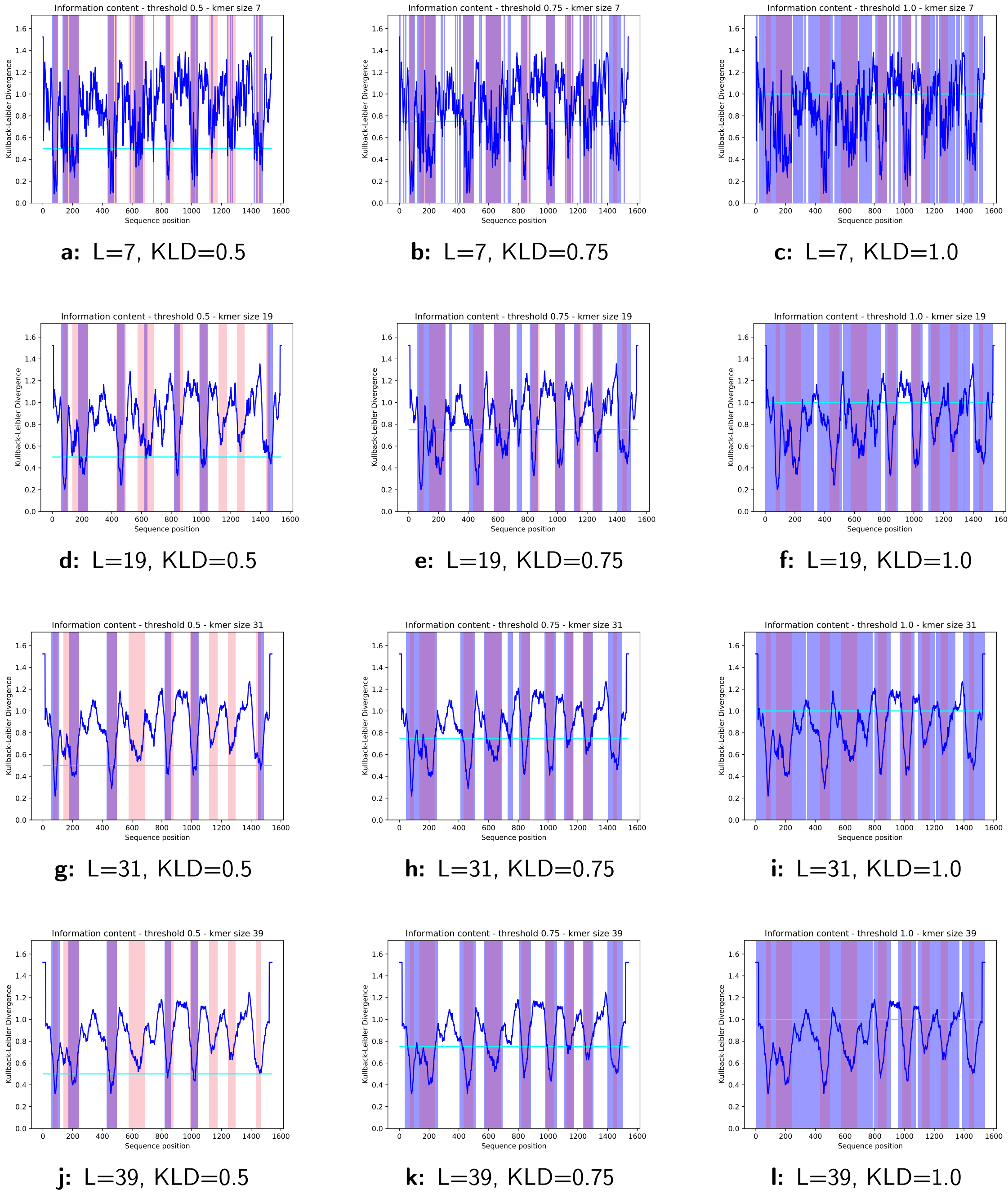
N The number of letters in the alphabet *ACTG*.

The KLD was averaged by grouping *"l"* columns. Finally, arbitrary thresholds were set (l of 7,19,31 and 39 and KLD of 0.5,0.75 and 1) to determine which regions information loss was lower and therefore could be potentially classed as evolutionary variable regions. Figure 4.6 shows the KLD values for the gene length, the calculated regions of interest (blue) and the already known variable regions (pale red) for comparison.

We next evaluated whether these newly defined regions were able to improve metagenomic taxonomic classification. To do so, the Silva Ref Trunc aligned SSU database version 138 was downloaded. The positions of the alignments were mapped to *E. coli* sequence accession CP007265.4699050.4700590. The previously described regions of interest, shown blue in figure 4.6, were then mapped to the alignment and the rest were masked out.

All the MGC data was analysed with Kraken2 for each masked database for each group of potential variable regions to determine the optimal limits.

The original MGC dataset contained 1426 organisms which were labelled as belonging to 35,857 nodes of the NCBI taxonomy tree. In figure 4.7 the number of correctly classified simulated reads (TND=0) increases with the number of regions included. In fact, the non-masked reference database (RefTrunc) was the one with the best results. The number of reads correctly identified was higher for those databases with higher coverage of the 16S rRNA gene. As observed previously, the number of mutations slightly decreased the number of correctly identified reads.



a: L=7, KLD=0.5

b: L=7, KLD=0.75

c: L=7, KLD=1.0

d: L=19, KLD=0.5

e: L=19, KLD=0.75

f: L=19, KLD=1.0

g: L=31, KLD=0.5

h: L=31, KLD=0.75

i: L=31, KLD=1.0

j: L=39, KLD=0.5

k: L=39, KLD=0.75

l: L=39, KLD=1.0

Figure 4.6: Potential database-specific variable regions of the 16S rRNA gene. Kullbac-Leiber Divergence was calculated for each position of the structural alignment of high quality 16SrRNA sequences in the Silva databases. The horizontal cyan line is the threshold below which regions could be described as variable, which are highlighted in blue. In red are the 16S rRNA gene variable regions. For KLD threshold of 1, the areas cover a rather large proportion of the genomic sequence, whereas KLD of 0.5 detects very little. The criteria of KLD of 0.75 is a middle ground and at the same time correlates better (not perfectly) with the variable regions already described in literature. When L is low (e.g. 7, **b**), the number of areas predicted is much higher, with a number of short or very short regions.

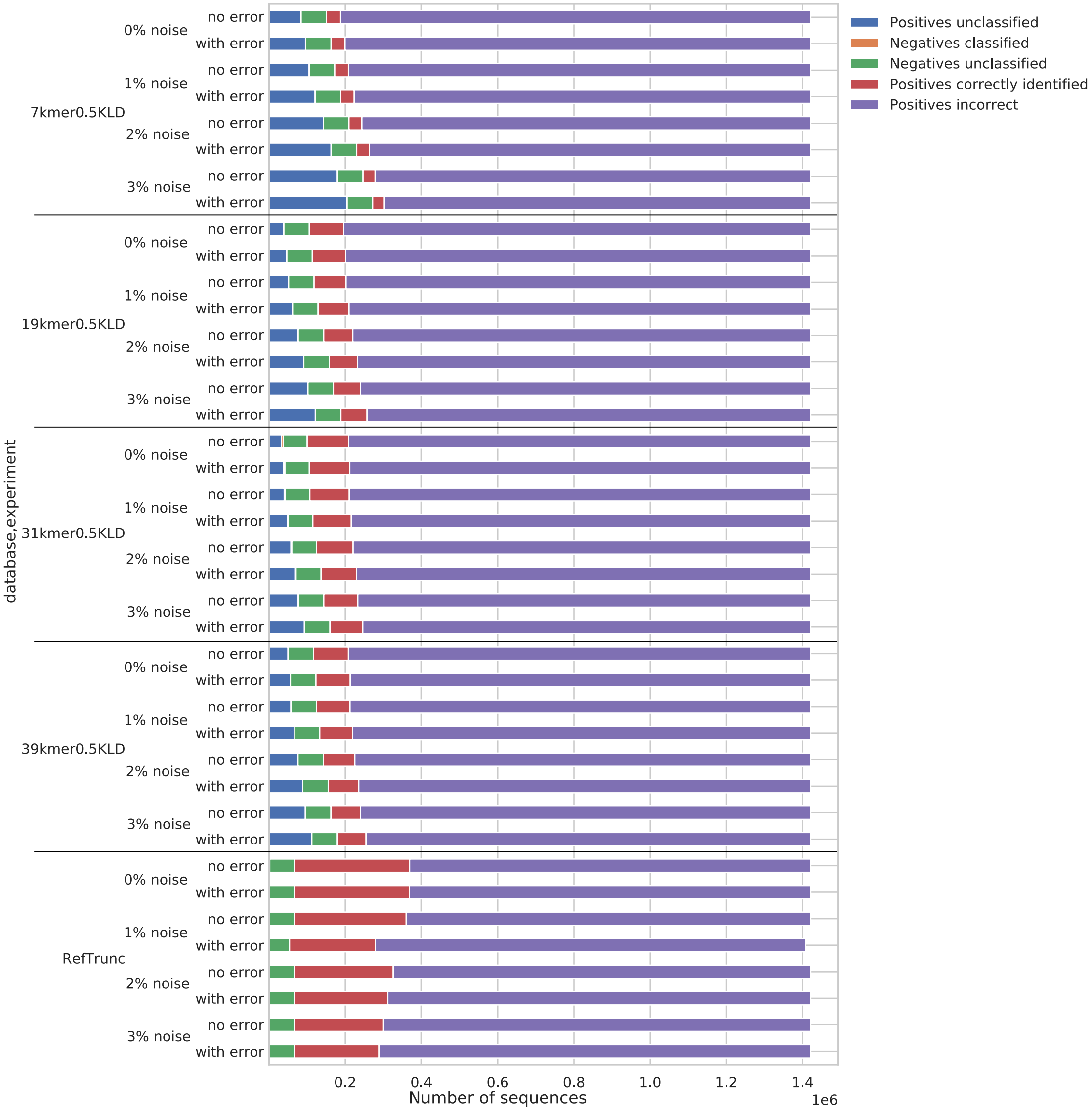


Figure 4.7: Number of sequences identified by masking newly defined variable regions.The MGC data was analysed with Kraken2 and Silva RefTrunc version 138. This database was masked according to the regions calculated by the IC criteria for several thresholds. The results are presented by the errors in the mock community. 'Positive' refers to the positive controls in the dataset, those that are present in the database, whereas 'negative' refers to those reads generated synthetically and are expected to be unclassified. The incorrectly identified sequences were labelled to some taxon, but not to the original organism. The number of unclassified reads was higher for those masked databases with very low coverage of the 16S rRNA gene, and also for those reads with a higher proportion of mutations. The full database resulted in a higher number of correct classification rates.

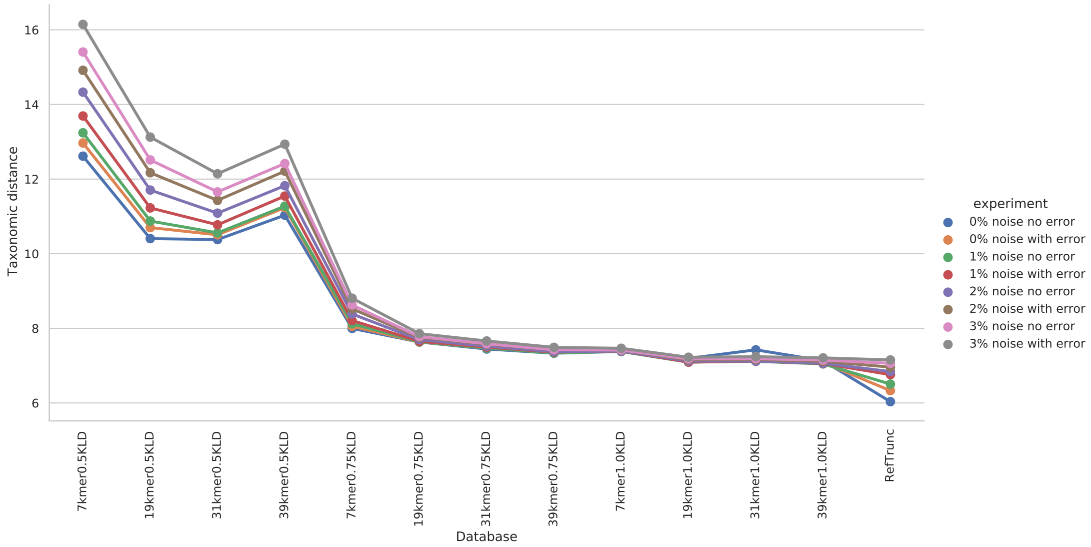


Figure 4.8: TND after masking newly proposed regions.After masking the RefTrunc database with the IC criteria regions, the average TND is high for a lower threshold of KLD, whereas for higher limits (0.75 and 1) the .TND is much lower. However, the best results were obtained when the data was analysed with the unmasked database.

TND was measured after masking out the newly defined potential conserved regions with the Silva Ref database. Those sets with regions of interest with a KLD threshold of 0.75 and 1.0 have the lower average TND (see figure 4.8). TND of the masked database with the regions of l-mer 39 and KLD of 0.5 increase compared to l-mer 31 counterparts. However, the best results were obtained with the unmasked reference database.

These results indicate that it is necessary to have full-length genomic sequences for taxonomic identification of short reads. Moreover, so far only genetic regions have been studied. k-mer based methods used genetic fragments of size k to classify sequences, which are found in varying frequencies across organisms. Finding a way to penalise those k-mers that are commonly in a database and giving more weight to those which are unique for a species might be a better solution. In the next section, linguistics methods were used to create a novel metagenomic taxonomy classifier.

4.3 Linguistic methods for taxonomic identification of metagenomic data

So far, the work presented was focused on the evolutionary conserved versus variable regions. Nevertheless, some k-mers can still be found in both type of regions.

Each reference database has a unique composition. K-mers are found in different proportions in each genome and across species. Organism specific k-mers are vital elements for taxonomic identification.

Term-Frequency Inverse Document Frequency (TF-IDF) is a widely applied method in linguistics which weights words according to their frequency in the body of the text or corpus. It is used for example to determine the type of document or find topics (e.g. sports, education, politics, etc). A document is composed of words, some occur multiple times in one or a few documents, whereas others are frequently observed in all or almost all. The first ones have more influence to discriminate the document compared to the later ones. TF-IDF It has been applied to distinguish plasmid material from metagenome data [Krawczyk et al., 2018], to determine lateral gene transfer [Cong et al., 2017], the community interaction and structure [Yan et al., 2017], and to identify long non-coding RNAs in combination with secondary structure [Madhavan and Gopakumar, 2018].

4.3.1 Genetic Fragments Score Aided Taxonomic Identification (GeFSATI)

The strategy coupled theTF-IDF method with a naive Bayes classifier to develop a method that taxonomically identifies sequencing reads up to the genus level by weighting k-mers, and it is called GEnetic Fragments Score Aided Taxonomic Identification (GeFSATI).

The TF-IDF needs to be redefined for the purpose of identifying organisms through their genetic material. In this new context, the *corpus* is the *reference database*, a *term* or a word is a fragment of size k or *k-mer*, and *documents* are all the genetic *sequences* belonging to the same taxon or clan.

These types of models can create an impractiably large matrix for metagenomics, which in some circumstances might be millions of k-mers by millions of species or subspecies, so a reduced dimensions' strategy was required. This method only identifies metagenomic sequences up to genus level, for two main reasons: firstly, because some phyla contain a large number of members and would generate huge model matrices, and secondly, and more importantly, the Silva database only curates up to genus level (see chapter 5 section 1.6.3).

Model generation

After cleaning the databases, and in order to generate the models, a two-step strategy was designed: the first is for classifying sequences at the phylum level and the second one is for classifying sequences at the genus taxonomic rank based on the predicted phylum. Figure 4.9 shows a schematic diagram of the model generation process. All the steps are detailed below.

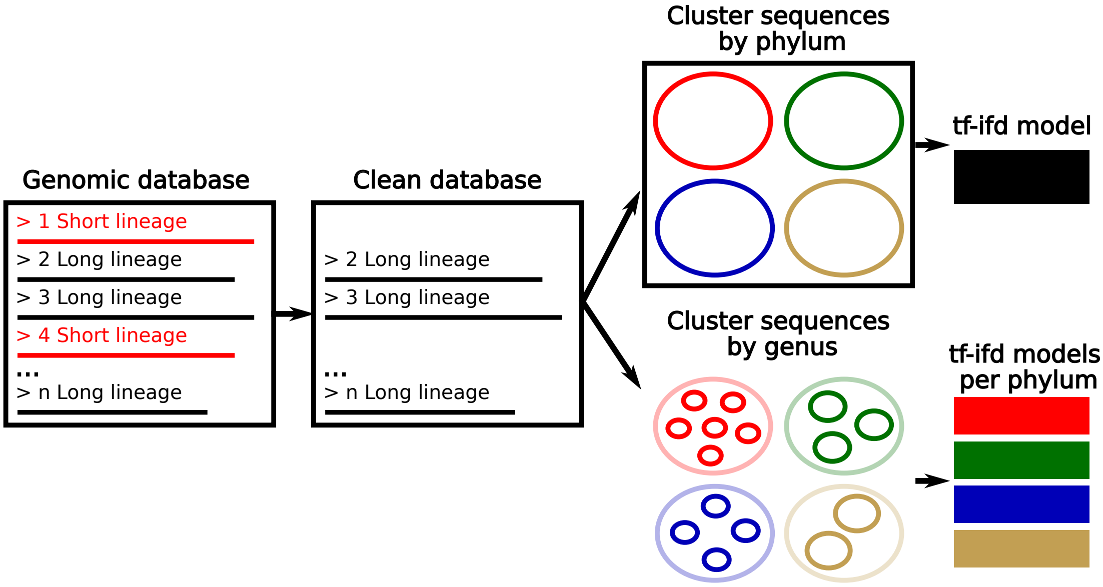


Figure 4.9: Building TF-IDF models. First, the database is cleaned by removing those sequences with lineages with less than 7 ranks, so keeping only those sequences that are likely to be highly curated annotations. Then sequences are clustered at the phylum level and a TF-IDF model is built. This model predicts only phylum. Next, for each phylum, sequences are grouped at the genus level and a TF-IDF model per phylum is calculated. So the classification works in two steps, first a phylum is predicted with the first model, and then genus is predicted by the corresponding model of the second stage.

Cleaning and clustering the reference database

First the reference database Silva NR99 version 138 with its own taxonomy, which contains 510,503 sequences, was cleaned by keeping those sequences which have the main ranks annotated (superkingdom, phylum, class, order, family and genus) and 383,513 sequences remained. Then sequences were clustered at the phylum rank and also at genus level for each of the 71 phylum, as figure 4.9 illustrates.

Bag of k-mers

Sequences were split in overlapping k-mers of size 9, ignoring fragments with ambiguous bases, as shown in figure 4.10. K-mers that are found in more than

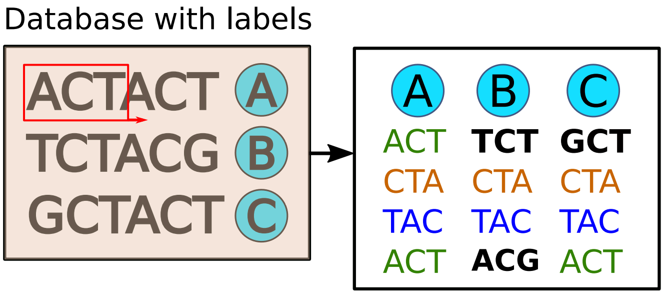


Figure 4.10: Bag of k-mers. Each sequence in the database is split into overlapping fragments of size k. In this example, k=3. Each sequence is taxonomically labelled, here represented with the capital letter inside the blue circle. Some fragments are present in all the sequences, whereas others are only present in some. Each k-mer is highlighted in a different colour. K-mers that are unique in the whole database are in bold and are expected to be the ones more informative and therefore should have higher weight during classification.

70% of the clans (phylum or genus accordingly) were considered not informative and consequently removed, except for some phyla that only contain 1 genus where it did not make sense to remove them.

Building TF-IDF models

For each of the sets, a TF-IDF matrix was calculated. All matrices were computed with the functions provided by scikit-learn [Pedregosa et al., 2011] (python package) as described in equations 4.4, 4.5. The tokenizer, to obtain the bag of k-mers, was built in-house.

$$tf - idf(t,d) = tf(t,d) \times idf(t) \tag{4.4}$$

$$idf(t) = log\tfrac{1+n_d}{1+df(d,t)} + 1 \tag{4.5}$$

Where:

tf-idf Term-frequency — inverse document-frequency.

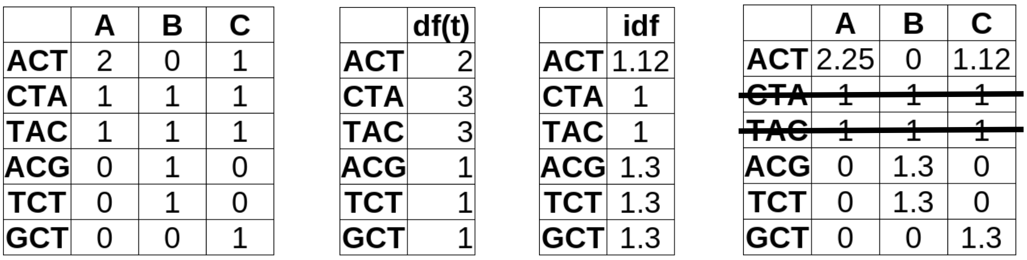
tf(t,d) k-mer count.

idf inverse document-frequency.

nd total number of clans.

df(d,t) number of clans that contain the k-mer t or clan frequency.

Figure 4.11 shows an example of how TF-IDF matrices are calculated for taxonomic identification. First, the frequency of each fragment of size k (rows) is counted per each clan or taxon (columns). The inverse document frequency was calculated and



	A	B	C
ACT	2	0	1
CTA	1	1	1
TAC	1	1	1
ACG	0	1	0
TCT	0	1	0
GCT	0	0	1

a: k-mer count

	df(t)
ACT	2
CTA	3
TAC	3
ACG	1
TCT	1
GCT	1

b: Clan frequency

	idf
ACT	1.12
CTA	1
TAC	1
ACG	1.3
TCT	1.3
GCT	1.3

c: Inverse document frequency

	A	B	C
ACT	2.25	0	1.12
~~CTA~~	~~1~~	~~1~~	~~1~~
~~TAC~~	~~1~~	~~1~~	~~1~~
ACG	0	1.3	0
TCT	0	1.3	0
GCT	0	0	1.3

d: Term Frequency - Inverse Document Frequency

Figure 4.11: Example of Term Frequency - Inverse Document Frequency. *a* - The frequency in each k-mer for each label is counted and stored in a matrix. *b* - The document or clan frequency is the number of taxonomic clans (in this example A, B or C) where each k-mer is found. *c* - The inverse document frequency is calculated according to the equation 4.5 for each k-mer. *d* - Finally the TF-IDF model is calculated according to the equation 4.4 and the k-mers that are present in all the documents are excluded from the final model.

multiplied by the count's matrix, which generated the TF-IDF model after removing those k-mers with frequency above the specified threshold.

Multinomial Naive Bayes model

Multinomial Naive Bayes is a supervised machine learning method which assumes independence between pairs of features (k-mers) given a class variable (taxon). It is implemented for multinomial distributed data, and it can be applied to counts and TF-IDF matrices.

For each matrix, a Multinomial Naive Bayes classifier model was created. The model with the sequences clustered at the phylum rank contains 1,262,680 features (k-mers).

Classification

The classification, shown in diagram 4.12, consisted of masking reads, predicting phylum followed by a prediction of genus and finally a final assignment step.

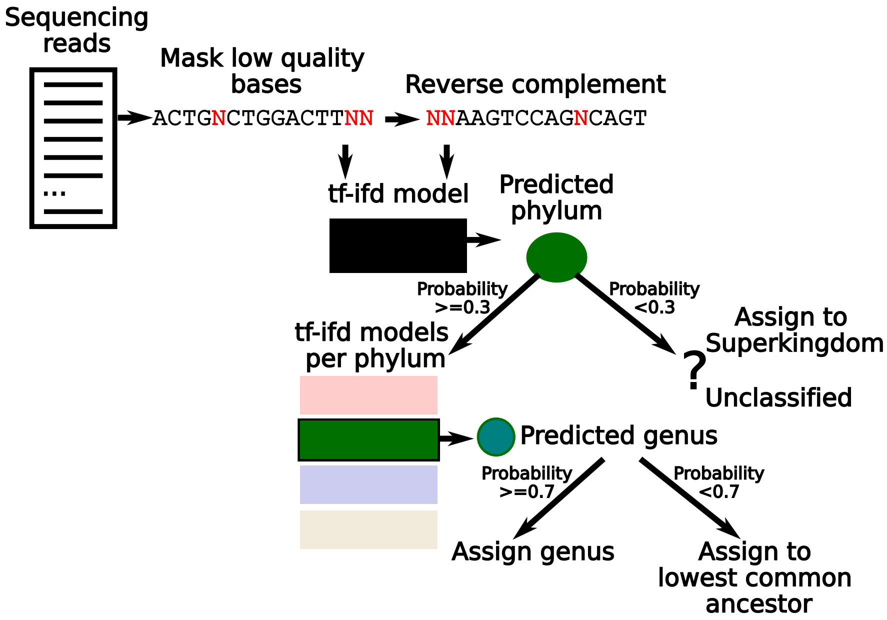


Figure 4.12: TF-IDF classification. Each sequencing read is first classified to the phylum level after masking the bases with low phred scores for the forward and reverse complement. Depending on the probability it can be assigned to the rank superkingdom, unclassified or a genus may be predicted from either forward or reverse complement sequence accordingly. If the probability at the genus rank is low, a lowest common ancestor approach is applied for the top *n* results whose probability is above the threshold.

Prior to the prediction steps, short read bases with low quality with a phred score inferior or equal to 20 were masked.

Then the phylum was predicted for both the forward and reverse complement, and the prediction with the highest probability of the two was chosen (if it was equal or superior to 0.3). Genus was predicted with the corresponding model and with either the forward or reverse sequence accordingly.

The assignment step for the predicted genus consisted of either labelling the sequence to the predicted genus (when the probability of the top result is equal or superior to 0.7), or the lowest common ancestor approach was applied for up to five of the top results whose probabilities add up at least 0.7. Where the previous conditions were not met, the phylum level was assigned.

When the probability for forward and reverse sequence was lower than 0.3, the probabilities of up to 5 of the top results were combined until they result in a minimum value of 0.5. The last common ancestor approach was applied to whichever resulted in the highest value. This meant- that sequences were assigned at the superkingdom level or above. Sequencing reads were *unclassified* where there is no common ancestor.

Results

The new approach, GeFSATI, was based on the TF-IDF linguistic method, which allows specific weighting for the unique k-mers (high scores) versus the common ones (low scores). Then a two stage multinomial naive Bayes classification was performed, first to classify up to phylum rank, and then up to genus level.

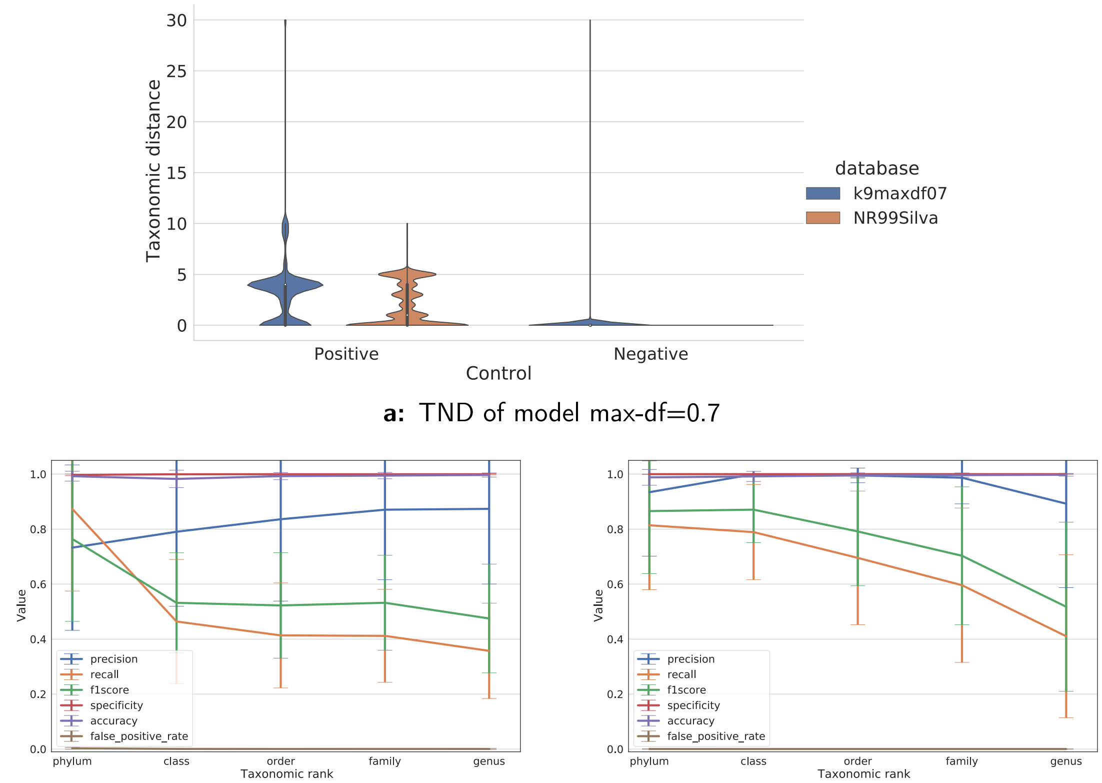
This new approach was tested with the soil sample replica 1 without noise and without sequencing errors from MGC. The results were compared to the most similar results available, which were obtained using Kraken2 and the full Silva NR99 database with its own taxonomy.

It took approximately 1.8 hours with 1 core in a machine Dell PowerEdge R440 Rack Server Xeon Silver 4110/128GB/8x2TB and up to 25Gb RAM to build the models and almost 15.3 hours and 10Gb RAM for the classification steps.

The taxonomic distance was calculated and is presented in figure 4.13.The bulk of the sequences were classified at 4 or 0 nodes away with GeFSATI, and 157 positive control sequences were unclassified. Kraken (in orange) classified more sequences at the lowest levels (TND = 0) and the next significant TND value is 5, and no sequences were unclassified. One negative control was labelled at the phylum level with GeFSATI.

Figure 4.13 show the metrics comparing the same sample with the method presented here and the other with Kraken2 and the full Silva NR99 database with its own taxonomy. GeFSATI's recall at the phylum level is higher than Kraken. At the genus level precision and recall are similar to Kraken, but precision and recall are lower at the rest of the ranks compared to Kraken with Silva with its own taxonomy as a reference.

To the best of my knowledge, this is the first time a TF-IDF strategy has been combined with a multinomial naive Bayes for metagenomic taxonomic identification.



a: TND of model max-df=0.7

b: Model max-df=0.7

c: Kraken NR99 Silva

Figure 4.13: Comparison of metrics by method. The classification works better with Kraken2 with Silva NR99 and its own taxonomy than the method presented here. The TND of GeFSATI shows that most of the positive control are classified at either 0 or 4 nodes away from their origin leaf. Whereas for Kraken Silva NR99 most of the sequences are classified correctly, and the second most abundant TND value is 5. The negative controls are all but one correctly identified by GeFSATI and all for Kraken. GeFSATI's present a better recall for phylum and similar values of precision and recall at genus level, but underperforms at the rest of the ranks.

4.4 Discussion

Genes contain conserved and variable regions across organisms, and it is important to understand how they affect taxonomic classification. In this chapter, two different approaches were investigated.

The first approach involved the selection of sequencing reads mapping to 16S rRNA gene variable regions defined in the literature. The best overall results were presented by the whole set. However, it can be argued that because databases are continuously increasing, the content of the reference varies and consequently the regions that may help to improve identification may be different.

This method, however, has the potential to be applied in other contexts, for example to remove PCR artefacts and contamination from the sequencing data, or to select reads that map onto specific genes or regions.

The second strategy, which consisted of masking out non-informative regions from each sequence, revealed that longer coverage of genes obtain better results.

Overall, the results confirm that the completeness of the gene in the reference database is essential to be able to identify short sequences more accurately and as close as possible to species level, as previously described in the literature [Johnson et al., 2019].

Kraken is a k-mer based method for metagenomic identification. Genetic fragments of size k are considered to be genetic signatures and are found in varying amounts. These fragments can be unique or common among species, and perhaps they may be the key to improving taxonomic classification.

A feature selection method with TF-IDF approach might have the power to obtain better results. The newly developed GeFSATI is capable of taxonomically identifying short sequences based on weighted k-mers.

This newly developed unigram type of model, based on scoring k-mers with the TF-IDF method, overcomes the dimensionality problem and the question of having enough information per clan (*topic* in linguistics). The strategy consisted of reducing dimensionality by selecting sequences with a high quality annotated taxonomy, a modest size of k, ignoring irrelevant k-mers, and splitting the classification into two steps: clustering them first at the phylum rank, and then at the genus rank per phylum, which means that in the vast majority of cases there is more than sequence (referred to as a document in linguistics terms) per taxonomic clan.

An improvement of our method compared to other similar approaches [Garbarine et al., 2011] is demonstrated by the fact that GeFSATI is capable of determining whether the short read is found in the forward or reverse complement strand in the reference database.

The reductionist strategy means that on one hand, it can handle better reference databases, sometimes at the cost of losing vital information for identification. More steps make it slower and potentially might produce less accurate results.

More investigation is needed to decrease TND values for GeFSATI and to obtain higher recall. More strict probability values at the phylum levels worsen the results, and relaxing probabilities at the genus level does not show much difference (data not shown), indicating that these probabilities might not be as reliable as expected.

The literature suggests that not necessarily longer k-mers improve classification. For example, combining TF-IDF with a Euclidean classifier reveals that accuracy is higher for shorter k-mers (comparison of 6 and 9) [Garbarine et al., 2011]. Another method [Şener et al., 2018], coupling TF-IDF with Latent Semantic Analysis (LSA) to identify similar metagenomic samples, also shows shorter k-mers perform better ($k<=10$) from a range up to 13. VirNet [Abdelkareem et al., 2020] is a tool specifically developed for virome identification. A TF-IDF method with k-mer sizes ranging from 1 to 20 was tested with a variety of classifiers (logistic regression, XGboost, AdaBoost, decision trees and random forest.), with an optimal k of 11 (88% accuracy). Therefore, the optimal value of k needs to be determined for GeFSATI for both types of models, at phylum and genera levels.

The multinomial naive Bayes classifier is a popular method in linguistics, it is easy to use, scalable, and works well for counts matrices or TF-IDF. Therefore, it is possible that another classifier might obtain better results. For example, Support Vector Machines is a supervised method widely used for text classification. It is claimed to be one of the best classifiers at the moment. It is effective in high dimensional spaces. However, there is the risk of over-fitting and does not directly provide probability estimates. Latent Dirichlet Allocation is an unsupervised algorithm designed to discover topics in a given document. This differs slightly from the TF-IDF approach. As an input, it needs the bag of words and the number of topics. It returns a document to topics matrix and topics to word matrix.

Another important factor is that the reference database contains full length genetic sequences (about 1500 bp), whereas the simulated sequencing reads used here were much shorter (151 bp). Therefore, because only a fragment of the real gene is observed, the proportions of k-mers will be different compared to the full length. By this argument, Bernoulli naive Bayes, which is binary (indicates only presence or absence of a k-mer) might be more appropriate.

At the moment, one of the main GeFSATI limitations is that, it has been implemented for single reads. It is expected in a future release to incorporate a paired-end mode option. Also, speed and memory efficient solutions need to be explored.

4.5 Conclusions

Evolutionary pressure can cause accumulation of mutations on certain genetic regions, but the taxonomic identification relies on the principles of the classifier. In this case, Kraken is k-mer based, and does not show any effect on selecting or masking informative regions for taxonomic identification. In general, metagenomics taxonomic classification needs to exploit this differences to improve taxonomic classification scores.

It is crucial that databases and the identification approaches contain genes as complete as possible, as well as having highly curated data including annotations.

In this chapter, a novel method has been developed to taxonomically classify metagenomic marker gene short reads. It scores overlapping k-mers according to their frequencies in the reference database, and is coupled with a multinomial naive Bayes classifier.

Chapter 5

Reference databases

5.1 Introduction

The composition of a reference database has a huge impact on taxonomic classification performance when used with the exact same tools and parameters [Méric et al., 2019, R. Marcelino et al., 2020, Dixit, 2021], from the genomic sequences to their corresponding taxonomic lineages. Yet, there is a lack of comprehensive studies on their effect for taxonomic identification of metagenomics data.

Some databases are specialised in one type of data, for example NCBI taxonomy only contains taxonomic lineages (see chapter 1, section 1.6.1 on page 8), whereas others curate both, like the Silva database which contains a selection of 16/18S rRNA sequences with their own curated taxonomy (see chapter 1, section 1.6.3 on page 11).

This chapter presents the impact databases have on taxonomic classification, depending on their genomic content and their corresponding lineages, focusing on the Silva and NCBI taxonomy databases.

