Metagenomics: Beyond the horizon of current implementations and methods

Lusine Khachatryan

This work is part of the research programme "Forensic Science" with project number 727.011.002, which is financed by the Dutch Research Council (NWO).

ISBN: 9789464020892
Cover Artwork: Alessandra Sequeira
Printing: GILDEPRINT, www.gildeprint.nl

Metagenomics: Beyond the horizon of current implementations and methods

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 28 april 2020
klokke 16:15 uur

door

Lusine Khachatryan
geboren te Jermuk, Armenië in 1990

Promotor:	Prof. dr. P. de Knijff
Co-promotor:	Dr. J. F. J. Laros
Leden promotiecommissie:	Prof.dr. A. Geluk Prof. dr. A. C. M. Kroes Prof. dr. J. N. Kok[1] Dr. T. Sijen[2]

1 Faculty of Electrical Engineering, University of Twente, Enschede, The Netherlands
2 Netherlands Forensic Institute, The Hague, The Netherlands

Моему Папочке

Matthew 7:7 New International Version

"Ask and it will be given to you; seek and you will find;
knock and the door will be opened to you"

Contents

Chapter 1

Introduction

1.1 Why metagenomics

METAGENOMICS is a new and rapidly developing branch of microbiology. In this chapter we will explain its advantages, list its possible applications and give an overview of the most valuable scientific findings in recent years that were made using mainly metagenomics approaches. Please note that the terms "microbes" and "microorganisms" in this chapter, as well as in the entire thesis, primarily refer solely to bacteria and archaea (another domain of prokaryotes distinct from bacteria).

We have all been taught about the importance of frequently washing our hands based on the unquestionable assurance stating that "microbes are everywhere". Though we often do not see them, we are well aware of their presence and possible harmful impact. However, not everyone can imagine that these little creatures, microbes, are the cornerstones of our biosphere.

Microorganisms are involved in a vast number of processes on our planet, making it a habitable and sustainable ecosystem [1, 2, 3, 4, 5]. They are key players in the biochemical cycling of elements such as carbon, nitrogen, oxygen and sulfur [6, 7, 8, 9, 10]. Most importantly, microbes can turn compounds that contain these elements into forms accessible by other organisms. Through billions of years of evolution, microorganisms became absolutely necessary symbionts for the majority of multi-cell life forms. Microbial communities are providing their hosts with the necessary vitamins, metals and nutrients [11, 12, 13]. They maintain digestion, flush out toxins and fight parasites (which are often microorganisms themselves) [14, 15, 16, 17, 18]. Besides being in a close symbiotic association with other life forms, microbes learned how to live in extreme environments where no other organisms can survive. In order to do so, microorganisms developed countless strategies allowing them to maintain their metabolism in the presence of for example severe temperatures, pressures, pH levels and combinations of these and other factors [19, 20, 21, 22]. The description of the roles of microbes in our biosphere would not be complete without mentioning their contribution to technology. Microorganisms are being utilized for fast and cheap food, drugs and chemical production, food fermentation, agricultural improvements, soil and water depollution, biological fuel and many other aspects that improve the quality of life [23, 24, 25, 26, 27, 28, 29, 30].

Investigation of microbes is extremely beneficent for humanity; it contributes to understanding the biochemical landscape of the biosphere, medicine, food production, farming, agriculture and many other fields.

Historically, microbiology - the study of microorganisms - was based on the description and comparison of organisms' morphological features, growth, and biochemical profiles [31, 32]. These techniques were applied to single organisms, grown separately as a pure culture without any ecological context. The invention of automated DNA sequencing in late 1970s allowed researchers to understand the genetic basis

underlying previous microbiological discoveries [33, 34]. It also became clear that the standard laboratory culture-based way of investigating microorganisms is restricted because of two main reasons: only a very small fraction of microorganisms has been found to be cultivable and functions performed by microorganisms are conducted within complex communities.
In 1985 the "great plate count anomaly" was discovered, the absolute majority of microorganisms that can be seen through the microscope cannot successfully be taken from the environment to laboratory cultivation [35]. The estimate was that only 0.1-1% of the total variability of microbiological species, habituating soil, can be grown under laboratory conditions. The cultivable fraction from some other environments can be thousands of times smaller. Furthermore, the organisms that can be cultivated, are not necessarily the most dominant or influential for a particular environment, but rather favoured by the cultivating conditions.
Metabolic functions performed by microorganisms are conducted within complex communities - microbiomes. The compositions of those communities are tailored to their particular environment and adapt swiftly to environmental change. Investigating the isolated separate members of such complicated entities as microbiomes often lead to incomplete and sometimes even incorrect conclusions, as the organisms' properties and behaviour within a community might differ drastically from those in a pure laboratory culture. Thus, the pure culture paradigm limits not only the number of organisms for studies, but also the understanding of microbes functioning as a whole. The shift from pure cultures to the community, from the individual to interaction, is the solution to the aforementioned problems.
Rapid improvements in sequencing techniques as well as deeper understanding of the microbial genome led to the origin of metagenomics - the direct genetic analysis of genomes contained within an environmental sample [36, 37, 38, 39, 40]. In pioneering metagenomics studies amplification of genes conserved among all microorganisms was conducted directly from an environmental sample, followed by cloning of the obtained amplicons into bacterial vectors and subsequent sequencing [41, 42]. The results were in agreement with the expectations: the reported biodiversity was much higher than the estimation obtained using the culture-based methods. These first revolutionary studies turned metagenomics into the most dynamic and quickly developing field within microbiology. Since then, the amount of metagenomics projects targeted on different environments has grown extensively, adapting different sequencing techniques, data types and bioinformatics algorithms which will be discussed in detail in the following chapters of this thesis.
As previously mentioned, microbial communities can be found practically everywhere on our planet. This provides metagenomics with unlimited options for scientific research. Metagenomics revolutionized the entire studies of microbial diversity and evolution by providing access to the "hidden phylogenetic composition of complex environmental microbial communities" [38]. The employment of metagenomics also allows functional and metabolic potentials of a particular metagenome to be

investigated. This all makes metagenomics a powerful tool, that can be used by researchers in an extensive range of projects.
The most popular and developed area of metagenomic studies is the investigation of microbiomes associated with other organisms, particularly human. The Human Microbiome Project (HMP, launched in 2008) and Integrative Human Microbiome Project (iHMP, launched in 2014) were announced as "a logical conceptual and experimental extension of the Human Genome Project", which stressed the importance of understanding human-microbe interaction [43, 44]. These projects received more than $170,000,000 in funding and contributed substantially to the understanding of the human microbiome with regards to health and disease, as well as contributed to developing diagnostics and treatment strategies based on metagenomics knowledge, association of particular communities with individuals and populations and correlations between the host genetics and microbiota [45, 46, 47, 48, 49, 50].
Studying microbial ecosystems in order to predict possible processes, changes and sustainability of particular environments is another popular topic in metagenomics. For example, various different studies contribute to understanding of how microorganisms maintain the atmosphere. Notably, it was shown that - contrary to the widely held belief - more than half of photosynthesis on our planet is performed by bacteria [51, 52]. Marine metagenomic investigations have shown that viruses are by far the most abundant group of marine life (both cellular and non-cellular), comprising approximately 94% of the nucleic-acid-containing particles [53]. The discovery of new microbial species and their functional and metabolic potential within a microbiome helps researchers to build better models for the microbiome-environment interaction, thus contributing to the microbial ecology field.
Exploring new metabolic pathways and discovering functional genes is the most important feature of metagenomics for technological uses. Genes isolated from soil metagenomes are successfully being used for the production of biofuels and for the tolerance of other microbiota to byproducts of biofuel production [30]. Various newly discovered biosynthetic capacities of microbial communities benefit the production of industrial, food and health products as well as contribute to the field of bioremediation [54, 55, 56, 57].
Last but not least, metagenomic projects can be implemented in various fields such as forensics [58, 59, 60, 61]. Mostly through skin microbiota, people leave marks on objects they touch and on the surfaces of houses they live in. Several studies have shown that human microbiota can be used to match touched subjects like computer keyboards or mobile phones and their owners [62, 63]. Recent research has shown a correlation between metagenomic DNA of household surfaces and the skin microbiome of its inhabitants [64, 65, 66, 67]. A number of studies were conducted for the identification of microbes associated with particular human cohorts, in order to use those microbes as signatures when analysing forensic traces [68].

The application area of metagenomics keeps expanding, challenging the scientific

community to try new sequencing techniques and to develop new bioinformatics tools and approaches for metagenomic data interpretation.

1.2 Metagenomics sequencing data

In this chapter we will introduce the most common types of data used in metagenomics, their advantages and disadvantages and possible sequencing platforms to acquire this data. This serves as a motivation behind the use of particular types of metagenomic data for each of the studies included in this thesis.
Technological advances in high-throughput sequencing enabling culture- and cloning-free microbiome analysis has led to a sharp growth of metagenomics studies in last 20 years. However, the data types used for the microbiome investigation remain quite conservative.

1.2.1 Amplicon sequencing data

The first datatype we will discuss is based on sequencing only one marker gene from each organism in the microbiome and performing the phylogenetic reconstruction of the microbiome content using this data. The most common target for such microbiome profiling is the 16S ribosomal (rRNA) gene. This approach was used in the pioneer metagenomics studies as well as for the major metagenomics projects such as Human Microbiome Project. The 16 rRNA gene is highly conserved among bacteria and archaea. The entire locus, which is about 1500 nucleotides long, contains conserved regions as well as 9 hypervariable regions (V1-V9) which are 30-100 base pairs long. Hypervariable regions provide phylogenetic signatures on different taxonomic levels. This important feature makes the 16S rRNA gene analysis prevalent for the classification of bacteria without the need for costly and elaborate phenotypic identification. Between the hypervariable regions of the 16S rRNA gene lie highly conserved sequences, which can be targeted by universal primers that can reliably produce the same sections of the 16S sequence across different taxa [69, 70, 71, 72].
Historically, both whole-locus and partial sequencing of the 16S rRNA gene was performed using the Sanger platform. However, since this approach is laborious, costly and has a low throughput, it was substituted first with 454-pyrosequencing and later with Illumina sequencing platforms. Presently, Illumina MiSeq is the most popular sequencing platform for 16S rRNA data due to its cost efficiency and improved community coverage in comparison to the 454-pyrosequensing platform. Recent studies suggest implementing full-length 16S rRNA gene sequencing by using the PacBio single molecule, real-time (SMRT) technology [73]. This approach is still questionable due to the high error rate of PacBio sequencing and requires large amounts of DNA for conducting the experiment.

The importance of the 16S rRNA gene for bacterial classification led to the existence of several curated databases designed to contain reference sequences and taxonomical classification exclusively for the 16S gene or its parts. The most well-known databases are the Ribosomal Database Project (RDB) [74, 75], SILVA [76] and GreenGenes [77]. These databases contain minor variations.
While 16S sequencing remains the most popular and routine procedure for metagenomics analysis, it has become clear that the method contains several biases, which might influence the final outcome of the analysis drastically. The level of conservation varies between different hypervariable regions [78]. Thus, the accuracy of the analysis based on the 16S rRNA sequencing directly depends on the choice of the hypervariable region or the combination of the regions. Various studies were done in order to identify the best hypervariable region suitable for the deep taxonomical analysis. However, their outcome was directly dependent on the type of microbiota used for the analysis and even on the choice of the sequencing platform. Recent studies [79, 80, 73] suggested using the sequence of the entire 16S rRNA molecule in order to solve this problem. However, this method is much costlier in comparison with the standard amplification of one or several variable regions. Whilst the 16S rRNA gene was considered to be a perfect phylogenetic marker before, there have recently been reports, showing that for certain taxa the 16S sequencing data analysis fails to differentiate between closely related organisms [81, 82]. Consequently, the search for and subsequent sequencing of other taxon-specific genes is required. Even the most popular and universal PCR primers cover the variability of the microorganisms unevenly and can lead to the incorrect analysis [83, 84]. Microorganisms might contain different numbers copies of the 16S rRNA gene and as a result negatively affects the abundance estimation within the metagenome [85]. Several tools [86, 87] have been developed for correcting this by using phylogenetic methods. However, the accuracy of its predictions have not been independently assessed [88]. Finally, the analysis of only the 16S rRNA gene can only provide the phylogenetic fingerprint of the microbial community, thus, missing its functional capacity. There are bioinformatics approaches are used to predict the functional landscape of the metagenome by using its phylogenetic fingerprint from 16S rRNA profiling (e.g. [89]). However, results obtained using these approaches are highly unreliable.

1.2.2 Whole genome sequencing data

The growing amount of evidence compromising the liability of the results obtained using only 16S rRNA data resulted in the popularity of whole genome shotgun sequencing (WGS) of metagenomics data [90, 91]. Though it used to be technically and computationally difficult, this technique is becoming more and more popular due to the advances in sequencing technologies, bioinformatics tools and approaches to deal with big data. The broad range of NGS platforms are available for WGS metagenomics sequencing, amongst them the popular platforms Illumina MiSeq and

HiSeq. The previously widely utilized the 454-pyrosequensing and the IonTorrent platforms are no longer popular due to their high cost and biases introduced during the sequencing process. Methods offering extremely long reads (PacBio and Oxford Nanopore) can be used for the WGS metagenomics sequencing as well[92, 93]. However, the price and the high DNA amount limitation in conjunction with the high error rate making these approaches available for only a limited number of projects. Therefore, PacBio sequencing is widely used in combination with Illumina sequencing to facilitate and improve the performance of the analysis for the most abundant metagenome inhabitants. WGS metagenomics data easily bypasses the biases introduced when using the 16S data as copy number variation or amplification of the marker gene. The obtained data allow a more detailed analysis of the studied microbiome, including species identification, functionality profiling and more precise abundance estimation. To perform the analysis the use of different databases or the combinations of databases can be utilized. However, it is important to note that performing the WGS sequencing is considerably more expensive in comparison with sequencing only the 16S rRNA. WGS data also require more extensive analysis. The estimation of the community complexity prior to the development of the WGS experiment is crucial, as the sufficient coverage of metagenome inhabitants is vital for the quality of the analysis results.

The question about the areas of the implementation of 16S and WGS data is still a topic of contention among researches. For each study it is important to find the data type that provides a comprehensive yet not excessive amount of information. The delicate balance between the analysis depth and the experiments costs is a direct consequence of understanding the advantages and the limitations of the data type, sequencing techniques and the properties of the metagenome.

1.3 Approaches used in metagenomics

Proper and accurate analysis of metagenomic data is crucial to reveal the information that a metagenome potentially provides. Most of the times during such analysis, researchers are trying to find an answer to three main questions "Who is in the metagenome?", "What are they doing?" and "What is the difference between two metagenomes?" In this chapter we will try to give an overview of common methods and techniques used to answer those questions.
Usually the analysis of every metagenomic dataset begins with reads preprocessing, which includes a quality check followed by identification and removing of low-quality sequences and contaminants. Preprocessing is performed by a set of standard tools such as FastQC [94], Cutadapt [95], BBDuk[1] and Trimmomatic [96]. In some

[1] tool of BBMap package, https://sourceforge.net/projects/bbmap/

cases, filtering against a host genome (e.g., human) is required, although many tools for downstream analysis already include this step.
The core process for each analysis of metagenomic data - called profiling or binning - is sorting the sequencing reads into genetically/functionally homogeneous groups. The key question is whether the profiling procedure should be performed by homology-based methods (comparing metagenomics reads to the known sequences), *de novo* (using DNA features alone), or as a combination of thereof. Let us review each of these profiling approaches.

1.3.1 Homology-based profiling

The vast majority of existing metagenomics binning approaches are homology-based and thus depend on the content of the sequences databases [97]. Using this group of methods allows researchers to find answers to all three questions that we listed above. Profiling is performed by comparison of sequencing reads to known genomes to find out which organisms are present in a particular microbiome and/or their possible functionalities. Comparison of profiles obtained for two different metagenomes (which will be discussed in section 1.3.1.3) allows us to address the level of their similarity.
The choice of homology-based metagenomics analysis workflow mainly depends on the sequencing data type. While Amplicon data analysis steps are rather standardized, the set of approaches designed for WGS metagenomic data analysis is much broader.

1.3.1.1 Amplicon metagenomic data profiling

The analysis of Amplicon metagenomic data will be discussed in the context of the most common marker gene - 16S rRNA (see section 1.2.1). 16S data can provide the researchers only with information regarding the metagenome taxonomical context. Preprocessed reads (see the beginning of section 1.3) are usually clustered into so-called 'Operational Taxonomic Units' or OTUs [98], based on sequences similarity. Each of the obtained clusters is intended to represent a taxonomic unit of a bacterial/archaeal species or genus depending on the sequence similarity threshold. Usually a similarity of 97% is utilized to distinguish bacteria and archaea at the genus level. After that, a representative sequence for each OTU is annotated using a 16S rRNA database, where OTU representative sequences without database hits are classified as "unknown". OTUs of unknown origin are usually discarded and the remaining OTUs are used to generate taxonomical and abundance profiles. Currently, there are two commonly used pipelines - Morthur [99] and QIIME [100] - that perform all of the steps listed above. Their main difference is the choice of the clustering approach for OTU formation: hierarchical clustering for Morthur and 'greedy' USEARCH [101] for QIIME (note that QIIME can be adjusted to work with

other clustering approaches, including the Morthur-specific one). The two methods also differ in the way they annotate OTU representative sequences, and they work with different databases.

1.3.1.2 WGS metagenomic data profiling

We will now switch gears and consider whole genome sequencing (WGS) data analysis. Preprocessed WGS reads can enter the binning procedure directly or be preliminarily assembled into contigs (longer contiguous sequences). The choice of assembly-based analyses versus direct binning of reads depends on the research question. Binning the contigs instead of reads has several advantages: higher reliability of the obtained classification and the possibility to correct profiles using the contigs co-abundances. On the other hand, the algorithms performing the metagenomic data assembly are still far from ideal: they often report chimeric (combining sequences from more than one genome) contigs and require information about the metagenome complexity *a priori*. In this chapter we will not discuss metagenomic data assembly methods, we assume that the downstream analysis is performed on sequencing reads directly after preprocessing.
The large number of tools available for the homology-based WGS metagenomics data analysis can be split into several groups using the following criteria: strategy for reads binning, possible database against which the search is performed, and the part of reads used for profiling (Table 1.1). Matching to the database (and thus binning) can be performed by various alignment tools (BLAST [102], DIAMOND [103], LAST [104], BWA [105], Bowtie 2 [106], BLAT [107], etc.) as well as by using k-mers (DNA sequences of length k). Alignment and k-mer searching can be performed on full-genome databases as well as on databases containing marker genes or genetic "signatures" (unique genomic regions) associated with different clades. While some metagenomics tools use the entire dataset, other prefer to perform binning only on reads with particular features (e.g., reads predicted to be part of 16S rRNA and coding sequences, CDS). Finally, a number of methods return one best match for every read, while others use the principle of Lowest Common Ancestor (LCA [108]) in situations when the same read got matches with a group of different references.
Despite the variety and broad use of homology-based metagenome profiling tools, reads binning provided by such approaches suffer from database incompleteness, since the majority of microbial species are still not sequenced.

1.3.1.3 Comparison of profiles obtained using homology-base techniques

Similarity levels among different metagenomes, answering the third question mentioned in the beginning of this chapter, can be retrieved using the profiles obtained during the homology-based analysis. Results of taxonomical binning can be used to compute two important quantities widely applied in environmental microbiology:

alpha and beta diversity. Alpha diversity represents taxonomical richness within a single microbiome and is often quantified by the Shannon Index [136] or the Simpson Index [137]. Beta diversity measures a similarity score between different microbiomes and can be calculated using simple taxa overlap or Bray-Curtis dissimilarity [138]. Phylogenetic distribution of taxa in metagenomics profiles also can be used to describe the diversity within and between communities. This method computes the alpha diversity as the cover of a phylogenetic tree by the taxa present in microbiome. Beta diversity is calculated as a proportion of phylogenetic tree shared between two microbiome profiles. The standard metric for the phylogeny-based measurements is UniFrac [139], which can be performed with the abundances of taxa considered (weighted UniFrac).

Method	Binning tool	Binning technique	Database
Kraken [110]	*k*-mer matching	All reads are classified. Each read is split into *k*-mers that are assigned to the database tree nodes using LCA principle. Each node is weighted by the number of *k*-mers mapped to the node. Leaf with the highest sum of weights on the path from root to leaf is used to classify the read.	Suitable for any database as long as the phylogeny within database is provided. Constructs a database that stores every *k*-mer for each reference genome.
MetaPhlAn [117]	Bowtie2	All reads are classified, but majority of them do not get any hits due to the database bias. Each read is assigned to the best hit.	Uses the database of clade-specific marker genes.
CLARK [115]	*k*-mer matching	All reads are classified. Read is assigned to the node with which it shares most of the *k*-mers.	Suitable for any database. Creates *k*-mer based database with all non-unique *k*-mers removed.

Table 1.1: To be continued on the next page

Method	Binning tool	Binning technique	Database
Centrifuge [111]	Comparison with FM-indexed genomes	All reads are classified. Each read is compared to all indexed genomes in the database.	Suitable for any database as long as the phylogeny within database is provided. Uses the Burrows-Wheeler transform [112] and an FM-index [113] to store and index the genome database. Combines shared sequences from closely related genomes using MUMmer [114].
GOTTCHA [116]	BWA *mem*	All reads are classified. Reads are split into non-overlapping 30-mers, that are used for the alignment.	Each 30-mer is assigned to the best hit. Suitable for any database. Preprocess the database, keeping only the genomic regions (signatures) that are unique to each reference.
MEGAN6 [109]	Alignment (BLASTX, DIAMOND, LAST)	All reads are classified. Reads are aligned to each sequence in the reference database. LCA principle is used to assign reads with multiple hits.	Suitable for any database as long as the phylogeny within the database is provided.
Kaiju [125]	BWT (modified) to the FM-indexed reference	Predicted protein-coding reads are classified. LCA principle is used to assign reads with multiple hits.	Uses NCBI BLAST non-redundant protein database

Table 1.1: To be continued on the next page

Method	Binning tool	Binning technique	Database
mOTU [122]	BWA	All reads are classified based on the results of comparison with 40 marker genes	Uses the database of 40 prokaryotic marker genes
MG-RAST [118]	BLAT	Only reads predicted (using FragGeneScan [119]) to be part of 16S rRNA or CDS are used for the analysis.	Bond to the set of custom databases (M5nr and M5nra)
EBI Metagenomics [128]	QIIME for 16S predicted reads, InterProScan [129] for predicted CDS	Only reads predicted (using rRNAselector [130] and FragGeneScan) to be part of 16S rRNA or CDS are used for the analysis.	Bond to the set of custom databases (GreenGenes, Pfam [131], TIGRFAMs [132], PRINTS [133], PROSITE patterns [134], Gene3d [135])
Quikr [123] and WGSQuikr [124]	*k*-mer matching (complete sequencing data profile to the database *k*-mer matrix)	All reads are classified. Solving the NNLS problem with variant of basis-pursuit denoising	Suitable for any database. Creates one *k*-mer-based matrix for the entire reference database
FOCUS [127]	*k*-mer matching (complete sequencing data profile to the database *k*-mer matrix)	All reads are classified. Uses non-negative least squares to compute the set of *k*-mer frequencies that explains the optimal possible abundance of *k*-mers in the analysed metagenome by selecting the optimal number of frequencies from the reference *k*-mer matrix	Suitable for any database. Creates one *k*-mer-based matrix for the entire reference database

Table 1.1: To be continued on the next page

Method	Binning tool	Binning technique	Database
Taxator-tk [120]	Local BLAST or LAST	All reads are classified. Local alignment for each read against the database is used to split the read into distinct segments and to determine a taxon for each segment. Taxon for the entire read is determined by the taxa assigned to its segments. All taxon assignments are performed using LCA principle.	Suitable for any database.
MetaPhyler [121]	BLASTX	All reads are classified, but majority of them do not get any hits due to the database bias. Each read is assigned to the best hit.	Uses the database of 31 marker genes.
TIPP [126]	All reads are classified. HMMER mapping	Mapping to the marking genes. SEPP phylogenetic placement	Using the database of 30 phylogenetic marker genes that span the Bacteria and Archaea domains

Table 1.1: The overview of popular metods for the homology-based analysis of metagenomic data

1.3.2 *De novo* profiling

De novo approaches for metagenomics binning try to solve the problem of missing taxonomic content: they are designed to classify reads into genetically homogeneous groups without utilizing any information from known genomes. Instead, they use only the features of the sequencing data (usually reads similarities or *k*-mer distributions) for classification. For example, the first step of homology-based profiling for 16S data, namely clustering sequences into OTUs, is nothing else but *de novo* profiling of a metagenomics dataset.
Due to their nature, *de novo* binning techniques cannot give an answer to the questions "Who is in metagenome?" and "What are they doing?". However, they can be

used for a metagenome complexity estimation, revealing the true composition diversity of a metagenome, which is usually underestimated during classical homology-based analyses.
There are several tools designed for de novo binning of WGS metagenomics data, which we will discuss in this section. One of them, LiklyBin [140], follows a Markov Chain Monte Carlo approach based on the assumption that the *k*-mer frequency distribution is homogeneous within a bacterial genome. This approach works well for very simple metagenomes with a significant phylogenetic diversity within the metagenome, but it cannot handle genomes with more complicated structures such as those resulting from horizontal gene transfer [141]. Another approach, AbundanceBin [142], works under the assumption that the abundances of species in metagenome reads are following a Poisson distribution, and thus struggles when analysing datasets where some species have similar abundance ratios. MetaCluster [143] and BiMeta [144] address the problem of non-Poissonian species distribution. However, for these tools it is necessary to provide an estimation of the final number of bins which cannot be done for many metagenomes without any a priori knowledge. Also, both MetaCluster and BiMeta use the Euclidian metric to compute the dissimilarity between *k*-mer profiles, which was shown to be easily influenced by stochastic noise in analysanalysed sequences [145]. Finally, one of the most recent approaches - MetaProb [146] - implements a more advanced similarity measure technique and can automatically estimate the number of read clusters. This tool classifies metagenomic datasets in two steps: first, reads are grouped based on the extent of their overlap. After that, a set of representing reads is being chosen for each group. Based on the comparison of the *de novo* distributions for those sets, groups are merged together into final clusters. Even though MetaProb outperformed other *de novo* binning approaches during the analysis of simulated data, it did not provide solid results when testing on real metagenomics data.
To conclude, *de novo* metagenomics binning remains a challenging task. However, a successful *de novo* technique would open up countless opportunities for the future of microbiology, due to the complete independence from reference databases.

1.3.3 Mixed profiling

After describing the set of homology-based and *de novo* approaches we would like to continue with the group of methods combining the features of reference-based and *de novo* profiling tools. Such approaches are recently gaining interest due to their indirect reliance on a reference database. These approaches use supervised training on known databases, to learn about differentiating sequence features in order to perform *de novo* reads binning. This enables metagenomics profiling for the reads that would not have any match with any known references. Supervised approaches can be trained using a various set of techniques, such as Interpolated Markov Models, Gaussian Mixture Models, Hidden Markov Models, mixtures of

variable-order Markov chains, naive Bayes classifier, Support Vector Machine and many others [147, 148, 149, 150, 151, 152, 153]. The training database can, as well as in case of classical homology-based techniques, consist of a complete genome or a set of marker genes. Features used for training are in most cases *k*-mers of a particular length, or a mixture of *k*-mers of different length. Sometimes species "signature" sequences and reads co-assurances can be used for model training.
The results of supervised classification techniques are still doubtful, since the content of the current reference databases utilized for the training differs from the true distribution of microbial species on our planet.

1.3.4 Reference-free comparison of metagenomics data

As was mentioned at the beginning of this section, there are three main questions the metagenomics studies. The first two can be answered only by using a reference-dependent analysis, whereas the third one, "What is the difference between two different metagenomes?" does not necessarily require any reference database. The group of methods allowing to determine the difference between two genetic datasets without comparing them to a known genetic reference are mostly based on reads overlapping between different samples, *k*-mer-mer counts and a comparison of the obtained profiles using various different metrics [154, 155, 156, 127, 157, 149, 158, 159, 160]. Some approaches for the reference-free comparison of metagenomics data work with results of mixed and *de novo* profiling, comparing the binning results obtained for the different metagenomes using the different variations of Bray-Curtis dissimilarity. For example, such analysis can be performed on 16S data by simple overlapping of OTUs derived from the different samples prior to the OTU annotation. This allows to preserve the data, that would be lost for the OTUs marked as 'unknown' during the annotation procedure. This dissimilarity measure, however, does not take into account the phylogeny of compared OTUs, which is provided, for example, by UniFrac (see section 1.3.1.1).

1.4 The outline of this thesis

As mentioned in the previous sections, the current field of metagenomics can be summarised by:

- Three main questions: "Who is in metagenome?" (or "How complex the metagenome is?"), "What are they doing?" and "What is the difference between two metagenomes?";
- Two popular techniques to generate metagenomic sequencing libraries: 16S and WGS;
- Two general approaches to analyse metagenomic data: reference-dependent and reference-free.

This research was dedicated to a better understanding of the limits of each of the analysis methods regarding different types of sequencing data. We also tried to perform the sequencing experiments using distinct sequencing platforms and protocols. To understand how far the boundaries of most popular analysis techniques, in combination with various data types, can be set we performed a number of studies. In Chapter 2 we discuss the taxonomic profiling quality obtained using 16S and WGS metagenomic data. During that research, we created a series of artificial bacterial mixes, each with a different distribution of species. These mixes were used to estimate the resolution of two different metagenomic experiments - 16S and WGS - and to evaluate several different bioinformatics approaches for taxonomic read classification.
We also tried to improve the analysis of metagenomics data in both directions: with and without using reference databases using both 16S rRNA and WGS data.
For the reference-free analysis of different NGS datasets, we developed a *k*-mer based method (kPal). We have shown that our approach can be used for two types of metagenomics analysis: to perform *de novo* reads binning within a single metagenome (Chapter 3) and to resolve the level of relatedness between microbiomes (Chapter 4).
Our approach in reference-based metagenomics was targeted to perform fast and accurate analysis for clinical samples that might contain more than one pathogen. We developed BacTag, a distributed bioinformatics pipeline for fast and accurate bacterial gene and allele typing using clinical WGS sequencing data. The reader can find more details about the algorithm behind this tool and its testing results in Chapter 5.
A general discussion, including a review on future perspectives in the field of metagenomics, can be found in Chapter 6.

Chapter 2

Taxonomic classification and abundance estimation using 16S and WGS - a comparison using controlled reference samples

L. Khachatryan[1], R. H. de Leeuw[1], M. E. M. Kraakman[2], N. Pappas[3], M. te Raa[1], H. Mei[3], P. de Knijff[1], and J. F. J. Laros[1,4]

1 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Department of Microbiology, Leiden University Medical Center, Leiden, The Netherlands
3 Sequencing Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
4 Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands

Forensic Science International: Genetics, 2020 46:102275 doi 10.1016/j.fsigen.2020.102257

2.1 Background

In recent years, metagenomics - the genomic analysis of microorganisms by direct extraction of DNA from an environmental sample - has become the most rapidly developing branch of microbiology [161, 162, 163]. The interest in metagenomics has grown drastically due to the expanding number of studies showing that the vast majority of microorganisms cannot be grown under laboratory conditions [35, 164, 165, 166]. The possibility of culture-free investigation of microbial biodiversity directly from an environmental habitat led to many amount of studies benefiting a wide range of fields such as human health [167, 168, 169, 170, 171], ecology [172, 173], agriculture [174, 175, 176], forensics [177, 178], food and drugs production [54, 55, 179]. Taxonomic profiling of metagenomic data is the key step during the data analysis, allowing researchers to understand the structure of a microbiome and to estimate the abundances of the organisms living in it. The main goal of this study is to compare different data types and methods for taxonomic profiling of metagenomic data sets with known abundance distributions of inhabitants.

The most common technique to investigate microbiome composition is amplicon-based sequencing of the 16S rRNA gene [180, 181]. This relatively short (~1500 bp) gene is universal among bacteria and archaea [70, 71]. There are in total nine hypervariable regions in the 16S rRNA gene that provide phylogenetic signatures on different taxonomic levels. Hypervariable regions are surrounded by highly conserved sequences, which are used for primer design. The analysis of 16S metagenomic datasets is usually performed in combination with one of several curated databases that contain annotated sequences of the 16S rRNA gene or its parts [182]. The most commonly used 16S-specific databases are RDB [74, 75], GreenGenes [77] and SILVA [76]. Analysis of 16S data is now routine for metagenomic-associated projects, though many studies demonstrated a number of biases associated with this type of data that make the validity of this approach questionable. Several reports stressed uneven coverage of microorganisms' diversity spectrum by common PCR primers for the 16S rRNA gene amplification [183, 184, 185, 186, 83, 84]. Second, the 16S rRNA gene does not have a correct phylogenetic relationship within particular taxa [81, 82]. The fact that bacteria and archaea might carry different copy numbers of the 16S rRNA gene in their genomes seriously influences a reliable abundance estimation after analysis of 16S data [187]. Additionally, the choice of a specific hypervariable region and the reference database for the subsequent analysis requires *a priori* knowledge about the investigated metagenome. Lastly, 16S data cannot be used to investigate the metagenome functional profile, nor does it provide any information about eukaryotic or viral members of the microbial community. The applicability of 16S data was shown for a set of forensic studies. For example, 16S data was successfully used for body fluid recognition [188] or matching between individuals' skin datasets and touched objects [62, 63]. The success of such analyses,

however, does not imply that a 16S-based analysis of all metagenomic data is reliable (or possible).

Apart from 16S, there are other methods that use rRNA genes to investigate microbial diversity. Among them are 23S, 5S, 12S and various combinations [189, 190, 191]. Other methods like the IS-pro approach use 16S-23S ribosomal interspace fragment lengths to analyse microbial communities [192]. Although these methods are very suitable for some specific tasks, they are not as widely applied as 16S. Several recent studies are based on targeting other genes in addition to 16S in order to determine the cell type of the forensic traces [193] or to perform skin sample identification using only microbial targeting genes [59, 194]. These studies also suggest that traditional 16S data is not always sufficient for a meaningful metagenomic analysis of forensic traces.

In recent years, the number of metagenomic studies based on the whole genome shotgun (WGS) sequencing data type has grown [90, 195, 196, 197, 198]. Among the main reasons for this are advantages in sequencing techniques allowing for the generation of sufficient number of high-quality reads for the WGS datasets, and bioinformatics algorithms to perform subsequent analysis of the big data. Though using WGS data avoids the biases introduced by 16S, it requires more computationally intense analysis, as well as higher sequencing costs.

While many studies in the field of forensics are based on the analysis of 16S data [199], "the capacity of WGS data of microbiomes to aid in forensic investigations by connecting objects and environments to individuals has been poorly investigated" [200]. Presently, WGS experiments are reserved for those studies for which analysis beyond the taxonomical assignment is required: investigating the microbiomes' functional profile, correlation between metagenome and host genome, search for the possible virulent genes, etc. The vast majority of taxonomical annotations is still performed by using only 16S data, despite all known disadvantages of the method [90]. One of the reasons for that is the lack of a well-performed benchmark study, comparing 16S and WGS data types. The vast majority of existing metagenomics benchmarks are created in order to evaluate the accuracy of various metagenomic profiles and comprise either only 16S [201] or only WGSdata [202, 203, 204, 205, 206, 117, 207, 208]. Existing benchmarks that can be used to compare 16S and WGS data types are *in-silico* created and based on a random set of bacterial species, lacking the information about whether or not the selected set of organisms might live together in the same environment [97]. One of the main goals of this study is the creation of a set of benchmarks allowing to compare the 16S and WGS data types using a set of in-vitro DNA mixes of bacteria species inhabiting skin.

Over the last decade, the number of different techniques for metagenomics data analysis has grown remarkably. The tools used for performing the taxonomical annotation, can be split into several groups based on the following criteria: strategy for reads assignment (alignment or matching based on the *k*-mers or sequences signatures); the database against which the search is performed; the proportion of

reads participating in the profiling (all reads, only one read per read group, only reads with particular features).

To investigate which type of metagenomic data is preferable for accurate taxonomic annotation, as well as to test which method of reads assignment yields more precise output, we created a series of bacterial mixes with known content. Each metagenomic mix incorporated 14 to 15 bacterial species belonging to 7 distinct bacterial genera. Each mix had a distinct distribution of the species abundances. For the analysis we selected two popular tools: Centrifuge [111] and MG-RAST [118]. These allow analysis of amplicon and WGS sequencing data and both perform the metagenome profiling by a comparison of sequencing data to a reference database. However, the strategies for metagenome profiling they exploit are different.

We did not include other popular tools for metagenomic analysis in this study as they either have a similar analysis strategy as the tools described above or are designed only for WGS or amplicon data analysis. In many studies, QIIME [100], objectively the most popular tool for amplicon data analysis, was shown to perform with the same accuracy as the MG-RAST pipeline for 16S rRNA sequencing data [209].

2.2 Materials and Methods

2.2.1 DNA extraction and concentration measurement

Laboratory pure cultures of 15 bacterial species that frequently inhabit human skin (Table 2.2) were grown with gentle shaking overnight at 37°C. Genomic DNA was isolated with the Easy-DNA™ gDNA Purification Kit (Invitrogen™ Thermo Fisher Scientific) using the standard protocol with ethanol precipitation [210]. RNA contamination was removed using RNase A (Roche) and the DNA was stored at 4°C. DNA concentrations were measured with the Qubit 3.1 Fluorometer (Invitrogen™).

2.2.2 Metagenomic mixes creation

Four bacterial mixes with known genome abundances were created for this research. In order to achieve the desired species abundances, the estimated genome size and the measured DNA concentration for each bacteria were used. One mix was created to have a uniform- and other three mixes an exponential ($\lambda = 1/6$, $\lambda = 1/2$ and $\lambda = 5/6$) distribution of species abundances. From here on, these mixes are referred to as EQ, EXP16, EXP12 and EXP56 respectively. Due to technical reasons, *Corynebacterium jeikeium* was included only in EQ. The remaining 14 species were used in all mixes.

Step	Temperature, °C	Duration, min	Cycles
Initial denaturation	95	3	1, hold
Denaturation	98	0.25	Ranged from 3 to 8
Annealing	59	0.5	depending on
Extension	72	1.5	sample
Final extension	72	5	1, hold

Table 2.1: PCR protocol for the WGS library preparation.

2.2.3 WGS sequencing library creation

DNA shearing was performed using the Covaris S2 sonicator (Covaris®) with the following settings: duty factor = 10%, intensity = 2.5, cycles/burst = 200, temperature = 6°C, total time, sec = 45. Size selection was performed on the sheared products with Ampure XP beads (Agencourt) to maintain insert size around 450 base pairs. Illumina sequencing libraries were prepared by ligating custom Illumina Truseq adapters with dual barcoding (10 base pairs) using the KAPA Hyper Prep Library Preparation kit (KAPA Biosystems, Inc.). To increase library yield, additional library amplification was performed with KAPA HIFI HotStart ReadyMix using the PCR protocol described in Table 2.1. To enable balanced pooling, sequencing libraries were quantified in duplicate by real time PCR using the KAPA SYBR®FAST qPCR kit. Quantification reactions were performed on a LightCycler®480 (Roche) using a dilution series of PhiX control library (Illumina) as standard [210]. After pooling the libraries, the final pool was quantified again using the same method to enable optimal loading of the flow cell.

2.2.4 16S sequencing library creation

Previously published [211] Primers and PCR-protocol for the amplification of V3-V4 region of the 16S rRNA were used. Illumina sequencing libraries were prepared by ligating custom Illumina Truseq adapters with dual barcoding (10 base pairs) using the KAPA Hyper Prep Library Preparation kit (KAPA Biosystems, Inc.).

2.2.5 DNA sequencing

Sequencing of WGS and 16S libraries was performed on the MiSeq®sequencer (Illumina) using v3 sequencing reagents according to the manufacturers' protocol with approximately 5% of PhiX control. This yielded one paired-end dataset with a read length of 299 bp per sample.

2.2.6 Bacterial genomes assembly

Sequencing reads for each bacterium were preprocessed using the Flexiprep quality control pipeline[1]. Post-QC reads were assembled by SPAdes Genome Assembler [212] with default settings.

2.2.7 Regression analysis

k-mer counting was performed using command count of the kPAL toolkit [213] with *k* set to 11. In case of the absence of the alternative DNA stand, *k*-mer profiles were balanced with balance command of the kPAL toolkit. Linear regression was done using the scikit-learn package for Python [214] with the *fit_intercept* parameter set to "False". The model training and prediction was performed using 5-fold Cross Validation.

2.2.8 Analysis using Centrifuge

Centrifuge is a popular tool that allows for fast classification of reads in a metagenomic sample using comparison of *k*-mers derived from each read to an indexed database. Centrifuge performs classification for all reads in a metagenomic sample independently using the following algorithm. For each read it creates a classification tree by pruning the taxonomy and only retaining taxa (including ancestors) associated with *k*-mers found in that read. Each node is weighted by the number of *k*-mers mapped to the node, and the path from root to leaf with the highest sum of weights is used to classify the read. A fast and effective comparison is achieved using the genome indexing technique, which is based on the Burrows-Wheeler transform [112] and the Ferragina-Manzini index [113]. To perform taxonomy assignment, Centrifuge requires an indexed database which is based on the reference database and its associated phylogenetic tree. A number of popular and regularly updated premade indexed databases are available on the Centrifuge website[2]. It is also possible to create a custom Centrifuge indexed database.
Metagenomic mixes samples were subjected to a QC-check using FastQC[3] (version 0.11.7). Leftover adapter removal and quality trimming of the reads was performed with cutadapt [95] (version 1.16, using options *--trim-n*, *--minimum-length* = 50 and *--quality-cutoff* = 20). The number of reads before and after each aforementioned step can be found in supplementary Table S1. High quality pairs of overlapping reads were merged with FLASH [215] (version 1.2.11, using option *--max-overlap*=300). For the subsequent taxonomic classification with Centrifuge, both merged reads and pairs of non-merged reads were used.

[1] Available online at http://biopet-docs.readthedocs.io/en/latest/pipelines/flexiprep/
[2] ftp://ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data
[3] Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Post-QC reads were analysed with Centrifuge (version 1-0-2-beta, default settings). Three different reference databases were used for the analysis: RefSeq database of complete genomes of bacteria and archaea [216] (downloaded as premade in April 2018 Centrifuge index); GreenGenes 16S sequences database (downloaded in June 2018) and SILVA 16S sequences database (downloaded June 2018). In order to make the content of reference databases comparable, sequences marked as eukaryotic were removed from SILVA database. Results obtained by Centrifuge were analysed using the Pavian interactive browser application [217].

2.2.9 Analysis using MG-RAST

MG-RAST is a web-based tool that allows the user to upload sequences and their metadata and download the analysis results. The MG-RAST pipeline creates a metagenomic profile by extracting rRNA and protein coding sequences. Gene calling is performed by the FragGenescan [119] algorithm, predicted protein sequences are clustered using UCLUST [101]. Potential rRNA genes are identified using BLAT [107] against a reduced version of the SILVA database and clustered with UCLUST. From each obtained cluster one representative sequence (the longest one) is chosen for the comparison with a reference database (M5nr58 [218] for proteins and combination of SILVA59, GreenGenes42 and RDP41 for rRNA analysis) using BLAT. All sequences from a particular cluster are assigned to the same taxonomic group as the clusters' representative. Thus, only rRNA genes and functional genes are used for the analysis of the metagenome, and the reads assignments are not independent. This strategy allows MG-RAST to perform taxonomic and functional profiling of metagenomic data. Finally, MG-RAST supports different metagenomic datatypes: genomic (including WGS and 16S) and transcriptomic. It also considers the metagenome origin, sequencing platform and many other features to tune the pipeline for a specific task. Raw reads of bacterial mixes samples were submitted to the MG-RAST Metagenomics analysis server under project number 85582. Paired reads merging and quality control was performed as part of the standard MG-RAST pipeline.

2.2.10 Taxa abundance estimation and results evaluation

Since the 16S amplification product has the same length among all bacterial taxa, no correction for genome length is needed when estimating relative abundances of the taxa. For the WGS samples however, normalization of read counts is required because of the differences in genome lengths. In order to perform correct taxa abundance estimations for taxonomic ranks higher than species, it is important to know how many reads assigned to that taxon belong to each species within the taxon. Both tools, Centrifuge and MG-RAST, assign reads to a node in the phylogenetic tree. Thus, reads assigned to a particular genus, for example, might belong to each of the species included to that genus as well as to the genus itself,

without species annotation. The main assumption of our approach for the estimation of taxa abundances is the following: all reads, assigned to the node higher than species level (regardless of whether or not they have species annotation), will be distributed among the species belonging to that node the same way as the reads with known species annotation. If the estimated abundances for species were known (in case of taxonomic annotation with Centrifuge), the procedure is trivial. When performing the analysis with MG-RAST the reads are classified only up to the genus level. In that case an equal distribution of reads among the species belonging to the particular genus was assumed.

2.2.11 Statistical and correlation analysis

Correlation analysis was performed using the Pearson correlation coefficient, pair wise comparisons were performed using the two-sided Mann-Whitney U test [219] and False Discovery Rate (FDR, a statistical approach used in multiple hypothesisto correct for multiple comparisons) control was performed using the Benjamini-Hochberg procedure [220]. We used the ratio of properly predicted taxa to all taxa predicted at that rank as a measure for the precision. Sensitivity was calculated as the ratio of properly predicted taxa to all taxa that were supposed to be present in the sample at that rank. F-scores (a measure of accuracy that considers both precision and sensitivity) were calculated as described in [221].

2.3 Results and Discussions

2.3.1 Individual bacterial genomes assembly

We sequenced and assembled the genomes of all 15 selected skin-associated bacteria individually. The total length of the assembly for each species was comparable to the length of the species references (Table 2.2 and section S1 of the Supplementary materials). For one species (*A. lwoffii*) there was no reference sequence available. Obtained assembly lengths as well as the DNA concentration measured for each bacterium were used to create four metagenomic mixes: one with equal and three with exponential ($\lambda = 1/6$, $\lambda = 1/2$ and $\lambda = 5/6$) distribution of bacterial species abundances. Taxa abundances were ordered from high to low as shown in Fig. 2.1.

2.3.2 Estimation of reference abundances

In order to estimate an abundance of an organism in terms of genome copies, the length of the genome and the lengths and (relative) copy numbers of any plasmids needs to be known. In the absence of a strain-specific reference sequence, *de novo* assembly of a single organism can be used to obtain these data [222]. In most common approaches [223], the coverage (and thereby the copy number) of contigs (see Supplementary Fig. S1) is not considered when estimating an assembly length, which leads to an inaccurate estimation of the organisms' genome length and thus influence the accuracy when creating bacterial mixes (see Supplementary Fig. S2 for a step-by-step explanation). Other factors, such as inaccuracy in DNA concentration measurement or mixing, can also lead to different abundances in the final bacterial mixes from those intended.

Since the content of all our metagenomic mixes is known and individual assemblies of all bacterial species were available, the intended distribution of bacterial abundances in the metagenomic mixes could be verified using the following approach. We used *k*-mer counts as a proxy for the number of genomes present in a pure (unmixed) sample. Using these counts, we are able to infer the relative contributions to a mixture. We use randomly chosen *k*-mers from the pure samples as profiles for the organisms, the same *k*-mers are used to make a profile of the mix and by linear regression, we estimate the contribution of each profile and thereby the contribution of each organism to the mix. For a more detailed description and a motivational example, see Section S1 and Figure S2 of the Supplementary materials. We calculated the 11-mer profiles for each bacteria using the contigs obtained after individual genome sequencing and assembly. Since profiles were calculated using contigs, we compensated for the absence of the reverse-complement DNA strand. We also calculated the 11-mer profiles of the WGS datasets of each of the metagenomic mixes, in these cases strand balancing was not applied. The 11-mer profiles were used to build a linear regression model in which the individual bacterial *k*-mer counts were

treated as independent variables and the k-mer counts of the metagenomic mix served as dependent variable.