5.1.1 Objectives

The main goals of this chapter were to:

- Determine the behaviour of taxonomic identification when organisms do not have genetic sequences present in the reference database.
- Establish potential biases of the reference databases.
- Characterise factors that might lead to taxonomic misclassification of short reads, depending on the content of the reference database.

Firstly, the effect of missing data from the reference database was explored with the *Chlamydia* dataset (described in chapter 2 section 2.3.1) and Kraken2.

Secondly, the impact of the quality of the genetic sequences present in the reference database for taxonomic classification was studied and critical and non-critical factors determined.

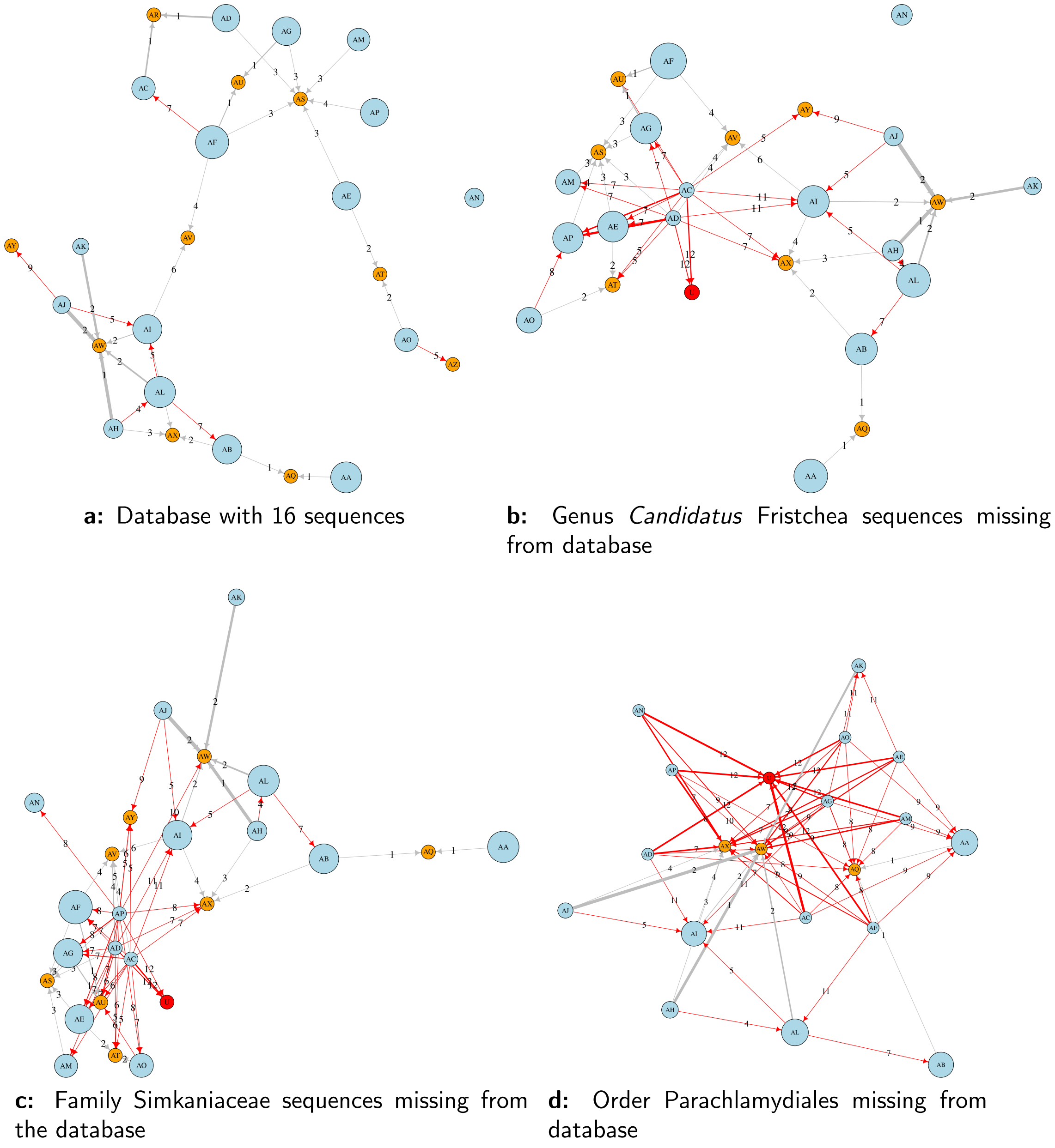
Finally, different taxonomic databases were compared with the exact same genetic content to understand how they affect classification.

5.2 Classification bias for incomplete reference databases

A small scale test was conducted with the *Chlamydia* dataset (described in chapter 3, section 3.3) to pinpoint weakness of taxonomic identification when sequences are missing from the reference database.

Thirty-two subset reference databases were generated by removing the sequences belonging to each taxonomic rank from the original 16 *Chlamydia* species and were compared to the full database as well as their corresponding taxons from the taxonomy database.

Figure 5.1 presents networks of classification for four of the previously described sub-databases: the first one contains the 16 original species included in the dataset; the second one, sequences belonging to the genus *Candidatus* Fristchea have been removed; the third one, sequences belonging to the order Parachlamydiales have been removed; and a fourth and final sub-database without the sequences belonging to the family Simkaniaceae. In general, taxonomic identification gets worse (red arrows) for higher taxonomic ranks missing from the reference.



a: Database with 16 sequences

b: Genus *Candidatus* Fristchea sequences missing from database

c: Family Simkaniaceae sequences missing from the database

d: Order Parachlamydiales missing from database

U – Unassigned
AA – Candidatus Amphibiichlamydia ranarum
AB – Candidatus Amphibiichlamydia salamandrae
AC – Candidatus Fritschea bemisiae
AD – Candidatus Fritschea eriococci
AE – Candidatus Metachlamydia lacustris
AF – Candidatus Rhabdochlamydia crassificans

AG – Candidatus Rhabdochlamydia porcellionis
AH – Chlamydia abortus
AI – Chlamydia pecorum PV3056/3
AJ – Chlamydia psittaci 84/55
AK – Chlamydia suis MD56
AL – Chlamydia trachomatis RC–L2(s)/46
AM – Criblamydia sequanensis

AN – Estrella lausannensis
AO – Neochlamydia hartmannellae
AP – Simkania negevensis Z
AQ – Candidatus Amphibiichlamydia
AR – Candidatus Fritschea
AS – Parachlamydiales
AT – Parachlamydiaceae

AU – Candidatus Rhabdochlamydia
AV – Chlamydiia
AW – Chlamydia
AX – Chlamydiaceae
AY – Criblamydiaceae
AZ – Simkaniaceae

e: Legend

Figure 5.1: Chlamydia dataset classification network when clades are missing from the database. Each node of the graphs represent a node in the taxonomic tree. Nodes representing species (leaves of the taxonomic tree) are shown in blue, while the rest are shown in orange with the exception of unclassified, which is in red. Species nodes have a size proportional to the number of correctly assigned sequences. Arrows from species nodes point to the nodes to which other sequences from that species have been assigned, with a width proportional to the number of sequences. The number by each arrow is the TND. Grey arrows are when sequences are classified in the same lineage whereas red ones indicate that sequences have been assigned to an incorrect lineage.

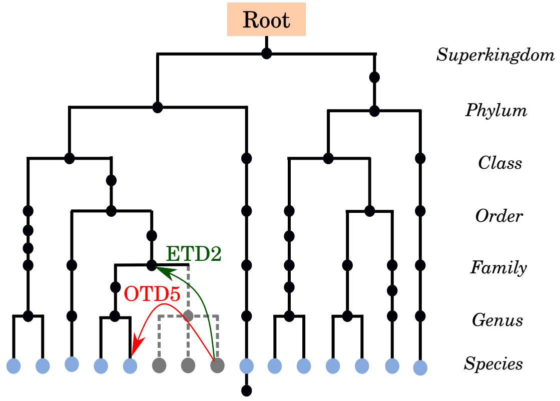


Figure 5.2: Taxonomic Distance Divergence. When sequences representing a taxonomic group, for instance a genus, do not have sequence representing them in the reference database, an ideal taxonomic profiler should assign them to the closest common ancestor present. The Taxonomic Distance Divergence is measured by subtracting the expected TND from the observed. In this example, TDD is equal to 3. In grey represents a genus whose sequences are missing in the database. It is expected (ETD) that the sequence will be classified in the rank immediately above, however it has been classified (OTD) to another species.

Knowing the missing fraction from each sub reference database, the Taxonomic Distance Divergence (TDD) can be measured as the difference between the expected TND and the observed TND. In the example in figure 5.2 this is calculated as 3. The expected TND is the number of nodes between the original node and the rank immediately above of the missing fraction of the taxonomic tree, or 0 if the organism has a sequence in the reference database.

When 16 sequences belonging to the chlamydial clan of the 16S rRNA gene as reference database, TDD is close to 0 for most species. For higher taxonomic ranks, e.g. order, TDD is larger, as shown in figure 5.3. Also, the genus *Chlamydia* TDD is higher than other genera, probably because this rank is the one that contains more representatives in the taxonomic tree (see figure B.3 in chapter 3 on page 142).

TDD analysis also reveals that the missing fraction of the reference database can lead to classification further in the taxonomic tree reads even when a clan (sequences belonging to a specific taxonomic rank) is present in the database. For example, it is expected that short simulated reads belonging to the species *Chlamydia abortus* to present TDD value of 0 when the order of Parachlamydiales is missing (its own order Chlamydiales is present), instead they classify to much further away taxa.

This confirms the findings of Marcelino et al [R. Marcelino et al., 2020] that metagenomic classification has a bias towards assigning reads to the sequences present in the reference database, ignoring potential missing fractions.

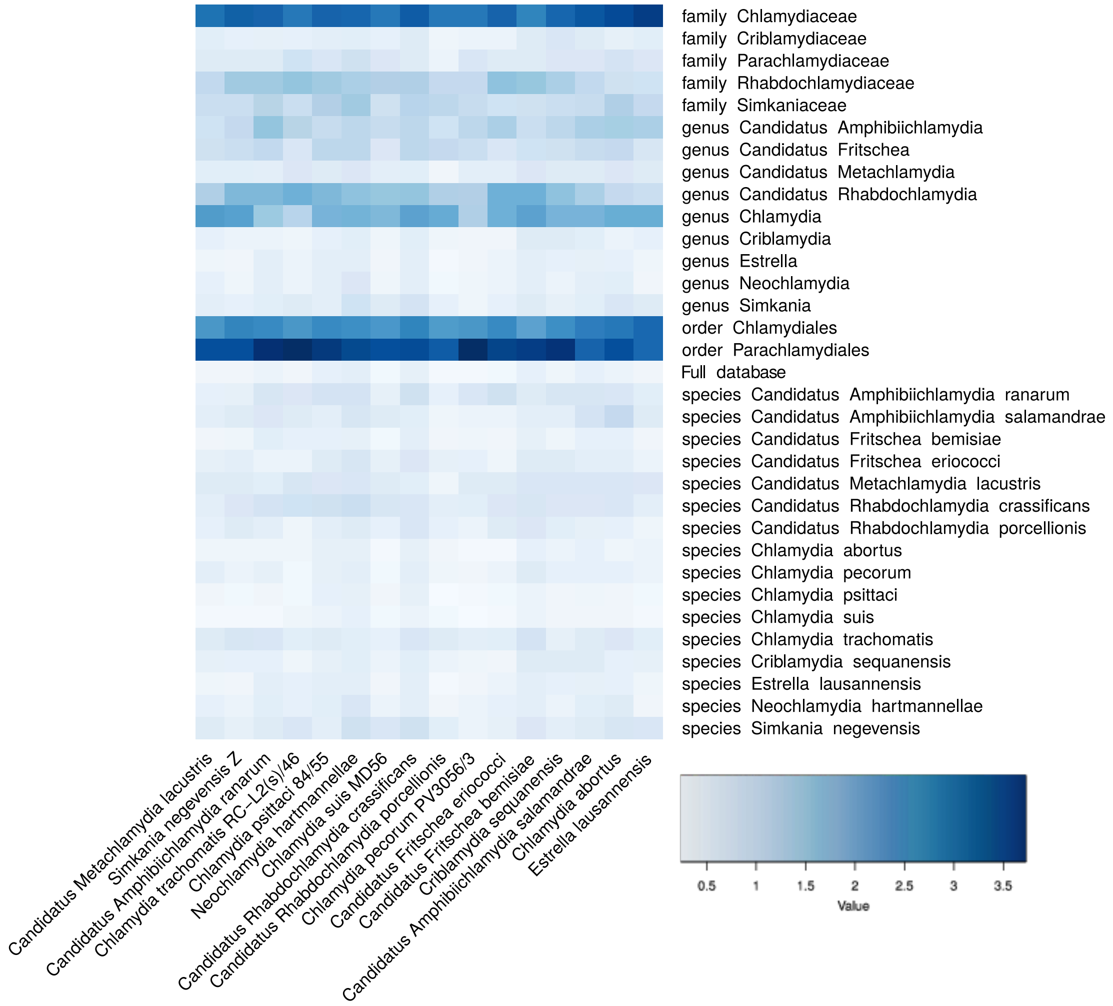


Figure 5.3: Taxonomic distance divergence of the *Chlamydia* dataset when the reference database is incomplete. Difference between the average TND per each species (x axis) when representatives of ranks are missing from the reference database (y axis). Note that Full database is when all the sequences are included in the reference database. In an ideal world, TDD should be 0. However, data shows that TDD is often above this desired level: it is low when a single species is missing, and high when higher taxonomic ranks are not represented in the database. This indicates that metagenomic classification is biased towards assigning reads to already known sequences and does not consider database incompleteness.

5.3 The effect of reference database choice on meta-taxonomic identification

The genomic content of the reference database is key for metagenomic taxonomic identification. The main goal is to compare the effect of the quality of the genomic sequences present in the reference database.

In order to study the behaviour of meta-taxonomic classification depending on the genetic content only, the MGC data, generated in chapter 3 section 3.5.1 on page 44, was analysed with Kraken2 and several subsets of the Silva database (described in chapter 1 section 1.6.3) mapped to NCBI taxonomy was used as reference: NR99, NR99Trunc, Ref, RefTrunc, Parc, ParcTrunc, ParcClean and ParcTruncClean (see chapter 2 section 2.2.2).

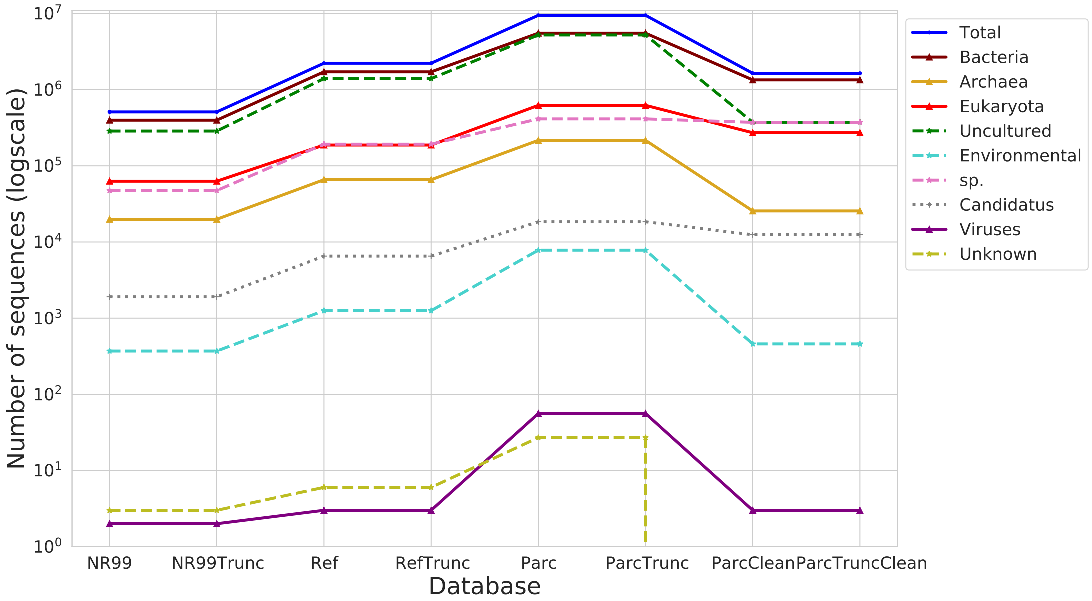
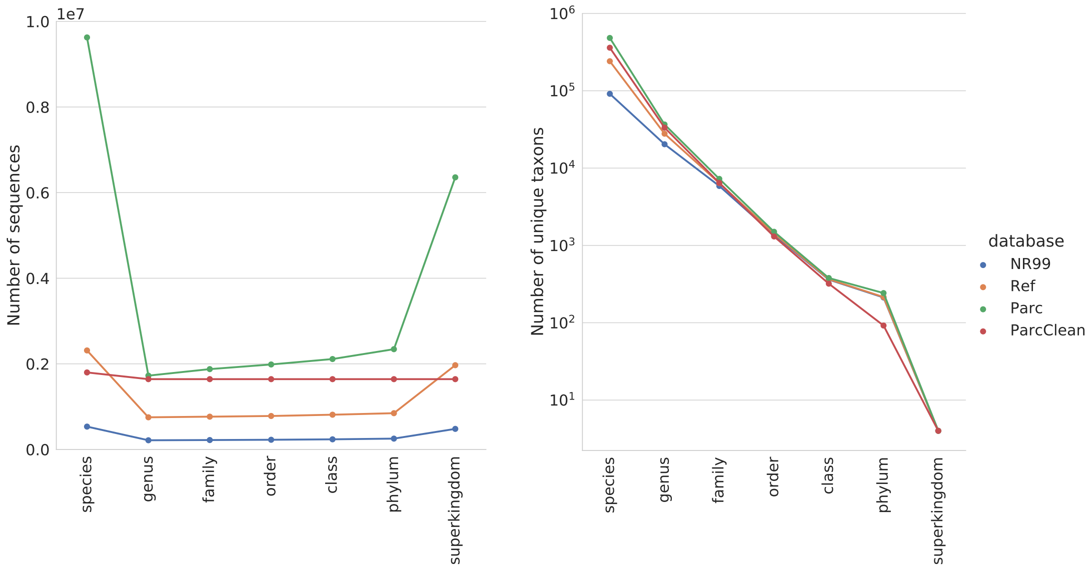


Figure 5.4: NCBI taxonomy content of the Silva SSU database, version 138.1. Number of sequences with each label. Dashed lines are labels at rank species, while solid lines are at superkingdom level (except for total, which are the total number of sequences). The vast majority of sequences are Bacterial, followed by Eukaryotes. There is a high number of uncultured sequences. Despite being a 16/18S rRNA database, there are between 2 and 51 sequences with viral lineages once mapped to NCBI taxonomy.

Understanding the content of the reference database and the data will help to interpret the results. As shown in figure 5.4, most of the sequences in the databases are bacterial. The taxonomic content of the truncated databases (noted as "Trunc") is exactly the same as their non-truncated counterparts. A high proportion of genetic sequences are labelled as environmental or uncultured, with poorly supported lineages. The number of sequences annotated for new species candidates ('sp.') is similar to the number of Eukaryotes present. Surprisingly, there are a few sequences classified as viral and viruses, even though these should not be present as they do not have the 16/18S rRNA gene. It was decided to include them anyway. This enabled the measurement of the effect of these types of sequences on the classification of metagenomic reads.

The number of sequences with annotated taxa at each taxonomic rank is shown in figure 5.5 as well as the number of unique taxon nodes. The Parc database has a much higher number of sequences annotated only at species and superkingdom ranks compared to the rest of subsets.



a: Number of sequences annotated per rank. **b:** Number of unique taxons per rank (logarithmic scale).

Figure 5.5: Silva database with NCBI taxonomy content by rank. **a** Number of sequences with annotation at each taxonomic rank by subset. The number of sequences annotated at ranks other than species and phylum is generally lower. Except for ParcClean, which is the same for all the ranks and Parc which contains disproportionally higher number of sequences with annotations at only species and superkingdom rank. This indicates that the curation of the taxonomic lineage is poor for a significant proportion of sequences. **b** shows the decreasing number of unique taxa per rank.

The vast majority of MGC simulated data have sequences representing the organisms in all the databases, as observed in figure 5.6. Twenty taxa are found in Ref and Parc only. One hundred and thirty-six organisms do not have any representative in the ParcClean database, and 237 are missing from NR99.

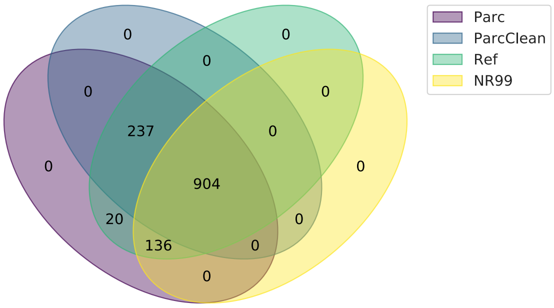


Figure 5.6: Number of taxa in common in the databases. Number of original taxons from the simulated data present in the different databases. Most of the taxa are shared across databases.

5.3.1 Mismatching taxonomic annotations

An unexpected discovery was made while mapping and comparing the Silva database taxonomy, which only contain 16S and 18S rRNA marker sequences, against the NCBI taxonomy database. Fifty-one sequences with virus species names and 29 as viral meta-genome were detected in the Silva database SSU Parc version 138.1 belonging to seven different lineages, of which 56 corresponding NCBI taxonomies were labelled as "viruses" superkindgom (see table B.1 in the appendix). Viruses do not contain 16S nor 18S rRNA genes. This leads to the worrying and challenging issue of detecting other types of inconsistent or wrong labelling for cases where it is not so obvious.

According to the Silva team[1], the species name of a sequence should be discarded when it does not match the rest of the taxonomic path. However, the sequences themselves should be kept as they provide unique genotypes [Robeson et al., 2021]. INSDC sequences and their taxonomic annotations are owned by the submitter, and include old legacy sequences, while the Silva database re-assigns taxonomic lineages according to their own system after selecting the 16/18S rRNA by machine learning approaches. Therefore, it is more likely that the latter annotation is correct, despite the species name being wrong.

The three examples illustrated in table 5.1 were investigated in greater detail. The first sequence (AF065755.1.676) was blasted using the web NCBI blast service with default parameters [2]. The vast majority of top results show similarities with the bacterial 16SrRNA partial sequence of Ochrobactrum species (shown in the table B.2 in the appendix), in agreement with the Silva taxonomic annotation. The sequence originated from a human clinical study [Martin et al., 1994] in 1994 where Cytomegalovirus was detected and confirmed. The symptoms described in the original paper are similar to the ones caused by organisms in the Ochrobactrum clade [Brady

[1] I communicated the mismatches found via email. In here there is a summary of their response.

[2] https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch &LINK_LOC=blasthome on 05/03/2021

Table 5.1: Examples of inconsistent Silva-NCBI taxonomic annotation in three Silva database genetic sequences. Silva database assigns the original species and subspecies names submitted by the author. This can lead to inaccurate taxonomic lineages (in red). This table only presents 3 examples. Their corresponding taxonomic lineage in the NCBI taxonomy database suggests that in some cases the annotation at the species level is in fact inaccurate (in green), or not so clear.

Silva accession	Silva taxonomy	NCBI taxonomy
AF065755 .1.676	Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiaceae; Ochrobactrum; Stealth virus 1	Viruses; Duplodnaviria; Heunggongvirae; Peploviricota; Herviviricetes; Herpesvirales; Herpesviridae; Betaherpesvirinae; Cytomegalovirus
AE006468 .4394688 .4396232	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella; Salmonella virus Fels2	cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella; Salmonella enterica; Salmonella enterica subsp. enterica; Salmonella enterica subsp. enterica serovar Typhimurium
BCRZ01001786 .50.1614	Eukaryota; Archaeplastida; Chloroplastida; Chlorophyta; Mamiellophyceae; Mamiellales; Micromonas; uncultured marine virus	Viruses; environmental samples; uncultured marine virus

and Leber, 2017] which are rare opportunistic human pathogens. It is possible that the sequence is the product of contamination during the DNA sequencing process.

For the second example, AE006468, 2 versions of the entry are found in the ENA archive. The original 2001 publication [Michael McClelland, 2001] sequenced the full genome of *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2. The genomic sequence contains presumably integrated viruses. One possibility is a potential issue with Silva's mapping to NCBI taxonomy ID for these type of cases.

The third sequence, BCRZ01001786, is searched by similarity (blastn), 122 hits are found, all against Eukaryota. The submitter study's [Nishimura et al., 2017] objective was to discover and characterise viruses in marine Eukaryotes. It is possible that the sequence might have been incorrectly mapped to the host 18S rRNA gene.

5.3.2 Quality filtered reference databases performance

The overall best performing databases were the quality filtered sub-databases, as figure 5.7 shows: databases filtered by well-annotated lineages (ParcClean and ParcTruncClean) followed by the NR99 and NR99Trunc, which are the strictest quality filtered set of Silva.

Next, individual ranks within each database were evaluated. To do this, the metrics of precision, recall, f1 score, false positive rate, accuracy and specificity were calculated for each taxonomic rank. As shown in figure 5.7 classification is better

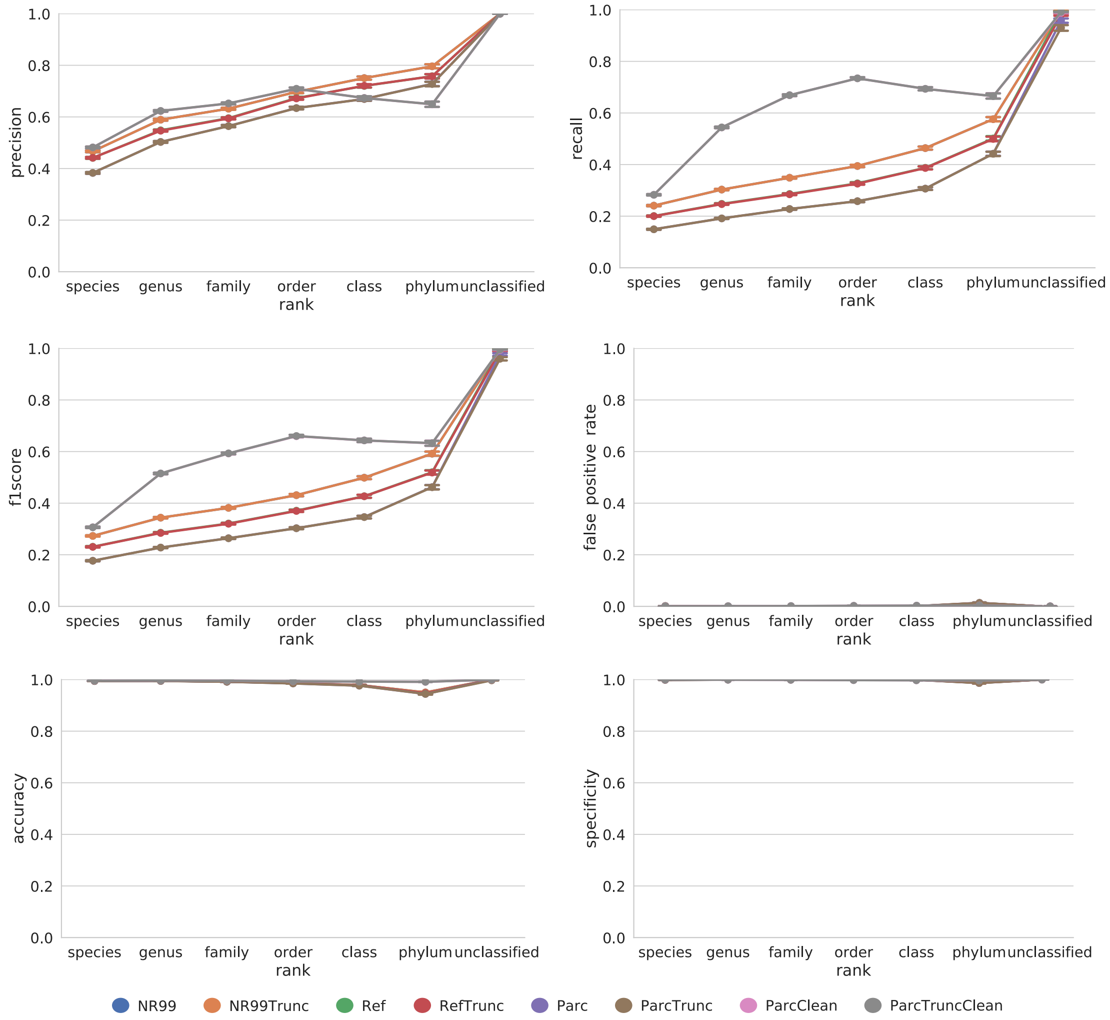


Figure 5.7: Metrics by database. Average of each metric with the corresponding confidence interval at 95%. The full databases and the corresponding truncated databases are overlapping. The classification varies significantly across the databases. The overall best performing are the cleaned versions of Parc and Parc Trunc. This might be due to the selection of better annotated lineages and removal of noisy sequences from the databases.

at higher ranks (e.g. Phylum) compared to lower ones (e.g. species). The recall and the f1 score are much higher for the ParcClean and ParcTruncClean databases compared to the others. However, precision for these two databases is a bit lower at the phylum level. The truncated databases and the corresponding non-truncated ones show identical performance (overlapping lines in figure 5.7).

In order to determine how close the classification between the reference databases and their truncated subsets were, the Spearman correlation of the scaled TND was calculated (see appendix figure B.7). A perfect correlation can be observed between the full database and their truncated counterparts (referred as "Trunc"). For this reason, subsequent TND plots are showing only the four non-truncated databases. It is also observed that ParcClean and ParcTruncClean are the ones that differ the most

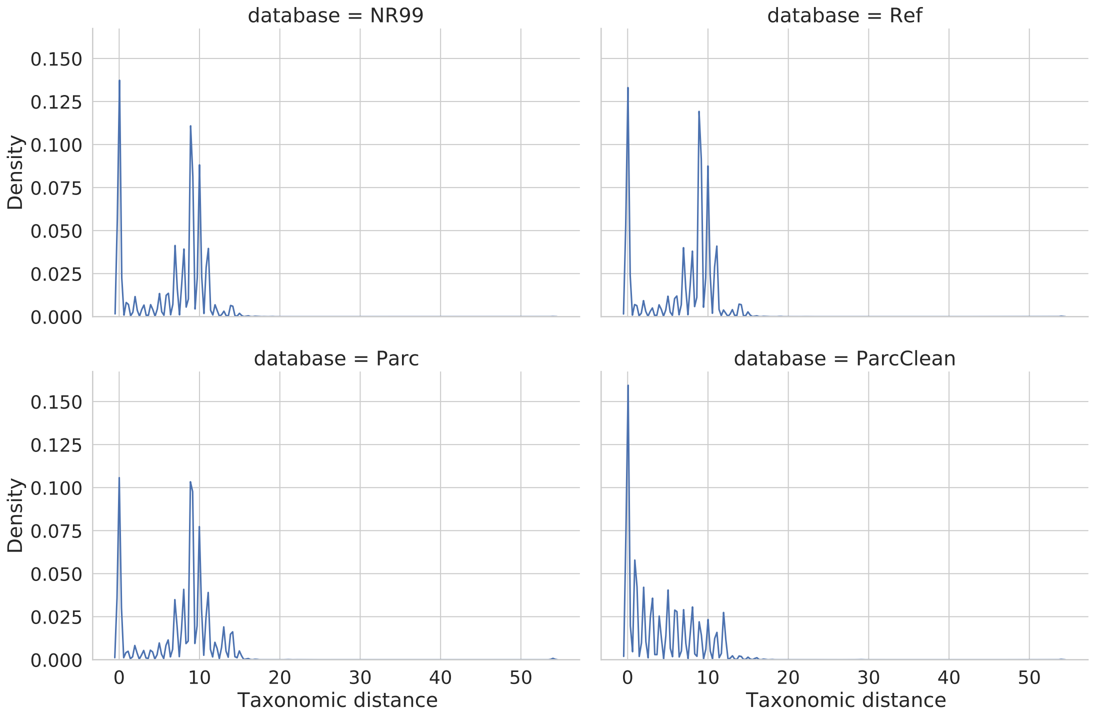


Figure 5.8: Kernel density estimation of the TND per database. The distribution of TND by database is quite similar in NR99, Ref and Parc, showing two relative maximum peaks, one at 0 and another one around 9. ParcClean present the best results with higher density of values closer to 0. This demonstrates that taxonomic annotation has a huge impact on metagenomic read identification.

from the rest of the databases, as previously seen.

To better understand differences of the resulting TND amongst databases, the density of TND was estimated. The Kernel Density Estimation (KDE) plot 5.8 shows that the TND distribution is almost identical for the databases NR99, Ref and Parc. The cleaned versions of the Parc database show a distinct profile with much lower TNDs, but there are still short simulated sequences classified far from their origin. This highlights the importance of highly curated taxonomic lineages associated to the genomic data.

5.3.3 Impact of short reads mutations and sequencing errors

The MGC data is formed of eight sets of different levels of mutations ranging from none to 3% or random mutations with or without sequencing errors. The exploration of the behaviour of these sets with the different databases will help to understand what happens when short reads fall within mutated regions, and also the effect of sequencing errors.

The sets without any mutations (0% mutations no sequencing errors) were not correctly identified in all cases, although presented better performance scores (see

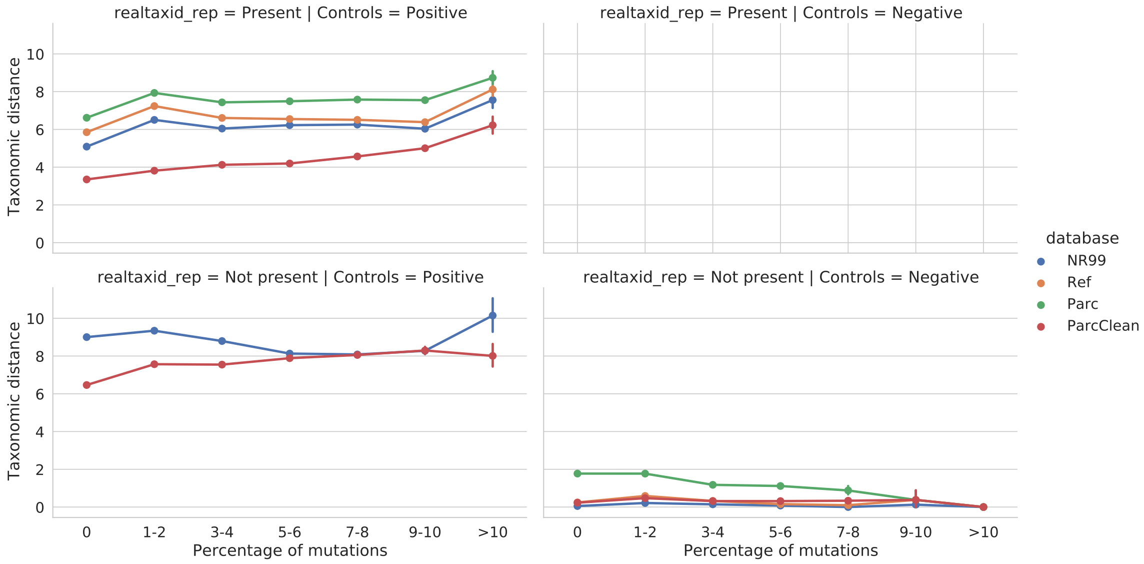


Figure 5.9: TND by database and proportion of mutations of controls depending on the presence of the original organism. Average TND with 95% confidence intervals separated by database, controls and proportion of mutations in the dataset. The TND increases when the original species genome is not present in the database. The ParcClean databaseTND is overall lower and rises with the proportion of mutations from the simulated data. However, this growing tendency is less clear for the rest of the databases.

appendix B figure B.8). Generally, f1 score values decrease with the accumulated proportion of mutations in the MGC data. However, the difference of f1 score values among the mutated datasets was lower for the Parc database compared to Ref and NR99 (as well as their truncated counterparts), and it is almost overlapping for ParcClean.

It was found that TND shows different tendencies with the number of accumulated mutations and errors depending on the type of control. The Spearman correlation for the positive controls ranges from -0.089 to 0.078 (see figure B.6 in the appendix). When the data were further separated by type of control and whether short reads have any representative in the reference database, as shown in figure 5.9, those whose representative is missing present higher overall values of TND, although without any correlation with the proportion of accumulated mutations and errors. The ParcClean database presents a lower TND compared to the rest of the databases studied here.

The negative controls without any mutations in the simulated data tended to classify around 2 nodes away from their original taxonomic IDs in the Parc database, and the distance decreased with a higher proportion of mutations. Whereas for the rest of the databases in the negative controls, TND was almost flat at 0. This suggests that more uncertainty may be introduced in the largest Parc database for classification purposes. Parc database is the largest studied here, with about 6 million sequences, and it is the Silva set with more relaxed quality controls.

5.3.4 The taxonomic classification success varies among samples

Each sample has a different taxonomic composition. Figure 5.10 shows that the classification success varies greatly per sample between the different databases. For example, while the f1 score of the ParcClean and ParcTruncClean databases for gut and plant is very high compared to the others, this is not so clear for the ice samples. The number of representatives of each taxon might have an influence and are this analysis is presented in the next section.

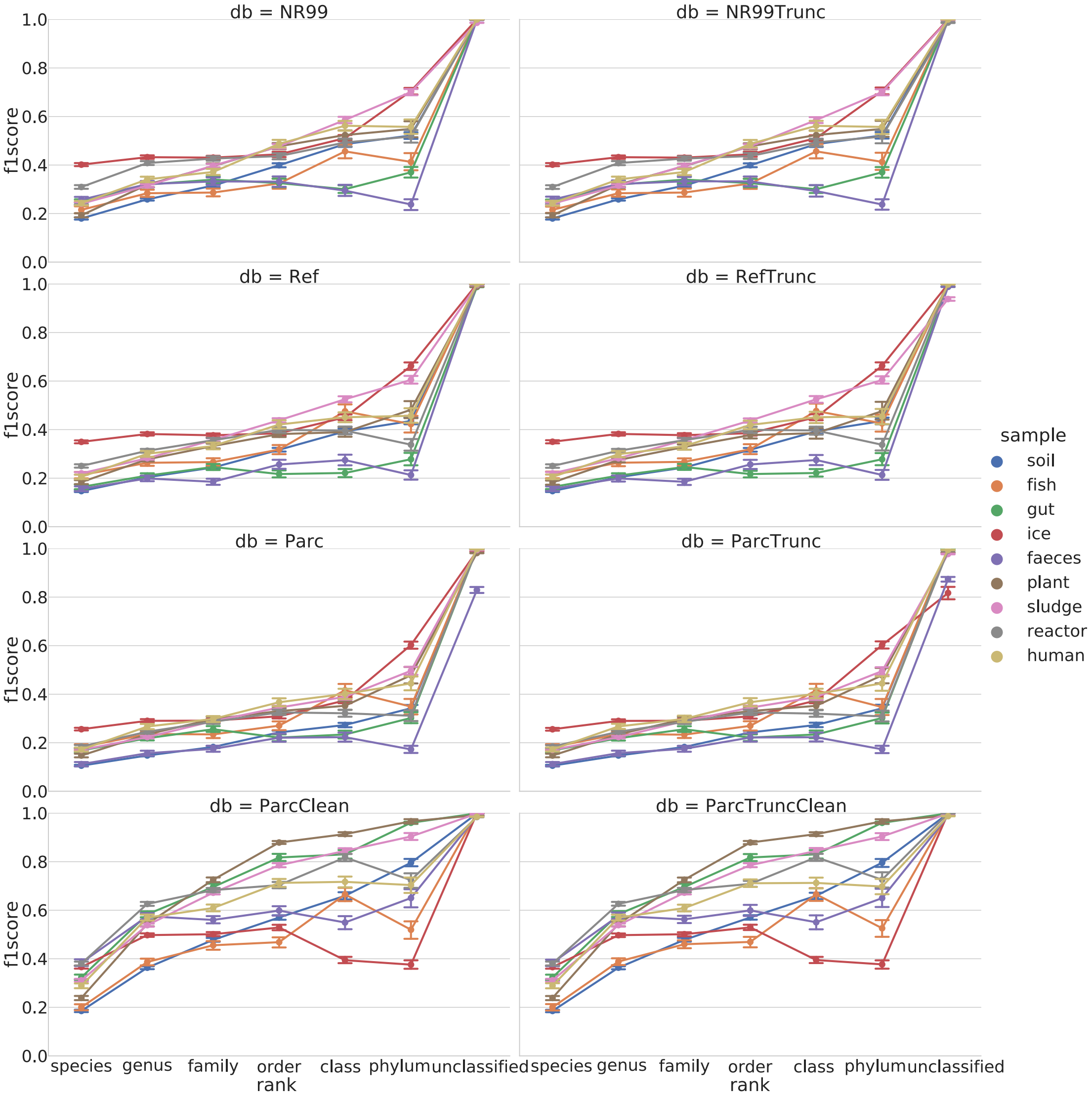


Figure 5.10: F1 score by sample and database. The f1 score is highly variable depending on the sample and the database. There are no differences between databases and their truncated counterpart (denoted as "Trunc"). The databases with lineages that are annoated at least at the 7 main taxonomic ranks ("cleaned") performed better overall, but not consistently. This is a clear example of the importance to evaluate the fitness of specific pipelines given the data object to study.

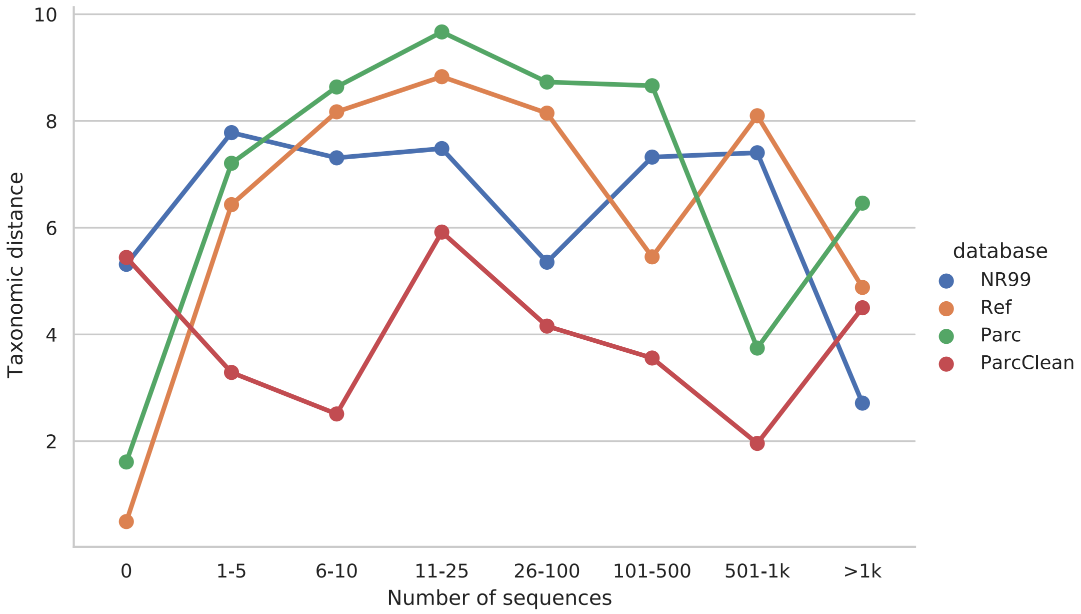


Figure 5.11: TND variation by the number of sequences representing each species. The number of sequences representing organisms in the different reference databases are grouped. Those species with "0" sequences included the negative controls. Some taxa simulated int he MGC is missing from the databases NR99 and ParcClean. The TND fluctuates, and overall ParcClean present lower values.

5.3.5 Number of sequence representatives per taxon

Each species may have one or more sequences present in the database, while some are clearly overrepresented like human pathogens, others might be entirely missing. The average TND for each organism is shown in figure 5.11 with organisms grouped by the number of sequences representing them in the database The data fluctuates and there is no clear tendency. The data analysed with the ParcClean database showed a different pattern from the rest. For the databases of Ref and Parc there is a maximum average TND for those species contains between 11 and 25 sequences. This result might mean that perhaps having one single representative is already enough for identification and adding more sequences does not necessarily help.

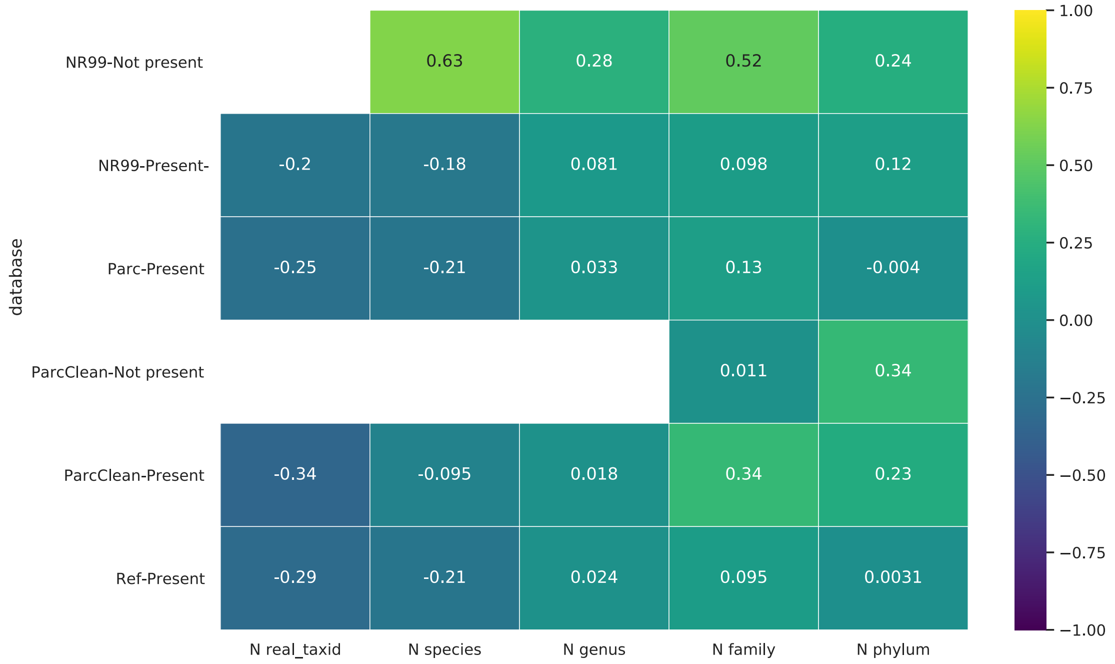
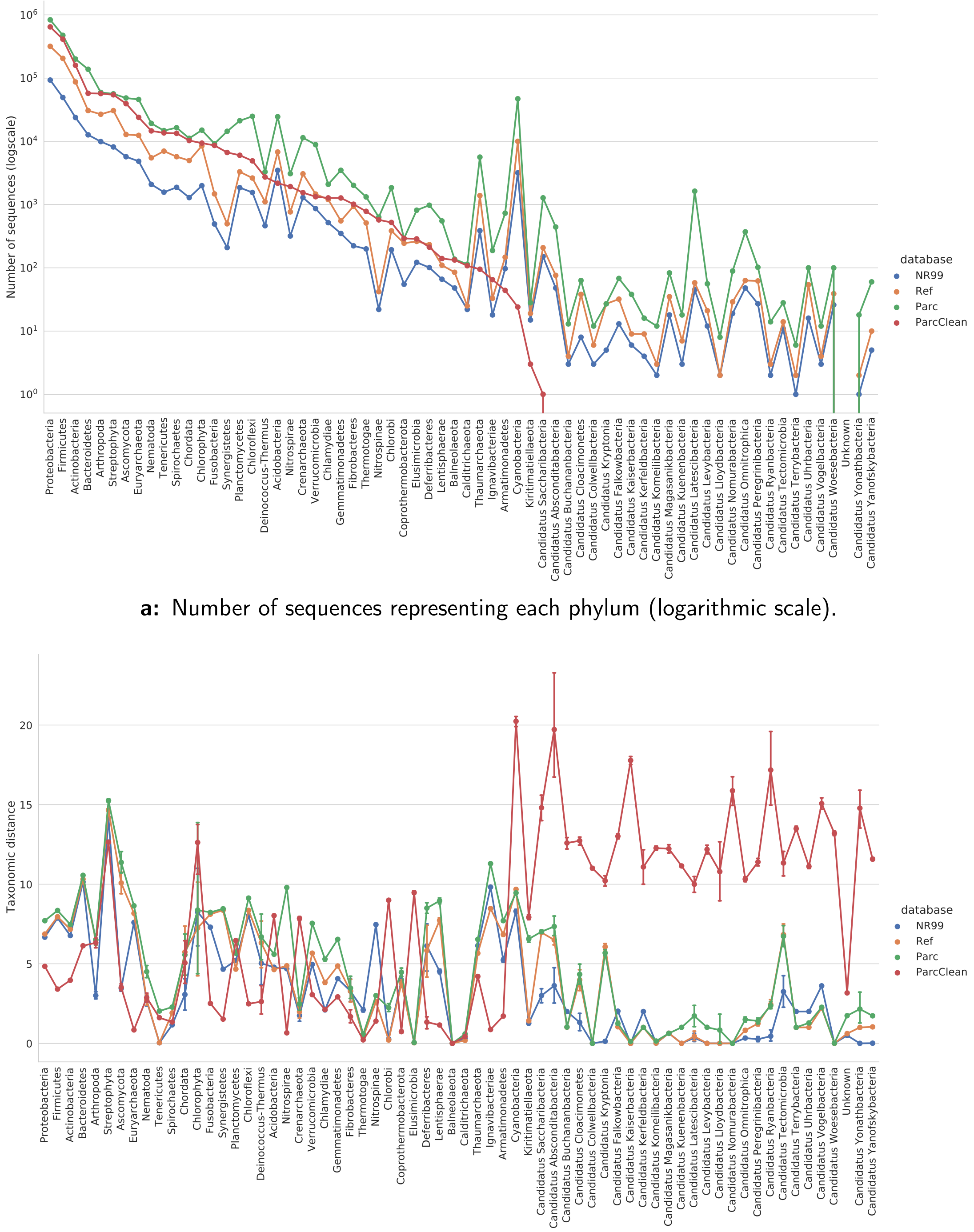


Figure 5.12: Spearman correlation of the TND with the presence or absence of the original taxon. N real taxid is when the original organisms are present, "N" means the number of sequences present at the different ranks. When the original organism is present it improves classification (negative correlation) and when it not present TND increases with the number of sequences. However, ParcClean database even when the original organism is present, the correlation is positive for the number of sequences at family and phylum levels.

The effect of the number of sequences at different ranks is studied depending on whether the original organism have any sequence present on the database. The Spearman correlation is calculated. The correlation shown in 5.12 is positive if TND increases with a higher number of sequences, and is negative if TND decreases with a higher number of sequences. A greater number of sequences representing species or subspecies improves classification (negative correlation) when the organisms are present in the database. Having more sequences at the genus level and above does not have any impact for taxonomic classification when the original organism is present except for ParcClean which presents positive correlation values for the number of sequences at the family and phylum level. The NR99 database without representatives presented the highest positive correlation values for species and family.

TND values fluctuate with the number of sequences at each phyla for all databases. Databases NR99, Ref and Parc TND tended to have lower values for those phyla with fewer representatives. Interestingly, it was observed that those phyla without any sequence representatives at this rank were labelled in taxonomic nodes much further away from their origin (right side of subfigure 5.13b). This observation led towards studying the classification for missing organisms from the database depending on how far the closest relative is, in detail in the next section 5.3.6.



a: Number of sequences representing each phylum (logarithmic scale).

b: Average TND by phylum.

Figure 5.13: TND by the number of sequence in each database per phylum. Phyla are order by desdendent number of reprsentatives at the ParcClean database. **a** shows the number of sequences per phyla and **b** shows the average TND. The TND fluctuates for all databases, but for NR99, Ref and Parc TND is generally smaller for those organisms with fewer sequences representing thier phylum. Those phyla without any sequences representing them tend to have much higher TND values, which is only observed for the ParcClean database.

5.3.6 Identification effect depending on the lowest common ancestor available

As already seen, when the original organisms have no sequence present in the reference database, classification tends to occur further away than might be expected (see section 5.2). Two databases have missing taxa from the original MGC simulated mock community: NR99 and ParcClean (see figure 5.6). This section aimed to investigate the effect on classification depending on how far the closest taxon is from the leaf.

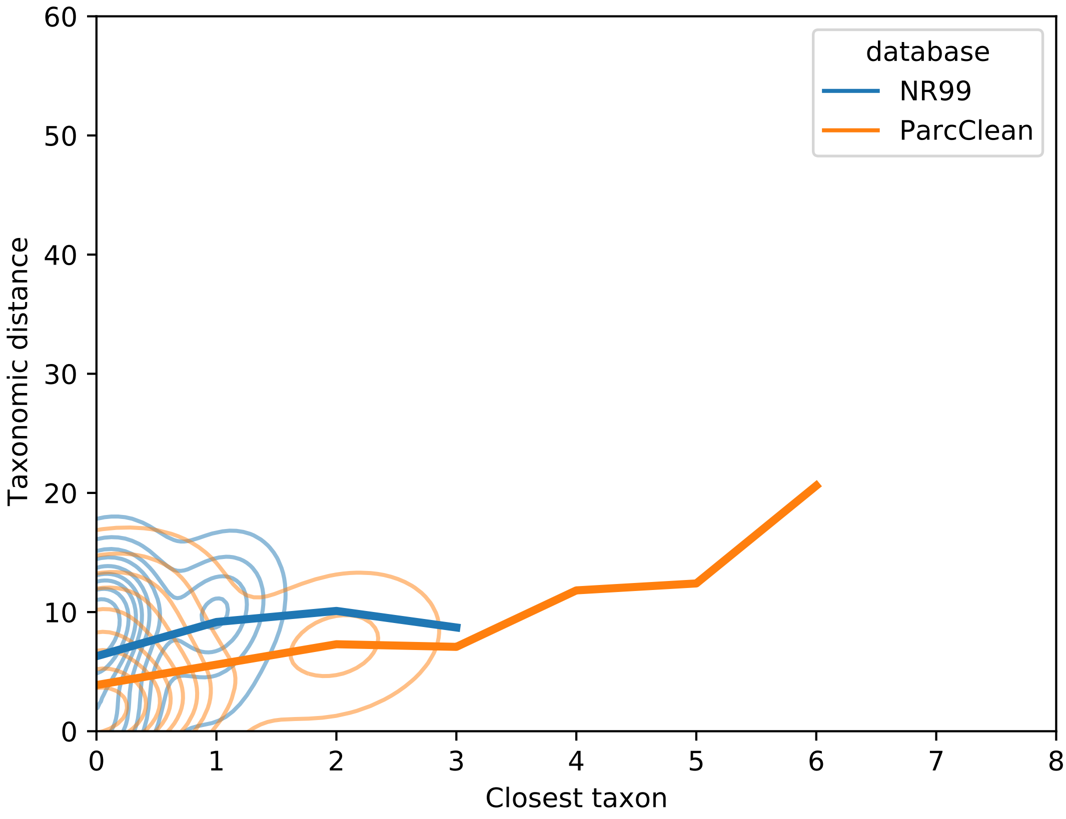


Figure 5.14: Relationship between the TND and closest taxon of the positive controls. The thick lines show the mean TND seen for each closest taxon. The contours represent the full distribution of observed values. In all cases the average TND is higher than might be expected (closest taxon). When the closest taxon is found 6 nodes away, the average TND is abut 3.5 times higher than expected.

The closest taxon is the node or rank[3] in a lineage which contain sequences representing it (similar concept to the expected TND introduced in section 5.2 and shown in figure 5.2). Imagine that the species and genus levels of a lineage do not have representatives, but the family does, then the family rank is the closest taxon. And it can be numerically expressed by counting the number of nodes between the

[3] "rank" and "node" are used as synonyms in this context.

original organism and the closest taxonomic rank.

The taxonomic identification of those organisms which are not present in the database tends to occur on average much further away than expected. For example, as figure 5.14 shows, the NR99 database whose closest taxon is 2 nodes away, the average TND is around 10, about 5 times higher than expected. The results also show that classification occurs much further away than expected when higher taxonomic ranks are completely missing, highlighting the importance of the need to use reference databases that are as complete as possible. The Spearman correlation is 0.31 for both database (see figure B.6 in appendix).

5.3.7 Number of nodes in the taxonomic lineages

Taxonomic lineages can have a highly variable number of nodes, referred to here as *lineage length*. Therefore, this might be a potential bias factor for classification for methods that are based on the lowest common ancestor, as Kraken is. Figure B.9 shows the number of sequences for each lineage length by database. ParcClean subset contains lineages from 8 taxa upwards, whereas for the others they start at 5 taxa. The longest lineages belong to Eukaryotes, whereas the shortest are less studied species, which include uncultured bacteria and environmental samples.

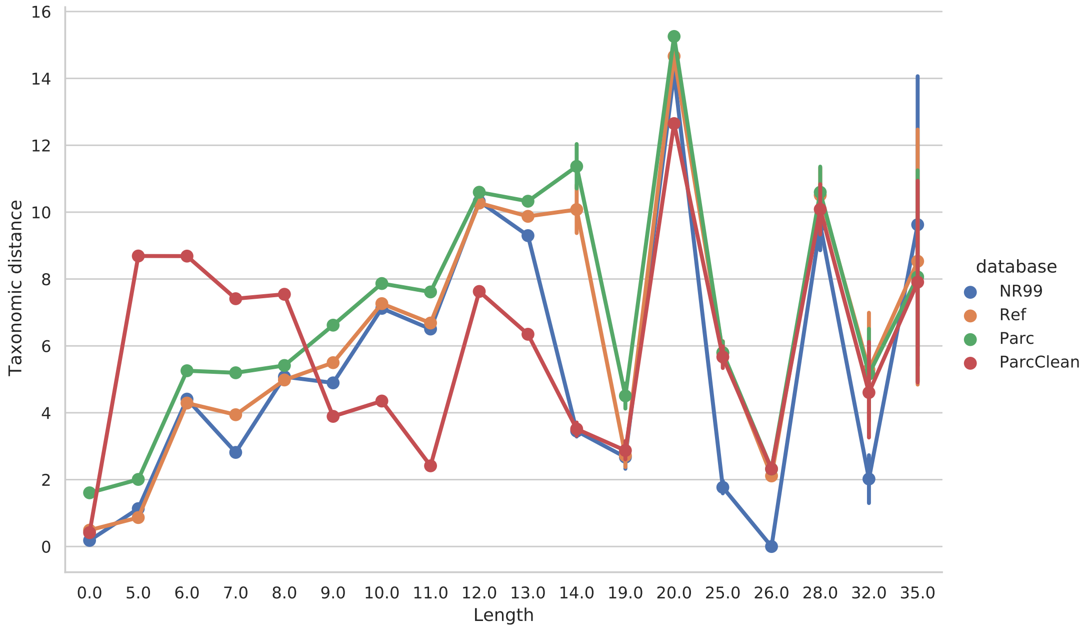


Figure 5.15: TND by lineage length. TND increases for Parc, Ref and NR99 for lineages with up to 12 nodes. The ParcClean has a much higherTND for those lineages which have between 5-8 nodes as they are missing from the reference database. For longer lineages from 14 nodes upwards, the trends vary greatly in all databases and no obvious patterns can be observed.

For the ParcClean database, lineages of length 5 to 8 have higher TND values

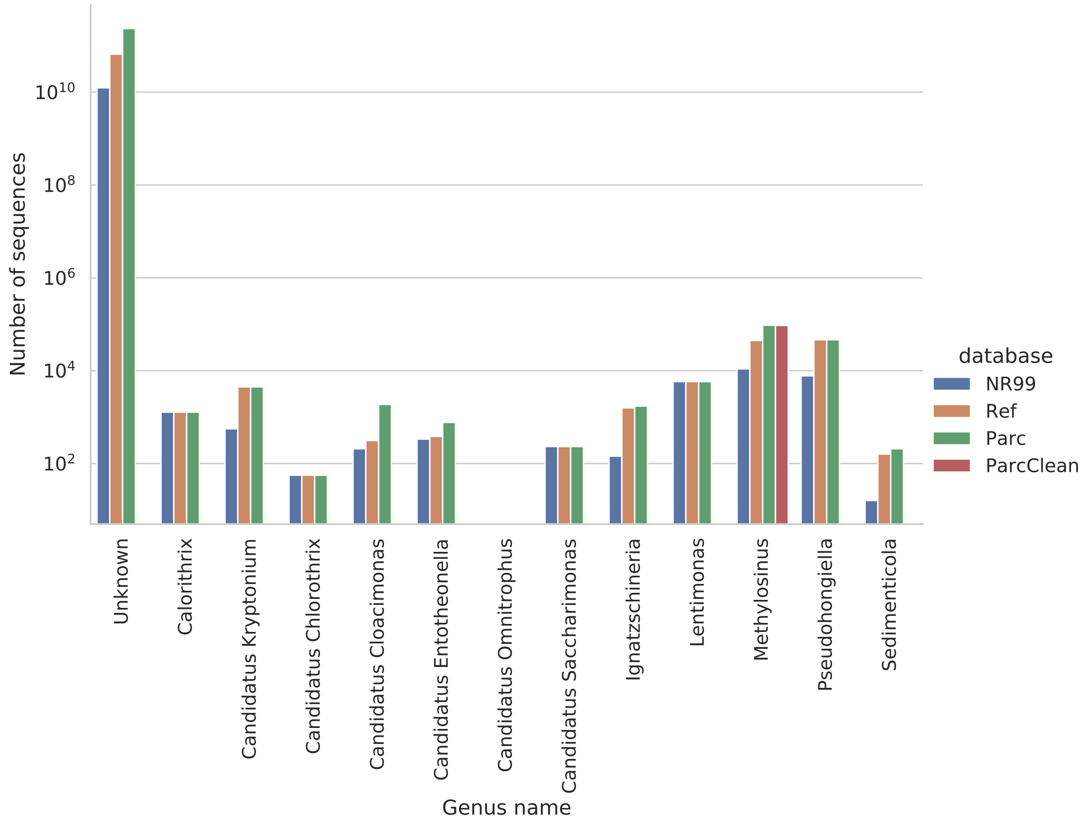


Figure 5.16: Number of sequences at genus level of taxonomic lineages with 8 or fewer ranks. From the lineages with up to 8 nodes, only 1 genus (Methylosinus) is present in the ParcClean database. Genus without assigned name (labelled here as "Unknown") are found in a large number of sequences (Parc, Ref, NR99).

compared to the other databases, as can be observed in figure 5.15. This is because these lineages do not have the original sequence present, except for the Methylosinus genus (see figure 5.16).

For those lineages of length 9 to 12, which contain a relatively large number of sequences (see figure B.9), the TND has an increasing tendency for NR99, Ref and Parc databases while ParcClean, is lower than the others but fluctuates as figure 5.15 shows. For longer lineages that have fewer sequences the average value of TND varies but is more similar across all databases.

Next, the relation between the number of nodes in the taxonomic lineages, the closest taxon and the sequencing modality (single or paired end) are modelled to find trends for the TND values.

Lineages of the positive controls were classified into three groups: length 5-11, 13-19 and >20. A linear regression model was fitted, separating the data by database and sequencing modality (figure 5.17).

For the two shorter groups, TND has an increasing tendency depending on how far away the closest ancestor is (the closest taxon) for ParcClean (both modes) and NR99

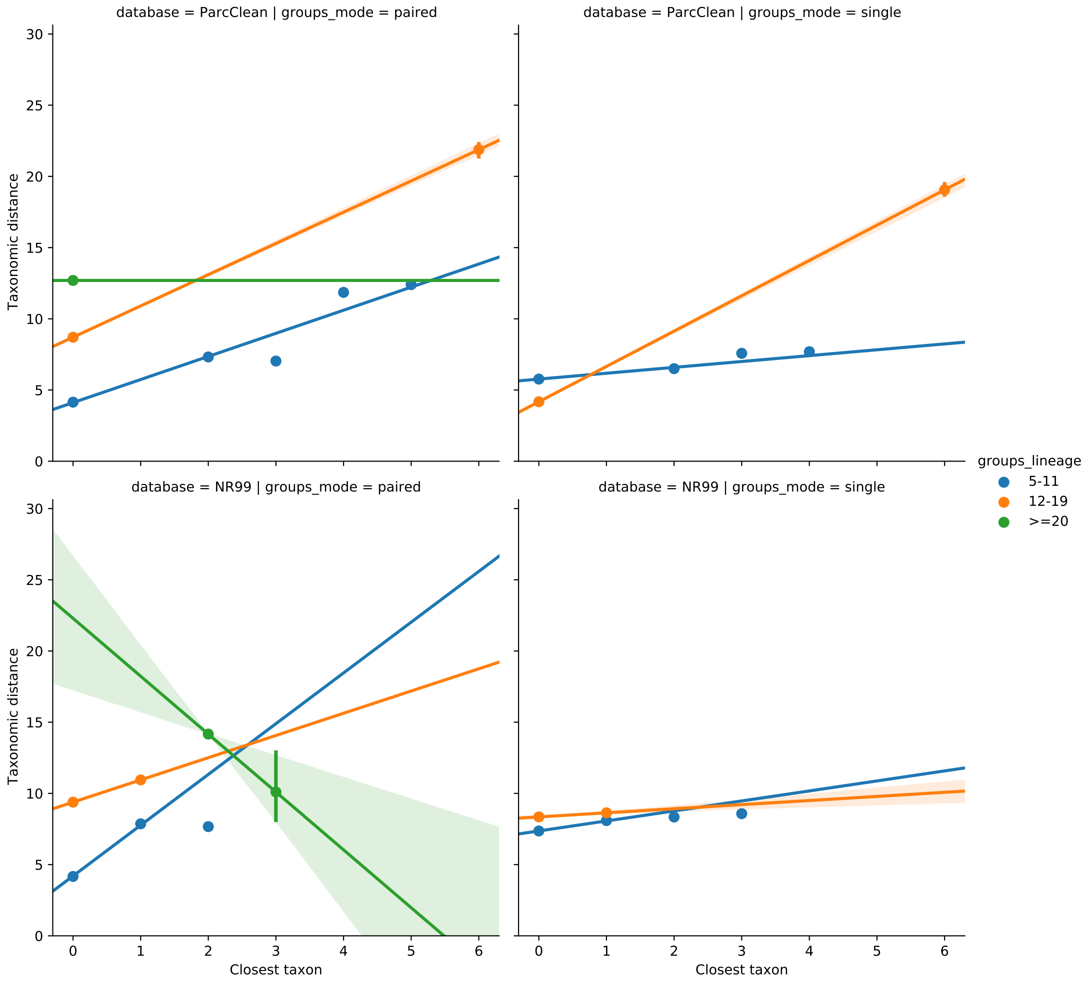


Figure 5.17: Linear regression of the closest taxon by lineage length and mode. Shorter lineages are presented in blue and orange. When lineages with 7 nodes or fewer are removed (ParcClean), the linear regression becomes flatter for the shorter taxonomies (blue) for both single and paired end data. Interestingly, lineages with 20 or more nodes have a negative correlation (although with high confidence intervals) for the paired data with NR99 database, meaning thatTND decreases with higher number of nodes per lineage.

paired-end. Interestingly, our data reveals that lineages with 20 or more ranks can be beneficial (decreased TND) for those organisms with far away common ancestors present in the ParcClean database paired-end data.

The results show that the removal of sequences with annotated lineages with less than 7 taxa improves classification for single data (top right plot on figure 5.17).

5.3.8 Effect of sequence length on taxonomic classification accuracy

For the simulated positive controls, the Spearman correlation of TND values associated with sequence length per database is almost 0, as the appendix figure B.6 shows. However, when the original taxon sequence is missing from the database, longer sequences tend to have a lower TND for the ParcClean database, with a Spearman correlation of -0.11, and for NR99, with a Spearman correlation of -0.23, as figure 5.18 shows.

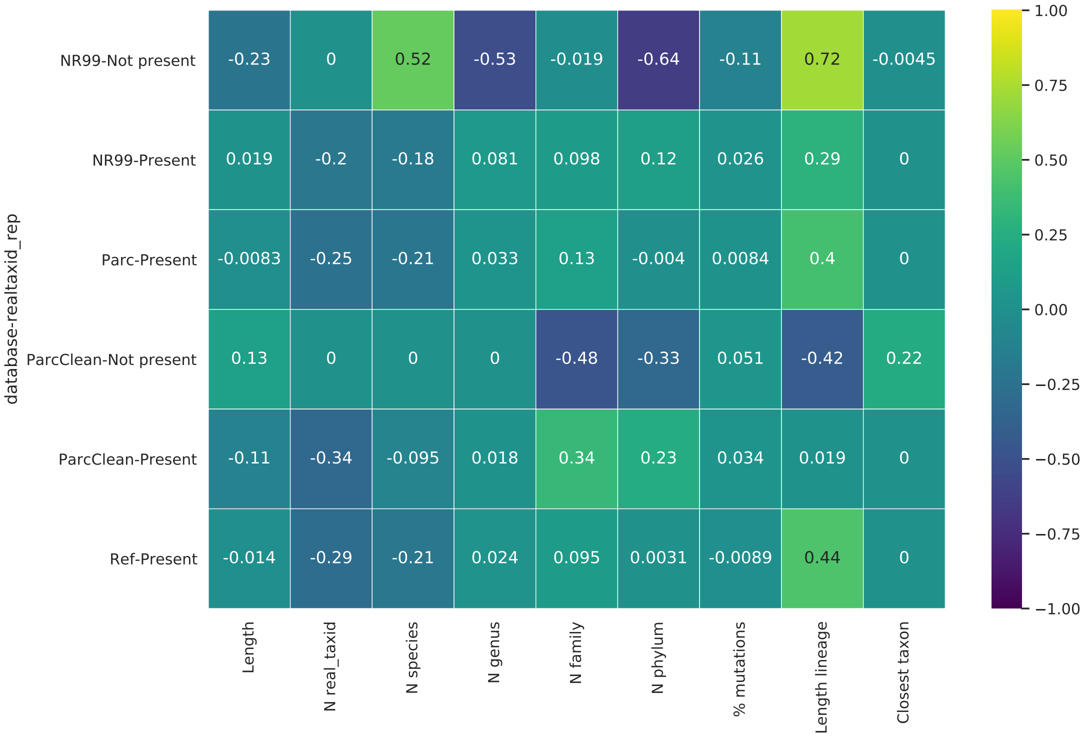
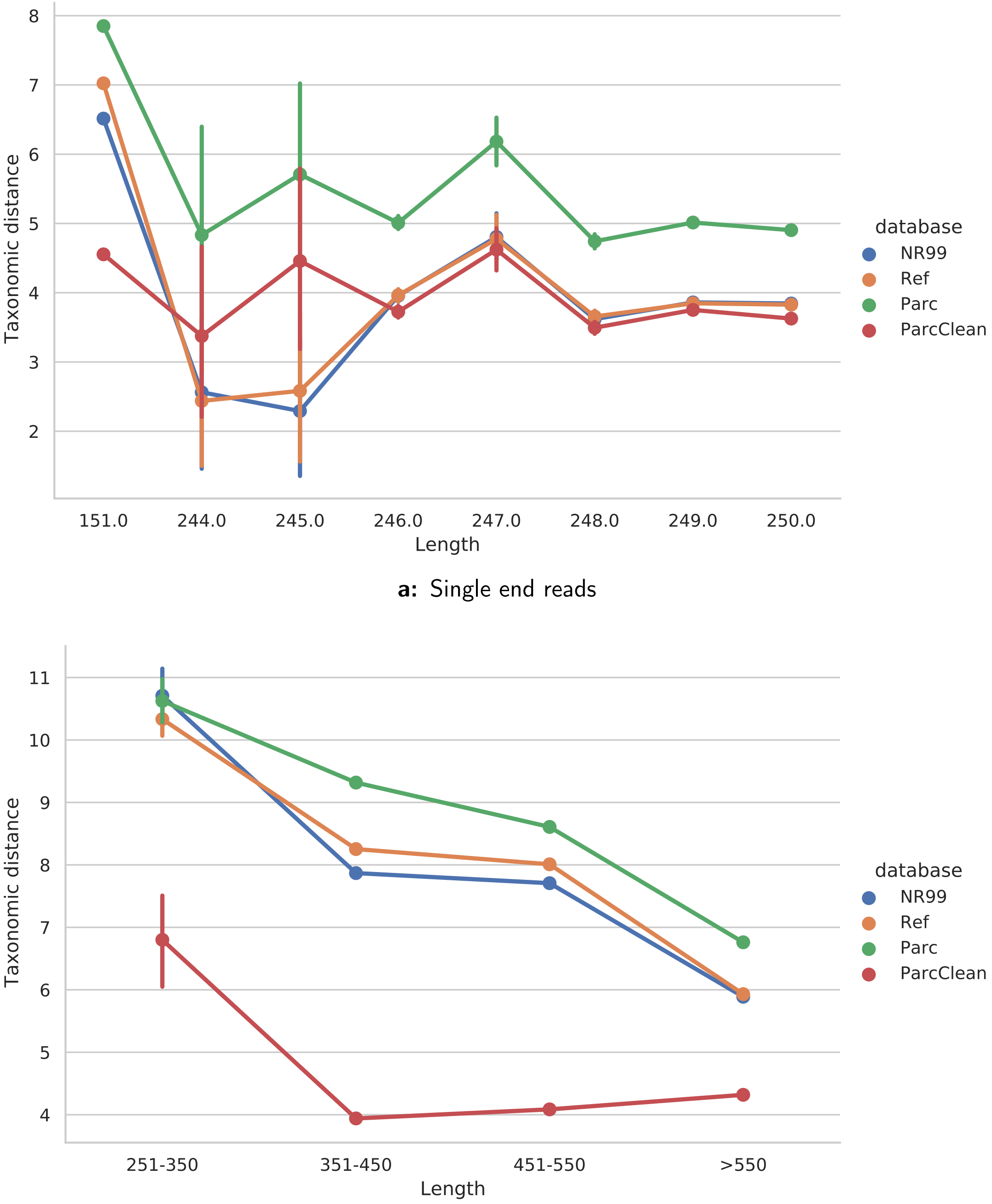


Figure 5.18: Spearman correlation by presence of the original taxon. The main factors that show negative correlation between TND and the databases, therefore overall better identification, are the number of sequences representing genus and phylum for NR99 when the original taxon is not present, and the number of family and phylum sequences as well as longer lineages for ParcClean when the taxon is missing from the database. Overall, longer lineages show a positive correlation with TND values, except for the already mentioned case of ParcClean.