Bacteria	Number of contigs	Accession number	Reference length, Mb	Assembly length, Mb
Acinetobacter johnsonii ATCC 17969	206	NZ_CP010350.1	3.51	3.88
Acinetobacter lwoffii ATCC 15309	180	NA	NA	3.44
Corynebacterium jeikeium ATCC 43734	234	NC_007164.1	2.46	2.6
Corynebacterium urealyticum ATCC 43042	99	NC_010545.1	2.37	2.35
Moraxella osloensis NCTC 10145	89	CP014234.1	2.43	2.58
Propionibacterium acnes ATCC 6919	26	NC_017550.1	2.49	2.55
Pseudomonas aeruginosa ATCC 10145	99	NC_002516.2	6.26	6.35
Staphylococcus aureus ATCC 29213	45	NZ_CP009361.1	2.78	2.72
Staphylococcus capitis ATCC 27840	52	NZ_CP007601.1	2.44	2.6
Staphylococcus epidermidis ATCC 12228	142	NC_00446	2.5	3.3
Staphylococcus haemolyticus ATCC 29970	770	NC_007168.1	2.69	2.86
Staphylococcus saprophyticus ATCC 15305	351	NC_007350.1	2.15	1.89
Streptococcus piogenes ATCC 19615	65	NZ_CP008926.1	1.84	1.82
Staphylococcus xylosus ATCC 29971	97	NZ_CP008724.1	2.52	2.74
Streptococcus mitis LMG 14552	49	NC_013853.1	2.76	2.83

Table 2.2: Bacterial species used for metagenomics mixes.

To verify the intended distribution of bacterial abundances in the metagenomic mixes, we use *k*-mer counts as a proxy for the number of genomes present in a pure (unmixed) sample. Using these counts, we are able to deconvolute a mixture. We use randomly chosen *k*-mers from the pure samples as profiles for the organisms,

the same *k*-mers are used to make a profile of the mix and by linear regression, we estimate the contribution of each profile and thereby the contribution of each organism to the mix. For a more detailed description and a motivational example, see section S1 and Figure S2 of the Supplementary materials. We calculated the 11-mer profiles for each bacterum using the contigs obtained after individual genome sequencing and assembly. Since profiles were calculated using contigs, we compensated for the absence of the reverse-complement DNA strand. We also calculated the 11-mer profiles of the WGS datasets of each of the metagenomic mixes, in these cases strand balancing was not applied. The 11-mer profiles were used to build a linear regression model in which the individual bacterial *k*-mer counts were treated as independent variables and the *k*-mer counts of the metagenomic mix served as dependent variable.

Since *k*-mer counts within one profile might be correlated, which violates the condition for using the regression analysis, we did not analyse the complete profile of 4,194,304 possible 11-mers. Instead we performed 1,000 iterations, in each iteration choosing 10,000 random *k*-mers and performing the regression analysis on that subset of *k*-mers. Thus, for each organism we got 1,000 estimations of its abundance in each mix. The result of this analysis is presented in Figure 2.1. Each boxplot shows the distribution of the organisms' abundances obtained from the regression analysis. The median model fit of the cross-validated models (measured using the R^2 coefficient of determination) for each mix was larger than 0.95, accuracy of the prediction (also measures using the R^2 but on the data that did not participate in the model training) ranged from 0.80 to 0.92 depending on the mix.

The regression analysis confirmed the distribution of bacterial abundances we aimed for (uniform distribution turning into the exponential one), though for some species (e.g., *S. haemoliticus* and *P. aeruginosa*), slight positive or negative deviations from the anticipated values were found. This can be caused by a number of factors such as inaccuracy in the DNA concentration measurement or DNA mixing, presence of large amounts of non-chromosomal DNA (e.g., plasmids) in the pool of bacterial DNA or inaccuracy in bacterial genome size estimation.

We use the results of this analysis as reference abundances for the experiments done in section 2.3.5.

2.3.3 Analysis of bacterial mixes using Centrifuge and MG-RAST

The mixes were sequenced on the Illumina MiSeq using WGS (samples EQ_WGS, EXP16_WGS, EXP12_WGS and EXP56_WGS) and 16S for V3-V4 region (samples EQ_16S, EXP16_16S, EXP12_16S and EXP56_16S) protocols. Information about read counts and QC statistics for each obtained dataset can be found in Supplementary table S1.

WGS and 16S samples obtained from our four metagenomic mixes were analysed with Centrifuge using the RefSeq complete bacterial genomes database. We per-

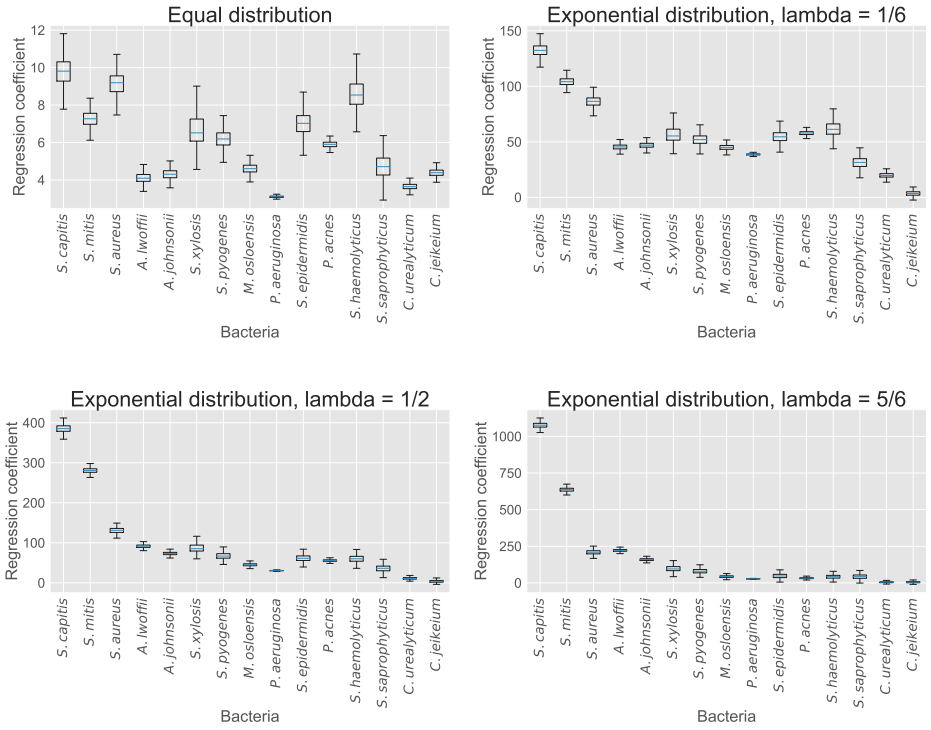

Figure 2.1: Regression analysis performed for metagenomic mixes to estimate relative abundances. Results for each mix are shown on a separate plot. Each boxplot represents the distribution of regression coefficients (vertical axes) obtained for the particular organism (horizontal axes), thus representing the distribution of bacterial abundances within that particular mix.

formed additional analysis for 16S samples using Centrifuge with the GreenGenes and SILVA reference databases.

All eight datasets (four WGS and four 16S) were submitted to the MG-RAST Metagenomics analysis server under project number 85582. RefSeq and GreenGenes databases provide taxonomic annotation down to the species level, while SILVA database as well as the databases used by MG-RAST are restricted to the genus level. Since the NCBI taxonomy and the taxonomy used by MG-RAST were different at the order level for our set of bacteria, we excluded annotation at the order level from further analysis.

2.3.4 Profiling accuracy without considering relative abundances

Because the content of the metagenomic mixes is known, we can verify how many of the reported taxa on each taxonomic rank are correct (true positive counts), how many are incorrect (false positive counts) and how many are missed (false negative counts).
Using these counts, both precision and sensitivity can be calculated. A perfect prediction is made if both precision and sensitivity equal one. As can be seen in Figure 2.2, both precision and sensitivity tend to increase in all cases with increasing taxonomic rank. For all 16S datasets analysed with Centrifuge, we observe that precision never reaches its maximum value, while for WGS datasets analysed with Centrifuge precision reaches its maximum already at the genus level. Interestingly, for 16S datasets analysed with MG-RAST, precision reaches its maximum at the genus level, but the sensitivity does not increase any further. For WGS datasets analysed with MG-RAST, sensitivity reaches its maximum already at the family level.
The accuracy of the classifications can be expressed using the F-score, which is calculated using precision and sensitivity. We tested whether the F-scores differed significantly for each pair-wise comparison using the Mann-Whitney U test and the Benjamini-Hochberg procedure for FDR control. The full table of p-values can be found in Supplementary Table S2, a summary of the results is shown in Figure 2.3. In most cases, the F-scores differ significantly when comparing WGS to 16S. At the same time, when comparing WGS datasets with different tools, a significant difference was observed only at the genus level.

2.3.5 Abundance assignment accuracy

Both Centrifuge and MG-RAST provide read counts for each reported taxon. We considered only reads that were assigned to the expected taxa and compared their relative abundances to the reference abundances.
Only Centrifuge, when using either the RefSeq or GreenGenes database, reported the taxonomic assignment down to the species level. In Figure 2.4, each metagenomic mix is shown as a separate graph with species listed on the horizontal axes and their relative abundances shown on the vertical axes. The black line represents the intended distribution of species abundances. The dark green line shows the mean reference abundances with the light green area representing ± 3 standard deviation around those means. The blue and red lines show the relative abundances obtained for 16S and WGS datasets respectively, with the solid blue line for the 16S analysis done using the RefSeq database and the dashed blue line using the GreenGenes database. As can be seen in Figure 2.4, the analysis of 16S data results in a considerable overestimation of abundance of *A. johnsonii*. Centrifuge failed to identify *A. lwoffii*, since there is no complete genome of that bacterium in the RefSeq database and it did not report any significant presence of *C. jeikeium* in

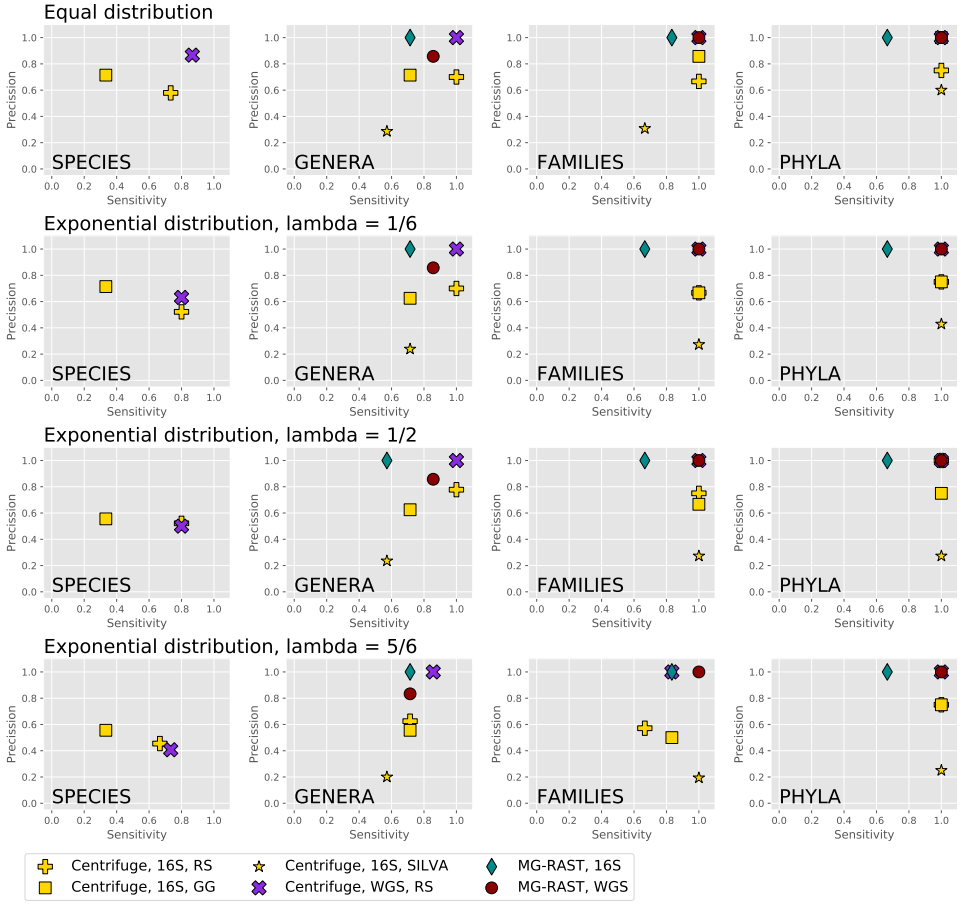

Figure 2.2: Precision vs. sensitivity of different profiling approaches. Results for each mix and taxonomic rank are shown separately, with sensitivity on the horizontal axes and the precision on the vertical axes. Each shape represents a combination of method, data type and reference database. RS - RefSeq database, GG - Greengenes database, S - SILVA database.

the exponentially distributed metagenomic mixes. Analysis of the 16S datasets using the GreenGenes database reported overestimated values for *S. epidermidis* and *A. johnsonii* and did not report the presence of nine out of fifteen bacteria because of their absence in the GreenGenes database.

We repeated the same analysis on three higher taxonomic ranks: genera, families and phyla. For all these three taxonomic levels we analysed the results of Centrifuge (Figure 2.5) and MG-RAST (Figure 2.6). As can be seen in Figure 2.5, the Centrifuge analysis of 16S datasets using different reference databases provided a similar biased

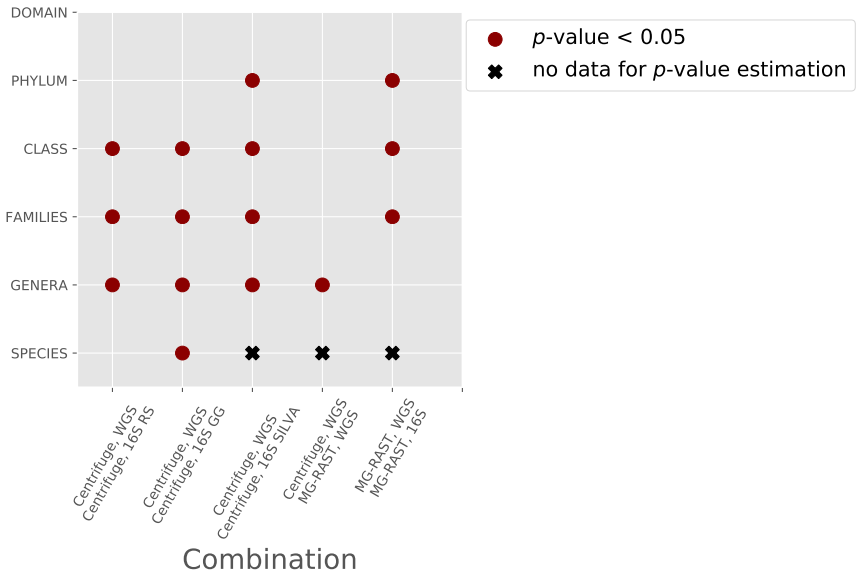

Figure 2.3: Comparison of F-scores (combination of precision and sensitivity) obtained from all four mixes for different combinations of methods, data type and databases. Red dots indicate a *p*-value below 0.05. Combinations of methods, data type and database are shown on the horizontal axis, taxonomic levels are shown on the vertical axis. RS - RefSeq database, GG - Greengenes database, S - SILVA database. Please note that data points are connected only to visualize the various types of distributions.

output, mostly due to an overestimation of the abundance of the Acinetobacter genus, Moraxelaceae family and Proteobacteria phylum. The dissimilarity with the reference abundances is especially pronounced at the phylum level. Results obtained for the WGS datasets with Centrifuge were concordant with the reference abundances with slight deviation for Acinetobacter genus, Moraxelaceae family and Proteobacteria phylum (Figure 2.5). It is interesting to note, that these taxa were also the major reason for disagreement between results obtained by Centrifuge for 16S datasets and reference abundances.

The results obtained for different 16S datasets by MG-RAST were not consistent (as is the case for Centrifuge) up to the phylum level. As can be seen in Figure 2.6, analysis of 16S datasets with MG-RAST reported many disagreements with reference abundances. The reasons of those disagreements are dataset- and taxonomy rank-specific. Results reported by MG-RAST became more or less consistent only at the phylum level, where they followed the same trend: overestimating the abundance of Firmicutes relative to that of Proteobacteria.

Abundances obtained after analysis with MG-RAST of WGS datasets were also following the reference results closely. There were, however, slight deviations from

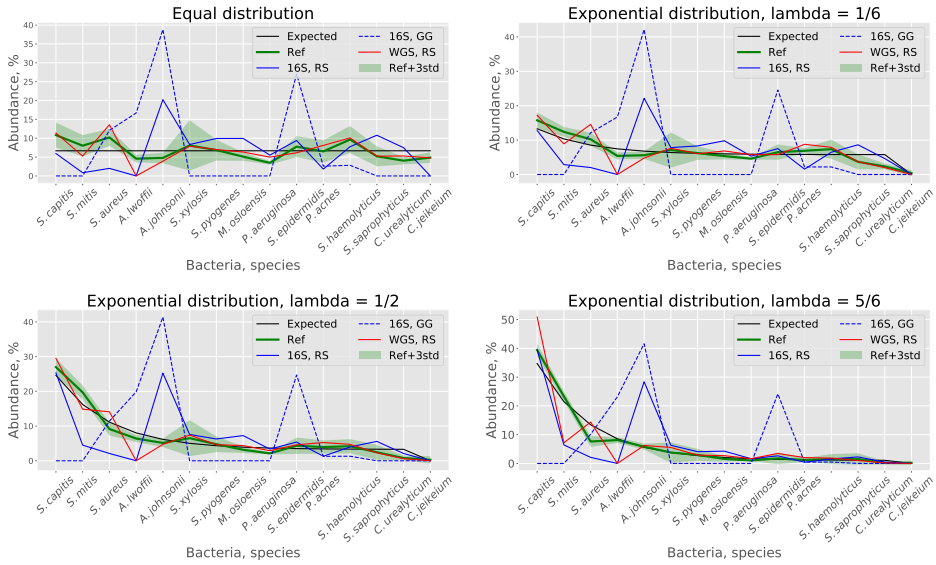

Figure 2.4: Comparison of relative abundances reported by Centrifuge (using two different reference databases) for WGS and 16S with relative abundances obtained from the regression analysis. Results for each mix are shown separately with the species' names on the horizontal axes and the relative abundance on the vertical axes. Ref - reference abundances, RS - RefSeq database, GG - GreenGenes database. Please note that data points are connected only to visualise the various types of distributions.

the reference abundances. These deviations were, like the results for 16S datasets, specific to taxonomy-rank and dataset.

In order to quantify the dissimilarity among the abundances provided by the different methods, datasets, reference databases and the results of regression analysis we calculated the absolute differences in abundances for each particular dataset and taxonomic rank. The averages of these values (from here on called the error rate) are reported in Figure 2.7. For the analyses of 16S datasets it is interesting to note that for Centrifuge the average error rate grew with the increase of the taxonomic rank in general. This was not the case for the error rate obtained for the 16S datasets using MG-RAST. We tested whether the average errors differed significantly for each pair-wise comparison using the Mann-Whitney U test and the Benjamini-Hochberg procedure for FDR control. The full table of p-values can be found in Supplementary Table S3, a summary of the results is shown in Figure 2.8. This analysis demonstrates that for all taxonomic levels the error rates in the abundance estimations provided by the analysis of 16S datasets (regardless of the method or reference database) are significantly different (higher) compared to the abundances reported for WGS datasets. We did not observe any significant difference in average error rate between

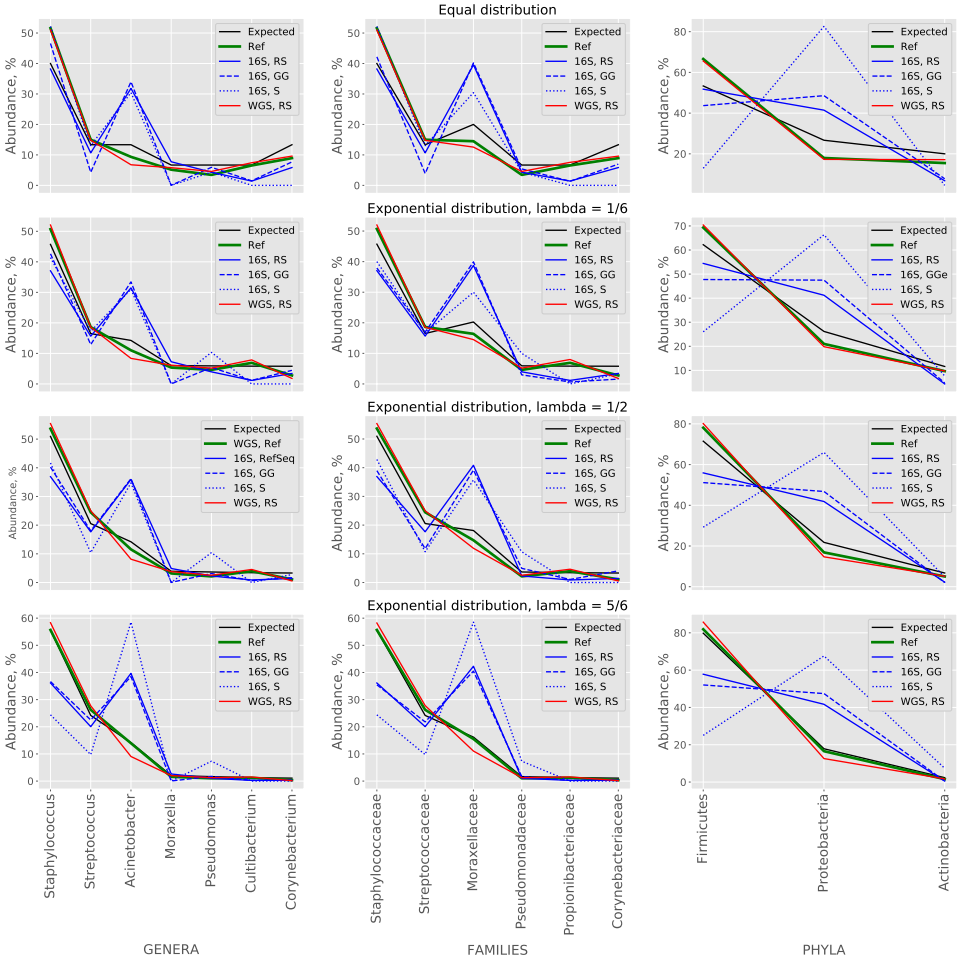

Figure 2.5: Comparison of relative abundances reported by Centrifuge (using three different reference databases) for WGS and 16S datasets on genera, orders and phyla levels with relative reference abundances. In the above grid of figures each row indicates the mix and each column indicates the taxonomic level. In each figure, the taxa are shown on the horizontal axes and the relative abundances are shown on the vertical axes. Ref - reference abundances, RS - RefSeq database, GG - GreenGenes database, S - SILVA database.

WGS datasets analysed with Centrifule and MG-RAST.