Overall, the main factors that correlate with an increase in TND (see positive correlation values in heatmap B.6) are longer taxonomic lineages (except for ParcClean database) and how far away the common ancestor (the closest taxon) is in the reference. And the most significant aid to labelling closer to reality is the number of sequences representing the subspecies or species in the database being used (N real taxID and N species in figure B.6).

It was observed that due to the nature of the simulated MGC data, the effect of sequence length on taxonomic identification can be classified into two categories: single end ($\leq$250 bp) and paired end (>250 bp). Figure 5.19 shows lower TND values for single end than for paired-end sequence data, but with paired-end performance decreasing with increasing length.

The mode of analysis is studied in more detail in the next section, where correlation of several of the factors presented so far are jointly investigated.



a: Single end reads

b: Paired end reads

Figure 5.19: TND by mode and length. The single end data belongs to 2 sets: soil and fish, and the paired end data contains the rest of the dataset described in table 2.2. The TND decreases more for the paired end data compared to singe end. However, the paired end data show much higher TND values and ParcClean clearly tends to assign taxa closer to the original species compared to the rest of the database. For the single end data, there is no clear pattern of a combination of read length-reference database that performs better that others.

5.3.9 Taxonomic classification difference with single and paired-end sequence data

As already observed in this chapter, classification of single and paired-end sequences behave differently. Figure 5.20 shows the density of the TND data for each sequence type and database. The single end data have lower average TND values for all databases except for ParcClean.

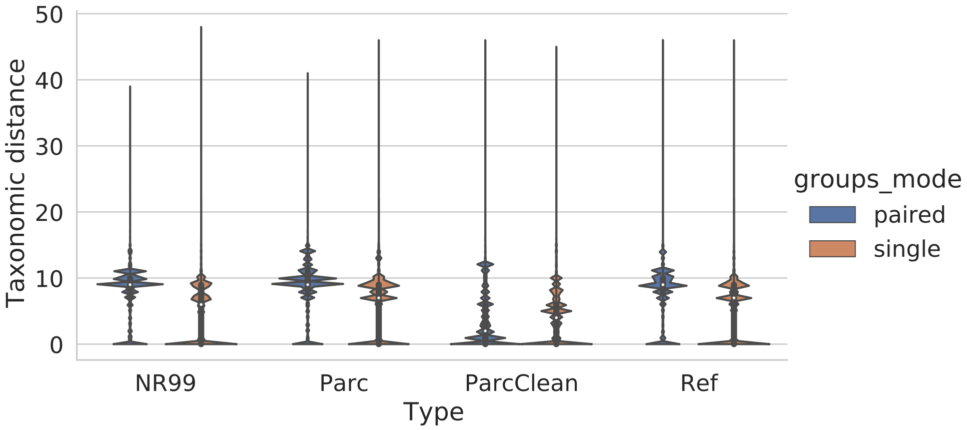
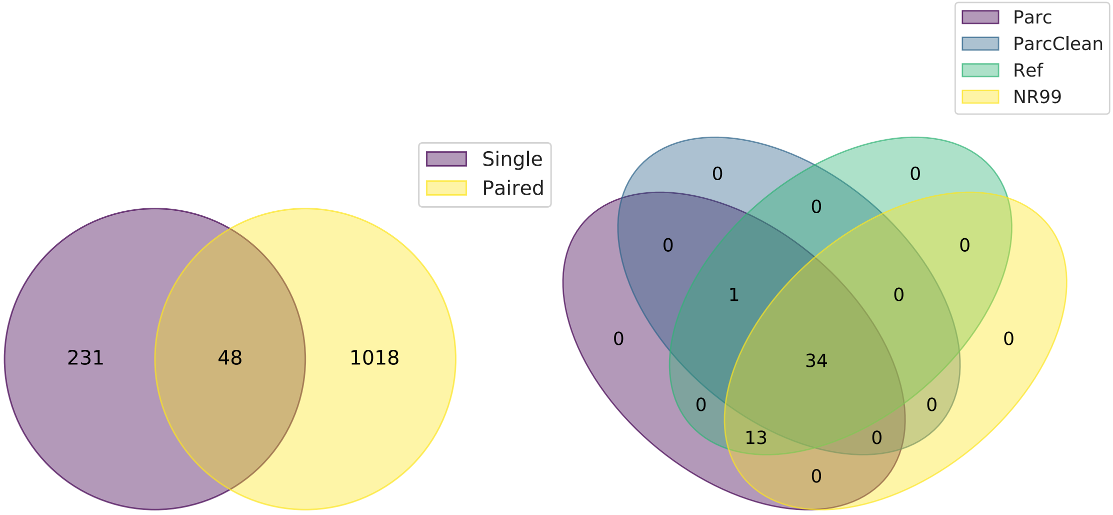


Figure 5.20: TND distribution single vs paired end by database. The TND is a bit lower for the singe end mode in all databases except ParcClean. This database also present overall lower TND.

This effect between the singe and paired-end was explored further. All the MGC data was split by mode (single or paired-end) in combination with multiple other factors previously described in this chapter. The correlation between TND values by database-type of data and multiple factors, including read length and number of representatives, was calculated (see appendix figure B.10). The main factors of the positive control data, whose taxon is present in the reference database, with negative correlation are the number of sequences per subspecies, species, genus, and longer lineages for single end data. Factors that increase TND are the number of sequences per phylum for the single-end and lineage length for the paired-end data.

There is a distinction for those reads which have sequences representing their taxon in the database compared to when it is missing. In the first case, some factors show opposite correlation between single and paired-end data, e.g. the number of genus representatives or the length of the lineage, others both type of data present a positive correlation, e.g. number of sequences representing a family. In the second case, TND differences between the type of data are less obvious.

Only two of the samples from the MGC generated data were single end reads (fish and soil) and represent distinct environments from the rest. From the total simulated data, forty-eight taxa are common for both, single and paired end data, as

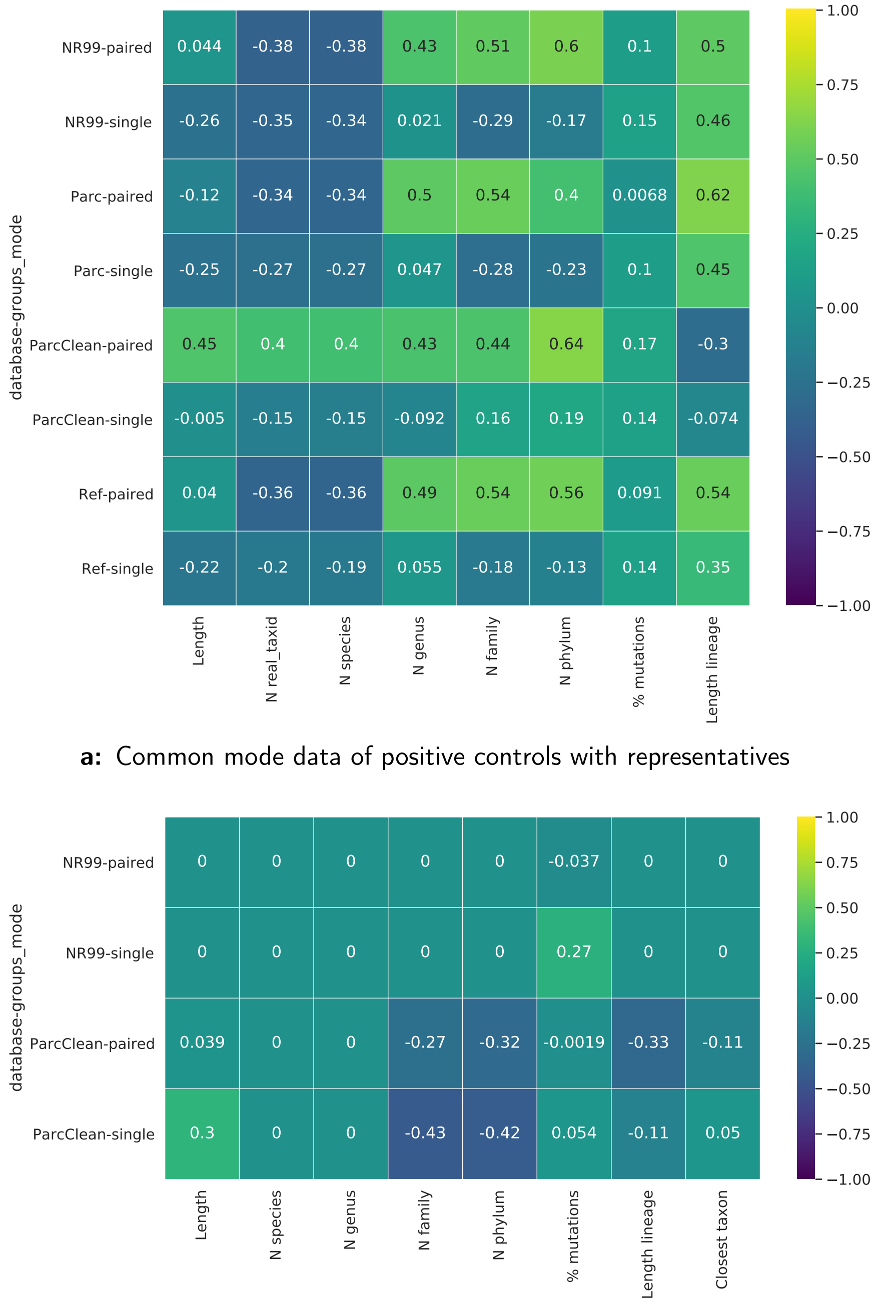


a: Taxa by mode

b: Common taxa by database

Figure 5.21: Venn diagram of the common taxonomic IDs between single/paired end data. The simulated MGC data contain 48 common taxonomic IDs (a) between single (fish, soil) and paired end. From the taxa in common in both modes, 34 are found in all databases (b).

figure 5.21 shows, with 3,499,240 simulated reads. However, this is a small subset of data belonging to different genomic regions and might still include biases. The Spearman correlation between the same previous studied factors and TND values for the different types of data was calculated and shown in figure 5.22. Most of the previous tendencies continue to be observed, for example single end data for number of species present in the database still is negative correlation, with exceptions e.g. the length of the lineage for both types, single and paired, show a positive trend (excluding ParcClean).



a: Common mode data of positive controls with representatives

b: Common mode taxa of positive controls without representatives

Figure 5.22: Spearman correlation of factors influencing taxonomic classification by mode and representatives. Spearman correlation of the TND of the positive controls where the original taxon is present (a) or not (b) in each database. **Common* refers to those taxa in common between paired and single end mode. N is the number of sequences per rank (species, genus, family, phylum) present in a given database. Length lineage is the number of taxa described in the NCBI database on 18 November 2020. The data shows almost no effect on TND with accumulated mutations. When representatives are present, ParcClean paired end mode show positive correlation except for lineage length. Except for ParcClean, the lineage length have some of the strongest positive correlations jointly with the number of sequences representing genus, family and phylum for the paired end mode. The ParcClean database show different tendencies compare to the rest, indicating the enormous effect of curating annotations of the genomic data used as reference. Also, the effect of some factors can be opposite in terms of taxonomic identification of short reads depending on whether the organism has any representative in the reference database.

5.4 The effect of taxonomic tree selection on classification

Here, the impact of the taxonomic annotation was studied by comparing the NCBI and Silva taxonomy with the same exact genomic content, the Silva NR99. As mentioned earlier, Silva database do not curate taxonomic lineages below genus rank. And therefore Kraken2, when building this reference database with default settings, it truncates any annotation below the genus. Therefore, the NCBI taxonomy was adjusted to the same ranks level to make them comparable. The maximum TND for the Silva taxonomic database is 30 and for NCBI is 52. The TND is normalised by the maximum-minimum method, where TND values were transformed to a decimal between 0 and 1, for each taxonomic database set.

The distribution of the normalised TND of the NR99 database with Silva versus NCBI taxonomy, figure 5.23, shows the massive improvement even with the same genomic content in the database.

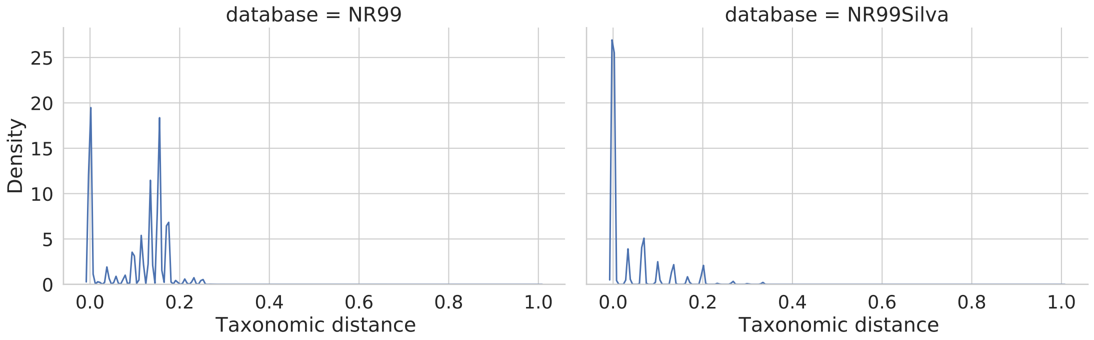


Figure 5.23: NCBI versus Silva taxonomy distance distribution. The data were analysed with Kraken2 and the same genotypic content, but different taxonomic annotations: NR99 used the NCBI taxonomy (truncated to genus) and NR99Silva used the Silva taxonomy. The normalised TND distribution show that most of the sequences have been classified correctly at genus level for the Silva taxonomy. The NCBI taxonomy performs poorly compared to Silva. This highlights the dependency on the curation of the taxonomic annotations (with the same genetic sequences).

5.5 Discussion

This chapter studied the effect of taxonomic identification depending on the data and metadata in the reference database.

Although the quality of the genomic sequences in the reference databases show almost negligible impact for classification, the larger reference databases Parc and ParcTrunc, which contain around 6 million sequences (compared to 0.8 to 2 million in the others), have the poorest accuracy metric values. This is in contrast with previous findings [Méric et al., 2019], where they improved taxonomic classification by adding extra sequences in their reference database. Improved taxonomic annotation is crucial for successful, accurate meta-taxonomic identification of sequencing data and can have a huge impact for downstream analysis [Robeson et al., 2021]. And therefore it is likely that larger databases with highly curated annotations perform better. For instance, taxonomic identification success improves when the Silva NR99 database is used with their own annotated taxonomy rather than with NCBI taxonomy database.

Databases are biased towards human-related microorganisms [Pollock et al., 2018], while others have not yet been sequenced or included in reference databases. Sometimes metagenomics projects can encounter organisms not present in the database. The factors that influence classification are distinct for these cases compared to when the organism has a representative, and also vary from database to database. In general, taxonomic identification occurs much further away than expected.

Other factors influencing the classification are the number of genomic sequences representing the organisms present in the sample: when there are no sequences representing a phylum rank it has a devastating effect, and too many can be equally confusing for the algorithm. However, our data shows negative correlation between the number of species and subspecies TND values.

The MGC was generated with a range of previously known samples to obtain a wide range of communities and obtain an idea of factors influencing taxonomic identification for a wide range of taxa. Single and paired end data are not equally represented, therefore the differences found between single and paired-end data require of further investigation.

In summary, after studying several factors that might influence the taxonomic identification of metagenomic reads, results show consistency in the fact the taxonomic annotation and its curation is vital to improve metagenomic read identification. The results confirm that Kraken2 is method resilient to mutations and sequencing errors. Also, having larger number of strains per species in the reference makes the identification more reliable. For the NCBI taxonomy, those lineages with longer number of nodes the taxonomic identification is less accurate.

Chapter 6

Discussion

Metagenomics can be applied to a wide range of fields. Each environment has a specific fragile microbial community equilibrium and can rapidly change. There are a number of factors influencing which microbes are detected in samples, from the experimental design to the taxonomic identification process [Knight et al., 2018]. The objective of this thesis was to establish the source of errors for 16S rRNA metagenomic taxonomic classification, and propose new suitable metrics and gold standards to allow comparable analysis, as well as applying this acquired knowledge to a real case study.

6.1 Benchmarking

Along the lines proposed by [Schlaberg et al., 2017] and given the general lack of validation of bioinformatics methods by individual laboratories, we developed a novel method that simulates synthetic data based on previously analysed samples. It includes a new approach for generating synthetic negative controls based on HMM profiles from known but unidentified sequences. It outputs the same number of reads as the original sequencing data which it is trying to emulate, with the same length profile and quality, based on the abundance profile from the method or combination of methods of choice for metagenomic taxonomic identification. We call this method My Goldstandard Community (MGC). It generates 3 replicas of 8 sets of simulated data with different levels of sequencing and mutation errors to support comprehensive parameters comparison and optimisation. This is a flexible benchmarking approach, as each laboratory tends to have preferences for certain types of tools, which are often based on compatibility with the IT system available, and also the nature of each project will present a different set of challenges. It allows the effect of different types of errors to be determined, as well as a comparison to any other pipeline parameter. However, an adequate choice of bag of templates needs to be made. For example, if the real sample sequenced amplicon data, or targeted certain taxonomic groups, it is essential to ensure the nature of the templates is of the same type. In other words, for the 16S rRNA data, any database that contains its genetic sequences would be appropriate to use (for instance Silva or Greengenes databases), but it does not make sense to choose a whole genome sequence database like RefSeq. It is essential that quantitative methods are used to evaluate the accuracy of metataxonomic sequence identification approaches, in much the same way that approaches are validated in wet-lab experimental research.

For benchmarking, it is essential to use appropriate and meaningful metrics [Capella-Gutierrez et al., 2017]. This thesis proposes a new set of meta-genetic specific metrics to complement the commonly used and generic precision and recall measures. Metagenomic reads can be classified at different nodes of the taxonomic tree of life, other than species or subspecies. These classifications are correct, but lack specificity. They involve either counting the number of nodes (TND, TNDSR, PND) or branch length (PBD) between the original microorganism and the assigned node by the classifier. A perfect identification in all cases has a value of 0 for all of them, and increases as the classification occurs in further away nodes in the taxonomic or phylogenetic tree.

These distance measures provide a better understanding of how specific the taxonomic identification has been. I believe they should be applied systematically in benchmarking studies to complement other metrics to obtain a much more reliable result.

The taxonomic node measures, presented here, can easily be applied to any taxonomic tree. The newly proposed phylogenetic measures should in principle be more precise compared to taxonomic node distances. Nevertheless, phylogenetic distances require a phylogenetic tree, which is often challenging to calculate for the highly diverse range of organisms commonly found in environmental samples. Additionally, in traditional metagenomics classification, sequences can be labelled to a node of the tree, which tend to be labelled with a taxonomic name and a taxonomic rank. Normally, phylogenetic trees do not have the same labels because classification is carried out purely based on genomic similarities, whereas taxonomies like NCBI use a combination of methods which include phenotypic observations, phylogentics of specific clans and manual curation, and it is hard to map the same type of annotations onto them. The GTDB has one of the most comprehensive phylogenetic trees and contains some phenotypical annotations. However, so far it is only available for bacteria and archaea, presented in two separate trees. But for metagenomics classification, it needs to go a step further and integrate more organisms. This is because frequently, the diversity in many environmental samples also includes eukaryotes, and this would avoid biased results from the metagenomic taxonomic classifier.

6.2 Source of errors for taxonomic identification

Several potential factors were investigated to understand which ones affected the process of identifying metagenomic reads. The number of representative sequences in the reference database at the genus, family, and phylum levels were shown to have very little effect on taxonomic classification. In contradiction to previously published work [Knight et al., 2018], the total number of errors, including mutations and sequencing errors, did not affect the classification performance of the k-mer based method used in this work, up to the maximum 3% of random noise and sequencing error tested. Kraken2 is a robust approach based on a sliding k-mer window, which works well even when sequences contain errors. Factors found that make classification worse were: incompleteness of the reference database or in other words, missing taxa, and long taxonomic lineages, which could be avoided by using only the 7 main taxonomic ranks, except when sequences with less accurate lineage annotation are removed from the reference database. This result is not surprising as Kraken uses a LCA approach. Finally, factors that appeared beneficial, reducing observed TND, were the number of species or subspecies which have several sequences present in the reference database, and high quality and curated taxonomic annotations associated with the genomic sequences.

Additionally, a group of differences arose between factors depending on whether single or paired-end data was employed. However, the experimental data was not designed to test this hypothesis and therefore these results should be investigated further before reaching any final conclusions. For example, the length of reads did not appear to have much effect when looked at globally, but when distinguished by single or paired-end mode, in both cases longer reads gave better results.

6.2.1 Reference databases

The effect of the reference database is one of the main and most influential factors for the taxonomic identification of metagenomic reads. Both the genomic quality and the taxonomic annotation associated are key.

The results presented in chapter 5, demonstrate how crucial highly curated taxonomic annotations are. Improving them or selecting only those sequences whose taxonomic and genomic content is of the highest quality possible ensures better metagenomic identification. Fortunately, none of the simulated short sequences have been labelled to the detected inconsistent viral lineages (no viruses were included in the MGC data as they do not have the 16S rRNA gene).

If INSDC reference databases, commonly used for metagenomic purposes, included quality annotations associated with the genetic regions then it would be possible to select sequences according to these criteria, and as a consequence metagenomics identification could be improved.

6.3 Genetic regions and k-mer frequencies

Different parts of the genome and genes evolve at different speeds, leaving some regions more evolutionary conserved than others, and in principle the hypervariable regions should have much more discriminatory power for species [Chakravorty et al., 2007]. However, our results do not seem to show this effect when studying simulated reads that map to these experimentally defined regions. Instead, our results indicate that conserved regions have as important a role as variable regions when it comes to taxonomic identification in our experimental conditions. One possible explanation could be due to the ongoing growth of the reference database, as new sequences might change the previously defined evolutionary regions, or it could be because the taxonomic classifier used is based on exact k-mer matching, and some k-mers might be shared between evolutionary regions, and that therefore this method would be insensitive to conserved and variable regions.

To test the first hypothesis, a set of new potential conserved and variable regions were defined for the Silva database with machine learning approaches, and the potentially conserved regions were masked from the reference database. The results showed that full length sequences are required for optimal identification.

The methodologies that I have developed can be used in research areas beyond those described here. For example, the method which first maps short reads against the 16S rRNA gene to select the ones of interest (chapter 4 section 4.2.1), has already been used in this thesis as a quality control to ensure the sequencing data in chapter A was targeted to the desired regions and removed any potential contaminants. The second method (chapter 4 section 4.2.2), has two parts, the first one consists of determining regions that evolve at different rates, and can be applied to any other gene or genetic region of interest for the same purpose.

To test the second hypothesis, a novel methodology for classifying metagenomic reads was developed. Each reference database has a distinct k-mer profile abundance. K-mers can be repeated within each sequence and across species, and a few can be unique to a certain species or taxonomic clan. Therefore, I developed a new tool called GeFSATI, which captures these properties for taxonomic classification. The idea was to apply the linguistic method of TF-IDF, which is capable of weighting k-mers according to their frequencies in the database, and couple it with a multinomial naive Bayes classifier. However, this approach presented challenges. First it needed to be adapted from linguistics terms to be useful for our purpose, and consequently, words were defined as fragments of size k or k-mers. Longer k-mers longer sizes of k are more specific for taxonomic identification. The second challenge was that the matrices required by the TF-IDF calculated for large k values can be intractable for many machines. A choice of $k = 9$ was decided for the following reasons: *(i)* it generates manageable matrices for our system given the choice of database, and *(ii)* some sequencing platforms (like Oxford Nanopore) can have about 10% of sequencing errors scattered along the read, therefore we hope that k-mers shorter than 10 will avoid a concentration of them in each fragment, whereas other technologies have much fewer error rates. Nevertheless, this was not sufficient to reduce dimensionality. The reference database had to be cleaned for high quality annotated sequences and a two-step approach was adopted, classifying first at a phylum level and secondly at the genus level. The multinomial naive Bayes is one of the most basic classifiers, but it has been described as working well with TF-IDF [Garbarine et al., 2011].

This newly created method for taxonomic classification proves that it is possible to adapt and apply linguistic methods for metagenomics purposes and that in terms of precision and recall, it performs in line with the currently available methods [Sczyrba et al., 2017]. GeFSATI however, requires optimisation for large-scale use, especially in terms of speed and memory efficiency. Other improvements would require to improve accuracy.

Source code for the relevant parts will be made available jointly with publications.

6.4 Cattle respiratory tract microbiome

A real case scenario of a metagenomic analysis was performed to characterise the microbiota present in different parts of the cattle respiratory tract. This was a prospective study to determine best sampling method as well as microbiome composition of healthy animals. To reduce costs, a 16S rRNA or metagenetic study was performed, which allow multiplexing.

In order to determine potential pathogens for BRD, first the microbiota of healthy animals needs to be determined as a source of comparison. However, while designing experiments two essential questions arose: one was about which DNA extraction method was best for this type of experiment, and the other one was to identify the best sampling method from three possible options (tissue, swab, and BAL). Therefore, two sets of experiments were necessary. While the first one was designed to find an extraction kit that works well for the three sampling methods to be tested, the purpose of the second experiment was to determine and compare the 'normal' microbiome of the bovine respiratory tract while comparing the sampling methods.

In the first experiment, some samples contained a high load of host DNA. The

Powersoil DNA kit was determined to work well for the different sampling methods, and at the same time it highlighted potentially similar results for different sampling methods, with major microbiota discrepancies due to either technical or batch or source related causes.

To determine the Differential Abundance (DA) taxa, different methods were explored. The assumptions by R packages of edgeR and DESeq (both designed for RNAseq data) are violated because the data is sparse, and therefore they were not tested for our data. Instead, it was decided to test using the Lasso method (general purpose) and the LEfSe approach (to determine metagenomic biomarkers). Compositional data analysis are the newly developing methods which are meant to be much more appropriate [Nearing et al., 2022].

The second experiment, with the double goal of establishing the healthy calf microbiome for lung lobes and nose and comparing sampling methods, revealed that tissue samples contained more diversity compared to the rest, and that for healthy animals the microbiome of each of the lung lobules was similar. Also, the nasal sample microbial community was found to be different from the lungs and therefore is perhaps not a good proxy for investigating changes in the lung due to disease.

At the phylum level it was observed that the population comprised two main dominant phyla which were either Tenericutes or Actinobacteria. A potential correlation between their relative abundances was also observed. Actinobacteria include *Mycobacterium*, commonly associated with respiratory infections. Also, in some samples the phylum of Proteobacteria was found to be dominant, instead of the two previous ones. This phylum contains *Pasteurellaceae* which is an opportunistic pathogen commonly associated with BRD. Firmicutes, which were the main phyla for a few of the samples, are commonly found in water, soil and gut, hence this might be the entry point to the lung.

6.5 Future work

More features could be integrated to MGC, for example instead of inserting random noise, the mutations introduced could be introduced by evolutionary mutation rates models.

To encourage the widespread adoption of benchmarking, it would be useful to make the MGC software more easily available (will be made public following relevant publications), through bioinformatics frameworks such as Snakemake or a Galaxy workflow, or more general installation methods such as a Docker container or a Debian package. Additional user-targeted documentation would also be helpful.

It would be useful to design experiments to study further the effect of the read length. Our results suggest that even though Illumina sequencing data is classified better with longer reads. My experimental design was for general benchmarking, and further investigation of this specific point would be interesting.

Further investigations are required to understand if the variable regions of the 16S rRNA gene have more discriminative power for metagenomics identification for similarity based methods than conserved regions. For example, regions V1-V2 are described as performing poorly for the phylum of Proteobacteria, whereas regions V3-V5 performance is good for the genus Klebsiella [Johnson et al., 2019]. This seems likely, since by definition they should be more sensitive to evolution than composition based classification. A good starting point, could be to reuse the same approaches as explained in chapter 4.

To improve GeFSATI, new features should be incorporated. For example, future releases should include compatibility with more sequencing platforms, including Illumina paired end mode. It needs to be explored if smaller k-mers provide at least similar results if not better, because this would help to reduce the RAM requirements. Also, a strategy should be implemented to try to improve accuracy results, especially at genus level. For example, a Bernoulli classifier could be tested. This classifier is binary, instead of taking into account weighted k-mers. The main reason is because, especially for short fragments platforms, the k-mer frequency observed may be different from the overall original gene, so this could potentially be a reason for misclassification. Speed also needs to be improved, so faster methods should be explored.

Once the microbiota of the healthy cattle is studied, and sampling methods tested, BRD samples should be sequenced and compared to healthy animals. The experimental design should consist of animals of approximately the same age, and with enough from each farm, so that statistical tests can be applied in order to identify potential pathogenic groups.

6.6 Concluding remarks

This work makes a contribution to end the lack of systematic validation of bioinformatics metagenomics pipelines by specific researchers to test the suitability of the methods with project specific data. MGC combined with the new quantitative measures, allows better evaluation of pipeline performance, which enables more informed choices about pipeline elements, leading to better overall results.

The methods developed in this thesis have potential beyond the scope of the specific applications presented.

GeFSATI not only has a future potential for metagenomic sequence identification, but is also a useful reminder that fruitful approaches to genomic data can be adapted from methods that have shown their value in natural language processing.

Appendix A

Applications

A.1 Motivation

Bovine Respiratory Disease (BRD) is a multifactorial disease caused by a complex of viral and bacterial pathogens that act individually or in concert. BRD—associated viruses, which include bovine herpes virus 1 (BHV-1; causative agent of Infectious Bovine Rhinotracheitis or IBR), bovine respiratory syncytial virus (BRSV), and bovine parainfluenza-3 virus (PI3V), are primary pathogens that can also pre-dispose calves to secondary bacterial infections. BRD-associated bacteria can also act as primary pathogens without any viral involvement and these include, *Pasteurella multocida*, *Mannheimia haemolytica*, *Histophilus somni*, and *Mycoplasma spp.* (especially *Mp. bovis*), while *Trueperella pyogenes* may be found in chronic cases. The susceptibility of calves to BRD is significantly affected by farm management practices, including inadequate colostrum intake, weaning-associated stress, high stocking density, change of feed, poor nutrition, transportation ('shipping fever') and poor ventilation. Increased susceptibility to BRD can result from BVDV infection-induced immunosuppression. However, Scotland has nearly eradicated BVDV and the rest of the UK (similar to some European countries) has embarked on eradication, so BVD is a less important contributory factor here

Many studies have been published to characterise the potential microbiota associated with BRD through high throughput sequencing. Most of them are on either samples from swabs or, Bronchoalveolar lavage (BAL) and only a few on tissue. They are normally focused on bacteria, because they are the principal cause of the pneumonia, and, commonly targeted to the 16S rRNA gene [Johnston et al., 2017, Hause et al., 2015, Davids et al., 2016, Holman et al., 2017, Zeineldin et al., 2017b, Holman et al., 2015]. It is necessary first to establish what a healthy respiratory microbiome comprises in terms of a 'normal' microbe population [McMullen et al., 2020]. However, these studies lack a comprehensive comparison and evaluation of different DNA extraction and sampling methodologies, which are the focus of this study.

A.1.1 Objectives

The main objectives are:

- Determine and establish a robust sampling protocol and DNA extraction method for the cattle lung.
- Compare sampling methods.
- Characterise the 'normal' microbiota present in different parts of the upper and lower respiratory tract of young calves.

The sampling was performed by the Respiratory Group team from the Moredun Research Institute, including Chris Cousens, Mark Dagleish, Mara Rocchi, David Longbottom and Alba Crespi. Sample processing and DNA extraction was performed by Mara Rocchi, Kevin Aitchison and Morag Livingstone. The sequencing was performed at the Liverpool Centre for Genomic Research. The bioinformatic analysis was performed by Alba Crespi.

A.2 Materials and methods

A.2.1 Samples for evaluating the DNA extraction methods

sampling

Two animal lungs were sampled post-mortem, one from adult cattle with mastitis and no other clinical signs and another one from a 3 to 6-month-old calf with no clinical signs. Three sampling techniques were used: swab, BAL and tissue. From the adult animal, swab and tissue samples were taken and from the young calf, swab and BAL.

Samples from swab, tissue and BAL were collected post-mortem (PM) as follows: the trachea was clamped first to stop blood or ruminal material to fall back into the lungs, then the heart and lungs were removed together from the thoracic cavity. Two bronchioles of the cranial bronchus were dissected. Swab were taken first at the entry into each separate section of the lung, then tissue samples were taken close to the entry and then at the distal end of the bronchiole. The bronchiole end was then clamped above where the tissue sample was taken (to prevent leakage/loss of BAL fluid) and 60 ml BAL fluid then introduced at the entry, tissue gently massaged, and then the fluid recovered and transferred to a sterile glass bottle. The samples collected were kept on ice until they were returned to the laboratory.

DNA extraction and sequencing

The DNA was extracted using the kits QIAamp cador pathogen, Powersoil DNA, Powerfecal DNA, DNeasy blood and tissue and QIAamp DNA microbiome following manufaturer's instructions. Then the quality of the DNA was checked and quantified by Nanodrop and Qubit.

The sequencing was performed with the platform Illumina MiSeq run v2, paired-end mode, 2x250 bp sequencing.

Bioinformatics analysis

All the samples were checked for quality, read lengths distribution, and 16S rRNA targeted region by applying the same selection of reads process as described in chapter 4 section 4.2.1. Samples were quality trimmed with sickle (q = 30) and filtered by length (<100bp). The taxonomic identification was performed with Kraken 1.1 and the Parc Silva database v.128.

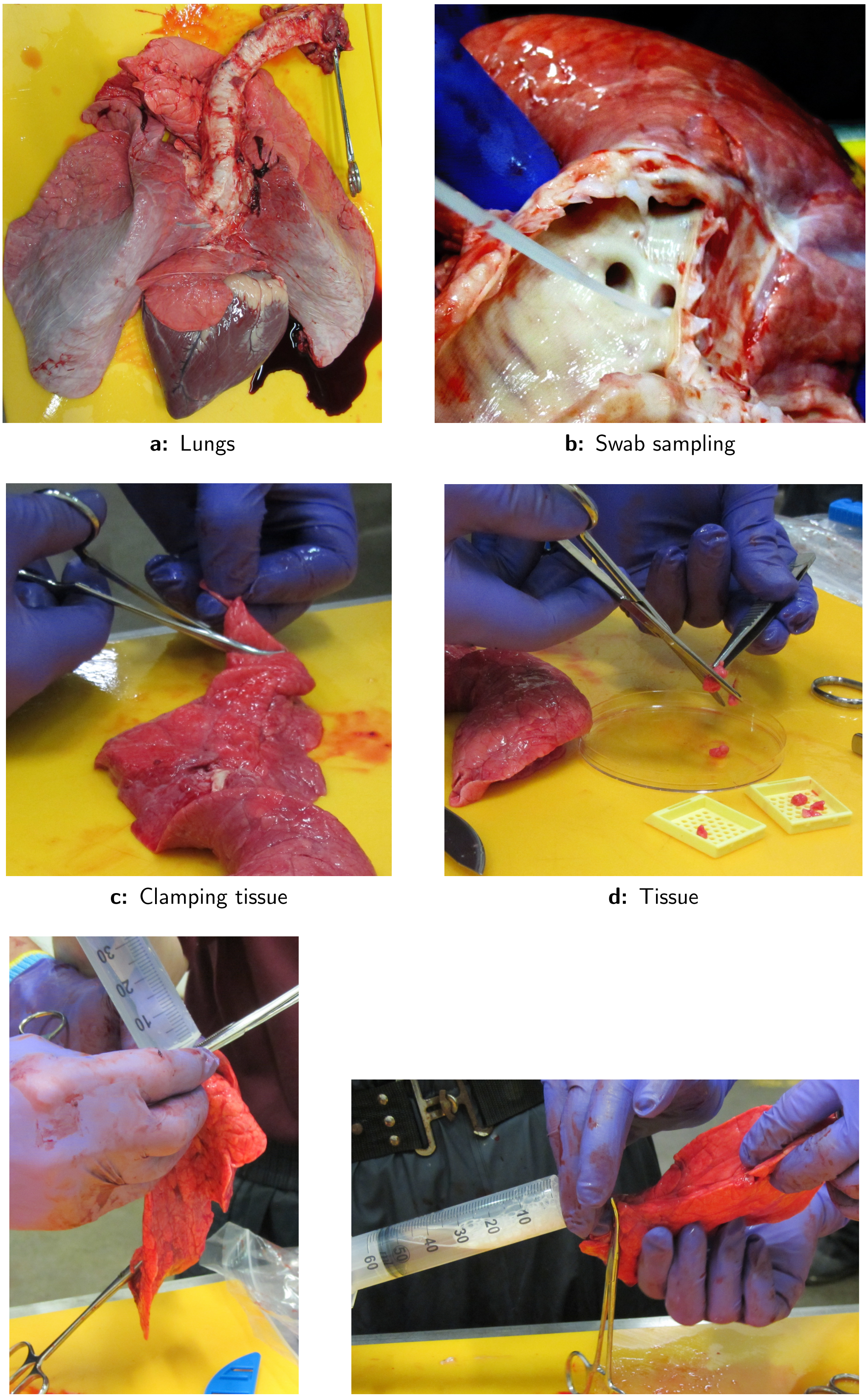
A.2.2 Samples for the characterisation of respiratory tract microbiome

Sampling and DNA extraction

Four 10-week-old male dairy calf, sourced from 2 different farms, were used for this experiment. Samples were collected PM. All the different lung lobes were sampled for BAL, swab and tissue (alveolar and peripheral), as detailed in section A.2.1 and shown in figure A.1. Swabs used for the nasal swabbing were of the "flocked" type and were stored in their sleeves after collection. Tissue samples were collected, trimmed and stored in histology cassettes which were immediately snap frozen in liquid nitrogen. Surgical instruments were changed between each sample to avoid sample cross-contamination. In total, 105 samples were collected for sequencing the 16S rRNA gene region V4 (PCR amplification for forward primer 5'TGCCAGCMGCCGCG-GTAA3' and reverse primer 5'GGACTACHVGGGTWTCTAAT3'). The DNA was extracted with the Powersoil DNA and sequenced by Illumina paired end sequencing at the Liverpool Centre for Genomic Research.

Sample were identified as follows: first a number identifying the sample, then the animal identification consisting of the letter "A" followed by a number, next is the sampling method (BAL for bronchoalveolar lavage, T for Tissue, S for swab) followed by a 2-letter code indication the lung lobule (RA right apical or cranial, RM right middle, RC right caudal, LA left apical or cranial, LC left caudal and AC accessory). For tissue samples there is an extra letter at the end which is a *P* for peripheral or *B* for bronchiole.

As the previous samples, they were sequenced with the platform Illumina MiSeq run v2, paired-end mode, 2x250 bp sequencing.



a: Lungs

b: Swab sampling

c: Clamping tissue

d: Tissue

e: BAL

f: BAL collection

Figure A.1: Lung sample collection. After extracting the lungs and heart (a), a swab sample was collected for each lobule (b). Then a small part of each lobule was clamped (c) for extracting a tissue sample (d). Finally Liquid was introduced in each lobule (e) and recovered with the same syringe (f).

Bioinformatics

Raw sequencing data received was checked for adaptors. A set of samples were checked to make sure that they mapped the target region V4 of the 16S region following the same steps as described in chapter 4 section 4.2.1. Reads were quality checked and trimmed with sickle[Joshi NA, 2011] (quality threshold 30 and minimum length of 100 base pairs).

Kraken [Wood and Salzberg, 2014] with Silva SSU Parc database v.132 was used for taxonomic assignment. Kraken has two modes, one for Illumina paired-end reads, and the other one for single reads which were used accordingly. The results were combined in a single file.

A.3 Evaluation of DNA extraction methods

The main goal of this experiment was to test the sampling protocols and DNA extraction methods to determine which produced the best nucleic acid recovery in terms of quantity and quality.

Five different extraction protocols were tested, and their main characteristic are listed in table A.1.

Table A.1: DNA extraction kits tested Main characteristics of the tested kits for DNA extraction. *NA: Nucleic Acid, **PowerSoil DNA: first follow PowerSoil total RNA isolation kit with DNA elution accessory kit.

Kit	**Sample target**	**NA***	**Carrier**	**bead beading**	**column**
QIAamp cador pathogen	blood, serum, plasma, body fluids, swabs and washes and tissue	viral RNA and DNA and bacterial DNA	Carrier RNA	glass beads	spin column
PowerSoil DNA**	Soil	RNA, DNA	No	Bead tube	RNA and DNA capture columns
QIAamp DNA microbiome	bacterial microbime from mixed samples	DNA	No	bead mill	UPC mini column
PowerFeacal DNA Isolation	stool and faeces	DNA	No	Dry bead tube	silica spin column
DNeasy blood and tissue	animal blood and tissue, rodent tails, ear punches, culture cells fixed cells, bacteria, insects	DNA	optional DNA and RNA carrier for small samples	No	mini spin column

Samples to be sequenced were selected based on the concentration, quality controls, purity and integrity of the genetic material, and are highlighted in red in the table A.2. Purity check was performed for traces of ethanol, phenol and protein by NanoDrop analysis. While the absorbance ratio at 260/280nm indicates purity of

Table A.2: Test samples DNA extraction summary. Experimental parameters of the different samples. Highlighed in gray are the sequenced samples. The number inside the parenthesis indicated the sample sequencing ID. *Conc.* - Concentration. *Purity I is the ratio of absorbance at 260/280nm, **Purity II is the ratio of absorbance at 260/230nm, acceptable values are green. *** Host content is only for the sequenced samples (greyed rows).

Sample	Animal	Kit	Conc. ng/ul	Purity I*	Purity II**	Conc. (b-actin) CT	number of reads	Host content ***
Swab A1	1	QIAamp cador pathogen	77.6	2.1	1.45	16.42		
Swab A2	1	QIAamp cador pathogen	231.3	2.2	0.84	17.31		
swab B **(2)**	1	Powersoil DNA	49	1.78	2.07	16.41	2,950,310	0.5%
swab B	2	Powersoil DNA	1.8	1.58	4.09	19.66		
Swab A1 + A2 **(1)**	2	QIAamp cador pathogen	117.2	2.02	0.62	13.78	3,234,003	70.0%
swab D	2	QIAamp DNA microbiome	6.4	2.02	0.32	24.52		
BAL A1	2	QIAamp cador pathogen	88.7	1.89	0.9	16.04		
BAL A2	2	QIAamp cador pathogen	312.4	2.01	1.1	11.01		
BAL B **(4)**	2	Powersoil DNA	0.7	1.26	0.39	29.97	3,439,655	72.0%
BAL D **(3)**	2	DNeasy blood and tissue + QIAamp DNA microbiome	20.7	2.13	0.2	11.91	2,129,391	32.0%
Tissue B 1st **(8)**	1	Powersoil DNA	1161.6	1.87	2.3	13.26	4,096,153	4.0%
Tissue B 2nd **(5)**	1	Powersoil DNA	172.2	1.95	2.15	16.18	1,051,518	4.0%
Tissue Dneasy	1	DNeasy blood and tissue kit	32.6	1.98	1.61	15.89		

DNA and should give a ratio of approx 1.8 for pure DNA (RNA gives a ratio of 2.0). If higher than 1.8 then indicates RNA contamination, the ratio of absorbance at 260/230nm indicates the presence of any unwanted organic compound such as phenol and should fall within the range 2-2.2. Higher values indicate contamination.

Representative samples of swabs, tissue and BAL were sequenced by Illumina paired-end sequencing at Liverpool Centre for Genomic Research. The V4 region of the 16S rRNA gene was sequenced in eight test samples, which included a positive (mixture of pathogens previously extracted with PowerFeacal) and a negative control (a swab which was 'waved in the air' at PM). The sequencing output generated reads between 19 and 250 bp long.

The analysis of taxonomic identification at the species rank, revealed that host sequences, *Bos Taurus*, were mislabelled as *Bos mutus*, which is wild yak. This can be observed in figure A.2, a word cloud of the most abundant species detected in the test samples. It was noted that human DNA was also detected. Therefore, an additional step was implemented in the pipeline to remove sequences belonging to mammals.

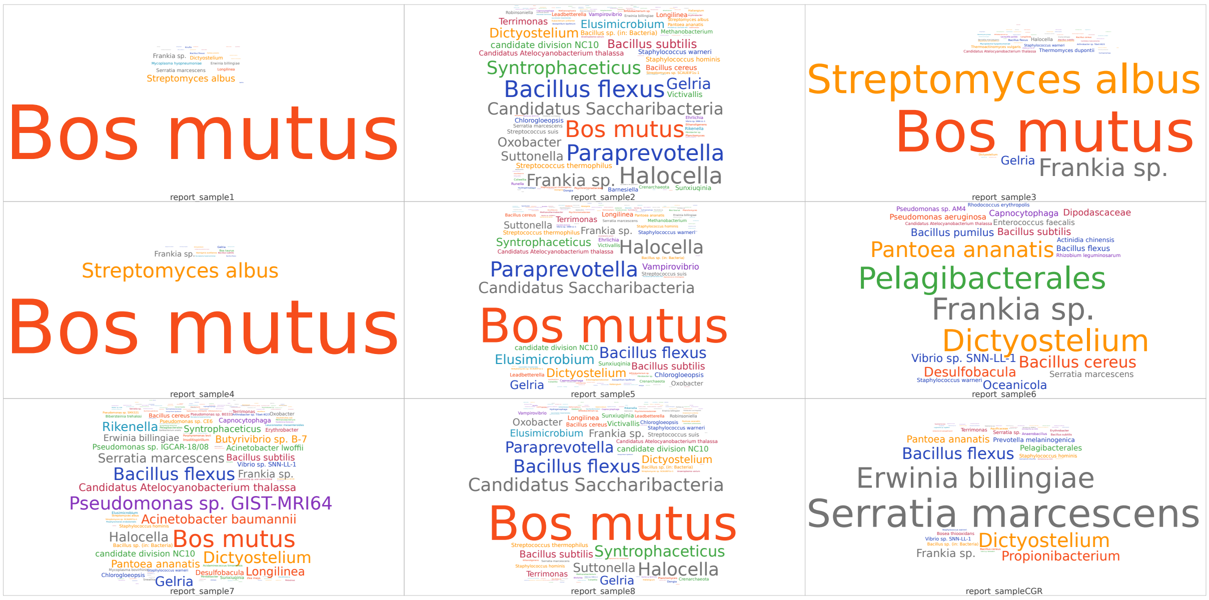


Figure A.2: Most abundant species in the test samples. The font size is proportional to the abundance. This word cloud reveals that host sequences have been mislabelled to *Bos mutus*, wild yak. Also host genome is found in a relatively high proportion of the sequenced data. SampleCRGNeg is a negative control from the sequencing.

The relative abundance by phylum is presented in figure A.3. Bacteroidetes, Tenericutes, Actinobacteria, Frimicutes and Proteobacteria were found to be the most abundant phyla in this set of test samples.

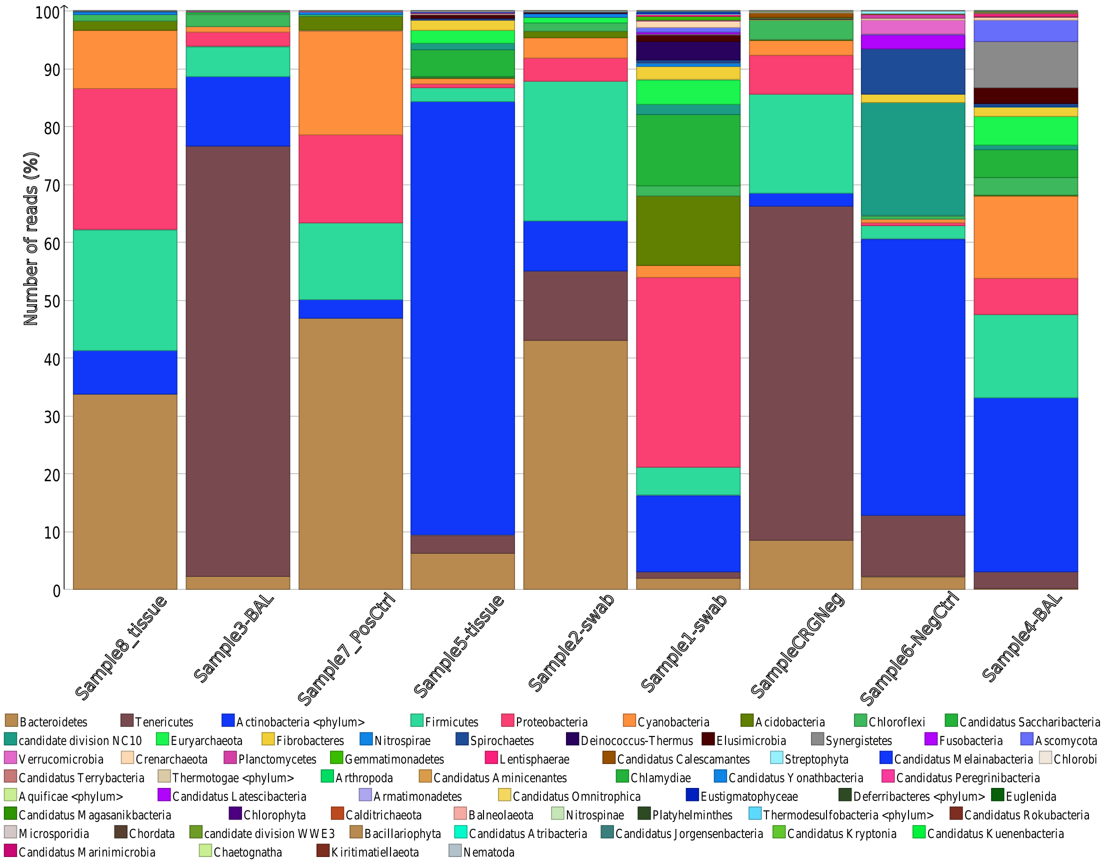


Figure A.3: Relative abundance of phyla for the test samples. Phylum legend is ordered according to abundance. The predominant phyla in the cattle samples are Bacteroidetes, Tenericutes, Actinobacteria, Firmicutes and Proteobacteria. SampleCRGNeg is the negative control of sequencing.

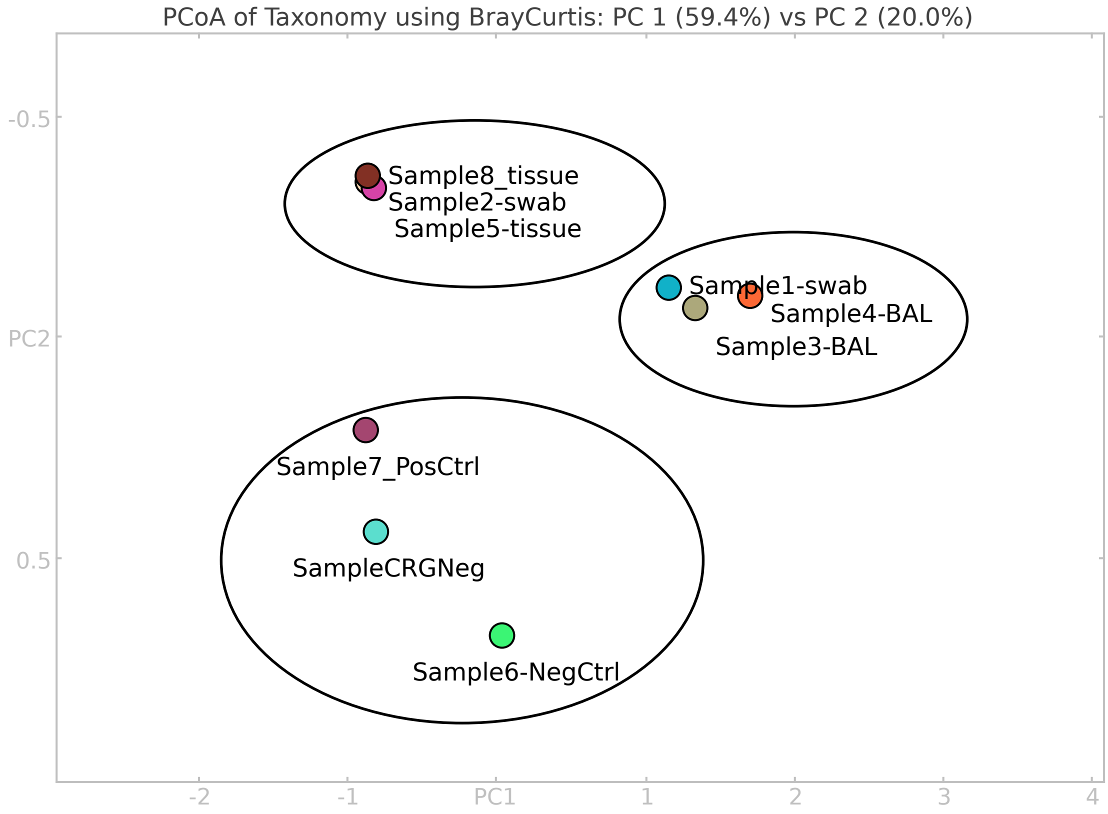


Figure A.4: Test samples PCA. The samples have 3 distinct clusters: 8,2 and 5 which belong to animal 1; samples 1,3 and 4 which belong to animal 2, and the controls.

A Principal Components Analysis (PCA) with Bray Curtis distance analysis was performed of the abundance profile obtained from Kraken, and it is shown in figure A.4. There are 3 clusters, one for samples 2, 5 and 8, which belong to animal 1, another one for samples 1, 3 and 4, which belong to animal 2 and a final cluster which contain the positive and negative control samples. Which indicates the major differences are due to microbiome being unique for each individual. And it also highlights potentially no difference for sampling methods.

Samples 1 and 3 were extracted with the kit QIAamp, whereas the remaining samples, apart from controls, were extracted with the Powersoil kit. Because of the small sample size, statistical tests were not performed to decide which kit works better in terms of microbiome abundance recovery. However, PowerSoil DNA showed the overall best purity results and works well for all the types of sampling to be tested next, as it can be observed in table A.2, and it was the kit of choice for the next experiment.

A.4 Characterisation of healthy calves' respiratory tract microbiota

To characterise the normal microbiota present in different regions of the upper and lower respiratory tract. Due to costs limitations, samples were collected by swab, tissue and BAL from 4 male calves that were approximately 10 weeks old and which were obtained from two different farms. The goals of this experiment were: *i* Determine the microorganism community in the lung of young cattle, *ii* Find differences between microbial composition in the different regions of the respiratory tract, and *iii* Compare the different sampling methods. At PM, animal 1 presented a runny nose, and about 5% consolidation in the right cranial lobe (indication of a respiratory infection), which indicates that this animal might not be used as a control sample.

Cattle lungs have six lobes, as depicted in figure A.5.

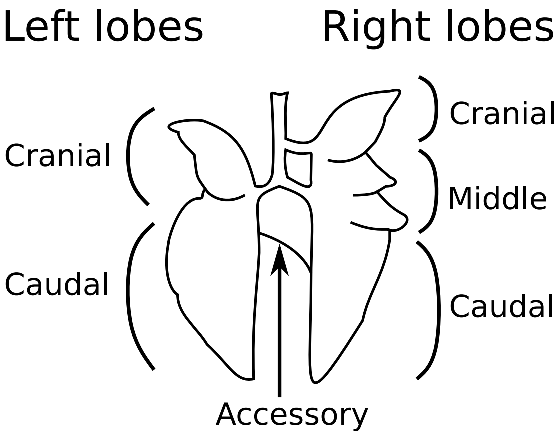


Figure A.5: Cattle lung lobes. Cattle lung contain 6 lobes: 2 on the left (cranial or apical and caudal), 3 on the right (cranial or apical, middle and caudal) and the accessory in between.

One hundred and five samples were collected for DNA 16S rRNA sequencing from three different sampling methods: swab, BAL, and tissue. A positive, which contained a pool of isolated pathogens, and negative controls, swab opened at PM, were included. Optimisation of the coupled amplification of the 16S V4 rRNA gene region and barcoding of multiple samples was conducted prior to sequencing.

A.4.1 Exploratory analysis

Prior to taxonomic assignment of the sequencing reads, the Canberra's distance of the raw reads was calculated with Simka [Benoit et al., 2016], which is a tool that analyses the k-mer composition or genomic composition. It does not require any previous taxonomic assignment.

The genomic content of the raw samples is compared to find out quickly similarities, and without the need of a taxonomic identification. Figure A.6 shows the

Canberra measure (distance between pairs) of the raw reads. This measure works well for highly dimensional data. At a first glance, the vast majority of BAL and swab samples cluster together, whereas tissue tends to be more separate.

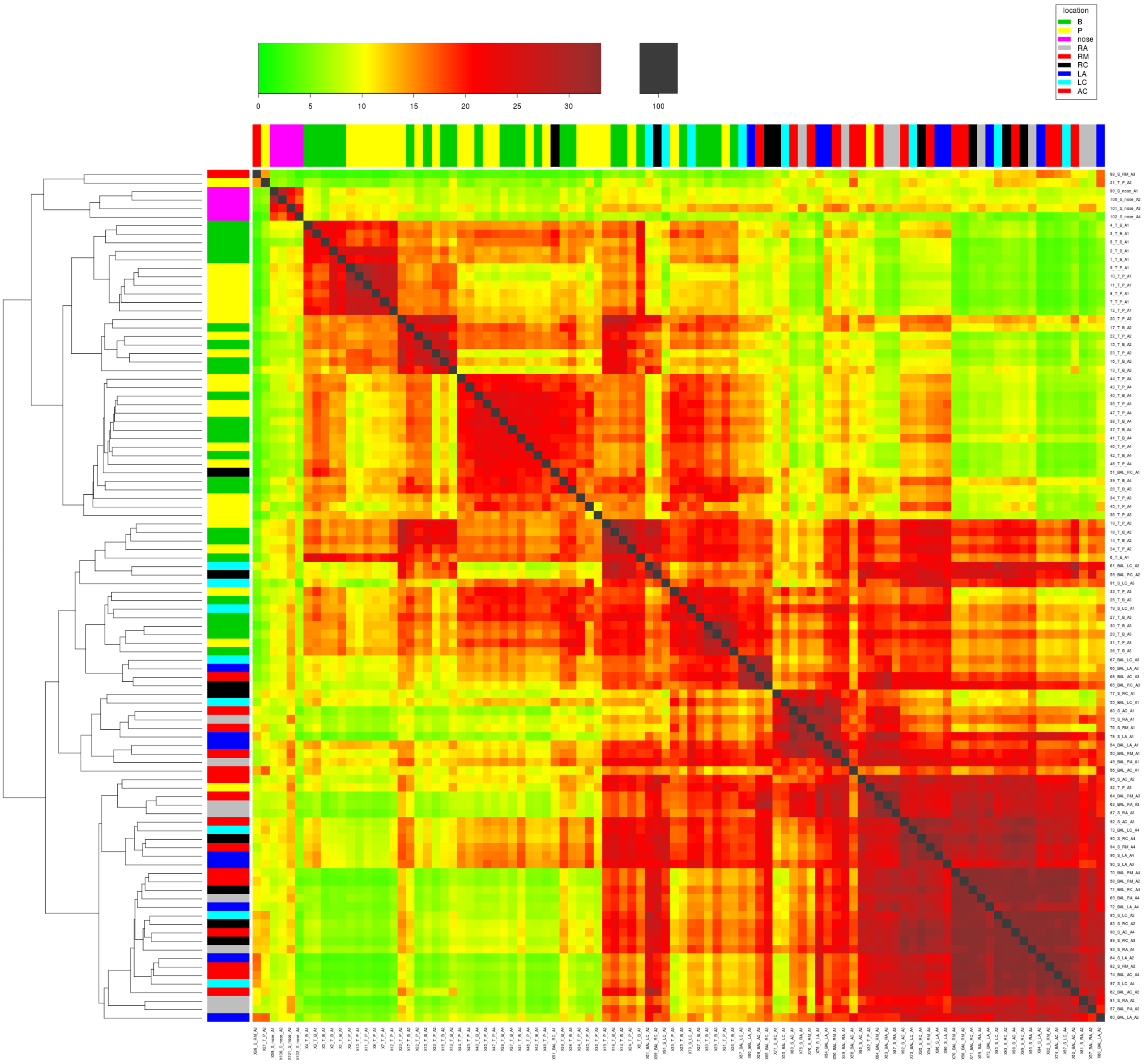


Figure A.6: Canberra's distance of the raw reads. Calculated with Simka. Tissue samples are mostly clustered together, and the vast majority of BAL and swab samples are clustered together.

A.4.2 Microbiome taxonomic composition

After quality filtering, read counts were found to range from 811 to 178,652, as shown in figure A.7. Next, sequences identifies as belonging to mammals were removed, which left the number of reads in the samples ranging from 521 to 178,079.

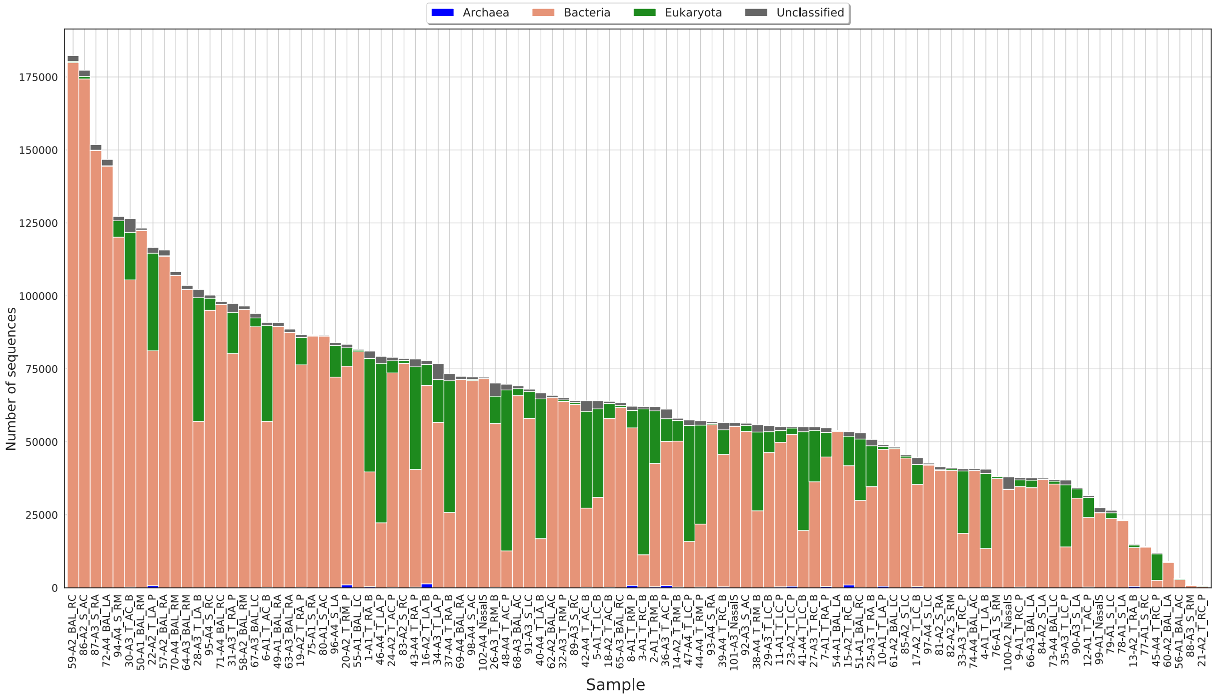
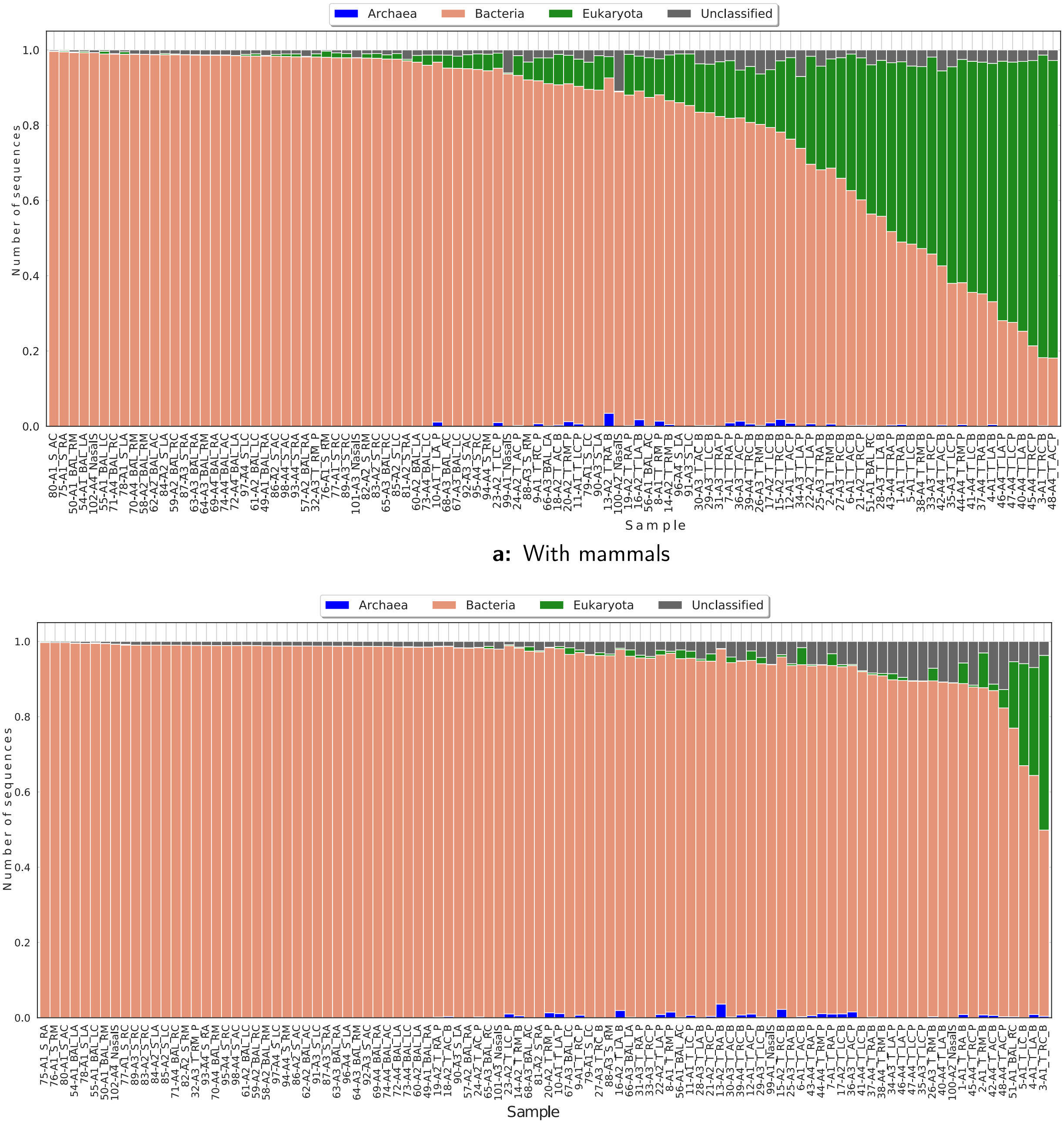


Figure A.7: Superkingdom absolute abundance. Total number of reads, descending order, per sample at the superkingdom level. After quality trimming and before removing reads belonging to mammals.

Most of the reads were identified as bacterial. However, samples contained up 79.58% of Eukaryotic DNA, as can be observed in figure A.8, and up to 78.5% of sequences were identified as mammalian origin. Samples from animal 1 contained up to 16.94% of fungi, whereas samples belonging to the other animals had less than 2%. Archaea were less commonly identified, with the sample with the highest proportion containing 3.44%.

Kraken classified host sequences as *Bos mutus* instead of *Bos taurus*. The results presented here were for the taxonomic ranks of superkingdom and phylum. As previously discussed in this thesis (see chapter 3, section 3.5) the most accurate results were found at these levels.



a: With mammals

b: Without mammals

Figure A.8: Superkingdom relative abundances.Relative abundance before and after removing sequences belonging to animals (sorted by bacterial abundance). There is a relatively small number of unclassified sequences. Most of the eukaryotes remaning after removing mammals, are fungi. Archaea represents a tiny fraction of the microbiome.

The most abundant phyla are Tenericutes followed by Actinobacteria, Proteobacteria, Firmicutes, Bacteroidetes and Ascomycota, as shown in figure A.9.

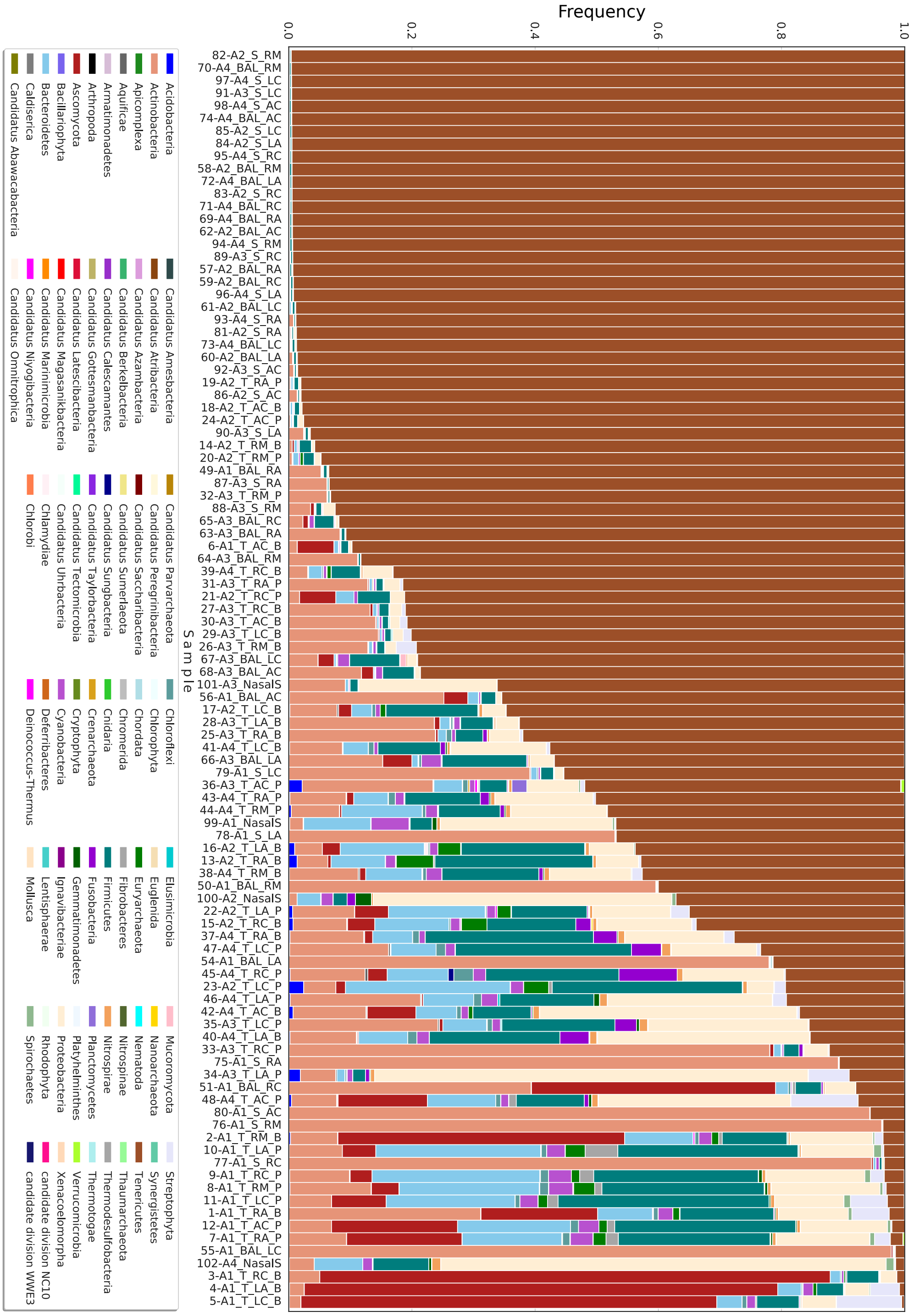


Figure A.9: Phylum relative abundance. The most abundant phyla in the samples are Tenericutes and Actinobacteria. Ascomycota (fungi) is present in animal 1 and some samples of animal 2.

A.4.3 Comparative analysis

Diversity

The alpha diversity from the Kraken assignment after removing mammalian sequences was calculated with the skbio. At species level was found to vary greatly across samples. Figure A.10 shows the alpha diversity by respiratory tract sampling point and sampling method. Tissue samples have the highest average alpha diversity values. BAL and swab samples present much lower values. These results are consistent across the lower respiratory tract. The nasal sample mean alpha diversity was found to be between the tissue and BAL samples.

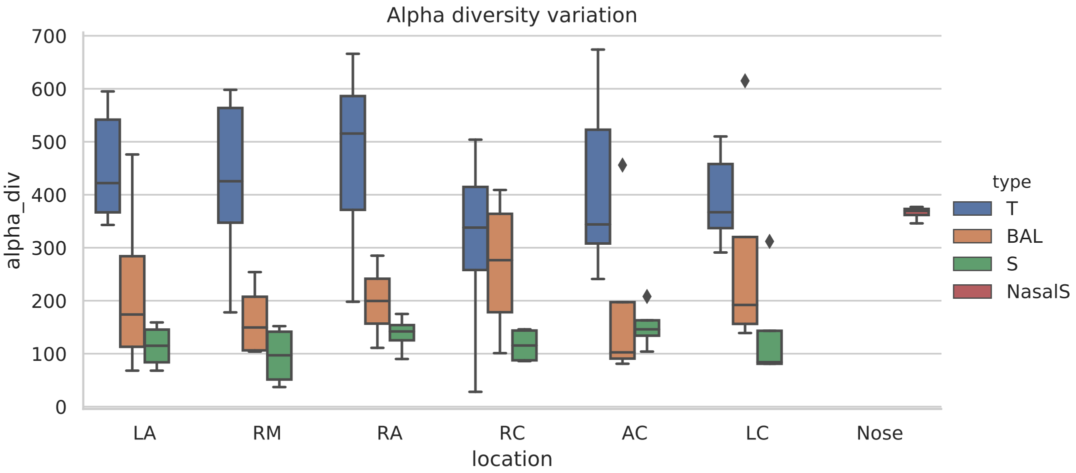


Figure A.10: Alpha diversity. Alpha diversity is grouped by respiratory tract location and sampling method. Tissue samples present higher average values, followed by BAL samples and swabs have the lowest alpha diversity values. R LA left apicak, RM right middle, RA rigth apical, RC right caudal, AC accessory, LC left caudal. Type refers to the sampling method: BAL, S - swab, T - tissue.