We compared the error rates reported by Centrifuge when using the three different 16S reference databases. Error rates observed in the analysis with RefSeq and GreenGenes databases were similar. Running the Centrifuge analysis using the SILVA database reported a much higher error rate. That might be a direct consequence of

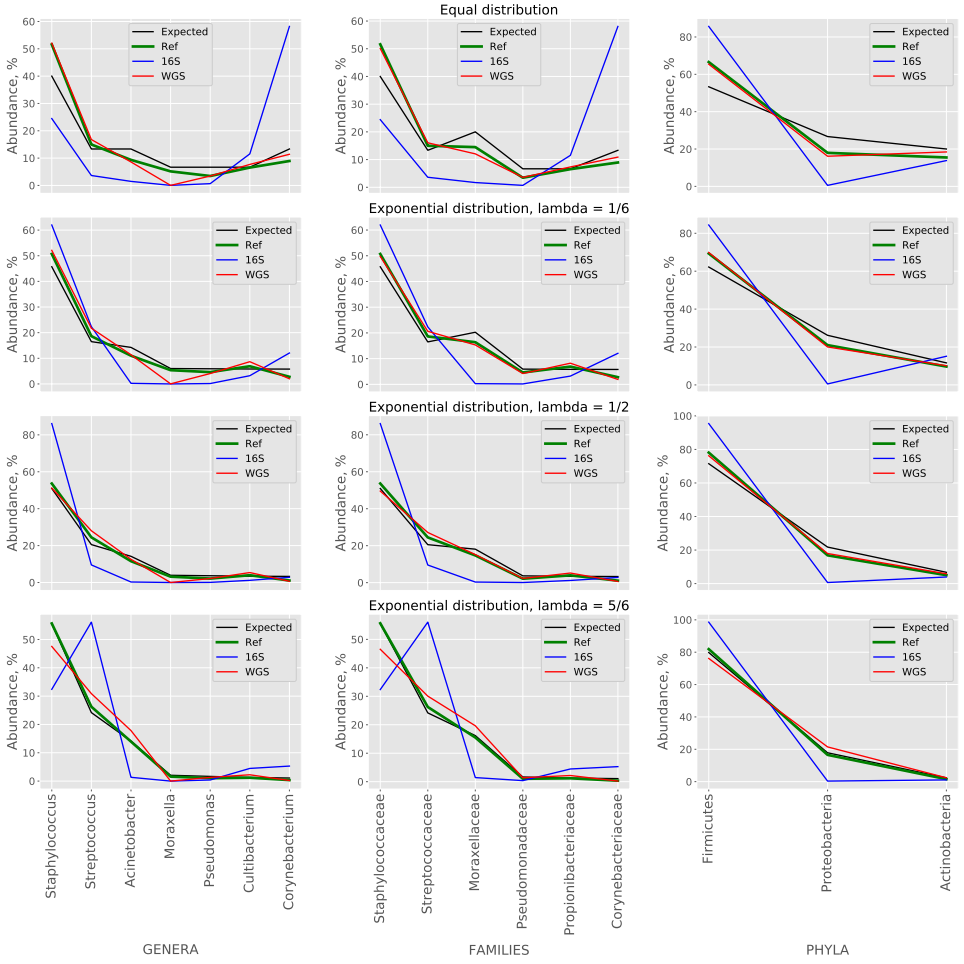

Figure 2.6: Comparison of relative abundances on reported by MG-RAST for WGS and 16S datasets on genera, orders and phyla levels with relative reference abundances. In the above grid of figures each row indicates the mix and each column indicates the taxonomic level. In each figure, the taxa are shown on the horizontal axes and the relative abundances are shown on the vertical axes. Ref - reference abundances

taxonomic annotation done using the SILVA database where a smaller proportion of reads was assigned to the expected taxa in comparison to other reference databases (see the section 2.3.4).

We also evaluated the similarity among the abundances obtained by employing distinct methods and databases using a correlation analysis. In Figure 2.9 the results of these comparisons are presented as a series of heatmaps.

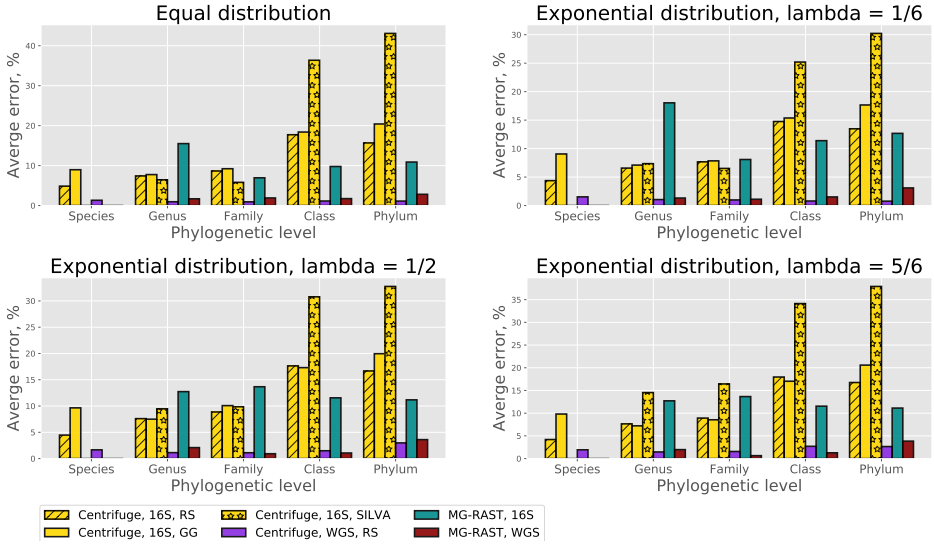

Figure 2.7: Average error between the abundances reported by the regression analysis and results obtained for different datasets, reference databases and tools. Results for each metagenomics mix are shown separately. Bar colours represent the tool, data type and reference database combination used for the analysis. The error rate (shown on the vertical axes) was calculated for each taxonomic rank (shown on the horizontal axes) separately. RS - RefSeq database, GG - GreenGenes database, SILVA - SILVA database

As can be seen from Figure 2.9, abundances obtained by the analysis of WGS data (Centrifuge and MG-RAST) for all datasets at all taxonomic levels positively correlate with reference abundances. Correlation of 16S analysis obtained using Centrifuge with the reference abundances becomes worse at higher taxonomic levels, which is the opposite for the 16S data results obtained using MG-RAST. The 16S data analyses obtained for Centrifuge and MG-RAST do not demonstrate positive correlation with each other.

2.4 Conclusions

In this study we created a series of bacterial mixes with known content in order to investigate which type of metagenomics data and reads assignment strategy yields better taxonomic classification. For each mix we generated WGS and 16S sequencing datasets and analysed them using Centrifuge with RefSeq, GreenGenes and SILVA reference databases and the MG-RAST metagenomics analysis server with M5nr and M5nra reference databases. We compared the results of all analysis done with Centrifuge and MG-RAST to the reference abundance profiles obtained from a regression *k*-mer-based regression analysis.

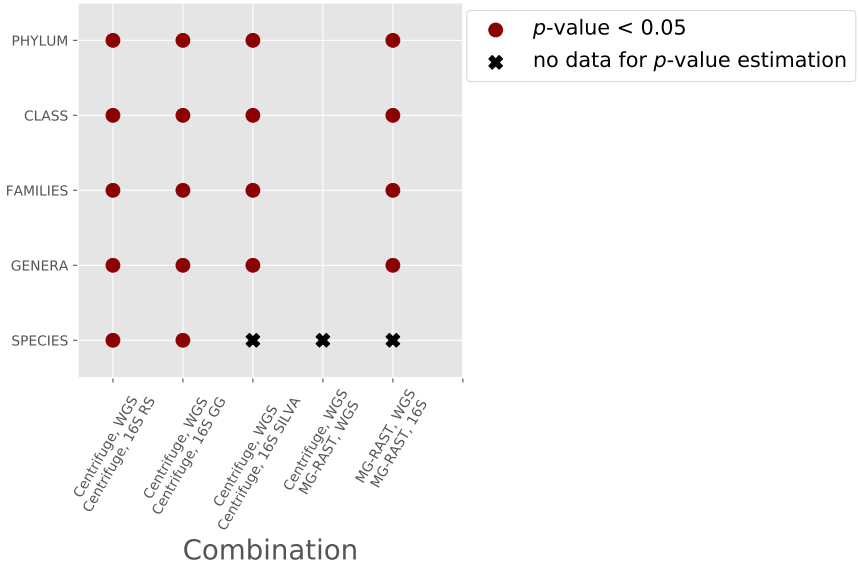

Figure 2.8: Comparison of average errors obtained from different mixes for different combinations of methods, data type and databases. Combinations of methods, data type and database are shown on the horizontal axis, taxonomic levels are shown on the vertical axis. Red dots indicate a *p*-value below 0.05. RS - RefSeq database, GG - GreenGenes database, SILVA - SILVA database

The results from both Centrifuge and MG-RAST show that WGS datasets provide much more accurate results in comparison to 16S-based methods. The analysis of WGS data displayed better coverage of all taxa expected to be present in the mixes on all phylogenetic levels, reaching the maximum accuracy already at the genus level for Centrifuge and at the family level for MG-RAST. On the other hand, results obtained for 16S-based data were often missing several taxa and/or had very high false-positive rate. Centrifuge analyses based on the 16S datasets were suffering from low precision, while MG-RAST analysis of the 16S datasets had low sensitivity. Abundance profiles obtained from WGS demonstrated much less disagreement with the expected abundances in comparison to the abundance profiles based on 16S data. This was shown using two different measurements: the average (per taxonomy rank) absolute difference between abundance profiles and by a correlation analysis. For 16S datasets analysed with Centrifuge, the deviation from the reference abundances, introduced at the species/genus levels, propagated further up the taxonomy which led to a greater difference with the expected outcome on the higher taxonomic ranks as well. In contrast, the analysis of 16S datasets performed by the MG-RAST pipeline demonstrated greater differences with the reference abundances on the lower taxonomic ranks in comparison with the higher ones. Our correlation anal-

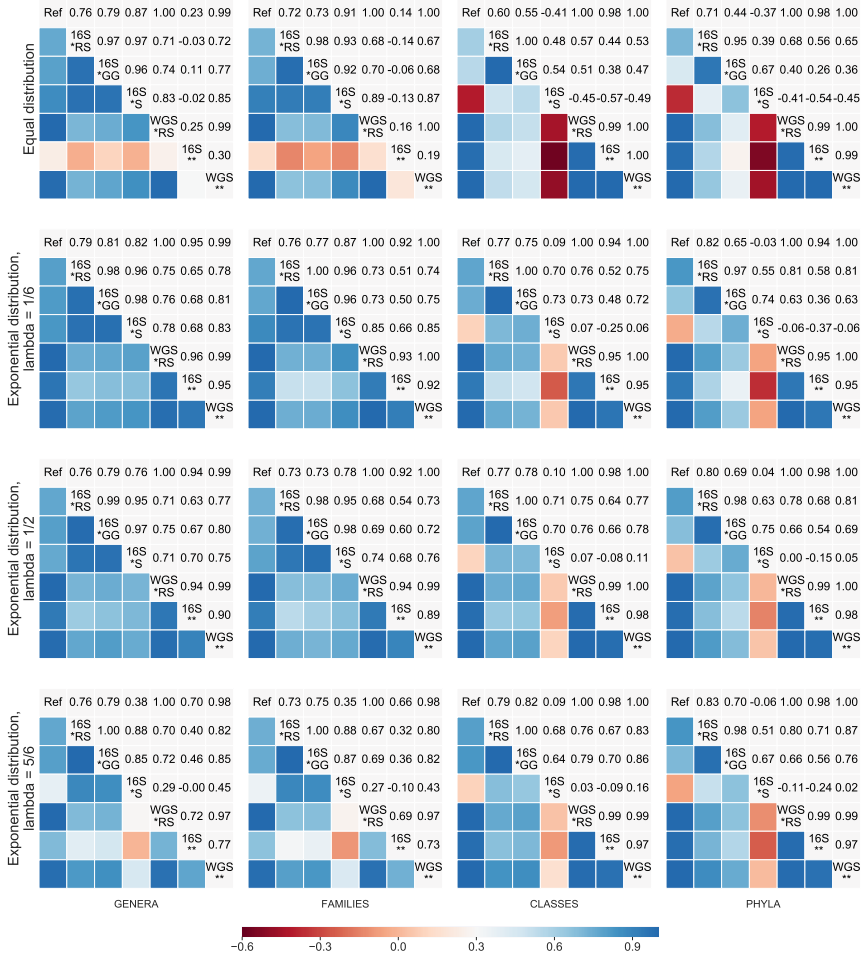

Figure 2.9: Correlation between the abundance profiles obtained for different combinations of method, datasets and reference database on distinct taxonomic levels. In the above grid of figures each row indicates the mix and each column indicates the taxonomic level. The combination of analysis type, dataset and reference database are shown on the main diagonal of the heatmap, with the lower triangle representing the correlation shown in colors and the upper triangle demonstrating the same data in the numeric representation. Ref - reference abundances, * - Centrifuge, ** - MG-RAST, RS - RefSeq database, GG 435 - GreenGenes database, S - SILVA database.

ysis shows that the agreement between the MG-RAST results of 16S datasets and reference abundances was growing with increasing taxonomic level.
Both tailor-made 16S databases (GreenGenes and SILVA) did not perform better than the RefSeq database when analysing 16S datasets using Centrifuge. The Centrifuge results using RefSeq and GreenGenes databases were correlated with a correlation

coefficient higher than 0.95 for all 16S datasets on each taxonomic rank starting with genus.
We conclude that WGS data is preferable for the study of metagenomic data, especially when the correct inhabitant abundances are required. We could not determine which of the explored methods for the taxonomic assignment of the WGS data provides a more accurate outcome. Centrifuge, however, has minor advantages in comparison to MG-RAST, such as a faster, deeper and slightly better reads classification, the possibility of local installation and use of custom databases and a more flexible tuning of the tools' settings. Among the investigated techniques for 16S metagenomic data analysis, MG-RAST demonstrated slightly better results in both reads assignment and abundance estimation, albeit only at higher taxonomic ranks. As previously quoted, "the capacity of WGS data of microbiomes to aid in forensic investigations by connecting objects and environments to individuals has been poorly investigated". In light of this, our results are especially important, as they demonstrate the inefficiency of routine 16S data to produce the accurate taxonomical profiling.
The synthetic metagenomes created in our study is restricted to DNA of bacteria that inhabit skin surface - a logical target for forensics analysis. However, human skin is also the environment with one of the most within- and between-individual diverse microbiota on the human body. The benchmark we created is rather small and simple as the diversity of microbial species living on the human skin surface is much larger than only 15 species [224]. The significant inaccuracy of the results obtained for 16S data in comparison with those for WGS data on a small and simple set of benchmarks can possibly question the accuracy of the previous 16S-based forensic studies, at least those done on skin-associated microbial communities.

2.5 Author Statements

2.5.1 Funding information

This research is financed by a grant number 727.011.002 of the Netherlands Organisation for Scientific Research (NWO).

2.5.2 Authors' contributions

Conceptualization LK and JFJL; Methodology RHdL, MtR; Resources MEMK; Software LK and NP; Investigation, Visualization, and Writing original draft LK; Supervision JFJL, PdK and HM; Funding Acquisition JFJL; Reviewing and editing LK, JFJL, HM, NP, PdK, RHdL and MEMK.

2.5.3 Acknowledgements

We would like to thank Guy Allard for the support with the assembly of bacterial genomes and Louk Rademaker for the feedback on this manuscript.

2.5.4 Conflicts of interest

The authors declare that there are no conflicts of interest.

2.6 Data Availability

- Reads obtained after the individual sequencing of each selected bacterial species and used for the genome assembly were upload to the Sequence Read Archive under the study SRP159200.
- Sequencing reads of the metagenomic mixes as well as the results of the analysis performed by MG-RAST can be downloaded from the MG-RAST server (project number 85582).
- The summary of results obtained using Centrifuge as well as the supplementary materials for this research are deposited on Figshare: https://doi.org/10.6084/m9.figshare.c.4217672

Chapter 3

Reference-free resolving of long-read metagenomic data

L. Khachatryan[1], S. Y. Anvar[1], R. H. A. M. Vossen[3] , and J. F. J. Laros[1,2,4]

1 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
3 Leiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands
4 GenomeScan, Leiden, The Netherlands

bioRxiv 2019 https://doi.org/10.1101/811760

3.1 Background

The analysis of metagenomic data is becoming a routine for many different research fields, since it serves scientific purposes as well as improves our life quality. Particularly, with the use of metagenomics a large step was made towards the understanding of the human microbiome and uncovering its real composition and diversity [225, 226, 227, 228, 229, 230]. The understanding of the human microbiome in health and disease contributed to the development of diagnostics and treatment strategies based on metagenomics knowledge [231, 232, 233, 234, 235, 236, 237, 238]. The study of microbial ecosystems allows us to predict the possible processes, changes and sustainability of particular environments [239, 240]. Genes isolated from uncultivable inhabitants of soil metagenomes are being successfully utilized, for example, in the biofuel industry for production and tolerance to byproducts [30, 241, 242]. Various newly discovered biosynthetic capacities of microbial communities benefit the manufacturing of industrial, food, and health products, as well as contribute into the field of bioremediation [54, 55, 56, 57].

Despite all the progress made in resolving genetic data derived from environmental samples, it is still a challenging task. Reads binning is one of the most critical steps in the analysis of metagenomic data. To estimate the composition of a particular microbiome, it is important to ensure that sequencing reads derived from the same organism are grouped together. Currently, alignment of DNA extracted from an environmental sample to a set of known sequences remains the main strategy for metagenomics binning [243, 244]. There is a full range of techniques allowing the comparison of metagenomic reads to a reference database. It can be performed using different metagenomic data types (16S or WGS) and various matching approaches (classic alignment or matching performed using *k*-mers or taxonomic signatures). Most of the time, the binning is performed for all reads in the database, but in some cases only a particular subset of sequencing data is selected for binning. Lastly, there is a wide spectrum of databases that can be used to perform the binning. The database might contain all possible annotated nucleotide/protein sequences, marker genes for distinct phylogenetic clades, sequencing signatures specific to particular taxa, etc. (see more detailed explanation in Chapter 1). The obvious downside of all listed strategies is the incapability to perform an accurate binning for the reads derived from organisms that are not present in the reference database.

Metagenomics binning was improved by alignment-free approaches, which can be split into two subgroups: reference-dependent and reference-independent methods. The tools from the first subgroup utilize existing databases to train a supervised classifier for the reads binning. Various techniques can be performed to achieve this goal: Support Vector Machine, Interpolated Markov Models, Gaussian Mixture Models, Hidden Markov Models [147, 148, 149, 151, 152, 153, 150]. Even though these approaches are reference dependent, they can be used to classify reads derived

from previously unknown species. However, the accuracy of reference-dependent methods will be always limited by the content of reference databases. The content of the current reference databases utilized for training differs from the true distribution of microbial species on our planet[245, 246, 247, 248, 249, 250, 251]. For some metagenomic datasets the amount of unknown sequences might be quite high [252, 253], thus using supervised classification tools based on known genetic sequences is questionable if this is the case.

Reference-independent approaches for metagenomics binning try to solve the problem of missing taxonomic content: they are designed to classify reads into genetically homogeneous groups without utilizing any information from known genomes. Instead, they use only the features of the sequencing data (usually *k*-mer distributions, DNA segments of length *k*) for classification. One of those tools, LickelyBin, performs a Markov Chain Monte Carlo approach based on the assumption that the *k*-mer frequency distribution is homogeneous within a bacterial genome [140]. This tool performs well for very simple metagenomes with significant phylogenetic diversity within the metagenome, but it cannot handle genomes with more complicated structure such as those resulting from horizontal gene transfer [141]. Another one, AbundanceBin [142], works under the assumption that the abundances of species in metagenome are following a Poisson distribution, and thus struggles analysing datasets where some species have similar abundance ratios. MetaCluster [143] and BiMeta [144] address this problem of non-Poisson species distribution. However, for these tools it is necessary to provide an estimation of the final number of clusters, which cannot be done for many metagenomes without any prior knowledge. Also, both MetaCluster and BiMeta are using a Euclidian metric to compute the dissimilarity between *k*-mer profiles, which was shown to be influenced by stochastic noise in analysed sequences [145]. Another recent tool, MetaProb, implements a more advanced similarity measure technique and can automatically estimate the number of read clusters [146]. This tool classifies metagenomic datasets in two steps: first, reads are grouped based on the extent of their overlap. After that, a set of representing reads is chosen for each group. Based on the comparison of the *k*-mer distributions for those sets, groups are merged together into final clusters. Even though MetaProb outperformed other tools during the analysis of simulated data, it was shown to perform not very well on the real metagenomicmetagenomic data data.

In this article we present a new technique for alignment- and reference-free classification of metagenomic data. Our approach is based on a pairwise comparison of *k*-mer profiles calculated for each sequencing read in a long-read metagenomic dataset, using the previously described kPAL toolkit [213]. It also performs unsupervised clustering to facilitate the identification of genetically homogeneous groups of reads present in a sample. The main assumption of our method is that after assigning the pairwise distances for all reads in the dataset, those belonging to the same organism will form dense groups, and thus the metagenome binning could be resolved using

density-based clustering. We developed an algorithm which automatically detects the regions with high density and hierarchically splits the dataset until there is one dense region per cluster. The approach is designed to work with long reads (more than 1,000 bp) since we calculate *k*-mer profiles for each read separately and shorter reads would yield non-informative profiles. We performed our analysis on long PacBio reads that were either simulated or generated from a real metagenomic sample. We have shown that despite the fact that PacBio data is known to have a high error rate, the approach successfully performed read classification for simulated and real metagenomic data.

3.2 Materials and Methods

3.2.1 Software

All analyses were done using publicly available tools (parameters used are listed below for each specific case) along with custom Python scripts which are stored in a Git repository[1].

3.2.2 PacBio data simulation

Complete genomes of five common skin bacteria were used to generate artificial PacBio metagenomes (see Table 3.1). The reads were simulated from reference sequences using the PBSIM toolkit [254] with CLR as the output data type and a final sequencing depth of 20. For the calibration of the read length distribution, a set of previously sequenced *C. difficile* reads [255] was used as a model.

3.2.3 Bioreactor metagenome PacBio sequencing

Bioreactor metagenome coupling anaerobic ammonium oxidation (Annamox) to Nitrite/Nitrate dependent Anaerobic Methane Oxidation (N-DAMO) processes [256] was used to generate WGS PacBio sequencing data.
Metagenome contained the N-DAMO bacteria *Methylomirabilis oxyfera* (complete genome with GeneBank Acsession FP565575.1 was used as a reference), two Annamox bacteria (*Kuenenia stuttgartiensis*, assembly contigs from the Bio Project PRJEB22746 were used as a reference and a member of *Broccardia* genus, assembly contigs of *Broccardia sinica* from Bio Project PRJDB103 were used as reference) and an archaea species *Methanoperedens nitroreducens* (assembly contigs from the Bio Project PRJNA242803 were used as a reference).
Bacterial cell pellets were disrupted with a Dounce homogenizer. DNA was isolated using a Genomic Tip 500/G kit (Qiagen) and needle sheared with a 26G blunt end

[1] Available at https://git.lumc.nl/l.khachatryan/pacbio-meta

needle (SAI Infusion). Pulsed-field Gel electrophoresis was performed to assess the size distribution of the sheared DNA. A SMRTbell library was constructed using 5μg of DNA following the 20 kb template preparation protocol (Pacific Biosciences). The SMRTbell library was size selected using the BluePippin system (SAGE Science) with a 10 kb lower cut-off setting. The final library was sequenced with the P6-C4 chemistry with a movie time of 360 minutes.

3.2.4 Reads origin checking

Reads were corrected using the PacBio Hierarchical Genome Assembly Process algorithm before being mapped to the genomes of the expected metagenome inhabitants genomes using the BLASR aligner [257] with default settings. The alignments were used to determine the origin of the reads. Reads that were not mapped during the previous step were subjected to the BLASTn [102] search against the NCBI database. The identity cut-off was set to 90, the (E)value was chosen to be 0.001.

3.2.5 Bioreactor metagenome PacBio reads assembly

The assembly of corrected PacBio reads was performed using the FALCON [258] assembler. The resulting contigs were mapped to the candidate reference genomes using LAST [104] with default settings. To determine the similarity cutoff for the mapping procedure, the curve representing the number of contigs versus the similarity to the reference genome was analysed. The first inflection point at (in case of mapping contigs to the *M. oxyfera* genome 12%), dividing the fast-declining part of the curve from the slow-declining part, was chosen as a threshold (See Supplementary materials for more details).

3.2.6 Binning procedure

For each read, the frequencies of all possible five-mers were calculated using the *count* command of the kPAL toolkit. The resulting profiles were balanced (a procedure that compensates for differences that occur because of reading either the forward or reverse complement strand) and compared in a pairwise manner by using the balance and matrix commands of kPAL accordingly, yielding a pairwise distance matrix. Normalization for differences in read length was dealt with by the scaling option during the pairwise comparison.
The resulting distance matrix, hereafter called the original distance matrix, was subjected to a multi-step clustering procedure. A schematic representation of this procedure can be found in Figure 3.1. Due to practical limitations (runtime), this analysis was restricted to a set of 10,000 randomly selected reads. This multi-step clustering procedure works recursively: it starts with the analysis of a set of reads and either reports the entire set as one cluster, or it splits the set into two subsets,

which are each analysed using the same procedure. The decision whether to split the set of reads into two subsets is made using the following approach. First, the pairwise distances for all reads in the set are extracted from the original distance matrix in order to construct the working distance matrix. After that, the dimensionality of the analysed set is decreased to three using the t-SNE algorithm [259] in order to reduce noise caused by outliers in the distance matrix. The reads, now represented by a point in three-dimensional space, are subjected to density-based clustering using the DBSCAN algorithm [260] with the default distance function.

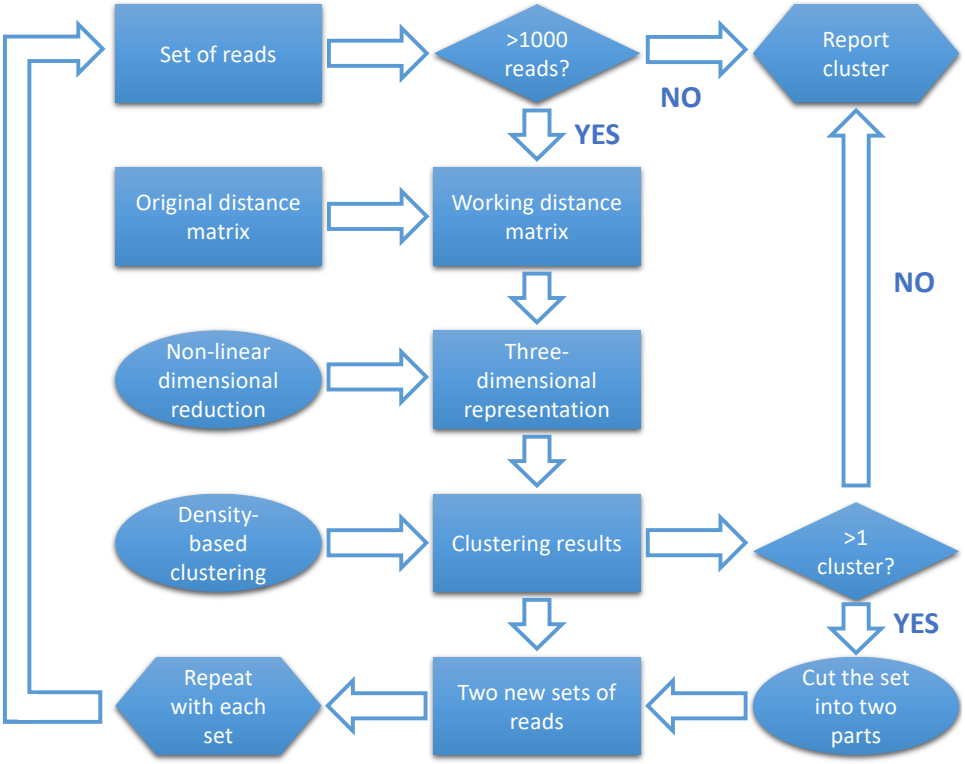

Figure 3.1: Schematic representation of the clustering procedure.

We choose the *MinPts* parameter of DBSCAN (the minimal amounts of points in the neighborhood to extend the cluster) to be either 1% of the size of the dataset for sets larger than 2,000 reads, or 20 for sets smaller than 2,000 reads. The number of clusters found by DBSCAN depends on the neighborhood diameter ε. When ε is too small, no clusters are reported since all points are isolated. On the other hand, when ε is too large all points are grouped into one cluster. Our algorithm

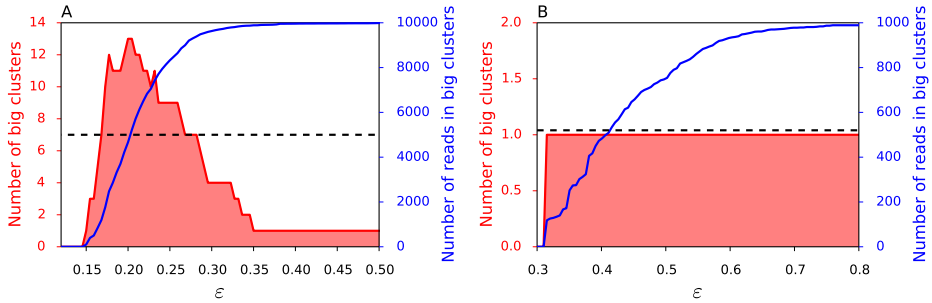

Figure 3.2: Density-based clustering analysis example. The data is clustered with DBSCAN with ε ranging from 0 to the value when 90% of the points are assigned to one cluster. When at least half of the data set is assigned to a dense cluster, the number of clusters is used to determine whether subdivision of the data set is required. Only if more than one cluster is identified at this point, the procedure is repeated recursively with two partitions of the data. The partitions are determined by using the largest ε that clusters the data into two clusters. In this example two datasets are shown: one that was further split into two partitions (A) and one that was reported as one dense cluster (B).

therefore performs a parameter sweep for ε, from the value providing zero clusters to the value with which 99% of the reads are grouped in one cluster for the chosen *MinPts*. The results of this parameter sweep are used to check the dependency of the number of dense clusters on a particular ε (only clusters larger than 100 points are considered) and how many points of the analysed set are included in the obtained clusters (Figure 3.2). If for some ε there are two or more clusters that together cover more than half of the total amount, the analysed set is divided into two new sets (Figure 3.2A). The analysed set is reported as one cluster if the aforementioned condition is not satisfied (Figure 3.2B), or when the size of the analysed set is smaller than 1,000 points.

The division is done using the following strategy. DBSCAN is performed using the optimal ε, yielding two dense clusters that serve as center points for two partitions. Each of the remaining unclassified points is assigned to the cluster containing the closest classified neighbor.

3.2.7 Classification for larger sets

Read classification for sets larger than 10,000 was performed in two steps. First, 10,000 reads (larger than 10 kb) were randomly chosen and classified using the algorithm described in section 3.2.6. After that, the pairwise distances between every unclassified read and every classified read were calculated using their 5-mer profiles. These distances were used to assign the unclassified read to the cluster containing the closest classified read.

3.2.8 Data avaliability

Sequencing reads of bioreactor metagenome were submitted to SRA under the study number SRP159147.
Supplementary materials were deposited on Figshare and available for downloading using the following link: https://doi.org/10.6084/m9.figshare.c.4218857.v1

3.3 Results

3.3.1 Reads classification in artificial PacBio metagenomes

To construct artificial metagenomes, we used simulated PacBio reads based on the genomes of five common skin flora bacteria together with so-called "noise" reads. These are reads from a PacBio sequencing data of an environmental metagenome [261] that were not assigned to the major inhabitant *K. stuttgartiensis* or other known organisms. They were added to represent low abundant species that are present in any typical metagenomic dataset.
We constructed four artificial PacBio datasets in this way, each containing 10,000 randomly selected reads (length > 9 kb) containing 0%, 5%, 10% and 15% noise reads, respectively. For the simplicity the number of simulated reads was adjusted to provide an equal abundance for each bacterium in the final metagenome (Table 3.1).

Reads origin	RefSeq AC	Genome length, Mb	Number of reads per dataset			
			0%	5%	10%	15%
S. mitis	NC_013853.1	2.1	1,246	1,183	1,121	1,059
P. acnes	NC_017550.1	2.5	1,443	1,371	1,298	1,226
S. epidermidis	NC_004461.1	2.6	1,448	1,376	1,304	1,231
A. calcoaceticus	NC_016603.1	3.9	2,236	2,125	2,013	1,901
P. aeruginosa	NC_002516.2	6.3	3,627	3,446	3,264	3,083

Table 3.1: Content of artificial metagenomics PacBio datasets.

We subjected each dataset to the classification procedure described in section 3.2.6. The reads in the resulting clusters were then classified according to their origin (See Supplementary Material for more data). In Figure 3.3, it can be seen that for each experiment we obtained five large clusters (> 1,000 reads) consisting mainly of reads belonging to the same species. For all three datasets containing noise reads we see the tendency of noise reads to be clustered with some fraction of *P. acnes* and *P. aeruginosa* reads. However, as can be seen from Figure 3.3 and Table 3.2, increasing the noise content leads to better isolation of these reads. Indeed, for dataset B (5% of the noise reads), the majority of noise reads were assigned to the cluster that is primarily occupied by reads belonging to *P. acnes* and *P. aeruginosa*. Increasing the noise content (dataset C and D in Fig. 4, 10% and 15% noise reads accordingly) led to the appearance of two clusters which contain mostly noise reads (Table 3.2, A).
We also see that with the increase of noise content, the fractions of *P. acnes* and *P. aeruginosa* reads included in the same clusters as the noise reads are dropping (Table 3.2, B). In conclusion, the more noise reads were added to the dataset, the better they were grouped together in one or two clusters (Table 3.2, A).

Dataset	5% noise	10% noise		15% noise	
Reads origin	Cluster 2	Cluster 2	Cluster 8	Cluster 6	Cluster 7
A					
noise	21.4	90.3	47.8	85.6	97.3
P. acnes	63.7	0.5	33.8	5.6	0
P. aeruginosa	10.4	1.3	19.1	8.9	0
B					
noise	91.8	55.9	39.9	45.0	50.8
P. acnes	99.6	0.2	22.3	3.6	0
P. aeruginosa	6.4	0.2	5.3	2.3	0

Table 3.2: Composition of clusters containing the majority of noise reads after the classification procedure for three artificial PacBio datasets. A - cluster composition; B - the percentage of reads with particular origin (noise, *P. acnes* or *P. aeruginosa*) included to the cluster within all reads of the same origin in the dataset. Clusters are grouped per dataset. Only organisms whose reads would occupy more than 90% of cluster content are shown.

3.3.2 PacBio sequencing of bioreactor metagenome

After sequencing and correction, we obtained 31,757 reads longer than 1kb for the bioreactor metagenome. The read length distribution for this dataset can be found in Figure 3.4. Reads were mapped to the genomes of the expected metagenome inhabitants. Since the groups of reads that we could map to the genomes of *K. stuttgartiensis* and *B. sinica* had a significant overlap (27%), we decided to combine reads mapped to the reference genomes of these two organisms in one group. We detected almost no (0.01%) reads that would map to the *M. nitroreducens* genome in the sequencing data, suggesting that this organism was either not present in the metagenome sample, or that its DNA could not be isolated reliably during the sample preparation.

Thus, we divided our reads into three groups: uniquely mapped on *M. oxyfera* (4,903 reads), uniquely mapped on *K. stuttgartiensis*/*B. sinica* (2,973 reads), and all remaining reads with unknown origin (75%, 23,881 reads). The reads with unknown origin were checked with the BLASTn software against NCBI microbial database, to find significant similarity to any known organism. However, only 334 reads (less then 2% of total number of checked reads) got hits; there were no organisms among the obtained hits reported more than 53 times.

3.3.3 Bioreactor metagenome PacBio read classification

For the reads originating from *M. oxyfera* and *K. stuttgartiensis*/*B. sinica*, we checked whether the data was clustered by origin. Since roughly 75% of this sequencing data is of unknown origin, we assessed whether the clustering results for reads with unknown origin is robust. To do this, we created five subsets using the bioreactor

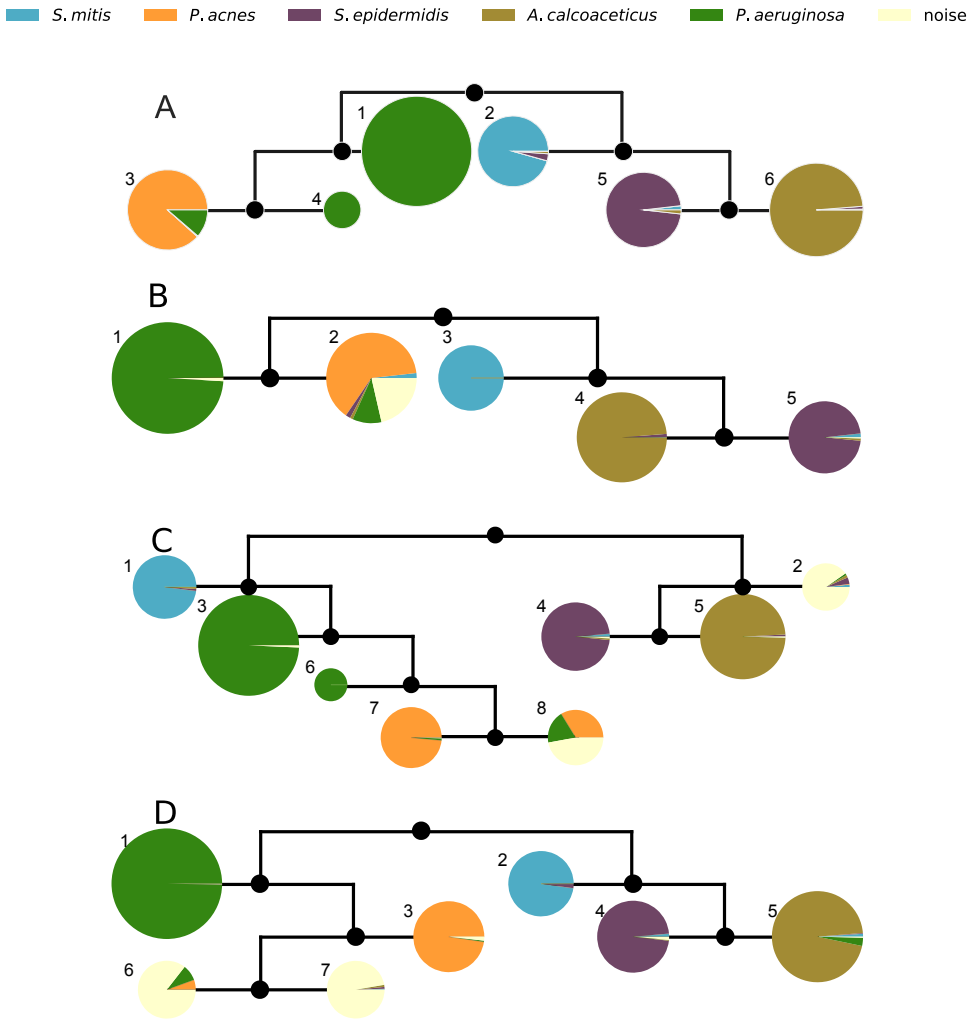

Figure 3.3: Classification recall for artificial PacBio metagenomes. Subsets that were subjected to the partitioning are shown as black circles, final clusters are represented as pie charts with the colour indicating the reads origin. The area of the pie chart corresponds to the relative cluster size. The cluster number is shown next to each pie chart. The results are shown for datasets with 0% (A), 5% (B), 10% (C) and 15% (D) of noise reads.

metagenome sequencing data. Each subset contains 10,000 randomly selected reads with length > 10 kb. After subjecting each subset to the classification procedure, we checked whether reads, shared by two subsets, are being clustered similarly. We compared all clusters from different subsets in a pairwise manner and marked two

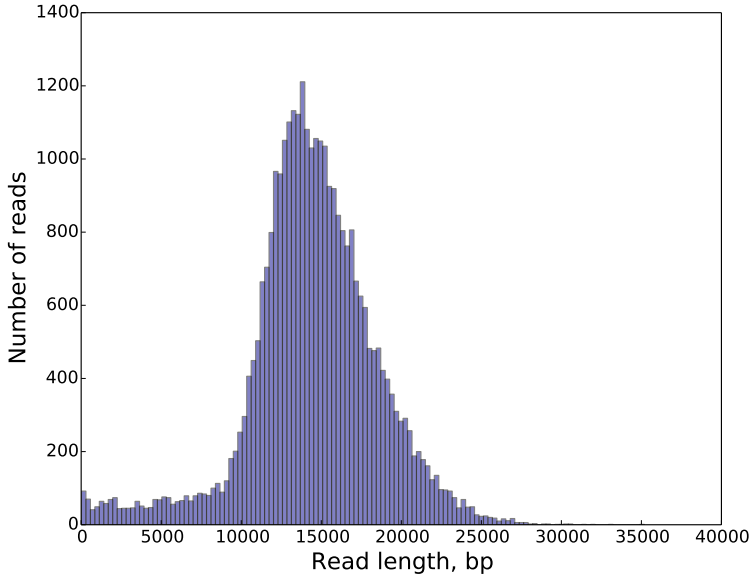

Figure 3.4: Bioreactor metagenome reads length distribution.

clusters 'similar' when they shared at least 25% of their content. On average, every pair of subsets shared 34% of their content. Thus, in case of perfect matching of clustering results, the pair of clusters from two different subsets should on average share 34% of their content. The 25% cutoff value was chosen to compensate for possible flaws introduced by clustering mis-assignments.
In Figure 3.5 this analysis is shown as a graph: each pie chart represents a cluster obtained for one of the subsets (with a subset number marked next to the pie chart). Clusters are connected if they were marked as similar and thus shared more then 25% of their content. We looked for sub-graphs, of size five for which all five nodes would be mutually connected. That would mean that all five clusters are coming from the different subsets and share a significant (at least 25% out of 34% possible) number of reads. These groups of clusters (here and after called the stable groups) represent reads that are clustered the same way regardless of the subset of reads selected. Clusters belonging to the stable groups are called the stable clusters. The proportion of reads in the stable clusters was comparable among datasets and equaled on average 64%. As displayed in Figure 3.5, we found seven groups of stable clusters. Four groups of stable clusters have clusters with more than 1,000 reads, and two of those four are represented by clusters enriched with *M. oxyfera* or *K. stuttgartiensis/B. sinica* reads. In Table 3.3 we display the content and the number of reported clusters after the classification procedure for each of the five subsets.
Once we estimated the robustness of the classification procedure, we selected the

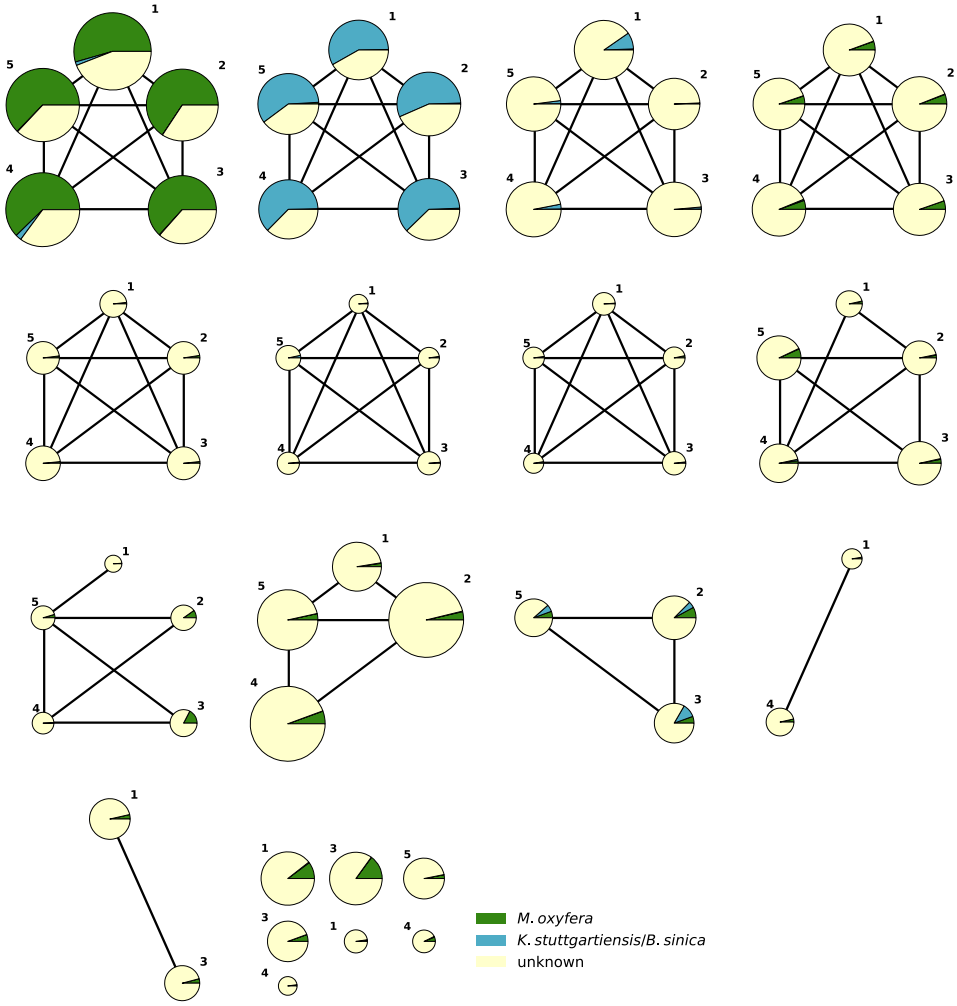

Figure 3.5: Comparison of classification results obtained for five bioreactor sub-datasets. The pie charts represent reported clusters for all sub-datasets coloured by the origin of reads in cluster. The pie chart area indicates the relative size of the cluster. The number next to the node denotes the sub-dataset, for which the cluster was obtained. Two clusters are connected with a node if they belong to two different sub-datasets and share at least 25% of their content. The groups of size five (the set of five fully connected pie-charts) represent groups of stable clusters.

Subset	1	2	3	4	5
Number of *M. oxyfera* reads	1,499	1,563	1,528	1,544	1,529
Number of *K. stuttgartiensis*/*B. sinica* reads	949	918	981	935	906
Clusters after the classification procedure	14	11	13	13	12
Big (>1,000 reads) clusters	5	5	5	5	5
% of reads in stable clusters	65.96	64.12	61.98	64.46	64.16

Table 3.3: Subsets information and clustering results.

subset that yielded the lowest number of clusters (subset 2, 11 clusters) for downstream analysis. The content of all clusters that were not reported as stable were merged into one cluster. Thus, the original 10,000 reads were spread among 8 clusters. These clusters were used as a classifier for the remaining 21,757 reads in the dataset (Table 3.4).

Cluster	Stable	Reads before extension	Reads after extension
1	Yes	403	1,038
2	Yes	168	528
3	Yes	1,133	3,204
4	Yes	1,540	5,151
5	Yes	1,004	3,337
6	Yes	181	506
7	Yes	1,983	6,459
8	No	3,588	11,534

Table 3.4: Results of bioreactor metagenome reads classification

3.3.4 Assembly of the bioreactor metagenome before and after reads binning

We assembled reads belonging to different clusters separately, and compared the resulting contigs with the results of the assembly of the entire dataset. The total number of contigs after assembly of the partitioned dataset was comparable to the amount of contigs obtained from the assembly of the entire dataset (Table 3.5). The same can be said about the total length of contigs and contigs length distributions (see supplementary materials). These results, showing that the database partitioning did not lead to the change of the contigs number or their lengths, can be seen as indirect evidence proving that our *k*-mer based binning of metagenome reads results in species-based clustering.

We compared the assembled contigs obtained for the entire and partitioned datasets to the reference genomes of *M. oxyfera*, *K. stuttgartiensis* and *B. sinica*. Even though

we could successfully map around 9% of the reads to the reference genomes of *K. stuttgartiensis* and *B. sinica*, we did not get contigs that could be mapped to these genomes. However, the contigs assembled from the entire and partitioned datasets did map to *M. oxyfera* genome. Only 91 out of 196 contigs obtained from the entire dataset assembly could be mapped back to the *M. oxyfera* genome covering 54% of its length. For the assembly of the partitioned dataset, 85 contigs were mapped to the genome of *M. oxyfera* in total, covering 52.65% of its length. The vast majority of those contigs (79, covering 51% of the *M. oxyfera* genome length) derived from the assembly of reads belonging to one cluster. Thus, our dataset partitioning binned the majority of contigs according to their origin.