Phyla composition across samples

There are a few dominant phyla in the samples. As figure A.11 shows, Tenericutes are present in most of the samples. However, it can also be observed that most of the samples have only one dominant phyla. Bacteroidetes and Firmicutes are in a few cases both found in relatively high proportions.

There seems to be an equilibrium between Actinobacteria and Tenericutes as dominant phyla, thus most samples only have one of them as dominant. In 3 tissue samples from animal 1 Ascomycota is dominant, probably it has replaced the previous dominant phyla, but in the remaining samples there is a combination of Proteobacteria, Firmicutes and Bacteroidetes instead of the Tenericutes population, which mostly occur in tissue and nasal samples.

Figure A.11 shows results clustered by sample, and it is colour coded by respiratory tract location (denoted as lobule), sampling method (type) and animal. The clusters appear to be somewhat more correlated with animal and sampling method.

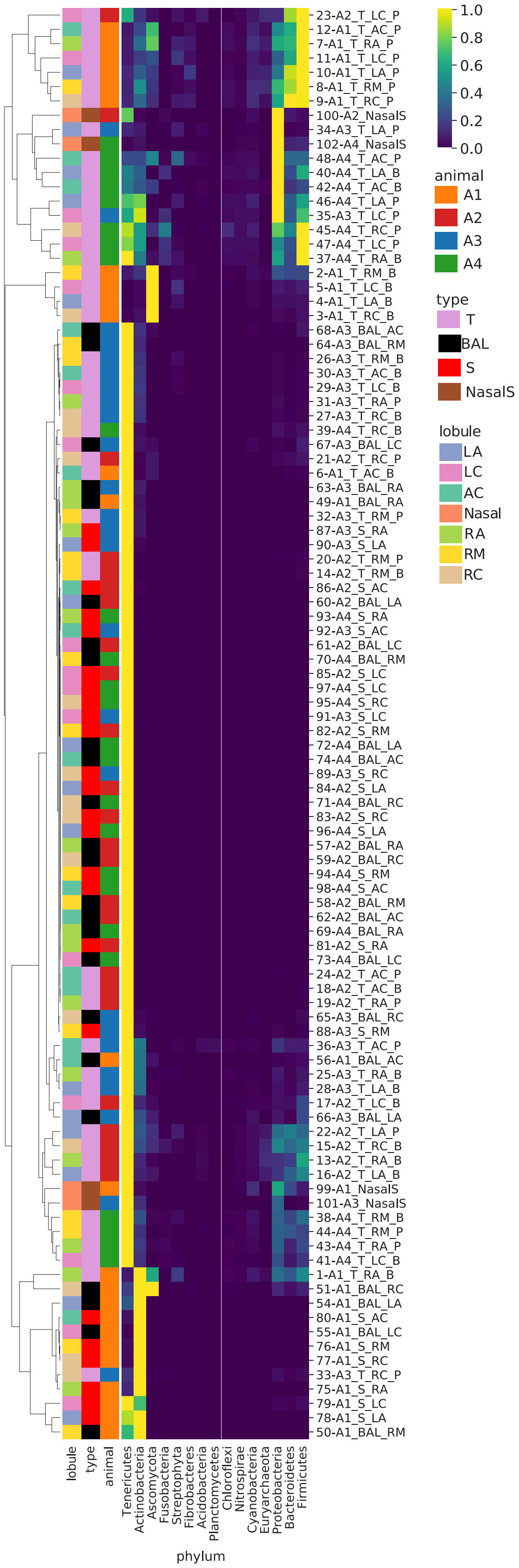


Figure A.11: Heatmap scaled phylum abundance. (Caption on next page.)

Figure A.11: (Previous page.) Data was grouped by sample and phylum. To the right it is color coded by lung location (lobule), sampling method (type) and animal. Type refers to sampling method: T for tissue, BAL for bronchoalveolar lavage, S for swab, and NasalS for nasal swab. Lobule indicates which part of the respiratory tract the sample is being taken from: LA left apical, LC left caudal, AC accessory, RA right apical, RM right middle, RC right caudal, and Nasal. The vast majority of samples have one or two dominant phyla. Samples cluster better for animal and sampling method than respiratory tract location.

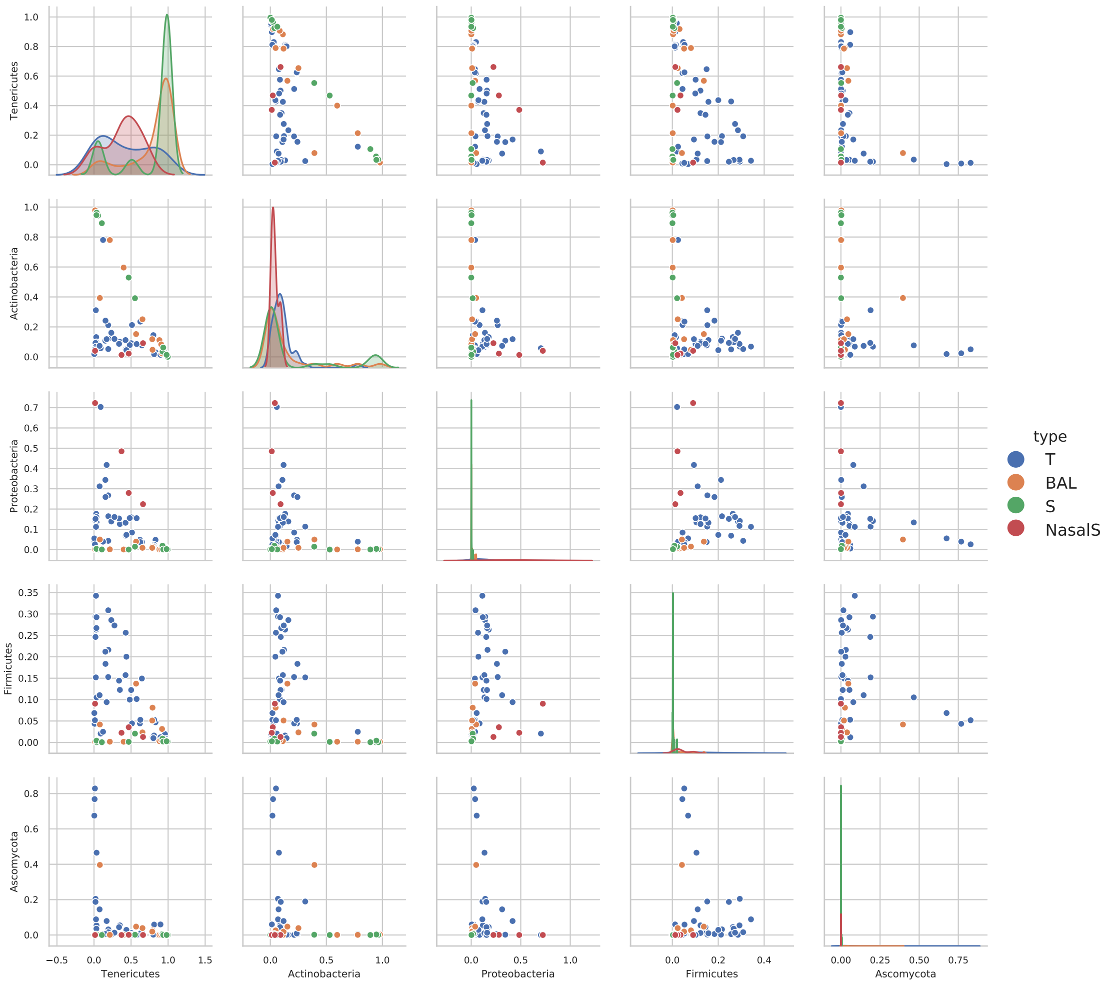


Figure A.12: Pairplot of the top 5 phyla by sampling method. The axis are at a different scale for each phylum. The plots in the diagonal show the density distribution for the phylum. Teneritues is dominant for swab and BAL samples. The nasal sample is composed mostly of Actinobacteria.

First, the top 5 most abundant phyla are compared by sampling method. As previously observed, the dominant phyla can change. Figure A.12 shows a potential correlation between Tenericutes and Actinobacteria, Proteobacteria and Firmicutes. The figure also shows the distribution of each of the 5 phyla per sampling method (subplots in the diagonal). Swab followed by BAL are the sampling methods that recover more Tenericutes. Actinobacteria distribution is similar, with a higher peak for nasal swab.

Next, the phyla detected at each lung location by sampling method is compared with the z-score metric (see figure A.13). This metric is useful to compare samples with different mean and standard deviation. The major differences were observed for sampling method and the score remains stable for each lobule for swab, tissue and BAL respectively. This shows that the recovered microbiome differs between each sampling method.

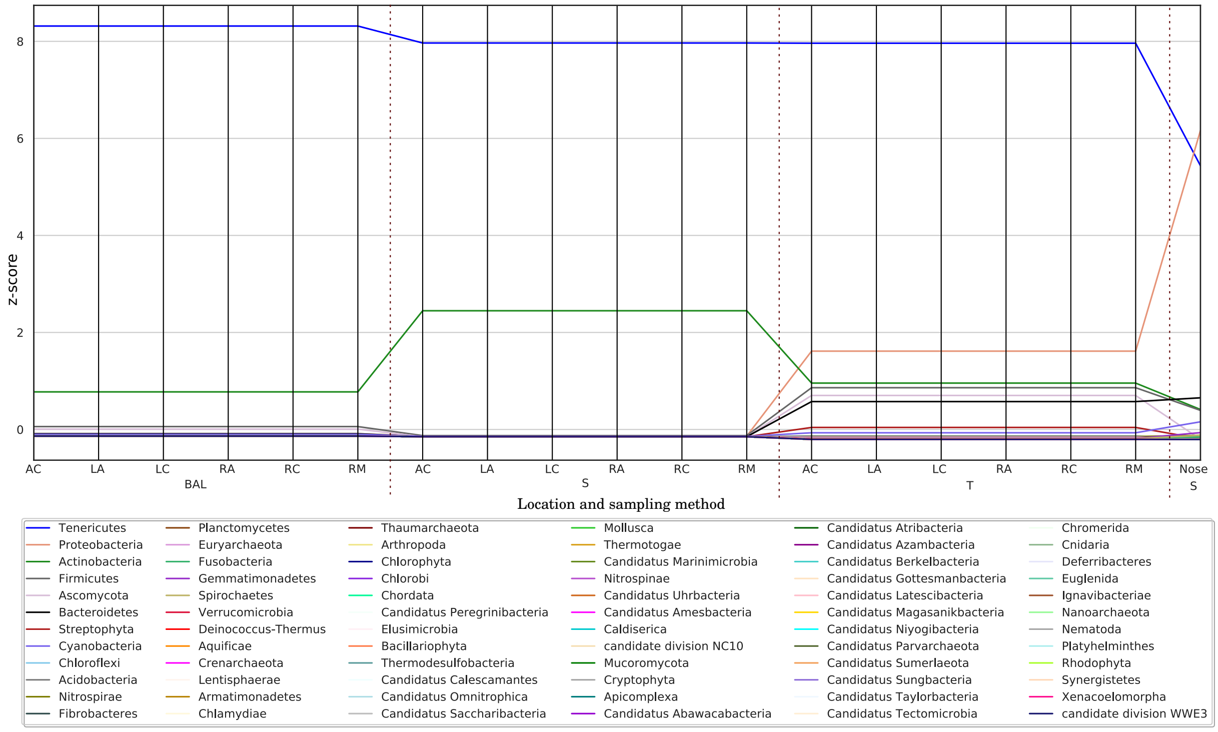


Figure A.13: Z-score by sampling method and location. Location refers to the lung lobule where the sample was taken from: AC - accessory, LA - left apical, LC - left caudal, RA - right apical, RC - right caudal, RM - right middle. Sampling method: BAL, S - swab, T - tissue. The z-scores values are presented for each pair of sampling method and respiratory tract location by phylum. Except for the nose, which has a distinct microbial community from the lower respiratory tract, the major differences observed are related to the sampling method and not to lung lobule location.

The Pearson correlation of the phylum relative frequency is shown in figure A.14. Most of the samples belonging to animal 1 do not correlate with the samples from the other animals. The sampling methods of swab and BAL results present high values of correlation, indicating the microbial community recovered by these two methods were very similar.

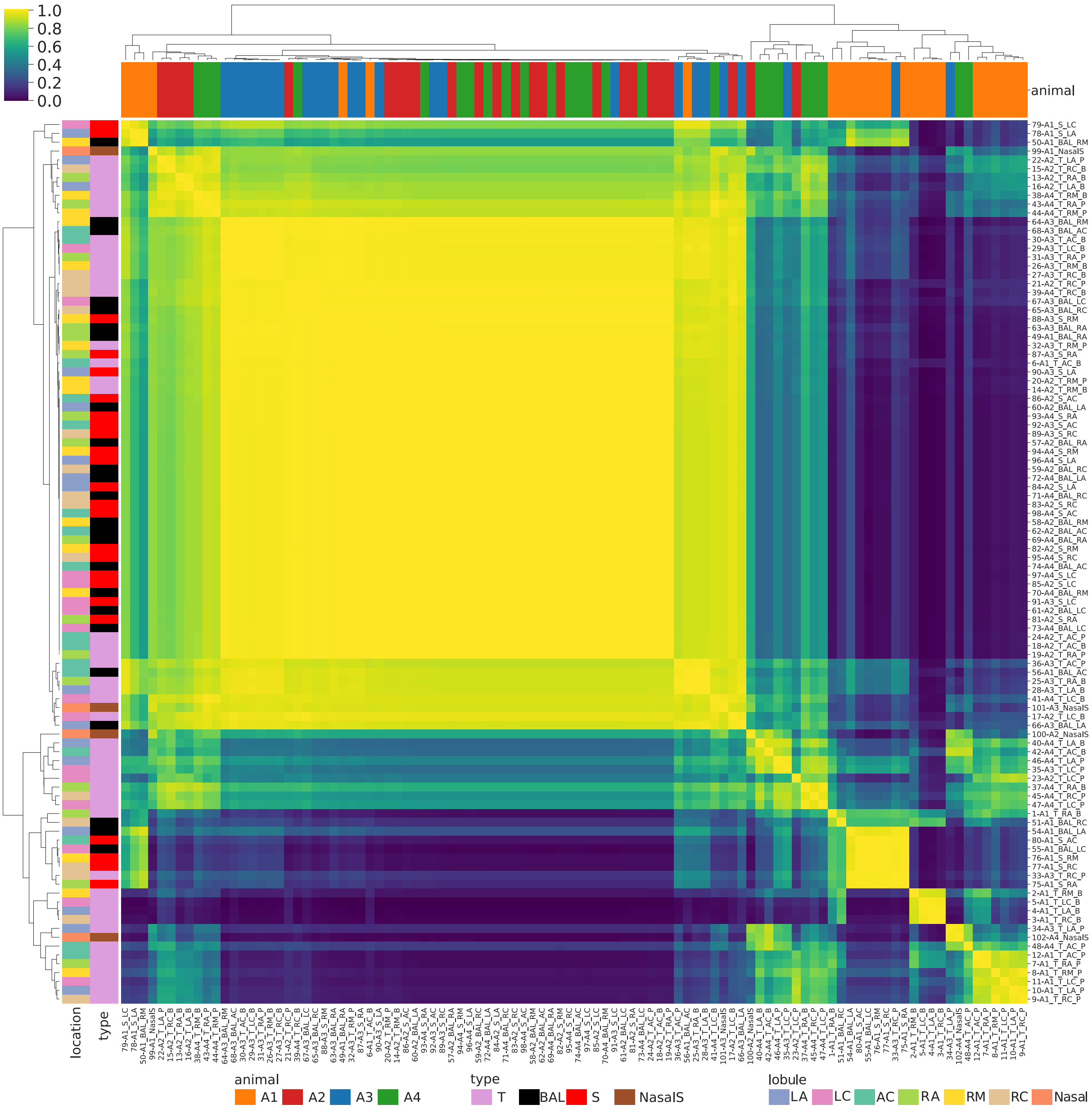


Figure A.14: Sample correlation of the phyla abundance. Location refers to the lung lobule where the sample was taken from: AC - accessory, LA - left apical, LC - left caudal, RA - right apical, RC - right caudal, RM - right middle. Sampling method: BAL, S - swab, T - tissue. The Pearson correlation is calculated from the phylum relative abundance. At the top animals are colour coded and at the left are the respiratory track sampling points (location) and the sampling method (type). There is no clear pattern of clustering according to location. BAL and swab samples tend to cluster together. Most of the samples belonging to animal 1 cluster together.

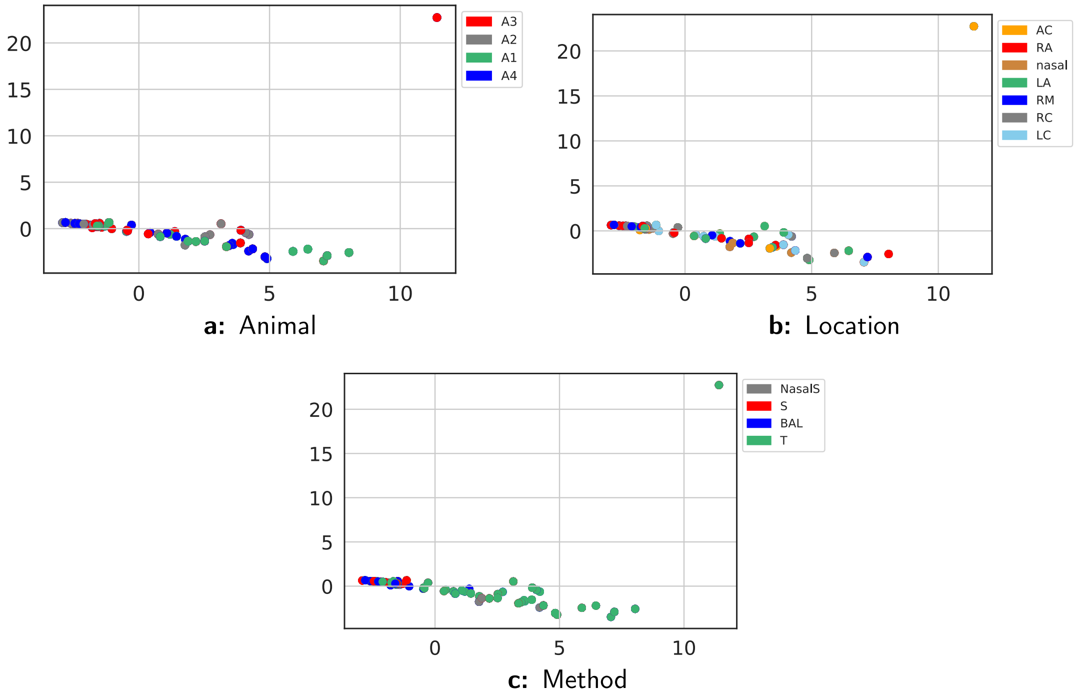


Figure A.15: Principal component analysis by animal, location and sampling method. The data is projected into the first two principal components. One sample belonging to animal tissue from the accessory lobule is far from any others. There is no clear grouping pattern for the different respiratory tract sampled locations. Most of the swab and BAL samples grouped together.

A PCA is plotted and coloured separately by sampling method, respiratory tract location, and animal, and is presented in figure A.15. Most of the BAL and swab samples grouped together, except for 2 belonging to animal 3 BAL left caudal and accessory lobules. There is no grouping pattern for the different respiratory tract sampling locations. Interestingly, one sample belonging to tissue of animal 3 accessory lobules appears isolated in the PCA.

A.4.4 Comparison of samples by taxon

Two different methods were applied to determine which taxa are differ in abundance by the different groups: the Lasso method and LEfSe [Segata et al., 2011] which were specifically designed for metagenomics using marker genes.

LASSO method

The Lasso method was applied to a matrix of 100 samples by 72 phyla (after removing mammalian sequences) to determine differences in taxa abundance amongst groups of samples. See chapter 2 on page 30. Nasal samples were excluded because there were too few for this method.

Table A.3: Differential abundance by sampling method. Phylum DA by sampling method. Lasso accuracy 65%. The Lasso coefficients presented here are the logarithmic probability of a phylum being more significant in one sampling method compare to the rest.

Phylum	Tissue	Swab	BAL
Candidatus Amesbacteria			1.634428e+04
Chlorobi			3.843308e+01
Ascomycota	0.17347906		
Bacteroidetes	2.91293347		
Chloroflexi	70.99154603		
Lentisphaerae	134.05896768		
Nitrospirae	199.93152836		
Streptophyta	3.01156983		

Table A.4: Differential abundance by animal. Lasso accuracy 70%. The Lasso coefficients presented here are the logarithmic probability of a phylum being more significant in one animal compare to the rest. Negative coefficients stand for the inverse correlations. For example Tenericutes is found less abundantly in animal 1 compared to the rest. Higher values are more significant than the ones closer to 0.

Phyla	animal 1	animal 2	animal 3	animal 4
Actinobacteria	1.1179408	-0.01213219		
Ascomycota	1.9867143			
Fibrobacteres	22.1525084			
Spirochaetes	13.6271250			
Tenericutes	-1.7011920			
Candidatus Magasanikbacteria		7541.04587362		
Euryarchaeota		25.56243057		
Candidatus Amesbacteria			12743.5748483	
Candidatus Calescamantes			6364.0992082	
Candidatus Marinimicrobia			97.2069565	
Chlorobi			996.2089632	
Chlorophyta			189.8931510	
Aquificae				31.8435685
Candidatus Peregrinibacteria				2071.3704385
Chordata				1614.8389986
Nitrospirae				156.6484326

The Lasso method did not detect any significant phyla over or underrepresented depending on the lung location, and a few phyla were detected when grouping by sampling method and animal. The Lasso coefficients presented in the tables A.3 and A.4 that are high are the more significant ones, and negative coefficients means inverse relationship, are expressed in log scale of the differentially abundant phyla by sampling method and animal, respectively. Except for Actinobacteria (in animal 1) and Ascomycota (tissue), which are in the top 5 most abundant phyla, the major differences were found in rarer phyla. Given the small size of the dataset per group available, and the sparsity of the data, results can vary greatly depending on the partition of train and test data. Nevertheless, it detects the fungal phylum of Ascomycota as more abundant in animal 1 compared to the rest of animals. This result coincide with the clinical observation of consolidation (as described previously).

Biomarker discovery with LEFSE

The Linear discriminant analysis effect size (LEfSe) [Segata et al., 2012] method, specifically designed for metagenomics, was applied per-sample normalising, subject is animal. The threshold on the absolute value of the logarithmic LDA score is 2.

Several combinations of class and subclass have been explored to identify any potential different abundant taxa.The correlations found are quite weak. No taxa were identified for to be DA for lung location, that indicates no difference in the microbiome across the lung lobules.

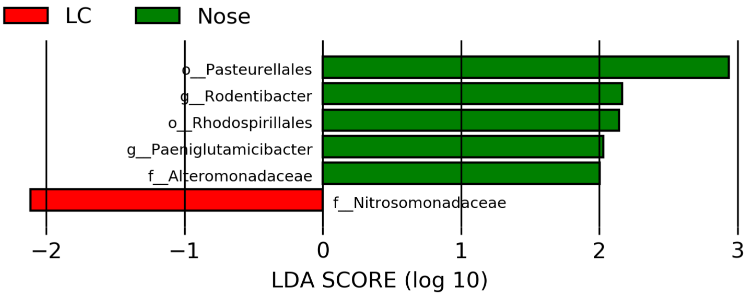


Figure A.16: Biomarkers by respiratory tract location and sampling method. LEfSe method with class lung location and subclass sampling method. A few taxons have been identified as more abundant in nose and the family of Nitrosomonadaceae is less abundant in the left caudal lobule.

The family of Nitosomonadaceae was found to be much less abundant in the left caudal lobule, and in nose the order of Pasturellales, genus of Rodentibacter, order of

Rhodospirilales, genus of Paeniglutamicibacter and family of Altermonadaceae were found to be more abundant (see figure A.16).

When the data is grouped for sampling method and location, a few DA taxa were found, as shown in figure A.17. However, the number of samples in each case is low and results might be biased.

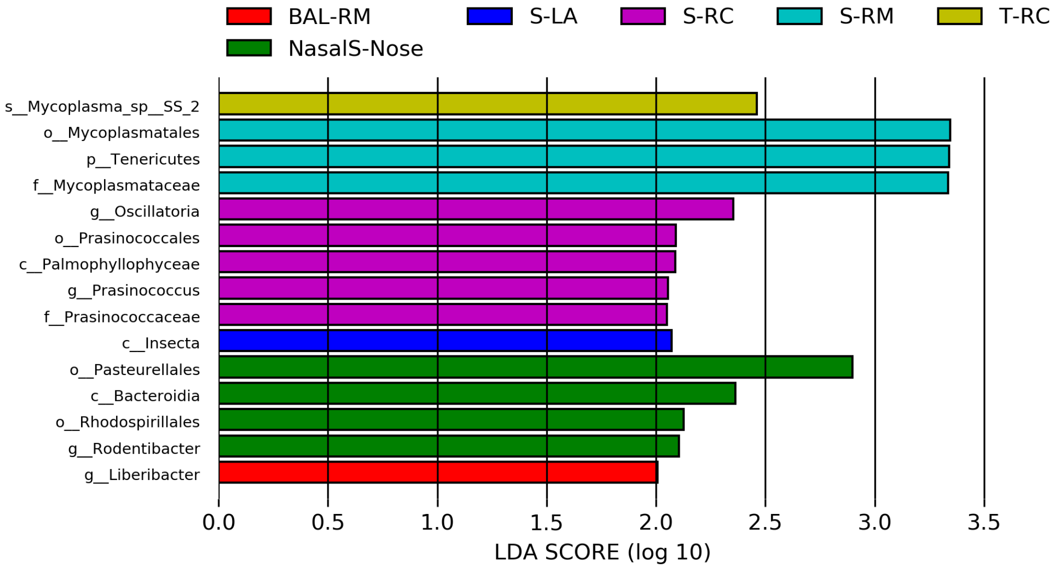


Figure A.17: Biomarkers by grouping sampling method and respiratory tract location. Lefse method with class grouped sampling method-lung location.

The species of Mycoplasma detected in tissue from the right caudal lobule and the taxa detected for swab samples right middle lobule all belong to the phylum Tenericutes.

The taxa found more abundantly in the swab samples of the right caudal lobule all belong to the phylum of Cyanobacteria (green or blue-green algae), probably came through fresh water.

Swab samples of the left apical lobule seem to be somewhat abundant in the class of insects. However, this is a surprising result, since no insect was observed during the sampling. As discussed previously in this thesis, the accuracy of Kraken at the class level is inferior compared to the phylum, it could possibly be a taxonomic assignment error. Although there is a small chance, it could have been picked up while grazing in the fields and managed to get into the respiratory tract.

The nose taxa biomarkers contain mostly the phylum of Proteobacteria, which include the Pasturellaceae and also Bacteroidetes, mainly found in soils, gut and on the skin of animals. The first ones, are typical commensal bacteria found in birds and mammals in the upper respiratory tract.

In the right middle lobule BAL samples, the LEfSe method detected the genus Liberibacter, which is associated with plant disease.

A.5 Discussion

The Powersoil DNA extraction kit works well for different types of sampling methods to characterise the respiratory tract microbiome.

The composition of the microbiota of both experiments described in this chapter were found to be different. This could be due to technical factors, e.g. different versions of databases, or other factors related to the age of the animals or other environmental factors to which each animal was exposed, which is the most likely one.

The samples from the experiment, to characterise the microbiota of the lungs, contained low numbers of sequence data. This might be indicative that the microbiome of young animal lungs is not yet fully established, and might be in the process of colonisation. The animal 1 microbiome was found to be different from the other animals according to the results. At the time of sampling, it was noted that there was a sign of respiratory infection in this animal.

Tenericutes, Actinobacteria, Firmicutes, Proteobacteria have previously been described as some of the most abundant phyla present in the healthy bovine lung [McMullen et al., 2020, Zeineldin et al., 2017a]. However, these same studies also found other highly abundant phyla, namely Fusobacteria and Bacteroidetes, which are not in the top 5 in the present work, which could again be a result of a number of possible environmental or other factors.

Tenericutes was the major phylum found. It contains Mycoplasma spp., which belongs to the clade of Mollicutes and are associated with respiratory infections in humans and other mammals. Because of their medical importance, mollicutes are overrepresented in the genomic databases. Tenericutes can adapt to extreme conditions. Their genome is reduced so much that they lack key functions like regulatory elements, biosynthesis of amino acids and many metabolic requirements, all of which they obtain from the host. However, they are a non-monophyletic clan, and recent studies show the boundary between them and Bacilli is unclear [Wang et al., 2020].

Actinobacteria are the main phyla in some of the animal 1 samples. They are ubiquitous and one of the largest bacterial phyla, which includes Mycobacterium, a species often associated with respiratory infections, e.g. tuberculosis, and leprosy in humans. They also play a central role in the carbon recycling and also produce secondary metabolites, some of which have antifungal properties. Some genera of this phylum are important pathogens in mammals, causing for example tuberculosis in humans (*Mycobacterium tuberculosis*) and in cattle (*Mycobacterium bovis*) [Ul-Hassan and Wellington, 2009].

Ascomycota is one of the largest phyla of Fungi. Some are adapted to extreme environments. This project targeted 16S rRNA gene, however it is a well known fact that the primers for such a gene capture some Eukaryota including fungi. Some of them are pathogenic and have an impact on animal and human health[Nalin N Wijayawardene et al., 2011]. They are identified in a large number of samples of animal 1, in which some clinical signs of infection were observed, and in a lower proportion in a few samples in animal 2. Both belonging to the same farm.

Proteobacteria, found more abundantly in tissue, is a phylum which presents a diverse phenotype, and include the majority of gram negative bacteria. They are often endosymbionts with eukaryotes: α-Proteobacteria are associated with eukaryotic cells, β-Proteobacteria are mainly plant pathogens and animal linked, and γ-Proteobacteria found in insects and vertebrates, including a wide range of animal and human pathogens and non-obligate symbionts [Stackebrandt, 2001].

The phylum Firmicutes is one of the most diverse [Seong et al., 2018]. They are commonly found in water and soil, and also in a mammal's gut, either present as commensals or pathogenic.

The Bacteroidetes phylum is found in many distinct environments, they are especially abundant and may play important roles in the gut [Hahnke et al., 2016]. Most of their species are anaerobic. This phylum is metabolically highly flexible, they can be simultaneously generalist and specialists, providing them with high adaptability to the constantly changing environmental conditions [Johnson et al., 2017]. They were specifically found in the tissue samples, and cattle potentially acquired them while grazing.

To determine the DA of taxa, different methods were explored. As discussed by Weiss et al. [Weiss et al., 2017]. Metagenomics taxa abundance is compositional, given it is sparse. Recently, there has been a booming in a new area of study applied to tackle this issue: Compositional data analysis (CoDa). This type of methods are based on the proportion of number of reads per taxa within a sample. Some of them are stringent and produce low number of significant abundant taxa. The comparison against other type, like DESeq, revealed that often the top hits are similar. However, these methods need to mature and be evaluated against gold standards, once established [Nearing et al., 2022].

Both the Lasso and LEfSe methods, applied to determine DA taxa, found commonalities in the data, for example no differences amongst the different lung lobules. The LEfSe method, which is specifically designed for metagenomics, although seems to work well, there is a debate about the rarefaction employed [Nearing et al., 2022].

The main differences found were amongst each individual sample, and for the combination of sampling methods and lung location. In the experimental conditions tested, tissue samples present a reasonable number of sequencing reads, and it also recovers more varied microbiome. Nevertheless, tissue samples are impossible to obtain from living animals, whereas swabs samples are easier to obtain except for deep into the lungs, and finally BAL is generally easy to get, although ethics approval might be required, which is often a lengthy process.

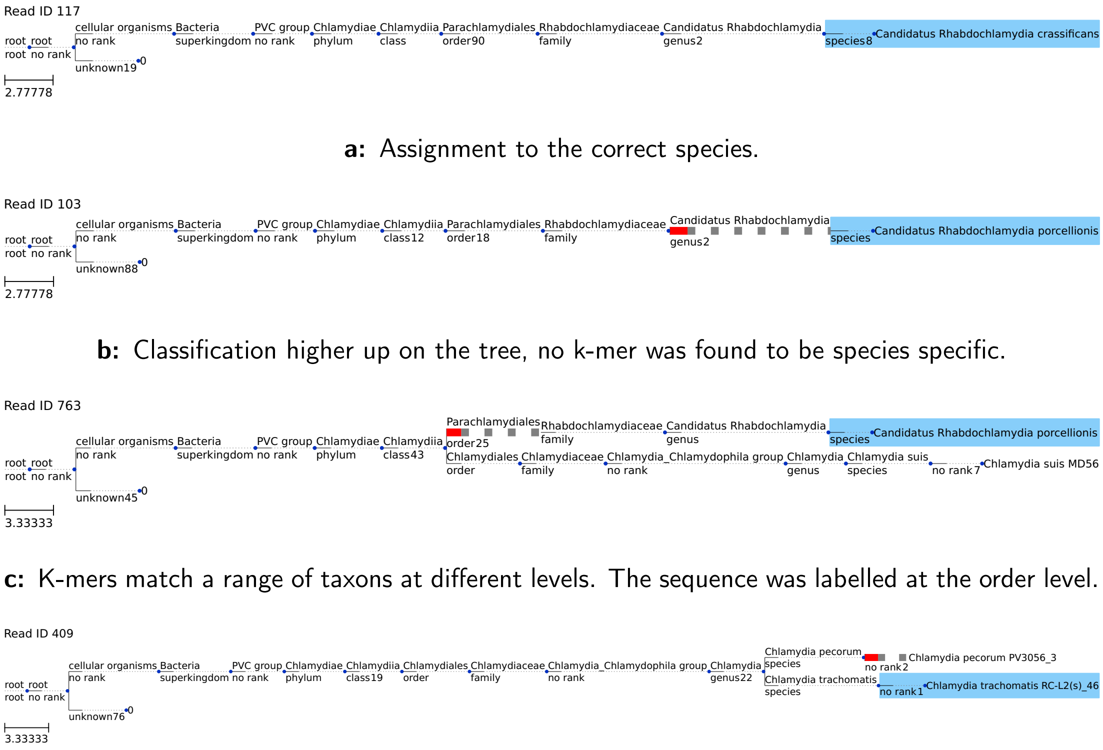
A.5.1 Future work

The nasal and lung lobule microbiota of young cattle has been determined. BRD is still causing major problems in many farms, and it would now be possible to determine pathogenic patterns.

In gut, the proportion of Firmicutes - Bacteroidetes has been proposed as an obesity biomarker [Magne et al., 2020]. However, the relationship, if any, with the respiratory tract remains unclear. Therefore, more work needs to be done to determine any potential correlation.

Appendix B

Supplementary figures and tables



a: Assignment to the correct species.

b: Classification higher up on the tree, no k-mer was found to be species specific.

c: K-mers match a range of taxons at different levels. The sequence was labelled at the order level.

d: K-mers match similar species and the sequence was wrongly assigned.

Figure B.1: kraken classification examples. Taxonomic trees where kraken found k-mers matching during the classification process. Above the lines there is the name of the clade and under there is the corresponding rank. Next to the rank, there is the number of matching k-mers (if any). Highlighted in blue is the original species. In red where the sequence was classified elsewhere that wasn't the origin. Note that in all the cases there are k-mers in the sequence that are not present in the original database. These correspond to the mutations introduced during the simulation.