Dataset assembled	Entire dataset	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6	Cl 7	Cl 8
Assembly length, bp	3,251,357	5,438	10,747	380,905	377,792	601,065	0	1,602,878	41,310
Contigs	196	1	1	28	30	47	0	71	2
Contigs mapped on *M. oxyfera* genome	91	0	0	9	1	2	0	71	2
Length of mapped contigs	1,842,182	0	0	132,863	11,945	21,105	0	1,497,132	17,013
M. oxyfera genome covered, %	54	0	0	1.2	0.1	0.15	0	51	0

Table 3.5: Results of entire and partitioned bioreactor sequencing data assembly and comparison of obtained contigs to the *M. oxyfera* genome. Cl - cluster.

3.4 Discussion

We described a new approach for efficient, alignment-free binning of metagenomic sequencing reads based on *k*-mer frequencies. Our method successfully classifies reads per organism of origin, for both simulated and real metagenomic data.
As shown in the results section, the approach was used to classify reads obtained by

PacBio sequencing of a real bioreactor metagenome. The absolute majority of the reads with known origin (*M. oxyfera* or *K. stuttgartiensis*/*B. sinica*) were clustered together per origin after pairwise comparison of their k mer profiles and subsequent density-based cluster detection. This result was robust, as we observed during the analysis of five subsets of the original PacBio sequencing data with overlapping content. The same experiment demonstrated that each subset provides a similar number of clusters. Reads with unknown origin had a tendency to cluster similarly among different subsets, again confirming the clustering consistency. Although the majority of reads in the analysed metagenome was of unknown origin, the results can be used to estimate the microbial community complexity for its most abundant inhabitants.

The binning of the bio-reactor metagenomic dataset had almost no influence on the results of the metagenome assembly. The number of contigs and their lengths obtained for the entire and partitioned datasets were comparable. This indicates that the *k*-mer based reads binning leads to the organism-based partitioning of metagenomic data. Furthermore, contigs, belonging to the same organism, were automatically grouped together when assembling the dataset subjected to the classification procedure. Thus, our *k*-mer based binning technique can be used to interpret metagenomic assembly results.

Performing the binning procedure on an artificially generated PacBio datasets lead to a reads classification per organism, even after adding reads with unknown origin (noise reads). Moreover, increasing the proportion of noise reads leads to a better separation between them and the reads with known origin. This observation supports the c*k*-merentral hypothesis of this research, namely that *k*-mer distances can be used to cluster reads of the same origin together once those reads provide sufficient coverage of the organisms' genome.

The main disadvantages of the current implementation of our method is the limited number of reads (10,000) that can be analysed. As mentioned before, reads, derived from the same organism, will cluster together, but this is possible only under the condition that the organisms' genome is sufficiently covered. Thus, the described technique is unsuitable for the analysis of metagenomes with a large number of inhabitants or when the inhabitants have large genomes, as 10,000 reads will not be enough to provide sufficient coverage. The depth of the classification that can be performed by the suggested method is still to be discovered.

We believe that adapting our metagenomics reads binning technique for larger sets of data and further investigation of its metagenome resolving capacity would allow to expand the current limits of microbiology in the future.

3.5 Author Statements

3.5.1 Funding information

This research is financed by a grant number 727.011.002 of the Netherlands Organisation for Scientific Research (NWO).

3.5.2 Acknowledgements

We would like to thank the group of Prof. Huub Op den Camp for the bioreactor metagenome material, Prof. Boudewijn P. F. Lelieveldt for the idea to perform dimensional reduction using t-SNE, and Martijn Vermaat for the help with coding.

3.5.3 Conflicts of interest

The authors declare that there are no conflicts of interest.

Chapter 4

Determining the quality and complexity of next-generation sequencing data without a reference genome

S. Y. Anvar[1,2], L. Khachatryan[1], M. Vermaat[1], M. van Galen[2], I. Pulyakhina[1], Y. Ariyurek[2], K. Kraaijeveld[2,3], J. T. den Dunnen[1,2,4], P. de Knijff[1], P. A. C. 't Hoen[1], and J. F. J. Laros[1,2]

1 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Leiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands
3 Department of Ecological Science, VU University Amsterdam, Amsterdam, The Netherlands
4 Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands

Genome Biology, 2014 15:555 doi 10.1186/s13059-014-0555-3

4.1 Background

DURING the past decade, DNA sequencing technologies have undergone notable improvements with great impacts on molecular diagnostics and biomedical and biological research. Today, next-generation sequencing (NGS) technologies can provide insights into sequence and structural variations by achieving unprecedented genome and transcriptome coverage. Despite molecular and computational advances, the fast growing developments in library preparation, sequencing chemistry and experimental settings are of concern as they can diversify the complexity and quality of sequencing data [262, 263, 264]. To address data quality, most strategies rely on basic statistics of the raw data, such as the quality scores associated with base calling, the total number of reads and average GC content. Technical artefacts are usually only spotted after mapping of reads to the reference genome. However, such approaches are prone to alignment biases and the loss of potentially valuable information due to the predisposed and incomplete reference genome sequences [265, 266, 267]. These biases are considerably more problematic in studies of microbiomes as the species diversity can be immense [268], whereas the evaluation of data complexity and quality is limited to the analysis of species for which a reference genome sequence is available.

Analysing the *k*-mer (DNA words of length *k*) frequency spectrum of the sequencing data provides a unique perspective on the complexity of the sequenced genomes, with more complex ones showing a greater diversity in unique sequences and repeated structures. Over- and under-represented *k*-mers have been associated with the presence of functional or structural elements (such as repetitive, mobile or regulatory elements), negative selection, or the hypermutability of CpGs [269, 270, 271, 272, 273]. Notably, the prevalence of functional elements and those caused by neutrally evolving DNA (including duplications, insertions, deletions and point mutations) is reflected in the modality (number of peaks) of the *k*-mer frequency spectrum [274, 275]. The modality of the human genome is also subjected to its function as all coding regions, including the 5′ untranslated regions (UTRs), exhibit a unimodal *k*-mer spectrum, while the introns, 3′ UTRs and other intergenic regions have a multimodal distribution [274, 275].

In recent years, *k*-mers have been used in a wide range of applications from the identification of regulatory elements to correction of sequencing errors, genome assembly, phylogeny analysis and the search for homologous regions [276, 277, 278, 279, 280, 281, 282]. It has also been shown that the characterization and comparative analysis of the *k*-mer spectrum can provide an unbiased view of genome size and structure, but it can also expose sequencing errors [283]. However, to our knowledge, most tools fail to accommodate for differences in library size and do not reliably expose problematic samples nor provide information on potential sources of variation in series of sequencing data. Here, we present a method, *k*-mer Profile

Analysis Library (kPAL), for assessing the quality and complexity of sequencing data without requiring any prior information about the reference sequence or the genetic makeup of the sample. The proposed method uses the distance between *k*-mer frequencies to measure the level of dissimilarity within or between *k*-mer profiles. Since most distance measures are susceptible to differences in library size, we have implemented a series of functions that ensure a more reliable assessment of the level of dissimilarity between *k*-mer profiles. Based on the same principle, kPAL can identify problematic samples, as their level of similarity reduces in the absence of a significant difference between the genome of the sequenced samples. In this work, we apply kPAL to four types of NGS data: 665 RNA sequencing (RNA-Seq) samples [284, 285], 49 whole genome sequencing (WGS) samples, 43 whole exome sequencing (WES) samples, and a series of microbiomes. We report the sources of technical and biological variation present in each set of NGS data, highlight a series of artefacts that were missed by standard NGS quality control (QC) tools, and demonstrate how the complexity of microbiomes is reflected in their *k*-mer profiles.

4.2 Materials and Methods

4.2.1 kPAL implementation

kPAL is a Python-based toolkit and programming library that provides various tools, many of which are used in this study. kPAL is an open-source package and can be downloaded [1] [2] [3]. kPAL can also be installed (including all prerequisites) through the command line using: pip install kPAL. Detailed documentation and tutorials are available [4]. For detailed a description of the kPAL methodology, refer to Additional file 1: Notes. The performance of kPAL, in terms of speed and memory usage, for generating and pairwise comparison of *k*-mer profiles is provided in Additional file 1: Figure S18.

4.2.2 Creating *k*-mer profiles

The *k*-mer profiles were generated using the index function built into kPAL. For all analyses *k* was set to 12 except when otherwise stated. To accommodate for the analysis of both sequencing reads and genome reference sequences, we have chosen to use the FASTA format as an input to kPAL. However, we provide a command-line tool to convert FASTQ files to the appropriate format [5]. For paired-end data, the

[1] *k*-mer Profile Analysis Library at GitHub repository https://github.com/LUMC/kPAL

[2] *k*-mer Profile Analysis Library at LUMC repository http://www.lgtc.nl/kPAL]

[3] *k*-mer Profile Analysis Library at official Python repository for open-source packages https://pypi.python.org/pypi/kPAL

[4] Online documentation for *k*-mer Profile Analysis Library http://kPAL.readthedocs.org

[5] Available from: https://git.lumc.nl/j.f.j.laros/fastools.

profiles for both reads were merged into a single *k*-mer profile using the kPAL merge function. For more information on performance, runtime and memory usage, see Additional file 1: Notes.

4.2.3 Measuring pairwise distances

The *matrix* function was used in combination with the *scale* and/or *smooth* options to measure the distance between two *k*-mer profiles. The pairwise distance between profiles was calculated using the multiset distance measure [286]. This measure was parameterized by a function that reflects the distance between two elements in a multiset, in this case the difference between frequencies of specific *k*-mers. The following function was used to calculate the distances after applying the *scale* and *smooth* options.

$$f(x,y) = \frac{|x-y|}{(x+1)(y+1)}$$

For further information about the procedure, refer to Additional file 1: Notes.

4.2.4 Calculating the *k*-mer balance

For all samples in this study, the balance between the frequencies of *k*-mers and their reverse complement were found using the *showbalance* function in kPAL (see Additional file 1: Notes). For all paired-end datasets, *k*-mer profiles were first merged and then assessed for their balance.

4.2.5 Statistical analysis

The distance matrices produced by the pairwise comparison of all samples were used to perform a hierarchical clustering and PCA in R and MATLAB, respectively. The mRNA analysis pipeline, QC and exon quantification procedure are described elsewhere [284, 285]. For the microbiomes, the hierarchical clustering was done using the distance matrices provided by the *k*-mer profile or UniFrac [139] analyses. Subsequently, the accuracy of the clustering arrangement was assessed based on the silhouette [287] and weighted kappa [288] measures.

4.2.6 Library preparation and sequencing

For WGS datasets, two separate library preparation protocols were used. The gDNA libraries for full genome libraries were prepared using the reagents from a True-Seq DNA Sample Prep Kit according to the manufacturers' instructions (TrueSeq DNA

Sample Preparation Guide, revision C; Illumina Inc., San Diego, CA) with minor modifications. After the ligation, the first protocol uses a gel-free method for samples instead of a gel step that was used for the second protocol. Furthermore, the number of PCR cycles in the PCR enrichment step differs between the two protocols (five and ten cycles, respectively). A High Sensitivity DNA chip (Agilent Technologies 2100; Santa Clara, CA) was used for quantification and samples were subsequently sequenced on an Illumina HiSeq 2000 sequencer at the same laboratory.

Libraries for the WES samples were prepared using the Agilent SureSelect Kit (Agilent Technologies, Santa Clara, CA), Nimblegen Capture Kit V2 or Nimblegen Capture Kit V3 (Roche NimbleGen Inc., Madison, WI), according to the manufacturers' instructions. A High Sensitivity DNA chip (Agilent Technologies 2100) was used for the quantification and the samples were subsequently sequenced on an Illumina HiSeq 2000 sequencer at the same laboratory.

The library preparation and sequencing of all RNA-Seq samples are described elsewhere [284, 285].

4.2.7 Pre-processing

FastQC was run for all samples prior to analysis to assess the quality of the data. However, none of the sequencing data was removed from the analysis as they all passed the FastQC quality measures. Reads were trimmed for low quality bases ($Q < 20$) using sickle[6] and cleaned up for adapters.

4.2.8 Alignment

Alignment to the human reference genome was performed for WGS and WES using Stampy [289], BWA [105] and Bowtie 2 [106] with default parameters. For the WES samples, the number of on-target reads was calculated using the BEDTools [290] intersect, BAM files and a BED track consisting of all targets according to the manufacturers' guidelines. Reads with no overlapping base were considered as off target. Basic alignment statistics (such as alignment rate, the fraction of properly paired reads, etc.) were extracted using SAMtools [291] *flagstat*. For WGS samples, the insert sizes were estimated using the Picard toolkit[7]. The number of base pairs that were soft clipped during the alignment was extracted from the SAM files using a custom script.

[6]Sickle: a windowed adaptive trimming tool for FASTQ files using quality https://github.com/najoshi/sickle

[7]Picard: a set of tools for working with next-generation sequencing data in the BAM format http://broadinstitute.github.io/picard/

4.2.9 SGA

Comparison QC and exploration of data properties were performed using the *Preqc* module of the SGA software. All analyses were performed according to SGA guidelines [283].

4.2.10 Data availability

For the WGS and WES data, the FASTQ and BAM files have been deposited at the European Genome-phenome Archive[8], which is hosted by the European Bioinformatics Institute, under the accession number [EGA:S00001000600]. In addition, all *k*-mer profiles are available under the same accession. For the RNA-Seq data, the *k*-mer profiles can be found online[9]. The FASTQ files and BAM alignments as well as different types of quantification are available in Array Express under accessions E-GEUV-1 (mRNA) and E-GEUV-2 (small RNA) for QC-passed samples and E-GEUV-3 for all sequenced samples[10] [11] [12]. Microbiomes were obtained from the 'Moving Pictures of the Human Microbiome' project [MG-RAST:4457768.3-4459735.3] [292].

[8] http://www.ebi.ac.uk/ega/
[9] http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-3/files/profiles/?ref=E-GEUV-3
[10] http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/
[11] http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-2/
[12] http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-3/

4.3 Results and Discussion

4.3.1 Principles of kPAL

We developed an open-source package kPAL, which provides a series of tools (such as distance calculation, smoothing and balancing) to investigate the spectrum of *k*-mers observed in a given NGS dataset (Figure 4.1 A and Additional file 1: Notes). The resulting *k*-mer profile holds valuable information on the complexity of the sequencing libraries and the sequenced genome(s). This is delineated in a graphical representation of the *k*-mer profiles, which plots the number of *k*-mers observed at each frequency. The complexity of genomic information is often reflected in the modality of this distribution, mainly due to repetitive and structural elements, and the context-specific composition of *k*-mers [271, 274, 275, 293]. First, *k*-mers are processed using efficient binary codes that facilitate a rapid reverse complement conversion and access to specific *k*-mers (Figure 4.1 B). Next, kPAL uses the distance between *k*-mer frequencies as a measure of dissimilarity between two *k*-mer profiles. In addition, calculating the correspondence between the frequencies of *k*-mers and their reverse complements aids in assessing the coverage balance between two strands of the sequenced library (Figure 4.1 C). Generally, *k*-mer profiles can be shrunk to a smaller *k* size using the *shrink* function to enable access to smaller *k*-mer profiles without the need to reprocess the sequencing data (Figure 4.1 D). However, it is important to note that large deviations from the original *k* size may obscure the true *k*-mer frequencies due to limited access to both ends of the sequencing reads (i.e., the last 12 nucleotides can be processed only once in a 12-mer profile whereas the same information is processed seven times in a 6-mer profile). To facilitate pairwise comparison of *k*-mer profiles and account for differences in library sizes, we have implemented complementary *scaling* and *smoothing* functions. Scaling *k*-mer frequencies to match the area under the curve of two profiles is a global normalization of the *k*-mer profiles. The smoothing function borrows the utility of shrinking and applies it locally to *k*-mers that have a frequency lower than a user-defined threshold, which results in local collapsing of those *k*-mers to a smaller size (i.e., *k*-1) until the threshold condition is met (Figure 4.1 E). For more information and a detailed explanation of kPAL features, see Additional file 1: Notes.

4.3.2 Setting *k* size

To identify which *k* provides the best specificity for a mixed sample of bacteria, the *k*-mer profiles from three modelled metagenomes consisting of 30 bacterial genomes from the Firmicutes and Proteobacteria phyla (in 100:0, 50:50 and 0:100 ratios from each phylum) were compared to ten randomly shuffled sequences (without changing the overall nucleotide composition). The optimal value for *k* is the one that best separates metagenomes from randomly permuted sets. The overall distance

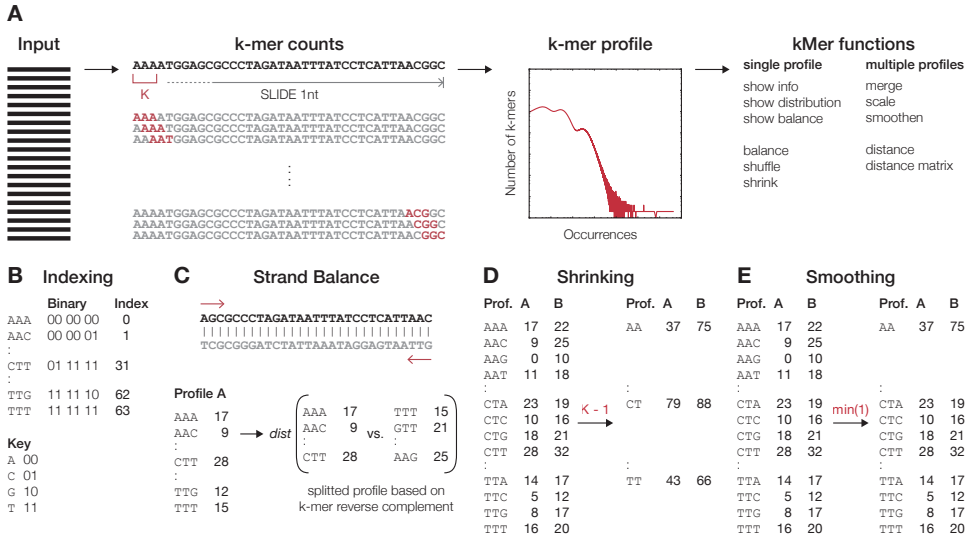

Figure 4.1: Schematic overview of main kPAL principles. (A) An overview of the procedure used by kPAL to assess the frequency of all *k*-mers within sequencing data. *k*-mers are identified and counted by a sliding window of size *k*. The *k*-mer spectrum can then be produced using the *k*-mer frequencies. The main functions of kPAL can be divided by their application to single or multiple profiles. For single *k*-mer profiles, general information about the number of nullomers, total number of counts, distribution of *k*-mer counts and balance between sequencing information from the plus and minus strands can be obtained with dedicated functions. If needed, profiles can be manipulated by the balance, shuffle and shrink functions. The balance function uses a sum of *k*-mers and their reverse complements to enforce balance between sequence information from the minus or plus strand. The shuffle function is designed to produce random *k*-mer profiles without changing the overall distribution of counts. (B) kPAL efficiently processes *k*-mers, as it encodes the sequences with a binary code using specific keys that can also facilitate a quick conversion to the reverse complement. Each nucleotide is represented by a binary code that is subsequently used to construct each *k*-mer. (C) The strand balance of a given *k*-mer profile is the overall distance measure between the frequency of the unique *k*-mer and its reverse complement. Thus, *k*-mer profiles are split into two sub-profiles that are reverse complements of each other and these are used to calculate the strand balance. (D) By design, kPAL can shrink *k*-mer profiles of size *k* to any smaller size. Counts from *k*-mers that share the first (n - 1) nucleotides are merged to collapse *k*-mer profiles to a size *k* - 1. (E) The smoothing function borrows the utility of shrinking and applies it locally to only *k*-mers that have lower counts than one defined by the user. Thus, for those affected, *k*-mer counts are merged and dropped to the size *k* - 1. The smoothing function accepts thresholds for the minimum, maximum or average counts of *k*-mers that share the first (n - 1) nucleotides but it also accepts user-defined functions. This process reiterates until the threshold condition is met. Prof., profile.

between *k*-mer profiles of the metagenomes and the corresponding randomly permuted sets starts to level off once *k* exceeds 10 (Additional file 1: Figure S1). A low amount of variation in distance between the *k*-mer profiles of metagenomes and their permuted sets indicates that the distance measure is generally robust and only changes according to *k*. Interestingly, the optimal separation coincides with the *k* for which the complete unimodal spectrum of frequencies (from those that are too rare to those that are highly recurrent) is observed (Additional file 1: Figure S2 A,B,C).

The human reference genome has a high complexity (described in Additional file

1: Notes), based on the multimodality of the *k*-mer profiles, which ranges from 9 to 15 (Additional file 1: Figure S3 A). In humans, $k = 11$ is the smallest value for which unique *k*-mers and nullomers (absent *k*-mers) are observed while genomic spectra for $k \geq 13$ start to lose their multimodality as they become too unique. Thus, $k = 12$ was used to give a relatively balanced number of nullomers, and unique and frequent *k*-mers. This allows for the identification of potential artefacts (mainly reflected by rare *k*-mers) as well as biological and contextual variations. Interestingly, the level of complexity varies between different types of genomic information (WGS, WES and RNA-Seq; see Additional file 1: Figure S3 B). In contrast to genomic sequences, the coding part of the human genome exhibits a unimodal profile, as shown before [274, 275]. The minor differences between the *k*-mer profiles of the exome and the transcriptome reference sequences are due to the number of shared coding regions between different transcript variants of the same gene. The transcriptome reference sequences generally exhibit higher counts for observed *k*-mers and lower numbers of nullomers introduced by exon-exon junctions. Moreover, the *k*-mer spectrum derived from sequencing data is in concordance with that of the reference (Additional file 1: Figure S3 C). The minor deviations from the unimodality of the exome and transcriptome data are mainly due to the capture performance (off-target reads introduce low-count *k*-mers that represent intronic and intergenic regions) and differences in the abundance of expressed mRNA.

In addition to the complexity of the genomic information, the sequencing depth contributes to the modality and the resolution of the *k*-mer spectrum derived from individual datasets. In RNA-Seq, we observed that the number of 12-nullomers correlates with the total number of reads per dataset (R = -0.80; see Additional file 1: Figure S4 A,B). The variation in the total read counts per sample is partly due to study design, as sequencing was performed in seven different laboratories [285]. Thus, the total number of 12-nullomers also varies between samples from different laboratories (Additional file 1: Figure S4 C). It is crucial to account for bias introduced by poor and variable coverage, as it may obscure the identification of factors that determine the complexity of the *k*-mer spectrum. One obvious solution would be to opt for lower *k* sizes (i.e., $k = 9$) at the expense of specificity. However, we propose the dynamic smoothing function, which is resilient towards coverage bias and does not sacrifice the specificity of the *k*-mer spectrum by choosing a smaller *k* (Additional file 1: Notes). This function only shrinks the *k*-mer profile locally when the counts do not pass predefined conditions (i.e., they fall below an acceptable threshold for *k*-mer frequencies). In the next section, we show how kPAL can be used to assess the quality of different types of sequencing data without relying on the availability of a well-characterized reference genome.

4.3.3 Evaluating data quality without a reference

Recently, we showed that performing a pairwise comparison of 9-mer (K9) profiles, without alignment to the reference sequence, can expose quality issues in RNASeq data [285]. The median of all pairwise distances for each sample correlated (R = -0.63) with the correlation measures obtained after alignment and quantification of exon expression levels, which are post-alignment measures often used for QC. Notably, some of the problematic samples (due to a high duplication rate and/or high rRNA content) could only be identified by an analysis of their *k*-mer profiles. However, kPAL scores could not separate all problematic samples. Thus, we performed these analyses for larger values of *k* to increase the specificity and investigate whether smoothing can remove biases introduced by variable sequencing depth between samples. For 12-mer (K12) profiles, the distance measures calculated after scaling only showed a much weaker correlation (R = -0.34) with the correlation measures obtained from the exon quantification of samples (Figure 4.2 A). They also displayed a broad distribution with no apparent clustering of known outliers (Figure 4.2 B). We also observed a variation between samples based on the laboratory in which the sequencing was performed, mainly reflecting the library size differences (Figure 4.2 C and Additional file 1: Figure S5 A). After smoothing the *k*-mer profiles, the *k*-mer pairwise distances were in good concordance (R = -0.62) with the correlation measures of the exon quantifications obtained after alignment (Figure 4.2 D). Smoothed K12 profiles exhibited a narrow distribution, having known problematic samples as only outliers (Figure 4.2 E). Importantly, the variation between laboratories was significantly reduced as the dynamic smoothing function can accommodate differences in library size (Figure 4.2 F and Additional file 1: Figure S5 B). These median pairwise distances were far less sensitive to differences in the total read counts per sample than distances obtained from scaled 9-mer and 12-mer profiles (R = -0.33, -0.67 and -0.83, respectively; Figure 4.2 G,H,I). Moreover, the number of known problematic samples that fall outside the 95% prediction bounds is improved to 11 (out of 12) in smoothed K12 distances compared to that of K9 and K12 (eight and five, respectively). The sample NA18861.4 has by far the highest distance to other samples in both K9 and smoothed K12 analyses (Figure 4.2 G,I). We have previously reported that this sample has a significant genomic DNA contamination since only 4% of reads mapped to exons [285]. This contamination can affect the complexity of the sequenced library as many reads represent the non-coding and repetitive regions of the genome. Whereas samples that passed the QC measures exhibited *k*-mer spectra that reflected the expected modality of the transcriptome (Additional file 1: Figure S6 A), the distribution of *k*-mer frequencies in NA18861.4 clearly mimicked that of the full human reference genome (Additional file 1: Figure S6 B).

We also addressed quality issues in WGS data. In our set of 49 WGS samples from nine individuals, pairwise distances between smoothed 12-mers clustered samples

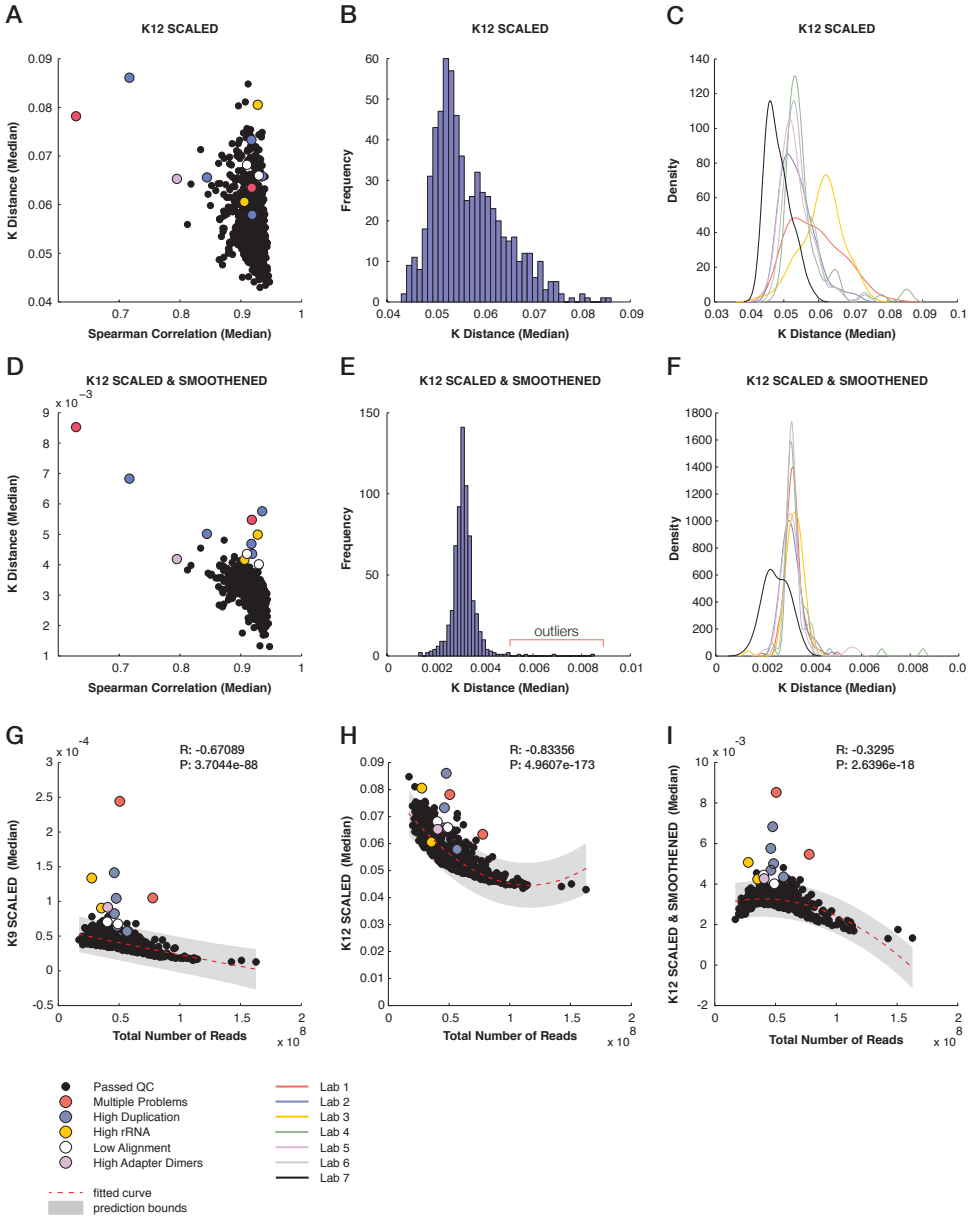

Figure 4.2: See the legend on the next page

Figure 4.2: See the figure on the previous page. Evaluating data quality for mRNA sequencing samples across different laboratories. (A) Scatter plot showing for each sample the median pairwise Spearman correlation for exon quantification and the median *k*-mer distance measures (K distance) after scaling. Problematic samples are highlighted in different colours. (B) Histogram of median K distance (scaled) for each individual sample. (C) Distribution of median K distance (scaled) for each sequencing laboratory (indicated by different colours). (D) Scatter plot of median pairwise Spearman correlation between exon quantification and K distance (smoothed and scaled). (E) Histogram of median K distance (smoothed and scaled) for each individual sample. (F) Distribution of median K distance (smoothed and scaled) for each sequencing laboratory (indicated by different colours). (G) Scatter plot of the total number of reads per sample versus the K distance of 9-mers (scaled). The poly2 fitted line and the 95% confidence intervals are indicated. (H) Scatter plot of the total number of reads per sample versus the K distance of 12-mers (scaled). (I) Scatter plot of the total number of reads per sample versus the K distance of 12-mers (smoothed and scaled). Lab, laboratory; QC, quality control.

into two main groups that represent the choice of the library preparation protocol (Figure 4.3 A). Within the cluster representing the first protocol, most datasets were further clustered on the individuals from whom the samples were obtained. Importantly, all datasets passed all the quality measures in the commonly used QC pipeline for NGS data, FastQC[13] . The alignment (99.7%), duplication rates (2.0%) and the overall GC content did not differ significantly between datasets (Figure 4.3 B,C,F). However, datasets differed in the percentage of properly paired reads (86.7% and 95.8%) and pairs mapping to different chromosomes (10.6% and 2.1% for protocol 1 and protocol 2, respectively) based on the choice of library preparation protocol (Figure 4.3 D,E). Pairs that mapped to different chromosomes did not cluster at specific loci but were distributed across the entire genome (Additional file 1: Figure S7). Moreover, the sequencing reads from the first protocol exhibited a bimodal and broader insert size distribution (Figure 4.3 G and Additional file 1: Figure S8 B). The enrichment of pairs that map to different chromosomes and the widening of the insert size distribution could indicate the presence of library chimeras (sequences derived from two or more different fragments). The number of soft clipping events (unmatched region of a partially aligned read, up to 80 base pairs long) during the alignment confirms the enrichment of library chimeras in samples that were prepared using the first protocol (Figure 4.3 H). We ruled out the influence of aligner as the results obtained from three different aligners (Stampy, BWA and Bowtie2) were in concordance (Additional file 1: Figure S9 A,B).

Library chimeras and erroneous bases can potentially introduce artificial *k*-mers and therefore enrich for rare features in the *k*-mer spectrum. This is supported by the *k*-mer profiles of the samples from the two library preparation protocols (Additional file 1: Figure S10). These artefacts can be detrimental to downstream analysis as the sequencing library partially represents artificial fragments.

In WES datasets, we identified four clusters after applying principal component analysis (PCA) on the distances obtained from a pairwise comparison of smoothed 12-mers (Figure 4.4 A). Principal component 1 (PC1) separated samples based on the rate of on-target reads (reads that map to the exons for which probes were designed).

[13] Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc

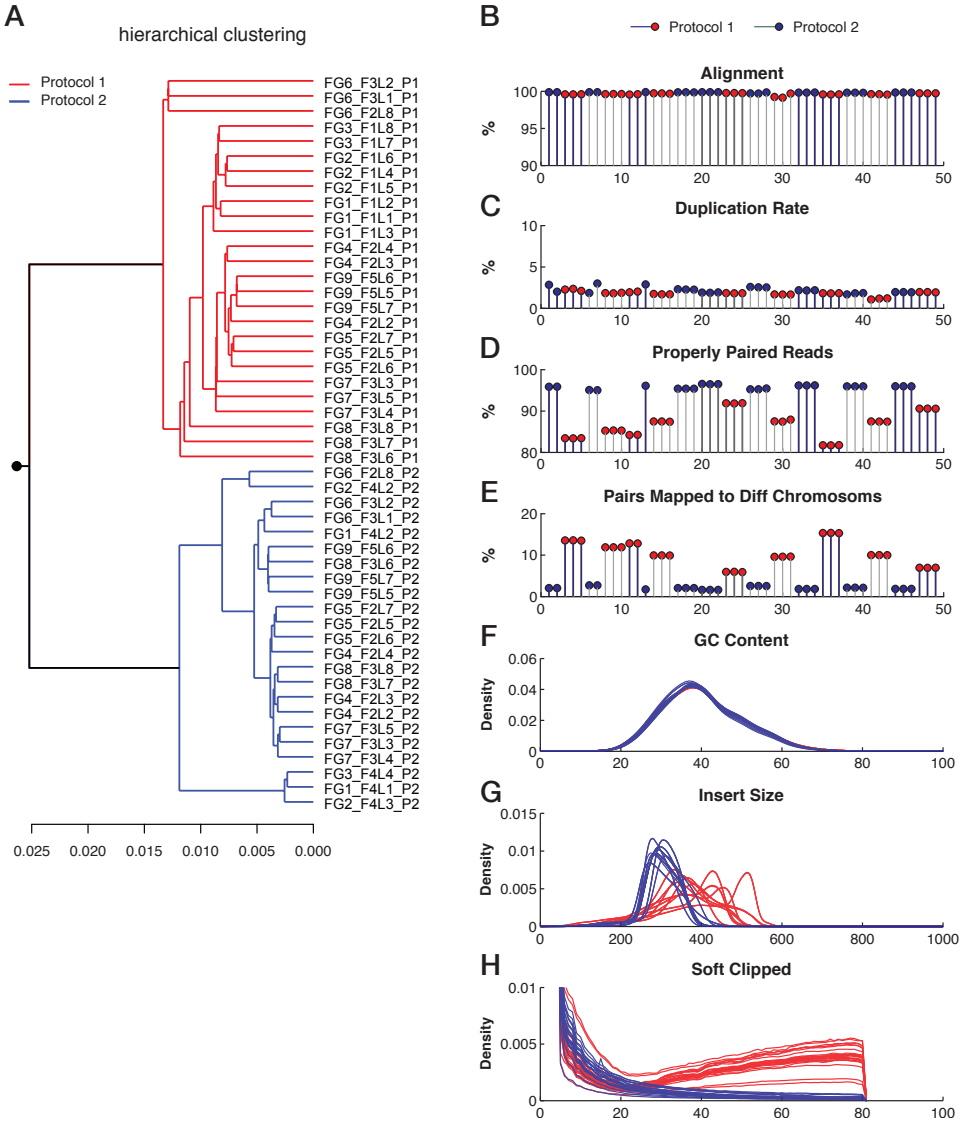

Figure 4.3: Data quality and the influence of library preparation protocol in whole genome sequencing data. (A) Hierarchical clustering of pairwise *k*-mer distance measures across WGS samples. Samples prepared using different protocols are indicated in different colours. (B) Percentage of aligned reads per sample. Black and grey bars separate samples from different individuals. Red and blue circles indicate the choice of library preparation protocol. (C) Percentage of duplicated reads. (D) Percentage of properly paired reads. (E) Percentage of paired reads that map to different chromosomes. (F) Distribution of average GC content per read. Samples prepared using different protocols are coloured accordingly. (G) Distribution of estimated insert size. (H) Distribution of the number of base pairs that are soft clipped from reads during the alignment. Diff, different; WGS, whole genome sequencing.

The low level of reads on target is the result of poor capture performance and not of low sequencing depth (Additional file 1: Figure S11 A,B). Interestingly, PC2 separates the successful WES datasets (69.9% ontarget reads, on average) based on the type of capture kit (Agilent or Nimblegen) that was used during the library preparation (Figure 4.4 A). The third principal component separates out a single failed dataset, WE10_F1L3_NIM. This dataset has multiple problems since the rate of ontarget reads is only 3.7% and the duplication rate is as high as 80%. The extreme level of duplication significantly affects the balance of coverage on the plus and minus strands of the reference genome. Therefore, the *k*-mer profile remains imbalanced since most *k*-mers and their reverse complements have different frequencies. While the hierarchical clustering concords with that of PCA, we observed another sub-clustering among failed samples in which samples with only 11.3% of reads on target were separated from those that exhibit an on-target rate of 49.8% (Figure 4.4 B). The influence of poor capture performance on *k*-mer profiles is evident from the *k*-mer frequency distributions, as those with poor capture performance begin to mimic that of the full genome (Additional file 1: Figure S12 A,B), due to an increase in the number of off-target reads. The multimodality of these spectra is the result of off-target reads that map to noncoding and repetitive regions [274]. Notably, samples that passed QC could be separated by the capture kit used during library preparation as a result of differences between the targeted regions of capture kits (Additional file 1: Figure S12 C).

The analysis of balance between the frequency of *k*-mers and their reverse complement can expose library biases and provide a measure for estimating an optimal sequencing depth to ensurebioreac comparable and sufficient coverage on both strands (Additional file 1: Notes). In human WGS datasets, the balance curve begins to level off as datasets exceed 400 million reads, which represents an approximately 12-times coverage of an entire human genome (Figure 4.5 A). Although the balance curve did not saturate in our WES set, we picked up WE10_F1L3_NIM as an outlier since the expected balance distance is roughly 0.015 for datasets with a comparable number of reads (Figure 4.5 B). This sample suffers from multiple problems. However, its extreme level of duplications (80%) contributes to the imbalanced coverage on the plus and minus strands (Additional file 1: Figure S13). In the RNASeq set, the change in balance begins to level off at the 140 million reads mark (Figure 4.5 C). Of course, this approach will not hold for strand-specific RNA-Seq runs. These data can now be used to assess whether an independent sequencing run has the expected balance distance and, thus, whether sufficient sequencing depth has been achieved.

4.3.4 Comparative analysis of kPAL performance

We benchmarked the performance of kPAL in the identification of problematic samples by comparing the QC analysis of kPAL on a subset of WGS, WES and RNASeq samples with results from the *Preqc* function of the recently developed

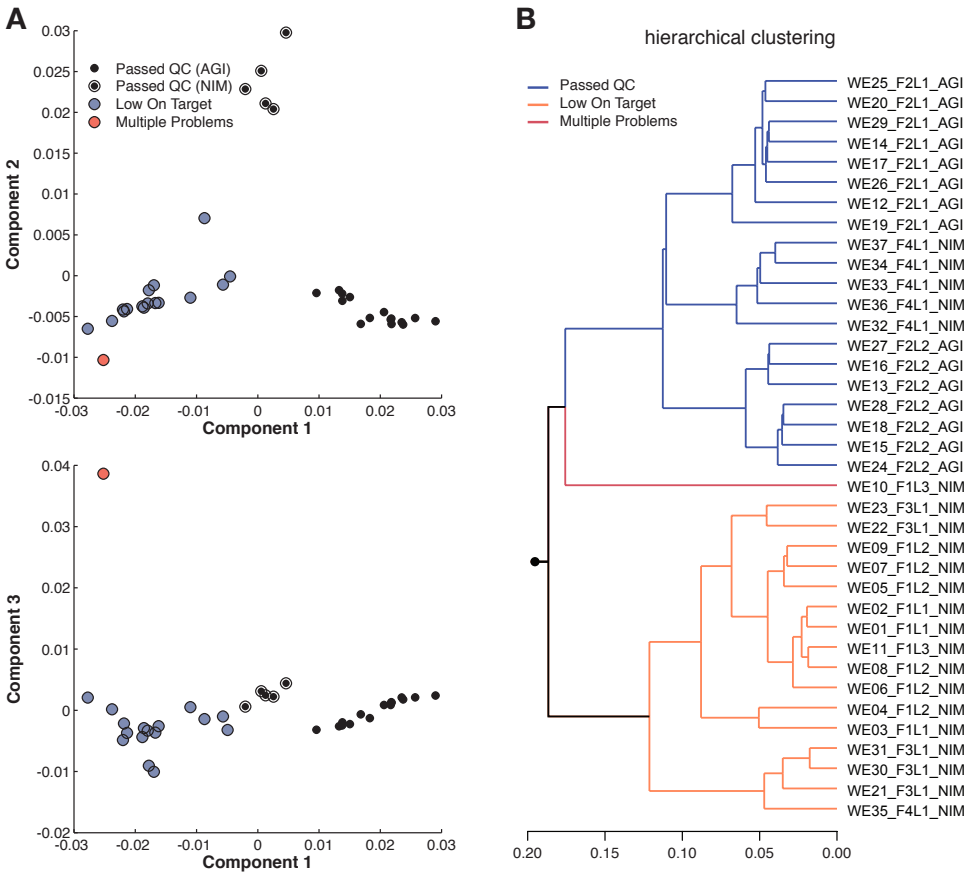

Figure 4.4: *k*-mer distances in whole exome sequencing data are associated with data quality and choice of capture protocol. (A) PCA of pairwise distance measures. Blue circles indicate samples with poor capture performance. The red circles highlight the WE10_F1L3_NIM sample, which suffers from multiple problems. Samples that passed the QC measures are indicated by different types of black circle based on the choice of capture kit (Nimblegen or Agilent SureSelect). (B) Hierarchical clustering of pairwise *k*-mer distance measures across WES samples. Different clusters are indicated by colour. AGI, Agilent SureSelect; NIM, Nimblegen; PCA, principal component analysis; QC, quality control; WES, whole exome sequencing.

k-mer based String Graph Assembler (SGA) [283]. SGA can estimate genome size, insert size distribution, repeat content and heterozygosity of a sequenced genome as well as the error rate and its potential consequence in *de novo* assembly. Unlike kPAL, SGA does not perform a pairwise comparison between *k*-mer profiles obtained from multiple datasets. Thus, we compared SGAs' performance to that of kPAL based on the identification of known problematic samples, using SGAs' estimated genome size, fragment size distribution and the overall error rate. A further evaluation of

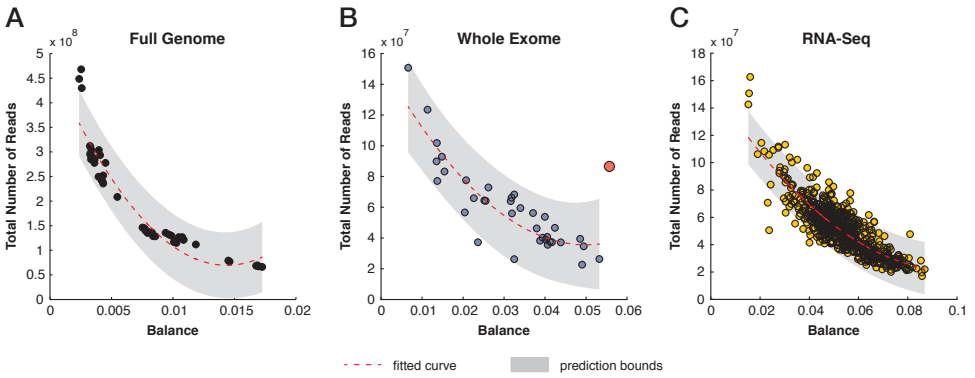

Figure 4.5: Detecting the balance in coverage depth of plus and minus strands in sequencing data. (A) Scatter plot of distance between the frequencies of *k*-mers and their reverse complement (balance) versus the total number of reads in WGS data. The poly2 fitted line and the 95% confidence intervals are indicated. (B) Scatter plot of balance versus the total number of reads in WES data. The red circle indicates an outlier with an extreme duplication rate and imbalance of coverage between the plus and minus strands. (C) Scatter plot of balance versus the total number of reads in RNA-Seq data. RNA-Seq, RNA sequencing; WES, whole exome sequencing; WGS, whole genome sequencing.

SGA on the selected datasets is presented in Additional file 1: Figures S14-S17.
In WGS data from the first sample (FG1), SGA confirmed the bimodal insert size distribution of libraries that were prepared based on the first protocol (Additional file 1: Figure S15). Moreover, sequencing data from the two library preparation protocols could be separated based on the position of the first occurring sequencing errors (Additional file 1: Figure S14 A). This is in concordance with kPAL results and the presence of a higher level of library chimeras that led to the introduction of artificial and rare *k*-mers.
The selected WES data consists of two samples with failed capture (WE01_F1L1_NIM and WE02_F1L1_NIM), one sample with multiple problems (WE10_F1L3_NIM), and four samples with acceptable quality that were prepared using Agilen or Nimblegen capture kits (WE13_F2L2_AGI, WE14_F2L1_AGI, WE36_F4L1_NIM and WE37_F4L1_NIM). SGA identified the problematic sample WE10_F1L1_NIM, which suffers from an extremely high duplication rate and a very low number of on-target reads (Additional file 1: Figure S14 B). The estimated genome size or duplication rate did not further assist in identifying problematic samples and the position of the first sequencing error seems to be obscured by the low coverage of off-target reads that may resemble erroneous sequences. Together, identification of problematic samples by SGA is less reliable for WES data than whole genome shotgun sequences.
For RNA-Seq data, we selected two samples that passed all quality measures (HG00096.1 and HG00108.7) and four failed samples with different underlying problems (HG00329.5: high duplication; NA12546.1: high rRNA; NA18858.1: poor alignment and NA18861.4: high genomic DNA contamination). SGAs' genome size

estimation is designed for WGS data and, therefore, applying SGA on RNA-Seq data should provide an estimate of the expressed part of the genome. Genomic DNA contamination artificially increases the expressed part of the genome and allowed SGA to identify NA18861.4 as a problematic sample (Additional file 1: Figure S14 C). SGA could not reliably identify HG00329.5 as a sample with an exceptionally high duplication rate (Additional file 1: Figure S14 C). Unlike kPAL, the SGA analysis could not identify the other problematic RNA-Seq samples.