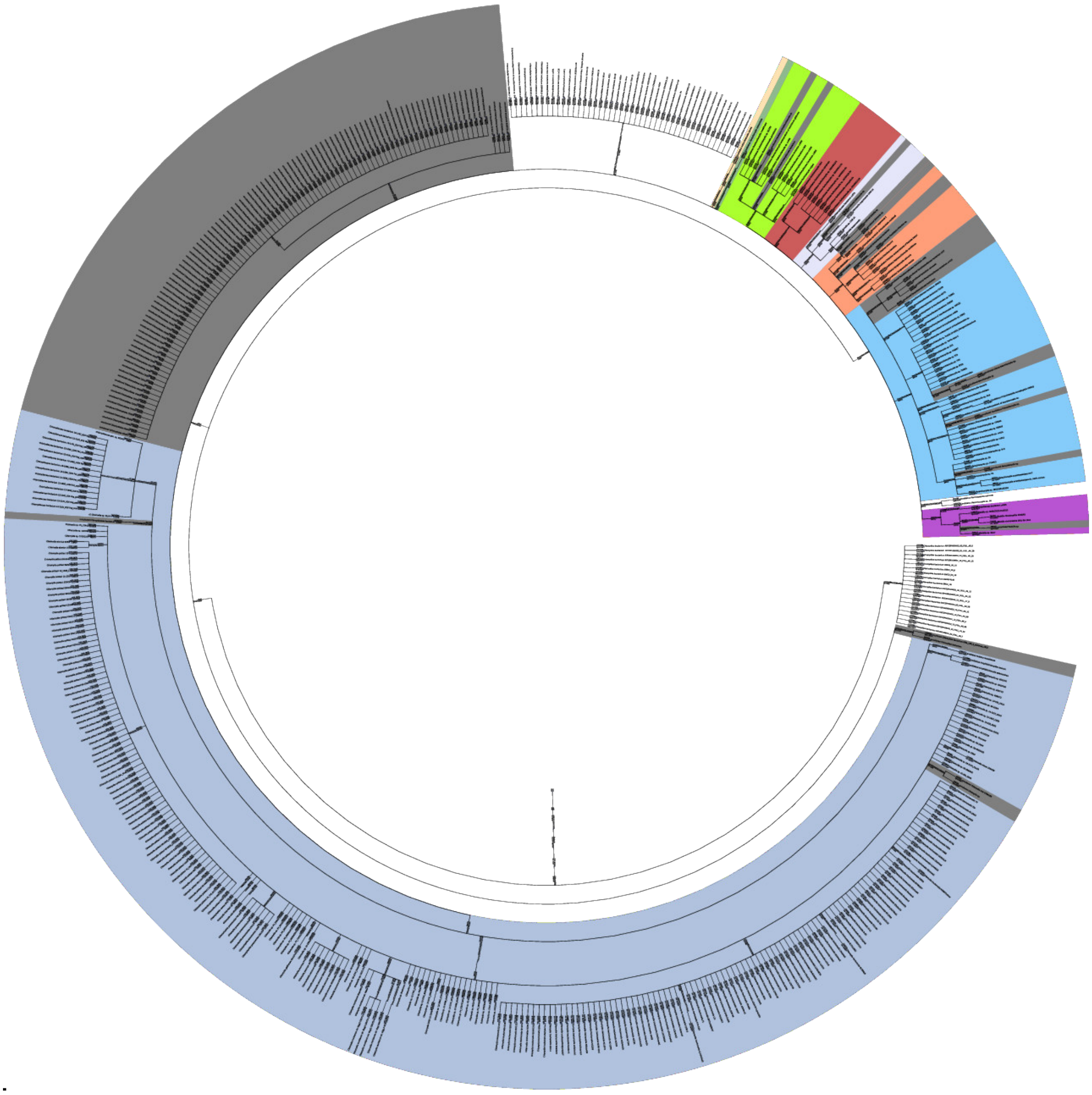


Figure B.2: Taxonomic tree of the Chlamydiae phylum. NCBI taxonomy of the Chlamydiae phylum which have sequences at the Silva database SSU RefNR99 version 128. Each order is highlited in a different color.

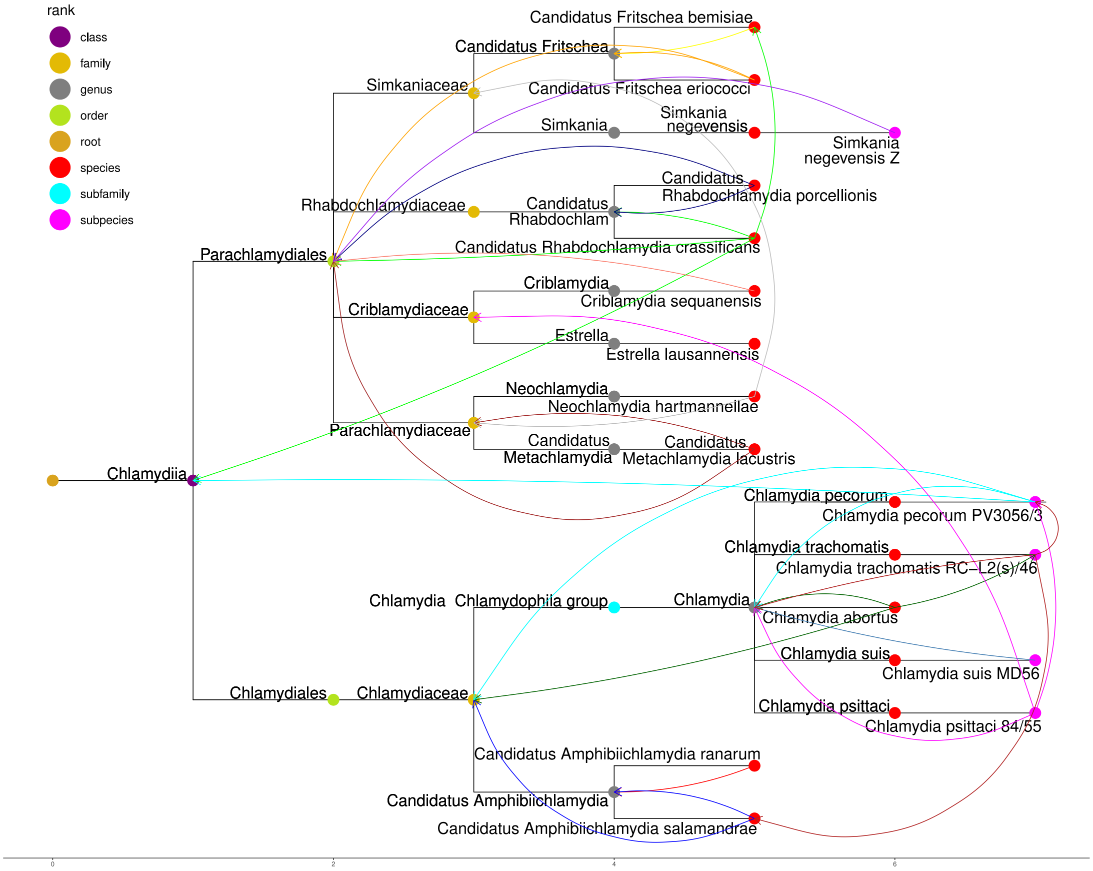
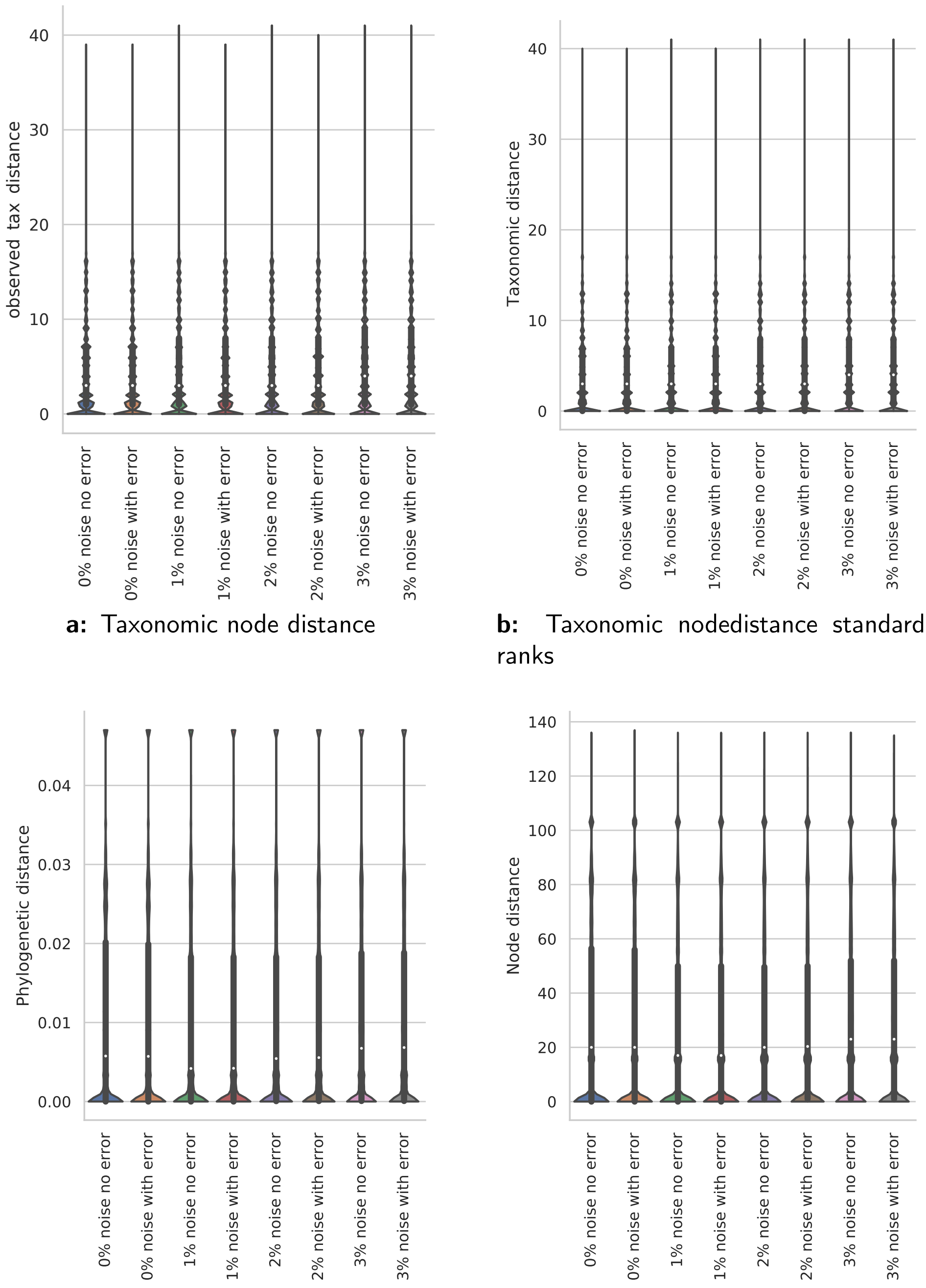


Figure B.3: Assignment to the wrong taxa. This diagram shows the taxonomic tree of the 16 selected species. A simple mock community is created and short reads simulated, and only the 16 sequences belonging to the selected species are used as reference database for the taxonomic classification. Under this condition, Kraken classifies the vast majority of short reads correctly. However, a few are not labelled to the correct species or subspecies. The black lines correspond to the taxonomic tree, taxonomic nodes (represented by dots) are coloured by ranks and the name is printed adjacently. The coloured curvy lines illustrate where the mislabelled sequences have been assigned to in the taxonomic tree. Arrows originate at leaves (corresponding to species or subspecies ranks). All the arrows from the same leaf have the same colour, and is different for each leaf. Some misclassified reads have been assigned to one or two taxonomic nodes away (e.g. species *Candidatus Amphibiichlamydia ranarum* assigned to the genus level of Candidatus Amphibiichlamydia, belonging to the same taxonomic clan). The great majority of misassigned taxonomies are assigned within the same taxonomic lineage, and only 9 sequences, all belonging to *Chlamydia psittaci 84_55* are completely labelled in the wrong lineage (order rank). We can also observe that sequences belonging to the same species are not consistently mislabelled to the same taxon. A possible cause could be the region which the short reads belong to and the number of errors.



a: Taxonomic node distance

b: Taxonomic nodedistance standard ranks

c: Phylogenetic branch distance

d: Phylogenetic node distance

Figure B.4: New distance metrics by experiment. The subfigures represent the distribution values for the new proposed measures by the different experiments (from no mutation to 3% background noise and with error mutations). TND is the number of nodes that differ from the original species of a read to where it has been classified to. Similarly TNDSR, is the same but only counting the main taxonomic ranks nodes. The PBD is the number of nodes in a phylogenetic tree that differ between the origin of a read to the assignment. Finally, the PND is the phylogenetic distance between original leaf and assigned node in a phylogenetic tree. Overall, the new metrics capture the average distance of misclassification. Kraken shows that the TND (**a**) and TNDSR (**b**) have almost imperceptible groth with the number of mutations introduced. PBD (**c**) and PND (**d**) are very similar across the experiments. The results indicate that kraken is robust to the introduction of mutations.

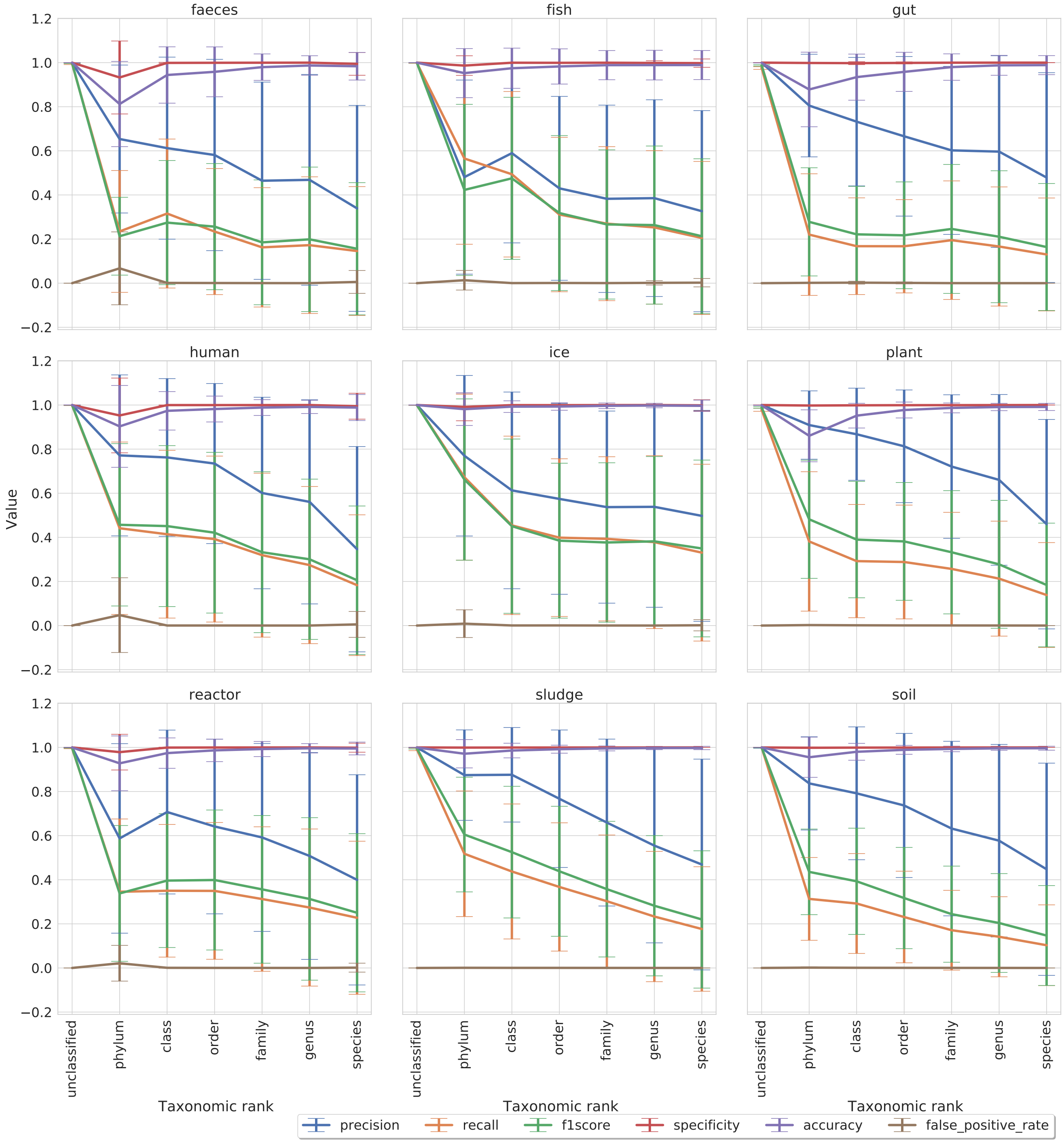


Figure B.5: Metrics by sample. The metrics diverge quite a lot amongst the samples. That is due to the content of the sample in respect of the database. While soil presents the the best overall performance reactor shows one of the poorest. This can be due to the fact it contains a number of Archaea but none are present in the reference database (MGC data).

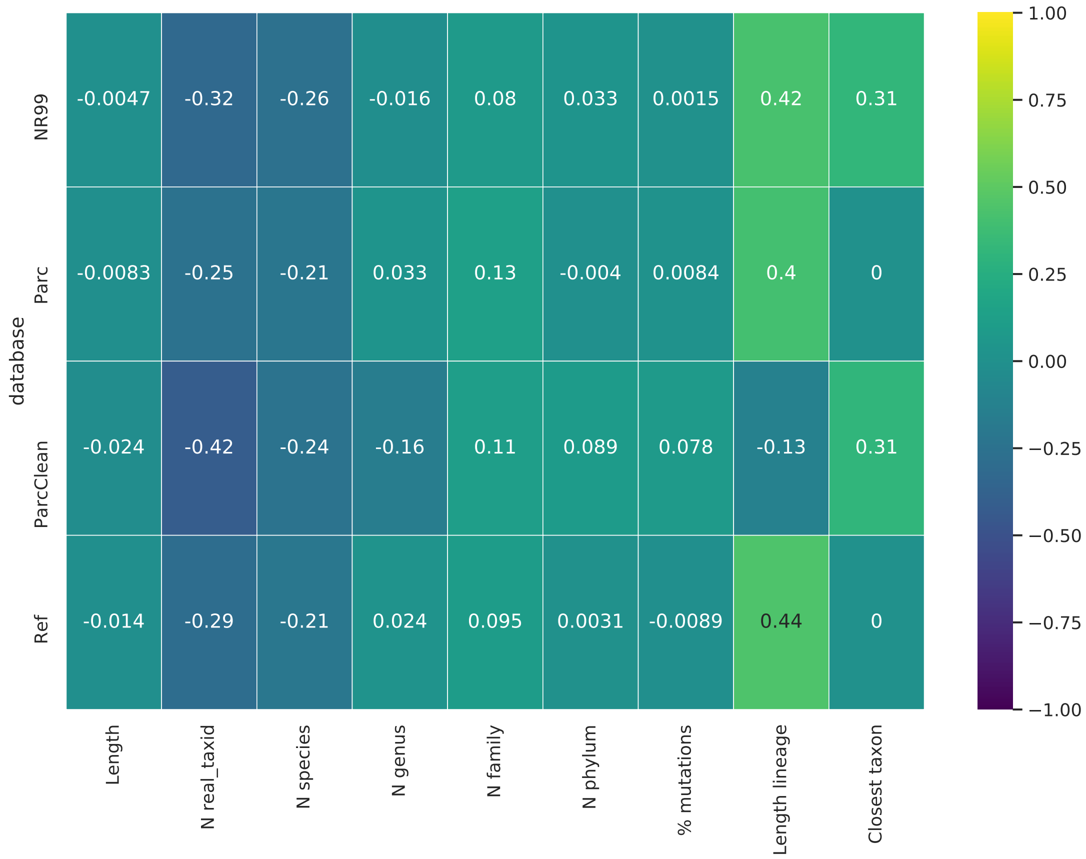


Figure B.6: Spearman correlation of the positive control data. "N" refers to the number of sequences at different taxonomic ranks, and "length" to the number of base pairs simulated. The number of sequences at the leaf of taxonomic trees are contributing to classify the short read closer. Longer number of nodes in a taxonomic lineage (length lineage) and further common ancestors (closest taxon) have a negative impact in the classification, and this increases TND

Table B.1: Viral labelled sequences in Silva taxonomy.

NCBI taxid	Accessions *	NCBI and Silva taxonomic lineages
36452	AF065755.1.676 AF191073.2766.3620	Viruses; Duplodnaviria; Heunggongvirae; Peploviricota; Herviviricetes; Herpesvirales; Herpesviridae; Betaherpesvirinae; Cytomegalovirus; Cercopithecine betaherpesvirus 5; Stealth virus 1 Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiaceae; Ochrobactrum; Stealth virus 1
99287	AE006468.4394688.4396232 AE006468.4196072.4197613 AE006468.289190.290733 AE006468.2800121.2801663 AE006468.3570470.3572013 AE006468.4100145.4101688 AE006468.4351143.4352686	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella; Salmonella enterica; Salmonella enterica subsp. enterica; Salmonella enterica subsp. enterica serovar Typhimurium; Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella; Salmonella virus Fels2
262728	CP002277.1779959.1781497 CP002277.1904650.1906188 CP002277.478016.479554 CP002277.357249.358787 CP002277.709269.710807 CP002277.1714279.1715817	Bacteria; Proteobacteria; Gammaproteobacteria; Pasteurellales; Pasteurellaceae; Haemophilus; Haemophilus influenzae; Haemophilus influenzae R2866 Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Pasteurellaceae; Haemophilus; Haemophilus virus HP2
1221328	CP008698.166510.168047 CP008698.30287.31824 CP008698.171512.173049 CP008698.90544.92081 CP008698.96400.97937 CP008698.9819.11356 CP008698.618990.620527 CP008698.930253.931790 CP008698.3155132.3156669 CP008698.160901.162438	Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus subtilis group; Bacillus subtilis; Bacillus subtilis subsp. subtilis; Bacillus subtilis subsp. subtilis str. AG1839 Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus virus SPbeta

1232554	CP007800.160901.162438 CP007800.930253.931790 CP007800.618990.620527 CP007800.9819.11356 CP007800.171512.173049 CP007800.90544.92081 CP007800.30287.31824 CP007800.3149861.3151398 CP007800.96400.97937 CP007800.166510.168047	Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus subtilis group; Bacillus subtilis; Bacillus subtilis subsp. subtilis; Bacillus subtilis subsp. subtilis str. JH642; Bacillus subtilis subsp. subtilis str. JH642 substr. AG174 Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus virus SPbeta
186617	BCRZ01001786.50.1614 BCSB01007630.22.1149 BCRW01002018.1.644 BCSF01026815.1.1043 BCSF01020973.1.1132	Viruses; environmental samples; uncultured marine virus Eukaryota; Archaeplastida; Chloroplastida; Chlorophyta; Mamiellophyceae; Mamiellales; Micromonas; uncultured marine virus
239364	AAMH01004094.1.727 AAMH01004484.1.831 AAMH01004487.1.820 AAMI01003277.1.886 AAMH01004561.8.906 AAMH01004365.1.787 AAMH01004441.1.808 AAMH01004478.1.865 AAMH01004442.1.808 AAMH01004568.21.905 AAMI01003326.6.853	Viruses; environmental samples; uncultured human fecal virus Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Xanthobacteraceae; uncultured; uncultured human fecal virus

1070528	MAVL01000973.46.795 MAVM01000094.8.913 MAVO01000069.1.831 MAVJ01000039.1.921 MAVK01000017.1.1314 MAVM01000102.4.886 MAVI01000089.4.740 MAVL01000555.91.1013 MAVJ01000057.1.884 MAVN01000025.87.1541 MAVG01000031.10.1072 MAVM01000112.1.859 MAVG01000032.238.1418 MAVO01000046.1.1190 MAVJ01000031.1.920 MAVK01000010.1.1037 MAVL01000618.164.963 MAVN01000046.1.1190 MAVO01000025.87.1541 MAVJ01000028.1.1047 MAVN01000069.1.831 MAVM01000062.1.1054 MAVL01000832.16.849 MAVM01000019.37.1451 MAVI01000061.1.849 MAVJ01000052.30.900 MAVI01000055.8.871 MAVI01000036.3.1003	unclassified entries; unclassified sequences; metagenomes; organismal metagenomes; viral metagenome Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Xanthobacteraceae; Bradyrhizobium; viral metagenome

* The accession number consist of the sequence id and then separated by dots thestart and end of the 16/18S rRNA sequence predicted by Silva. In the taxonomies column, in black are the NCBI and in green Silva.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len
Stealth virus 1 clone 3B43 T3	Stealth virus 1	1227	1227	99%	0.0	100.00	814
Ochrobactrum pituitosum strain NIHHS108 16S ribosomal RNA gene, partial sequence	Ochrobactrum pituitosum	1114	1114	99%	0.0	96.59	1329
Stealth virus 1 clone 3B43, genomic sequence	Stealth virus 1	1112	1112	99%	0.0	96.74	3620
Ochrobactrum rhizosphaerae strain L35 16S ribosomal RNA gene, partial sequence	Ochrobactrum rhizosphaerae	1109	1109	99%	0.0	96.59	1346
Ochrobactrum sp. strain FA75 16S ribosomal RNA gene, partial sequence	Ochrobactrum sp.	1109	1109	99%	0.0	96.59	1439
Ochrobactrum sp. strain H140 16S ribosomal RNA gene, partial sequence	Ochrobactrum sp.	1109	1109	99%	0.0	96.59	1412
Ochrobactrum rhizosphaerae strain Sample$_{8}516SribosomalRNAgene,partialsequence$	Ochrobactrum rhizosphaerae	1109	1109	99%	0.0	96.59	1344
Ochrobactrum rhizosphaerae strain Sample$_{2}016SribosomalRNAgene,partialsequence$	Ochrobactrum rhizosphaerae	1109	1109	99%	0.0	96.59	1344
Ochrobactrum sp. strain PN2-B07P1-15 16S ribosomal RNA gene, partial sequence	Ochrobactrum sp.	1109	1109	99%	0.0	96.59	1049
Ochrobactrum sp. strain PN2-B04P2-22 16S ribosomal RNA gene, partial sequence	Ochrobactrum sp.	1109	1109	99%	0.0	96.59	1053
Ochrobactrum sp. strain PN2-B04P2-17 16S ribosomal RNA gene, partial sequence	Ochrobactrum sp.	1109	1109	99%	0.0	96.59	1062
Ochrobactrum pituitosum strain Saad11 16S ribosomal RNA gene, partial sequence	Ochrobactrum pituitosum	1109	1109	99%	0.0	96.59	834
Ochrobactrum grignonense strain BN 16S ribosomal RNA gene, partial sequence	Ochrobactrum grignonense	1109	1109	99%	0.0	96.59	930
Ochrobactrum sp. strain RPTAtOch1 16S ribosomal RNA gene, partial sequence	Ochrobactrum sp.	1109	1109	99%	0.0	96.59	1346

Table B.2: Blast results sequence AF065755.1.686.

Top 15 results of web NCBI blastn search with default parameters. The first hit is the original sequence, taxonomically annotated as viral. The rest of the hits with high coverage and identity match partial 16S rRNA sequences of the bacterial species Ochrobatrum.

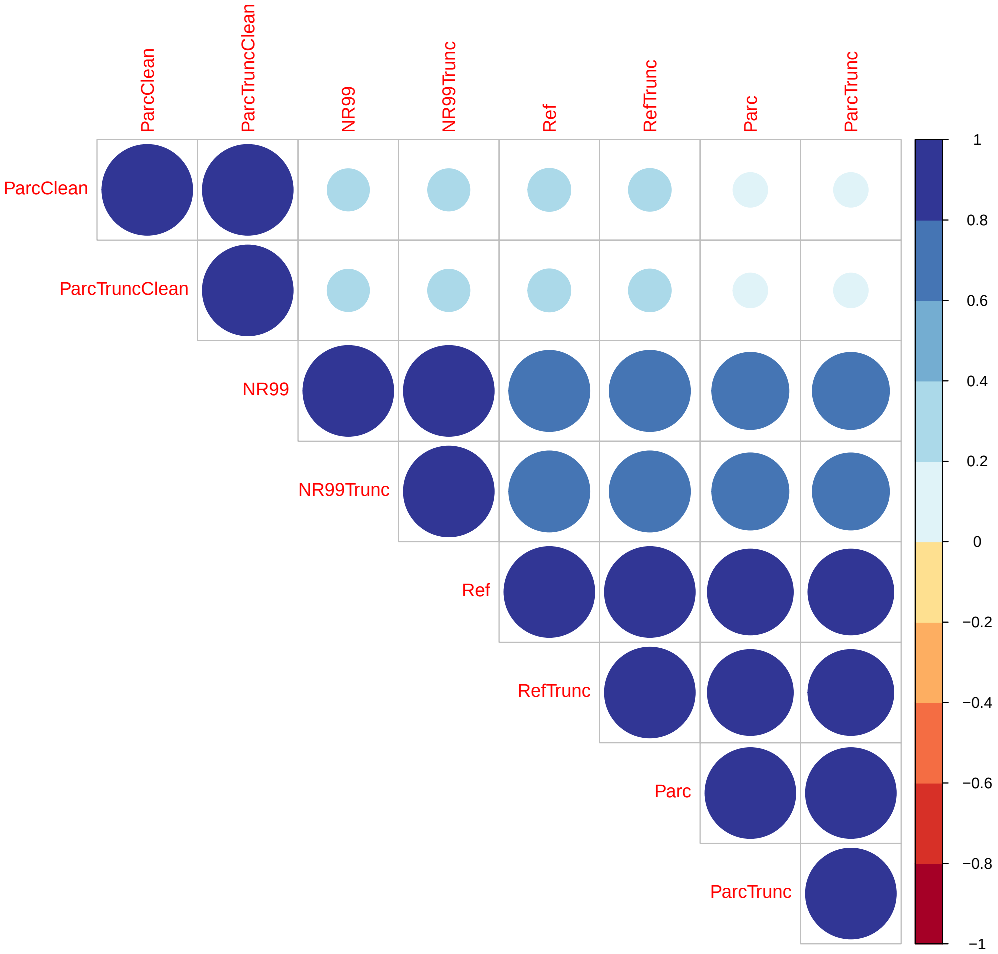


Figure B.7: Spearman correlation of TND by database. Similarity on classification success based on the scaled TND of the MGC data. ParcClean and ParcTrunc, which are the best performing, have lower correlation with the rest and identical one to the other. NR99 and NR99Trunc with intermediate correlation values.

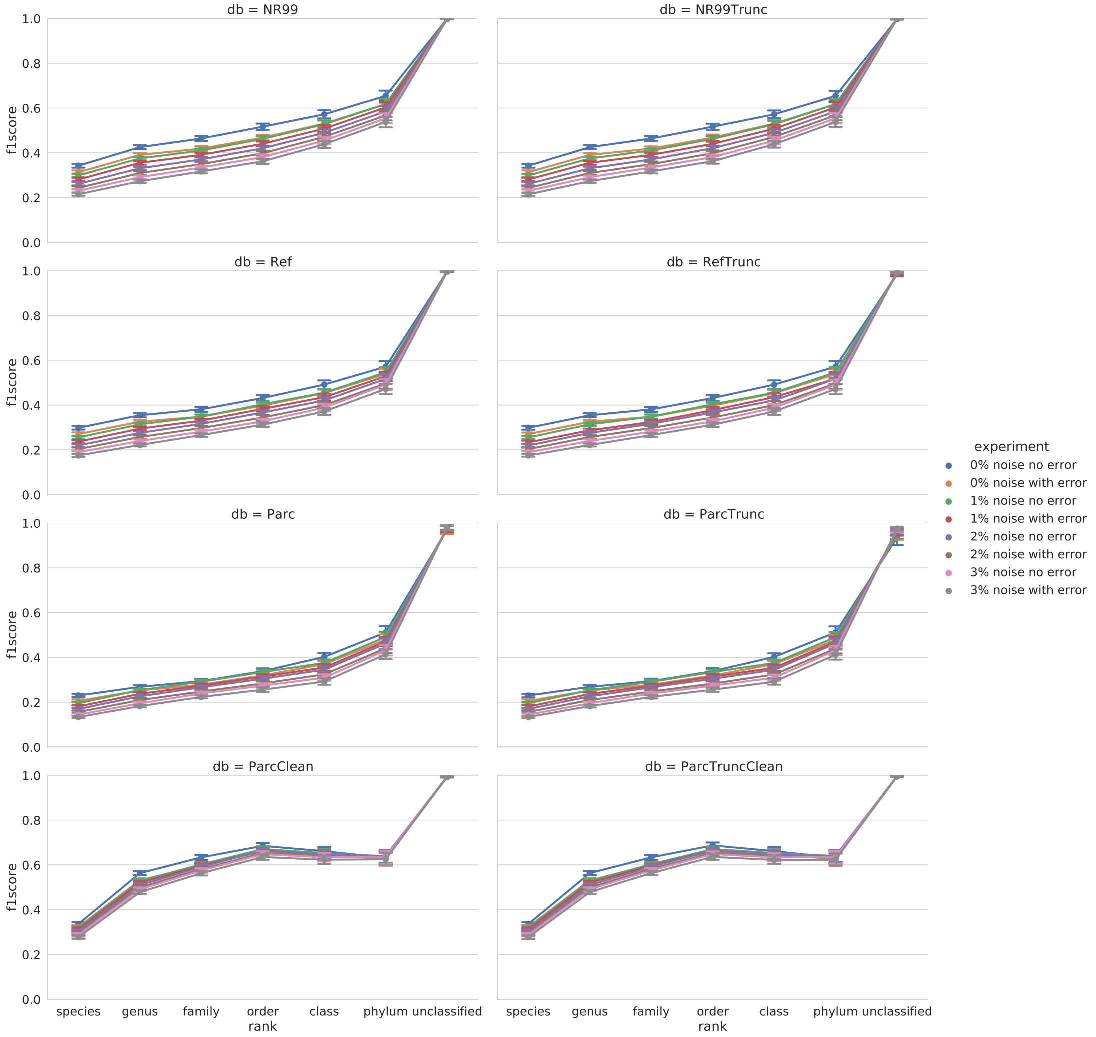


Figure B.8: F1 score by accumulated mutations in the simulated sets per database. The data was MGC. The experiments are the same sets with different number of mutations. There are two sources: the first are randomly introduced mutations in the reads, ranging from 0 to 3%, and the second are the mutations linked to the quality of sequencing. Kraken2 is not capable to 100% identify short reads without mutations (blue line, "0% noise no error"), which are also present in the databases. In general, there is a tendency for less accurate results with increased number of mutations introduced. Higher number of sequences present in the database seem to contribute to resilience of the number of mutations in the reads. Interestingly, results show almost overlapping f1 scores for the clean Parc databases.

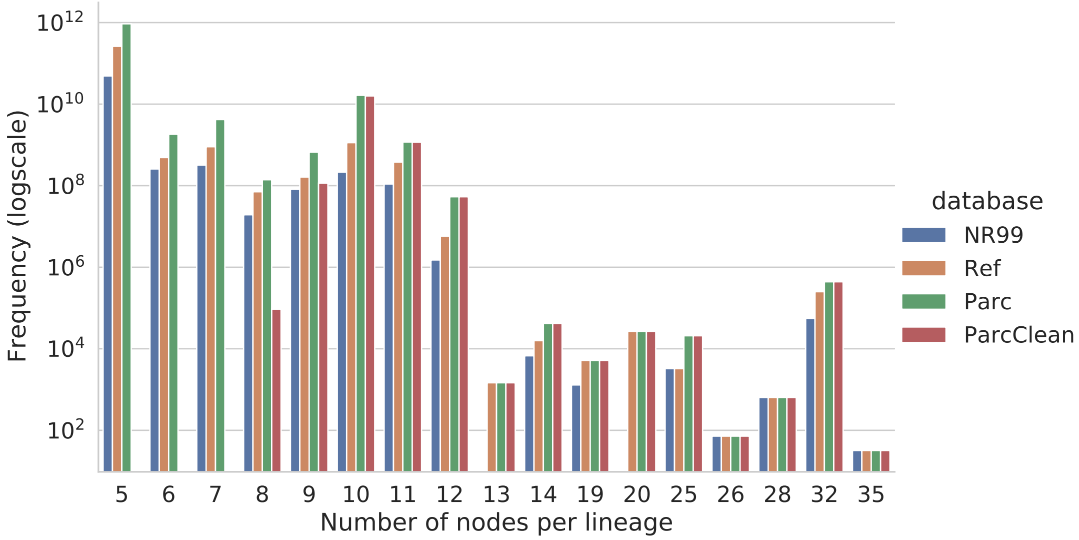
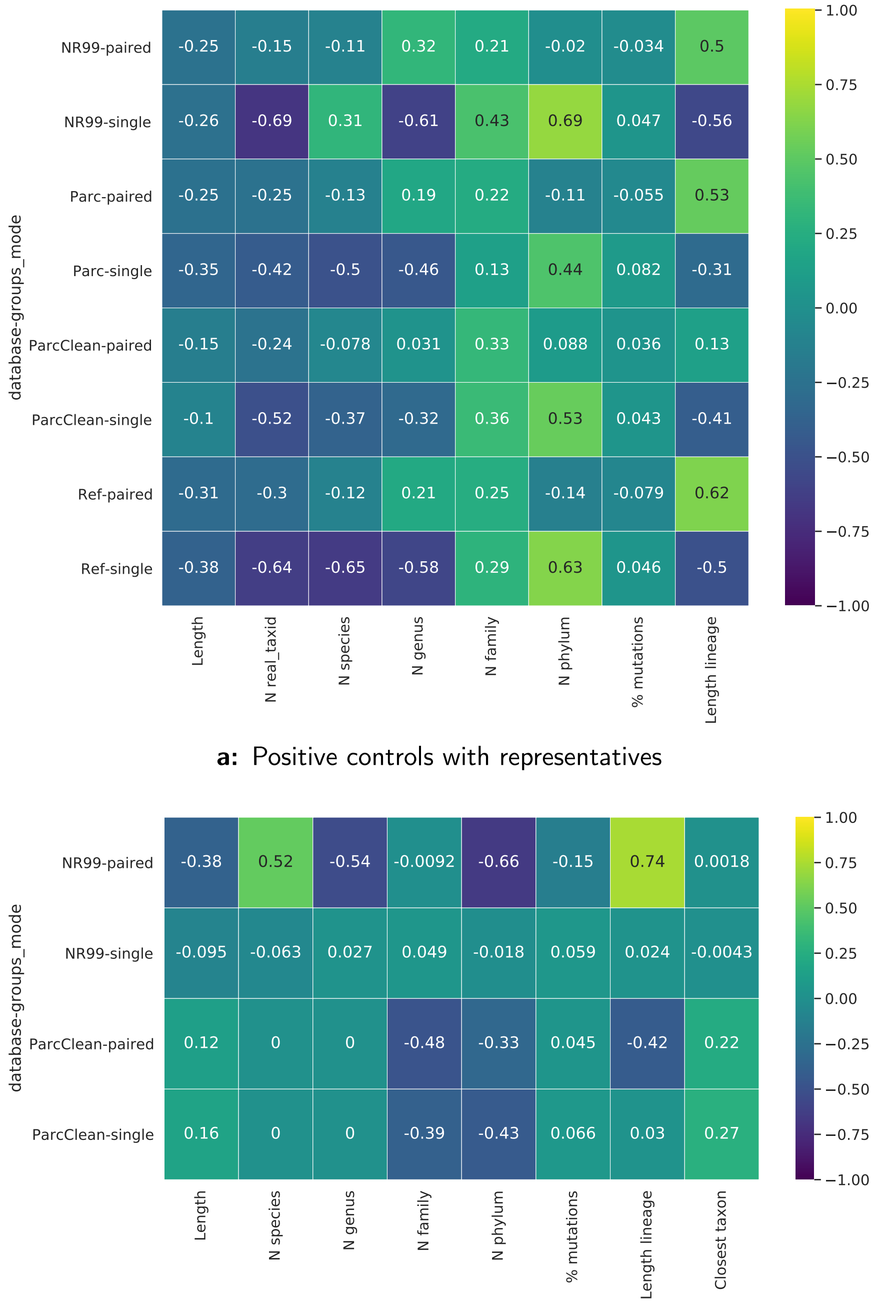


Figure B.9: Number of sequences per lineage length. Each database contains a different subset of taxa, which have varying lineage length. ParcClean does not contain lineages with 7 or less annotated ranks. The most common number of nodes per lineage is 5. The longest lineages contain 35 nodes.



a: Positive controls with representatives

b: Positive controls without representatives

Figure B.10: Spearman correlation of factors influencing taxonomic classification of positive controls by single or paired-end data and representatives. Spearman correlation of the TND of the positive controls depends on the presence *a* or absence *b* of the original taxon in the databases. N is the number of sequences per rank (species, genus, family, phylum) present in a given database. Lineage length is the number of taxa described in the NCBI database on 18 November 2020. Also, the correlation values present differences depending on the type of data, single or paired-end. Generally, the number of sequence representing a phylum in the database show a positive correlation in both cases, whereas in many other cases there is an opposite trend. For those taxa without representatives in the database, no factor clearly influences classification for single-end data in the NR99 database. However, for the NR99 and paired-end data the read length , number of genera and number of phyla improves identification whereas longer lineages and more sequences in the database at species level make things worse.

Bibliography

[Abdelkareem et al., 2020] Abdelkareem, A. O., Khalil, M. I., Elbehery, A. H., and Abbas, H. M. (2020). Viral Sequence Identification in Metagenomes using Natural Language Processing Techniques. *bioRxiv*.

[Afshinnekoo et al., 2017] Afshinnekoo, E., Chou, C., Alexander, N., Ahsanuddin, S., Schuetz, A. N., and Mason, C. E. (2017). Precision metagenomics: Rapid metagenomic analyses for infectious disease diagnostics and public health surveillance. *Journal of Biomolecular Techniques*, 28(1):40–45.

[Almeida et al., 2018] Almeida, A., Mitchell, A. L., Tarkowska, A., and Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience*, 7(5):1–10.

[Alves et al., 2016] Alves, J. M. P., de Oliveira, A. L., Sandberg, T. O. M., Moreno-Gallego, J. L., de Toledo, M. A. F., de Moura, E. M. M., Oliveira, L. S., Durham, A. M., Mehnert, D. U., Zanotto, P. M. d. A., Reyes, A., and Gruber, A. (2016). GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data. *Frontiers in Microbiology*, 7(March):1–15.

[Andrews, 2010] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

[Angly et al., 2012] Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., and Tyson, G. W. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic acids research*, 40(12):e94.

[Arita et al., 2021] Arita, M., Karsch-Mizrachi, I., and Cochrane, G. (2021). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 49(D1):D121–D124.

[Bagheri et al., 2020] Bagheri, H., Severin, A. J., and Rajan, H. (2020). Detecting and correcting misclassified sequences in the large-scale public databases. *Bioinformatics*, 36(18):4699–4705.

[Benoit et al., 2016] Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., and Lemaitre, C. (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2(e94).

[Bharti and Grimm, 2021] Bharti, R. and Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, 22(1):178–193.

[Bharucha et al., 2020] Bharucha, T., Oeser, C., Balloux, F., Brown, J. R., Carbo, E. C., Charlett, A., Chiu, C. Y., Claas, E. C. J., Goffau, M. C. d., Vries, J. J. C. d., Eloit, M., Hopkins, S., Huggett, J. F., MacCannell, D., Morfopoulou, S., Nath, A., O'Sullivan, D. M., Reoma, L. B., Shaw, L. P., Sidorov, I., Simner, P. J., Tan, L. V., Thomson, E. C., Dorp, L. v., Wilson, M. R., Breuer, J., and Nigel Field (2020). STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies. *The Lancet. Infectious Diseases*, pages 19–20.

[Brady and Leber, 2017] Brady, M. T. and Leber, A. (2017). *Less Commonly Encountered Nonenteric Gram-Negative Bacilli*. Elsevier Inc., fifth edit edition.

[Breitwieser et al., 2019] Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 20(4):1125–1139.

[Callahan et al., 2017] Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, 11(12):2639–2643.

[Capella-Gutierrez et al., 2017] Capella-Gutierrez, S., Iglesia, D. d. l., Haas, J., Lourenco, A., Fernández, J. M., Repchevsky, D., Dessimoz, C., Schwede, T., Notredame, C., Gelpi, J. L., and Valencia, A. (2017). Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking. *bioRxiv*, (September):181677.

[Chakravorty et al., 2007] Chakravorty, S., Helb, D., Burday, M., and Connell, N. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods*, 69(2):330–339.

[Chen et al., 2019] Chen, C. Y., Tang, S. L., and Chou, S. C. T. (2019). Taxonomy based performance metrics for evaluating taxonomic assignment methods. *BMC Bioinformatics*, 20(1):1–11.

[Cole et al., 2014] Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., and Tiedje, J. M. (2014). Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(D1):633–642.

[Cong et al., 2017] Cong, Y., Chan, Y. b., Phillips, C. A., Langston, M. A., and Ragan, M. A. (2017). Robust inference of genetic exchange communities from microbial genomes using TF-IDF. *Frontiers in Microbiology*, 8(JAN):1–11.

[Cusack et al., 2003] Cusack, P., McMeniman, N., and Lean, I. J. (2003). The medicine and epidemiology of bovine respiratory disease in feedlots. *Australian veterinary journal*, 81(8):480–487.

[Davids et al., 2016] Davids, M., Hugenholtz, F., Martins dos Santos, V., Smidt, H., Kleerebezem, M., and Schaap, P. J. (2016). Functional Profiling of Unfamiliar Microbial Communities Using a Validated De Novo Assembly Metatranscriptome Pipeline. *Plos One*, 11(1):e0146423.

[De Simone et al., 2020] De Simone, G., Pasquadibisceglie, A., Proietto, R., Polticelli, F., Aime, S., J.M. Op den Camp, H., and Ascenzi, P. (2020). Contaminations in (meta)genome data: An open issue for the scientific community. *IUBMB Life*, 72(4):698–705.

[DeDonder and Apley, 2015] DeDonder, K. and Apley, M. (2015). A literature review of antimicrobial resistance in Pathogens associated with bovine respiratory disease. *Animal Health Research Reviews*, 16(2):1–10.

[DeSantis et al., 2006] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072.

[Dixit, 2021] Dixit, K. (2021). Benchmarking of 16S rRNA gene databases using known strain sequences. *Bioinformation*, 17(3):377–391.

[Eddy and Durbin, 1994] Eddy, S. R. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088.

[Edgar, 2010] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

[Escobar-Zepeda et al., 2015] Escobar-Zepeda, A., Vera-Ponce de León, A., and Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics*, 6(December):348.

[Espejo and Plaza, 2018] Espejo, R. T. and Plaza, N. (2018). Multiple Ribosomal RNA operons in bacteria; Their concerted evolution and potential consequences on the rate of evolution of their 16S rRNA. *Frontiers in Microbiology*, 9(JUN):1–6.

[Fu et al., 2012] Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

[Garbarine et al., 2011] Garbarine, E., Depasquale, J., Gadia, V., Polikar, R., and Rosen, G. (2011). Information-theoretic approaches to SVM feature selection for metagenome read classification. *Computational Biology and Chemistry*, 35(3):199–209.

[Gardner et al., 2019] Gardner, P. P., Watson, R. J., Morgan, X. C., Draper, J. L., Finn, R. D., Morales, S. E., and Stott, M. B. (2019). Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ*, 7(e6160):1–19.

[Gershwin et al., 2015] Gershwin, L. J., Van Eenennaam, A. L., Anderson, M. L., McEligot, H. A., Shao, M. X., Toaff-Rosenstein, R., Taylor, J. F., Neibergs, H. L., and Womack, J. (2015). Single Pathogen Challenge with Agents of the Bovine Respiratory Disease Complex. *Plos One*, 10(11):e0142479.

[Glassman and Martiny, 2018] Glassman, S. I. and Martiny, J. B. H. (2018). Broad-scale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere*, 3(4).

[Glöckner et al., 2017] Glöckner, F. O., Yilmaz, P., Quast, C., Gerken, J., Beccati, A., Ciuprina, A., Bruns, G., Yarza, P., Peplies, J., Westram, R., and Ludwig, W. (2017). 25 years of serving the community with ribosomal RNA gene reference databases and tools. *Journal of Biotechnology*, 261(June):169–176.

[Grissett et al., 2015] Grissett, G., White, B., and Larson, R. (2015). Structured Literature Review of Responses of Cattle to Viral and Bacterial Pathogens Causing Bovine Respiratory Disease Complex. *Journal of Veterinary Internal Medicine*, 29(3):770–780.

[Hahnke et al., 2016] Hahnke, R. L., Meier-Kolthoff, J. P., García-López, M., Mukherjee, S., Huntemann, M., Ivanova, N. N., Woyke, T., Kyrpides, N. C., Klenk, H. P., and Göker, M. (2016). Genome-based taxonomic classification of Bacteroidetes. *Frontiers in Microbiology*, 7(DEC).

[Harrison et al., 2021] Harrison, P. W., Ahamed, A., Aslam, R., Alako, B. T., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., Holt, S., Ibrahim, T., Ivanov, E., Jayathilaka, S., Kadhirvelu, V. B., Kumar, M., Lopez, R., Kay, S., Leinonen, R., Liu, X., O'Cathail, C., Pakseresht, A., Park, Y., Pesant, S., Rahman, N., Rajan, J., Sokolov, A., Vijayaraja, S., Waheed, Z., Zyoud, A., Burdett, T., and Cochrane, G. (2021). The European Nucleotide Archive in 2020. *Nucleic Acids Research*, 49(D1):D82–D85.

[Hause et al., 2015] Hause, B. M., Collin, E. A., Anderson, J., Hesse, R. A., and Anderson, G. (2015). Bovine rhinitis viruses are common in U.S. cattle with bovine respiratory disease. *PLoS ONE*, 10(3):1–12.

[Holman et al., 2017] Holman, D. B., Klima, C. L., Ralston, B. J., Niu, Y. D., Stanford, K., Alexander, T. W., and McAllister, T. A. (2017). Metagenomic Sequencing of Bronchoalveolar Lavage Samples from Feedlot Cattle Mortalities Associated with Bovine Respiratory Disease. *Genome Announcements*, 5(40):01045–17.

[Holman et al., 2015] Holman, D. B., Timsit, E., and Alexander, T. W. (2015). The nasopharyngeal microbiota of feedlot cattle. *Scientific Reports*, 5(15557):1–9.

[Hornung et al., 2019] Hornung, B. V., Zwittink, R. D., and Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiology Ecology*, 95(5):1–7.

[Howe et al., 2021] Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Ridwan Amode, M., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., Giron, C. G., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O. G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J. G., Marugán, J. C., Maurel, T., McMahon, A. C., Mohanan, S., Moore, B., Muffato, M., Oheh,

D. N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M. P., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., de Silva, N., Flint, B., Frankish, A., Hunt, S. E., IIsley, G. R., Langridge, N., Loveland, J. E., Martin, F. J., Mudge, J. M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S. J., Cunningham, F., Yates, A. D., Zerbino, D. R., and Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891.

[Hu et al., 2021] Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811.

[Imelfort et al., 2014] Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P., and Tyson, G. W. (2014). GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603.

[James et al., 2021] James, G. L., Latif, M. T., Isa, M. N. M., Bakar, M. F. A., Yusuf, N. Y. M., Broughton, W., Murad, A. M., and Abu Bakar, F. D. (2021). Metagenomic datasets of air samples collected during episodes of severe smoke-haze in Malaysia. *Data in Brief*, 36:107124.

[Jeske and Gallert, 2022] Jeske, J. T. and Gallert, C. (2022). Microbiome Analysis via OTU and ASV-Based Pipelines-A Comparative Interpretation of Ecological Data in WWTP Systems.

[Johnson et al., 2017] Johnson, E. L., Heaver, S. L., Walters, W. A., and Ley, R. E. (2017). Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes. *Journal of Molecular Medicine*, 95(1).

[Johnson et al., 2019] Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., and Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):1–11.

[Johnston et al., 2017] Johnston, D., Earley, B., Cormican, P., Murray, G., Kenny, D. A., Waters, S. M., McGee, M., Kelly, A. K., and McCabe, M. S. (2017). Illumina MiSeq 16S amplicon sequence analysis of bovine respiratory disease associated bacteria in lung and mediastinal lymph node tissue. *BMC Veterinary Research*, 13(1):118.

[Joos et al., 2020] Joos, L., Beirinckx, S., Haegeman, A., Debode, J., Vandecasteele, B., Baeyen, S., Goormachtig, S., Clement, L., and De Tender, C. (2020). Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. *BMC Genomics*, 21(1).

[Joshi NA, 2011] Joshi NA, F. J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33).

[Kajale et al., 2021] Kajale, S., Jani, K., and Sharma, A. (2021). Contribution of archaea and bacteria in sustaining climate change by oxidizing ammonia and sulfur in an Arctic Fjord. *Genomics*, 113(1):1272–1276.

[Kalvari et al., 2018] Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D., and Petrov, A. I. (2018). Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, 46(November 2017):D335–D342.

[Khodakova et al., 2014] Khodakova, A. S., Smith, R. J., Burgoyne, L., Abarno, D., and Linacre, A. (2014). Random whole metagenomic sequencing for forensic discrimination of soils. *PLoS ONE*, 9(8).

[Kim et al., 2011] Kim, M., Morrison, M., and Yu, Z. (2011). Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods*, 84(1):81–87.

[Knight et al., 2018] Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L. I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., Zaneveld, J. R., Zhu, Q., Caporaso, J. G., and Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422.

[Kodama et al., 2018] Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., and Takagi, T. (2018). DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Research*, 46(D1):D30–D35.

[Krawczyk et al., 2018] Krawczyk, P. S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, 46(6).

[Lagesen et al., 2007] Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9):3100–3108.

[Liang et al., 2020] Liang, Q., Bible, P. W., Liu, Y., Zou, B., and Wei, L. (2020). DeepMicrobes : taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics*, 2(1):1–13.

[Lindgreen et al., 2016] Lindgreen, S., Adair, K. L., and Gardner, P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6(19233).

[Liu et al., 2012] Liu, K. L., Porras-Alfaro, A., Kuske, C. R., Eichorst, S. A., and Xie, G. (2012). Accurate, rapid taxonomic classification of fungal large-subunit rRNA Genes. *Applied and Environmental Microbiology*, 78(5):1523–1533.

[Luz Calle, 2019] Luz Calle, M. (2019). Statistical analysis of metagenomics data. *Genomics and Informatics*, 17(1).

[Madhavan and Gopakumar, 2018] Madhavan, M. and Gopakumar, G. (2018). A tf-idf based topic model for identifying lncRNAs from genomic background. *Proceedings of the ACM Symposium on Applied Computing*, pages 40–46.

[Magne et al., 2020] Magne, F., Gotteland, M., Gauthier, L., Zazueta, A., Pesoa, S., Navarrete, P., and Balamurugan, R. (2020). The firmicutes/bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients? *Nutrients*, 12(5).

[Mangul et al., 2019] Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K. M., Distler, M. G., Zelikovsky, A., Eskin, E., and Flint, J. (2019). Systematic benchmarking of omics computational tools. *Nature Communications*, 10(1):1–11.

[Markowitz et al., 2012] Markowitz, V. M., Chen, I. M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N., and Kyrpides, N. C. (2012). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, 40(D1):115–122.

[Martin et al., 1994] Martin, W. J., Zeng, L. C., Ahmed, K., and Roy, M. (1994). Cytomegalovirus-related sequence in an atypical cytopathic virus repeatedly isolated from a patient with chronic fatigue syndrome. *American Journal of Pathology*, 145(2):440–451.

[Mavromatis et al., 2007] Mavromatis, K., Ivanova, N., Barry, K. W., Shapiro, H. J., Goltsman, E., McHardy, A. C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., and Kyrpides, N. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500.

[McIntyre et al., 2017] McIntyre, A. B., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N. A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R. R., Rosen, G. L., and Mason, C. E. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, 18(1):1–19.

[McMullen et al., 2020] McMullen, C., Alexander, T. W., Léguillette, R., Workentine, M., and Timsit, E. (2020). Topography of the respiratory tract bacterial microbiota in cattle. *Microbiome*, 8(1):1–15.

[Méric et al., 2019] Méric, G., Wick, R., Watts, S., Holt, K., and Inouye, M. (2019). Correcting index databases improves metagenomic studies. *bioRxiv*, page 712166.

[Meyer et al., 2021] Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P. T., Cristian, A., Dabrowski, P. W., Darling, A. E., Egan, R., Eskin, E., Georganas, E., Goltsman, E., Gray, M. A., Hansen, L. H., Hofmeyr, S., Huang, P., Irber, L., Jia, H., Jørgensen, T. S., Kieser, S. D., Klemetsen, T., Kola, A., Kolmogorov, M., Korobeynikov, A., Kwan, J., LaPierre, N., Lemaitre, C., Li, C., Limasset, A., Malcher-Miranda, F., Mangul, S., Marcelino, V. R., Marchet, C., Marijon, P., Meleshko, D., Mende, D. R., Milanese, A., Nagarajan, N., Nissen, J., Nurk, S., Oliker, L., Paoli, L., Peterlongo, P., Piro, V. C., Porter, J. S., Rasmussen, S., Rees, E. R., Reinert, K., Renard, B., Robertsen, E. M., Rosen, G. L., Ruscheweyh, H.-J., Sarwal, V., Segata, N., Seiler, E., Shi, L., Sun, F., Sunagawa, S., Sørensen, S. J., Thomas, A., Tong, C., Trajkovski, M., Tremblay, J., Uritskiy, G., Vicedomini, R., Wang, Z., Wang, Z., Wang, Z., Warren, A., Willassen, N. P., Yelick, K., You, R., Zeller, G., Zhao, Z., Zhu, S., Zhu, J., Garrido-Oter, R., Gastmeier, P., Hacquard, S., Häußler, S., Khaledi, A., Maechler, F., Mesny, F., Radutoiu, S., Schulze-Lefert, P., Smit, N., Strowig, T., Bremges, A., Sczyrba, A., McHardy, A. C., and 1Computational (2021). Critical Assessment of Metagenome Interpretation - the Second Round of challanges. *Bioarxiv*.

[Meyer et al., 2022] Meyer, F., Fritz, A., Deng, Z. L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C. T., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P. T., Cristian, A., Dabrowski, P. W., Darling, A. E., Egan, R., Eskin, E., Georganas, E., Goltsman, E., Gray, M. A., Hansen, L. H., Hofmeyr, S., Huang, P., Irber, L., Jia, H., Jørgensen, T. S., Kieser, S. D., Klemetsen, T., Kola, A., Kolmogorov, M., Korobeynikov, A., Kwan, J., LaPierre, N., Lemaitre, C., Li, C., Limasset, A., Malcher-Miranda, F., Mangul, S., Marcelino, V. R., Marchet, C., Marijon, P., Meleshko, D., Mende, D. R., Milanese, A., Nagarajan, N., Nissen, J., Nurk, S., Oliker, L., Paoli, L., Peterlongo, P., Piro, V. C., Porter, J. S., Rasmussen, S., Rees, E. R., Reinert, K., Renard, B., Robertsen, E. M., Rosen, G. L., Ruscheweyh, H. J., Sarwal, V., Segata, N., Seiler, E., Shi, L., Sun, F., Sunagawa, S., Sørensen, S. J., Thomas, A., Tong, C., Trajkovski, M., Tremblay, J., Uritskiy, G., Vicedomini, R., Wang, Z., Wang, Z., Wang, Z., Warren, A., Willassen, N. P., Yelick, K., You, R., Zeller, G., Zhao, Z., Zhu, S., Zhu, J., Garrido-Oter, R., Gastmeier, P., Hacquard, S., Häußler, S., Khaledi, A., Maechler, F., Mesny, F., Radutoiu, S., Schulze-Lefert, P., Smit, N., Strowig, T., Bremges, A., Sczyrba, A., and McHardy, A. C. (2022). Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nature Methods*, 19(4):429–440.

[Michael McClelland, 2001] Michael McClelland, Kenneth E. Sanderson, J. S. S. W. C. P. L. L. C. S. P. J. A. M. D. F. D. S. H. D. L. S. L. C. N. K. S. A. H. N. G. R. K. W. (2001). Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature*, 413:852–856.

[Miralles et al., 2020] Miralles, A., Bruy, T., Wolcott, K., Scherz, M. D., Begerow, D., Beszteri, B., Bonkowski, M., Felden, J., Gemeinholzer, B., Glaw, F., Glöckner, F. O., Hawlitschek, O., Kostadinov, I., Nattkemper, T. W., Printzen, C., Renz, J., Rybalka, N., Stadler, M., Weibulat, T., Wilke, T., Renner, S. S., and Vences, M. (2020). Repositories for taxonomic data: Where we are and what is missing. *Systematic Biology*, 69(6):1231–1253.

[Mitchell et al., 2020] Mitchell, K., Mandric, I., Brito, J., Wu, Q., Knyazev, S., Chang, S., Martin, L. S., Karlsberg, A., Gerasimov, E., Littman, R., Hill, B. L., Wu, N. C., Yang, H., Hsieh, K., Chen, L., Shabani, T., Shabanets, G., Yao, D., Sun, R., Schroeder, J., Eskin, E., Zelikovsky, A., Skums, P., Pop, M., and Mangul, S. (2020). Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biology*, page 21:71.

[Mukherjee et al., 2016] Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., Thomas, A. D., Ali, R., Sharma, K., Kyrpides, N. C., and Reddy, T. B. K. (2016). Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic acids research*, 45(October 2016):gkw992.

[Nalin N Wijayawardene et al., 2011] Nalin N Wijayawardene, Kevin D Hyde, and Dong-Qin Dai (2011). Outline of Ascomycota. *IMA Fungus*, 2(1):4.

[Naranjo-Ortiz and Gabaldón, 2019] Naranjo-Ortiz, M. A. and Gabaldón, T. (2019). Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biological Reviews*, 94(6):2101–2137.

[Nawrocki and Eddy, 2013] Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.

[Nearing et al., 2022] Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M., Wright, R. J., Dhanani, A. S., Comeau, A. M., and Langille, M. G. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, 13(1).

[Ng et al., 2015] Ng, T. F. F., Kondov, N. O., Deng, X., Van Eenennaam, A., Neibergs, H. L., and Delwart, E. (2015). A metagenomics and case-control study to identify viruses associated with bovine respiratory disease. *Journal of Virology*, 89(March):00064–15.

[Nishimura et al., 2017] Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., Blanc-Mathieu, R., Yamamoto, K., Hingamp, P., Sako, Y., Sullivan, M. B., Goto, S., Ogata, H., and Yoshida, T. (2017). Environmental Viral Genomes Shed New. *Ecological and Evolutionary Science*, 2(2):00359–16.

[O'Leary et al., 2016] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):733–45.

[Ondov et al., 2011] Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1):385.

[Parks et al., 2020] Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, 38(September).

[Parks et al., 2018] Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10).

[Peabody et al., 2015] Peabody, M. A., Van Rossum, T., Lo, R., and Brinkman, F. S. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, 16(1).

[Pearman et al., 2020] Pearman, W. S., Freed, N. E., and Silander, O. K. (2020). Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC bioinformatics*, 21(220):1–15.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourgh, V., Vnaderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal ofMachine Learning Research*, 12(85):2825–2830.

[Pereira et al., 2020] Pereira, R., Oliveira, J., and Sousa, M. (2020). Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of Clinical Medicine*, 9(1):132.

[Pérez-Cobas et al., 2020] Pérez-Cobas, A. E., Gomez-Valero, L., and Buchrieser, C. (2020). Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*, 6(8):1–22.

[Pfeiffer et al., 2018] Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., and Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, 8(1):1–14.

[Piombo et al., 2021] Piombo, E., Abdelfattah, A., Droby, S., Wisniewski, M., Spadaro, D., and Schena, L. (2021). Metagenomics approaches for the detection and surveillance of emerging and recurrent plant pathogens. *Microorganisms*, 9(1):1–19.

[Pollock et al., 2018] Pollock, J., Glendinning, L., Wisedchanwet, T., and Watson, M. (2018). The Madness of Microbiome : Attempting To Find Consensus. *Applied and Environmental Microbiology*, 84(7):1–12.

[Quast et al., 2013] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F. O., and Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596.

[R. Marcelino et al., 2020] R. Marcelino, V., Holmes, E. C., and Sorrell, T. C. (2020). The use of taxon-specific reference databases compromises metagenomic classification. *BMC Genomics*, 21(1):1–5.

[Reimer et al., 2019] Reimer, L. C., Vetcininova, A., Carbasse, J. S., Söhngen, C., Gleim, D., Ebeling, C., and Overmann, J. (2019). BacDive in 2019: Bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Research*, 47(D1):D631–D636.

[Robeson et al., 2021] Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., and Bokulich, N. A. (2021). RESCRIPt: Reproducible sequence taxonomy reference database management. *PLoS Computational Biology*, 17(11).

[Salter et al., 2014] Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12(87):1–12.

[Santiago-Rodriguez et al., 2017] Santiago-Rodriguez, T. M., Fornaciari, G., Luciani, S., Toranzos, G. A., Marota, I., Giuffra, V., and Cano, R. J. (2017). Gut microbiome and putative resistome of inca and italian nobility mummies. *Genes*, 8(11).

[Sarkar et al., 2021] Sarkar, A., Harty, S., Moeller, A. H., Klein, S. L., Erdman, S. E., Friston, K. J., and Carmody, R. N. (2021). The Gut Microbiome as a Biomarker of Differential Susceptibility to SARS-CoV-2. *Trends in Molecular Medicine*.

[Sayers et al., 2020] Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2020). GenBank. *Nucleic Acids Research*, 48(D1):D84–D86.

[Schlaberg et al., 2017] Schlaberg, R., Chiu, C. Y., Miller, S., Procop, G. W., and Weinstock, G. (2017). Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Archives of Pathology & Laboratory Medicine*, 141(6):776–786.

[Schliep, 2011] Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics (Oxford, England)*, 27(4):592–3.

[Schoch et al., 2020] Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., and Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database : the journal of biological databases and curation*, 2020(2):1–21.

[Sczyrba et al., 2017] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., Demaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiutė, M., Hansen, L. H., Sørensen, S. J., Chia, B. K., Denis, B., Froula,

J. L., Wang, Z., Egan, R., Don Kang, D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y. W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M. D., Lingner, T., Lin, H. H., Liao, Y. C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H. P., Göker, M., Kyrpides, N. C., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nature Methods*, 14(11):1063–1071.

[Segata et al., 2011] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(6):R60.

[Segata et al., 2012] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–4.

[Şener et al., 2018] Şener, D. D., Santoni, D., Felici, G., and Oğul, H. (2018). A Content-Based Retrieval Framework for Whole Metagenome Sequencing Samples. *Journal of Integrative Bioinformatics*, 15(4):1–11.

[Seong et al., 2018] Seong, C. N., Kang, J. W., Lee, J. H., Seo, S. Y., Woo, J. J., Park, C., Bae, K. S., and Kim, M. S. (2018). Taxonomic hierarchy of the phylum Firmicutes and novel Firmicutes species originated from various environments in Korea. *Journal of Microbiology*, 56(1):1–10.

[Seppey et al., 2020] Seppey, M., Manni, M., and Zdobnov, E. M. (2020). LEMMI: A continuous benchmarking platform for metagenomics classifiers. *Genome research*.

[Söhngen et al., 2014] Söhngen, C., Bunk, B., Podstawka, A., Gleim, D., and Overmann, J. (2014). BacDive - The Bacterial Diversity Metadatabase. *Nucleic Acids Research*, 42(D1):592–599.

[Soverini et al., 2019] Soverini, M., Turroni, S., Biagi, E., Brigidi, P., Candela, M., and Rampelli, S. (2019). HumanMycobiomeScan: A new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples. *BMC Genomics*, 20(1):1–7.

[Stackebrandt, 2001] Stackebrandt, E. (2001). Bacterial Biodiversity. *Encyclopedia of Biodiversity: Second Edition*, 1:307–316.

[Steinegger and Salzberg, 2020] Steinegger, M. and Salzberg, S. L. (2020). Terminating contamination : large-scale search identifies more than 2 , 000 , 000 contaminated entries in GenBank. *Genome biology*, 21(115):1–12.

[Sun et al., 2021] Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A. P., Vázquez-Baeza, Y., Parida, L., Kim, H. C., Knight, R., and Liu, Y. Y. (2021). Challenges in benchmarking metagenomic profilers. *Nature Methods*, 18(6):618–626.

[Sunagawa et al., 2013] Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. a., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., Rasmussen, S., Brunak, S., Pedersen, O., Guarner, F., de Vos, W. M., Wang, J., Li, J., Dore, J., Ehrlich, S. D., Stamatakis, a., and Bork, P. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods*, 10(12):1196–1199.

[Sweeney et al., 2019] Sweeney, B. A., Petrov, A. I., Burkov, B., Finn, R. D., Bateman, A., Szymanski, M., Karlowski, W. M., Gorodkin, J., Seemann, S. E., Cannone, J. J., Gutell, R. R., Fey, P., Basu, S., Kay, S., Cochrane, G., Billis, K., Emmert, D., Marygold, S. J., Huntley, R. P., Lovering, R. C., Frankish, A., Chan, P. P., Lowe, T. M., Bruford, E., Seal, R., Vandesompele, J., Volders, P. J., Paraskevopoulou, M., Ma, L., Zhang, Z., Griffiths-Jones, S., Bujnicki, J. M., Boccaletto, P., Blake, J. A., Bult, C. J., Chen, R., Zhao, Y., Wood, V., Rutherford, K., Rivas, E., Cole, J., Laulederkind, S. J., Shimoyama, M., Gillespie, M. E., Orlic-Milacic, M., Kalvari, I., Nawrocki, E., Engel, S. R., Cherry, J. M., Team, S., Berardini, T. Z., Hatzigeorgiou, A., Karagkouni, D., Howe, K., Davis, P., Dinger, M., He, S., Yoshihama, M., Kenmochi, N., Stadler, P. F., and Williams, K. P. (2019). RNAcentral: A hub of information for non-coding RNA sequences. *Nucleic Acids Research*, 47(D1):D221–D229.

[Sweeney et al., 2021] Sweeney, B. A., Petrov, A. I., Ribas, C. E., Finn, R. D., Bateman, A., Szymanski, M., Karlowski, W. M., Seemann, S. E., Gorodkin, J., Cannone, J. J., Gutell, R. R., Kay, S., Marygold, S., Dos Santos, G., Frankish, A., Mudge, J. M., Barshir, R., Fishilevich, S., Chan, P. P., Lowe, T. M., Seal, R., Bruford, E., Panni, S., Porras, P., Karagkouni, D., Hatzigeorgiou, A. G., Ma, L., Zhang, Z., Volders, P. J., Mestdagh, P., Griffiths-Jones, S., Fromm, B., Peterson, K. J., Kalvari, I., Nawrocki, E. P., Petrov, A. S., Weng, S., Bouchard-Bourelle, P., Scott, M., Lui, L. M., Hoksza, D., Lovering, R. C., Kramarz, B., Mani, P., Ramachandran, S., and Weinberg, Z. (2021). RNAcentral 2021: Secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220.

[Tan et al., 2019] Tan, G., Opitz, L., Schlapbach, R., and Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, 9(1):1–7.

[Tatusova et al., 2014] Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., and Tolstoy, I. (2014). RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Research*, 42(D1):553–559.

[Tizioto et al., 2015] Tizioto, P. C., Kim, J., Seabury, C. M., Schnabel, R. D., Gershwin, L. J., Van Eenennaam, A. L., Toaff-Rosenstein, R., Neibergs, H. L., and Taylor, J. F. (2015). Immunological Response to Single Pathogen Challenge with Agents of the Bovine Respiratory Disease Complex: An RNA-Sequence Analysis of the Bronchial Lymph Node Transcriptome. *PloS one*, 10(6):e0131459.

[Ul-Hassan and Wellington, 2009] Ul-Hassan, A. and Wellington, E. (2009). *Encyclopedia of Microbiology (Third Edition)*. Academic Press, third edition.

[Velsko et al., 2018] Velsko, I. M., Frantz, L. A. F., Herbig, A., Larson, G., and Warinner, C. (2018). Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research. *mSystems*, 3(4):00080–18.

[Vilanova et al., 2015] Vilanova, C., Iglesias, A., and Porcar, M. (2015). The coffee-machine bacteriome: biodiversity and colonisation of the wasted coffee tray leach. *Scientific Reports*, 5:17163.

[Wang et al., 2020] Wang, Y., Huang, J. M., Zhou, Y. L., Almeida, A., Finn, R. D., Danchin, A., and He, L. S. (2020). Phylogenomics of expanding uncultured environmental Tenericutes provides insights into their pathogenicity and evolutionary relationship with Bacilli. *BMC Genomics*, 21(1):1–12.

[Weber et al., 2019] Weber, L. M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P. P., Boulesteix, A. L., Saeys, Y., and Robinson, M. D. (2019). Essential guidelines for computational method benchmarking. *Genome Biology*, 20(1):1–12.

[Weiss et al., 2017] Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27.

[Wheeler and Eddy, 2013] Wheeler, T. J. and Eddy, S. R. (2013). Nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29(19):2487–2489.

[Wood et al., 2019] Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *bioRxiv*, pages 1–13.

[Wood and Salzberg, 2014] Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46.

[Yan et al., 2017] Yan, J., Chuai, G., Qi, T., Shao, F., Zhou, C., Zhu, C., Yang, J., Yu, Y., Shi, C., Kang, N., He, Y., and Liu, Q. (2017). MetaTopics: An integration tool to analyze microbial community profile by topic model. *BMC Genomics*, 18(Suppl 1):1–5.

[Yang et al., 2016] Yang, B., Wang, Y., and Qian, P.-Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*, 17(135).

[Yarza et al., 2014] Yarza, P., Yilmaz, P., Pruesse, E., Oliver Glöckner, F., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R., and Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews. Microbiology*, 12:635–645.

[Yilmaz et al., 2014] Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glöckner, F. O. (2014). The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research*, 42(D1):643–648.

[Zeineldin et al., 2017a] Zeineldin, M., Lowe, J., de Godoy, M., Maradiaga, N., Ramirez, C., Ghanem, M., Abd El-Raof, Y., and Aldridge, B. (2017a). Disparity in the nasopharyngeal microbiota between healthy cattle on feed, at entry processing and with respiratory disease. *Veterinary Microbiology*, 208(February):30–37.

[Zeineldin et al., 2017b] Zeineldin, M. M., Lowe, J. F., Grimmer, E. D., de Godoy, M. R. C., Ghanem, M. M., Abd El-Raof, Y. M., and Aldridge, B. M. (2017b). Relationship between nasopharyngeal and bronchoalveolar microbial communities in clinically healthy feedlot cattle. *BMC Microbiology*, (17):138.