4.3.5 Detecting data complexity

The complexity of sequencing libraries is reflected in the *k*-mer spectrum as *k* frequencies often represent functional or structural elements of the associated genome. For metagenomes, the abundance of different bacteria diversifies the frequency of *k*-mers, which can be used to differentiate microbiome communities. To investigate the application of kPAL in the comparative analysis of microbiomes, we first simulated a series of metagenomes with different copy number for three closely related bacterial genomes: *Bifidobacterium animalis* subspecies *lactis* (NC_017834.1), *Bifidobacterium animalis* subspecies *animalis* (NC_017867.1) and *Bifidobacterium adolescentis* (NC_008618.1). The selected genomes have a comparable genome size of approximately 2 Mbp. The level of homology between *Bifidobacterium animalis* subspecies *lactis* and *Bifidobacterium animalis* subspecies *animalis* is estimated to be between 85% and 95% [294]. The genomes of these bacteria are represented in copies of 6:0:0, 3:3:0 and 2:2:2. The distances from a pairwise comparison of 10-mer profiles show an interesting pattern (Figure 6 A). Within the three-dimensional space of individual species, datasets with six copies of a single genome lie within a main triangular space bounded by the absolute minimum distance to their corresponding species. The second triangular space holds datasets that have three copies of two genomes while the dataset with two copies of all genomes sits in the middle of the three-dimensional space (Figure 4.6 A). The relatedness of these datasets relies on the number of rare *k*-mer s that could differentiate the abundance of different species within each set.

Next, we explored the capability of kPAL in resolving the composition of a more complex series of simulated metagenomes. Without considering the phylogeny, 30 bacterial genomes were selected from both the Firmicutes and Proteobacteria phyla and used to construct 31 datasets where the first set comprises 30 genomes from the Firmicutes phylum. The sequence content of each set was subsequently shifted to the Proteobacteria phylum by single genome substitutions (Additional file 1: Table S2). Thus, the 31st dataset consists of 30 genomes from only the Proteobacteria phylum. After performing the pairwise distance comparison on 10-mer profiles, datasets were plotted based on their distance to each phylum (Figure 4.6 B). Notably, the order of the datasets concords with the number of genomes from each phylum. Although the modelled metagenomes do not reflect the true relative abundance of

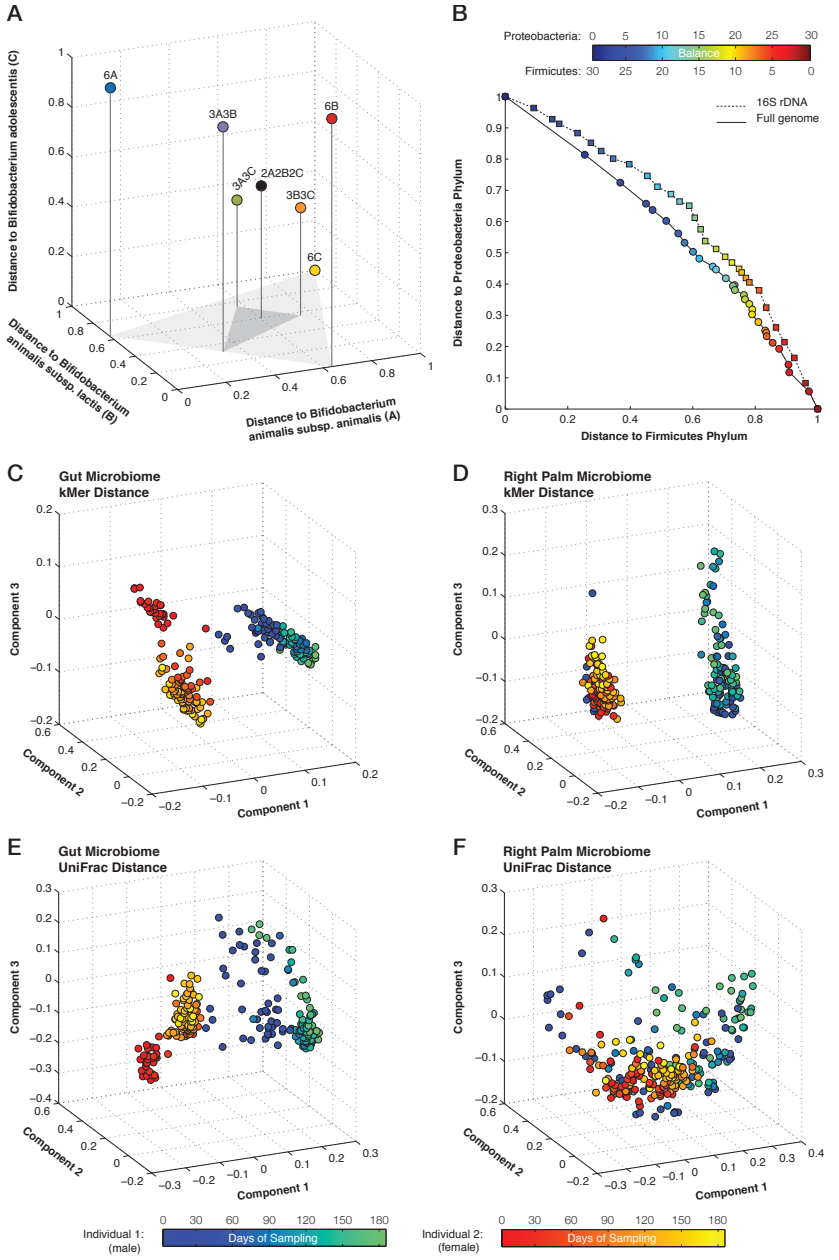

Figure 4.6: See the legend on the next page

Figure 4.6: See the figure on the previous page. Resolving the level of relatedness between microbiomes. (A) Three-dimensional scatter plot of the *k*-mer distance measures for a series of metagenomes with different copy number of three closely related species. (B) Scatter plot *k*-merof the relative distance between Firmicutes and Proteobacteria phyla. Each data point represents a metagenome with a differing number of species from each phylum. Data points are colored according to the number of species from each phylum. (C) PCA plot of pairwise *k* distance measures for gut microbiomes. Data points are colored based on the origin of the sample (male in blue and female in red) and time. (D) PCA plot of pairwise *k*-mer distance measures for right-palm microbiomes. (E) PCA plot of pairwise UniFrac distance measures for gut microbiomes. (F) PCA plot of pairwise UniFrac distance measures for right-palm microbiomes. PCA, principal component analysis.

these bacteria, they allow us to assess whether kPAL can resolve the level of similarity between a series of modelled metagenomes. Distances between *k*-mer profiles generated on the 16S rDNA also confirm the relative similarity of datasets with a slightly smoother transition. This is mainly due to the limited amount of genomic information that is available in 16S rDNA and different rate of evolution compared to the entire genome.

We used the previously pcolourublished data by Caporaso *et al.* [292] to evaluate further the performance of kPAL in resolving microbiomes. The gut and right-palm microbiomes of a male individual and a female individual were sequenced over a period of 6 months. For this analysis, we only included samples that were collected on the same day from both individuals (122 gut microbiomes and 128 right-palm microbiomes). Furthermore, we also excluded 14 samples that were classified as being mislabeled using a random forest classifier as described by Caporaso *et al.* [292]. Pairwise distances were calculated for samples from each body part using kPAL (using 10- mer profiles) and UniFrac [139], which relies on the characterization of operational taxonomical units and inferred phylogeny. UniFrac parameters were set to those specified in the original paper [292]. The agreement between the expected clusters (based on the origin of samples) and that obtained from distance matrices was estimated using the weighted kappa index (Kw). PCA analysis of *k*-mer distance matrices from gut (Figure 4.6 C) and right-palm (Figure 4.6 D) microbiomes revealed that samples from each individual could be separated using the kPAL approach (Kw = 0.95 and 0.82, respectively). In addition, PC2 and PC3 indicate that temporal changes in the microbiomes of each individual influence the relative distances between datasets. We also noticed that datasets from the first 12 days of right palm microbiomes from the male individual cluster with female samples. This can be caused by possible contamination or sample swapping. Gut microbiomes could also be resolved using UniFrac (Figure 4.6 E), with Kw = 0.94. Concordant to the kPAL results, PC2 and PC3 jointly order samples based on the sampling day. However, UniFrac failed to differentiate right-palm microbiomes based on their origin (Kw = 0.47) with no apparent pattern corresponding to the day on which samples were collected (Figure 4.6 F).

4.4 Conclusions

The continued decrease in sequencing costs and technological development have overtaken our ability to assess the quality of data and the complexity of sequencing libraries robustly. For instance, many QC steps that are essential for accurate downstream analysis of NGS data are often neglected in the absence of a reliable reference genome. In addition, NGS data are always subjected to some degree of technical and run-to-run variation, which can hamper the interpretation of the genetic makeup of the sequenced sample. As shown here, variations introduced during library preparation can have a significant influence on the complexity and quality of the sequencing data.
So far, *k*-mer profiles have been used in a wide range of applications, such as the identification of regulatory elements, error correction of sequencing reads, identification of point mutations, whole genome assembly, searches for homologous regions and phylogenetic analysis [276, 277, 278, 279, 280, 281, 282, 295, 296]. A number of *k*-mer analysis tools are capable of efficiently generating *k*-mer profiles (such as Jellyfish [297] and khmer [298]), and the recent work of Simpson [283] proposes a novel method to estimate the repeat content, genome size, heterozygosity of the sequenced genome, insert size distribution and estimated level of erroneous reads in sequencing data using a *k*-mer approach.
Although SGA provides valuable information on the genetic makeup and quality of sequencing data, it cannot reliably identify outliers from a series of NGS data or provide information on potential sources of variation. Thus, in the absence of a well characterized reference sequence, there is an urgent need for tools that can characterize potential biases such as sample swapping, library chimeras, high duplication rates and potential contamination.

In this work, we introduce a new strategy for determining the quality and complexity of a variety of different NGS datasets without any prior information about the reference sequence. The kPAL package consists of a variety of tools to generate *k*-mer frequencies and enables pairwise comparisons. kPAL measures the level of similarity between multiple NGS datasets, based on the genomic information that is shared between them.
We show that kPAL outperforms pre-alignment QC tools (such as FastQC) in reliably exposing samples that suffer from poor capture performance, contamination, enrichment of library chimeras or other types of artefact. Even though the last step in assessing data quality by FastQC involves the analysis of overrepresented 5-mers, FastQC fails to identify problematic samples due to the low *k*-mer size and the way *k*-mer profiles are processed. In contrast, tools that rely on aligned reads (such as RNASeQC [299] and the Picard toolkit) can expose the majority of these technical artefacts, though some of them still require a thorough and vigorous assessment to be identified. The *Preqc* feature of SGA performs well on WGS data and can

precisely estimate insert size distribution and expose erroneous reads. However, the performance of SGA on other types of NGS data, such as WES and RNA-Seq, is less reliable since it was originally developed for pre-processing, error correction and *de novo* assembly of whole genome sequences. The lack of a pairwise comparison and accommodation for differences in library size limits the application of SGA in quality assessment and measuring the level of dissimilarity between *k*-mer profiles of sequenced samples.

The unique feature of kPAL is its ability to account for biases introduced by differences in sequencing depth between samples to expose outliers and problematic samples and that, like SGA, it does not rely on prior information. Potential applications of this strategy are to determine the quality of sequencing data, estimate the sequencing depth required for *de novo* assembly projects and identifying sequencing reads that represent the uncharacterized regions of the genome of a given species.

Most microbiome studies have focused on phylogenetically informative markers such as 16S rDNA to reveal the relative composition and diversity of the metagenome in question (reviewed in [268, 300]). Despite the efficiency of such approaches, amplicon-based studies lack the ability to provide a genome-wide characterization of microbiomes. Moreover, sequencing errors and the presence of library chimeras can hamper the analysis of microbiomes using conventional tools, as only a handful of reads may be produced from any given fragment. This results in unreliable operational taxonomical units, which are often used in microbiome studies.

The advantage of our approach is that it can potentially discriminate between different species of a common phylum by relying on sequence content beyond the resolution of 16S rDNA sequences. We show that the similarity of microbiomes based on their composition and diversity can be revealed using kPAL, which is purely founded upon the sequencing data alone. In contrast, although UniFrac could reliably resolve rather stable gut microbiomes, it struggled with resolving highly diverse and dynamic microbiomes, such as those obtained from skin (i.e., the palm). We show that kPAL is sensitive to temporal changes in microbiomes and can potentially be used for a wide range of applications, such as forensic DNA fingerprinting. It is important to note that further developments are required for reliable assessment of temporal changes in a microbial community using the kPAL approach. Although kPAL does not provide a biological reason for the sources of variation within and between datasets, it opens the way to a more accurate and unbiased determination of the quality and complexity of genomic sequences.

4.5 Appendix

Supplementary file **Additional file 1: Supplemental notes, figures and tables** is accessible online:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4298064/bin/13059_2014_555_MOESM1_ESM.pdf

4.6 Abbreviations

- kPAL: *k*-mer Profile Analysis Library;
- Mbp: Megabase pairs;
- NGS: Next-generation sequencing;
- PCA: Principal component analysis;
- PCR: Polymera se chain reaction;
- QC: Quality control;
- RNA-Seq:
- RNA sequencing;
- SGA: String Graph Assembler;
- UTR: Untranslated region;
- WES: Whole exome sequencing;
- WGS: Whole genome sequencing.

4.7 Competing interests

The authors declare that they have no competing interests.

4.8 Authors' contributions

SYA, LK, MV, MvG, IP and PACtH performed the analyses. SYA and JFJL designed the study and JFJL developed the tool. KK and YA performed the wet-lab experiments and the sequencing. SYA, PACtH, JTdD, PdK and JFJL coordinated the study. SYA drafted the manuscript that was subsequently revised by all co-authors. All authors read and approved the final manuscript.

4.9 Acknowledgements

We thank Dr Jelle Goeman and Dr Erik W. van Zwet for their help, advice and input. This work was partially supported by the European Communitys' Seventh Framework Program (FP7/2007-2013) GEUVADIS (grant 261123), the Center for Medical Systems Biology and the Center for Genome Diagnostics in the Netherlands.

Chapter 5

BacTag - a pipeline for fast and accurate gene and allele typing in bacterial sequencing data

L. Khachatryan[1], M. E. M. Kraakman[4], A. T. Bernards[4], and J. F. J. Laros[1,2,3]

1 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
3 GenomeScan, Leiden, The Netherlands
4 Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands

BMC Genomics, 2019 20:338 doi 10.1186/s12864-019-5723-0

5.1 Background

In order to understand and predict the pathogenic impact and the outbreak potential of a bacterial infection, knowing the species responsible for this infection is not sufficient. Bacterial virulence is often controlled on the sub-species level by the set of specific genes or sometimes even alleles, leading to the necessity of diverse treatment strategies for infections induced by the same bacterial species [301, 302, 303, 304, 305]. For example, antibiotic resistance is one of the most well-known examples where slight variations in a gene can lead to a vast collection of antibiotics resistance profiles within one taxonomic group [306, 307]. Furthermore, different alleles of the same gene can be responsible for distinct adhesion and invasion strategies, reactions to the immune response of the infected organism and toxin production [308, 309]. Besides its relevance for understanding virulence, finding the alleles of specific genes also contributes to a more accurate bacterial classification. One of the most popular methods for subspecies bacterial typing, MultiLocus Sequence Typing (MLST), is based on determination of the alleles of multiple housekeeping genes [310, 311]. Knowing the allele combination allows to identify so called Sequencing Type (ST) of the organism, which is often associated with the important pathogens' attributes such as infection potential [312, 313, 314] or the ability to cause disease in human by transmitting from their animal reservoirs [315, 316, 317]. MLST typing is crucial for the epidemiological studies as it provides fast and accurate identification of geographical dispersal of pathogens and even reveals the migration patterns of the host organism [318, 319].

Despite the importance of the gene and allele typing in the bacterial genomes, there is no "gold standard" method to perform it. For a long time, the presence of particular virulent genes was detected using phenotypic markers such as serotyping [320]. Unfortunately, the set of genetic features that can be revealed using only the phenotype is very limited. Among other restrictions of this group of methods are the inability to grow certain fastidious pathogens in laboratory conditions as well as the extensive delay in cultivation and identification for slowly growing pathogens [321, 322, 323, 166, 324]. In particular cases, the gene and allele identification problem can be solved by using PCR or microarrays with gene- and allele specific primers or probes [325, 326, 327]. TThese types of methods are much faster and more reliable in comparison to the phenotype-based approaches. However, for the vast majority of genes it is impossible to generate primers or probes that would perform the allele discrimination due to the high similarity among sequences of alleles. Thus, PCR based typing often needs additional analysis, for example, a restriction fragment length polymorphism typing [328, 329] which elaborates the analysis process. PCR-based gene and allele typing most of the time has to be "tailor-made" for the particular group of organisms and the gene of interest. The rapid growth of newly discovered bacteria together with the high mutation rate of some

genes causes the necessity of constant changes in the existing PCR-protocols.

With the improvement of high throughput sequencing techniques and the development of associated bioinformatics software, it became possible to identify the allele variations directly from Whole Shotgun Genome Sequencing (WGS) data by comparing sequencing reads to the reference sequences of the known alleles of the gene of interest in the curated database. Currently, most of the curated and publicly available databases suitable for the gene typing are designed for subspecies classification using the MLST principle. These databases contain variable alleles of housekeeping genes and MLST schemas, associated with those housekeeping genes, for more than 60 bacterial species [330]. There are several tools that perform MLST by aligning assembled WGS data to each sequence in the linked database and reporting the alleles of housekeeping genes with the highest similarity to the provided data [331, 332]. The most recent tools for automated MLST performs the analysis on raw WGS data, as the assembly step is included in its pipeline ([333, 334]). Finally, stringMLST software [335] performs allele identification by comparing the *k*-mer profiles of raw sequencing data to the *k*-mer profiles of sequences in the MLST database. This strategy allows to speed up the analysis process drastically, yet the accuracy of the method is lower in comparison with alignment-based ones [336].

Though the WGS-based methods for gene and allele typing potentially requires less effort than any laboratory technique, it has some disadvantages and room for improvement. First of all, the time-consuming separate alignment of WGS data to each sequence in the database can be substituted with a faster algorithm. Furthermore, most of the existing bioinformatics tools for MLST do not provide an option to optimize the analysis settings, which means that the user cannot control, for example, parameters of reads mapping. Finally, it is also not possible to perform the analysis using a database or MLST schema that is not associated with the tool.

In this paper we present BacTag (**B**acterial **T**yping of **a**lleles and **g**enes) - a new pipeline, designed to rapidly and accurately detect genes and alleles in sequencing data. Due to the database preprocessing prior to the analysis, BacTag providing a solid and more detailed basis for downstream in comparison with similar tools while retaining the same accuracy. Additionally, our method performs gene and allele detection slightly faster than its current analogs. Our pipeline was successfully tested on both artificial (*E. coli*, *S. pseudintermedius*, *P. gingivalis*, *M. bovis*, *Borrelia spp*. and *Streptomyces spp.*) data and real (*E. coli*, *K. pneumoniae*) clinical WGS samples, by preprocessing the corresponding MLST databases and by performing the subsequent typing. This method is publicly available at https://git.lumc.nl/l.khachatryan/BacTag.

5.2 Materials and Methods

5.2.1 Pipeline implementation

The user interface is implemented in Bash, the processing modules are written in GNU Make. Bash allows for user interaction and files maintenance, while GNU Make makes the pipeline suitable for parallel high-performance computing. The pipeline consists of two parts: database preprocessing and sequencing data analysis. Both parts contain modules that include published tools and the scripts from our Python library. The pairwise sequence alignment is performed by the *aln* command from fastools [1]. Artificial paired end Illumina FASTQ formatted reads are created by the *make_fastq* local command of sim-reads[2]. Reads are mapped to a reference sequence with BWA *mem* [105]. Alignment sorting and indexing are performed by SAMtools [291]. Potential PCR duplicates are removed using SAMtools *rmdup* command. The SAMtools *mpileup* utility is used to summarize the coverage of mapped reads on a reference sequence at single base pair resolution. Variant calling is performed by the *call* command of BCFtools [337]. To verify whether the called variants for each allele really correspond to the allele sequence, the *vcf-consensus* command of VCFtools [338] is used. Comparison of two VCF files boils down to reporting the number of variants sites that are not equal for both files. Programming languages and software versions used for pipeline construction can be found in Supplementary Table S1. The user may specify parameters for artificial reads generation (by default read length, insert size and coverage are equal 50 nucleotides, 100 nucleotides and 40 respectively), the BWA *mem* and SAMtools *mpileup* utilities for both database preprocessing and sequencing data analysis parts separately. It is also possible to set the ploidy (by default this is one) of the sequencing data, which will be concidered during the variants calling in the analysis part of the pipeline.

5.2.1.1 Database preprocessing

The database preprocessing workflow is shown in Figure 5.1. We designed the pipeline such that all independent processes are performed in parallel, which reduces the calculation time.
The user provides the database that consists of alleles grouped by genes of interest. Optionally, the user can provide the 5?- and 3?-flanking regions for each gene, otherwise, every allele will be flanked on both sides with a fifty-nucleotide long poly-N sequence. That is done in order to prevent the coverage drop at the end of sequence during the sequencing data mapping. In the first step of the preprocessing stage, the sequences of all alleles belonging to the same gene are aligned in a pairwise manner, yielding the Levenshtein [339] distance for each pair of alleles.

[1] Available from: https://git.lumc.nl/j.f.j.laros/fastools. Accessed 27 Oct 2018.
[2] Available online at https://git.lumc.nl/j.f.j.laros/sim-reads

These distances are used to select the allele with the smallest average distance to all other sequences as the gene reference. In the same step the quality of the provided database is checked: it is reported when the same sequence is provided for multiple alleles or when one allele sequence is a subsequence of another. Once the quality report is created, the user can fix the original database when needed. In the next step, artificial Illumina paired end reads are created based on the sequence of each allele.

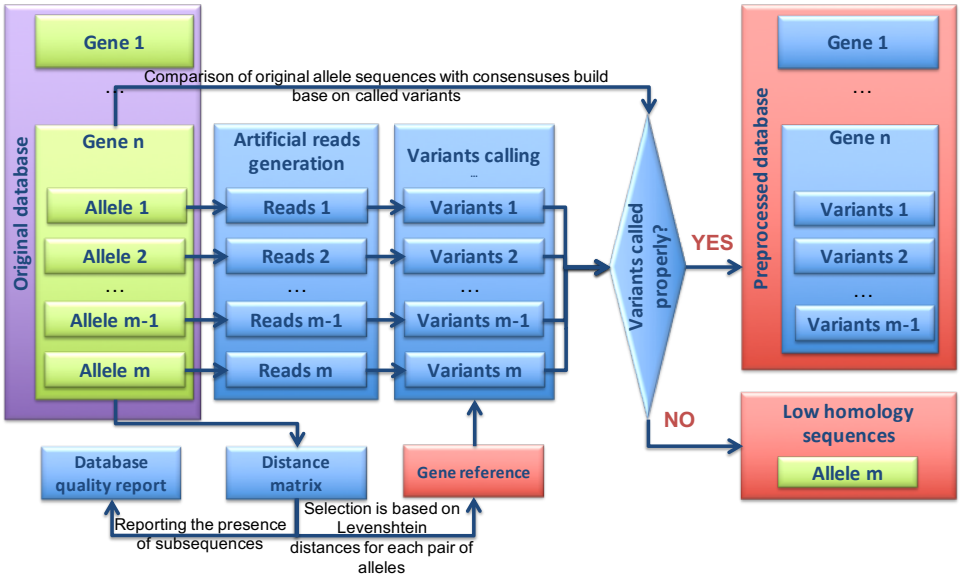

Figure 5.1: Schematic representation of the database preprocessing. All of the processes are illustrated for one gene. Calculations for several genes are done independently in parallel.

Reads are mapped to the selected gene reference, the alignment map file is sorted and indexed, after which the coverage of mapped reads on the reference sequence at a single base pair resolution is summarized and stored in a BCF file, which is used for variants calling. Variants are stored in a VCF file and further subjected to a quality check to verify whether they really correspond to the allele sequence. If variants defining alleles' sequence were not properly called, allele is reported and assigned to the so-called low similarity group of sequences. The low similarity group contains sequences for which the variants were not called correctly during the database preprocessing when using the centroid reference. I.e., for these alleles, the centroid is not an appropriate reference and therefore these sequences should be considered to be references themselves. In the final step the references of all genes are concatenated into one FASTA file, which further serves as the database reference.

5.2.1.2 Sequencing data analysis

The data analysis workflow can be found in Figure 5.2. To initiate the analysis,

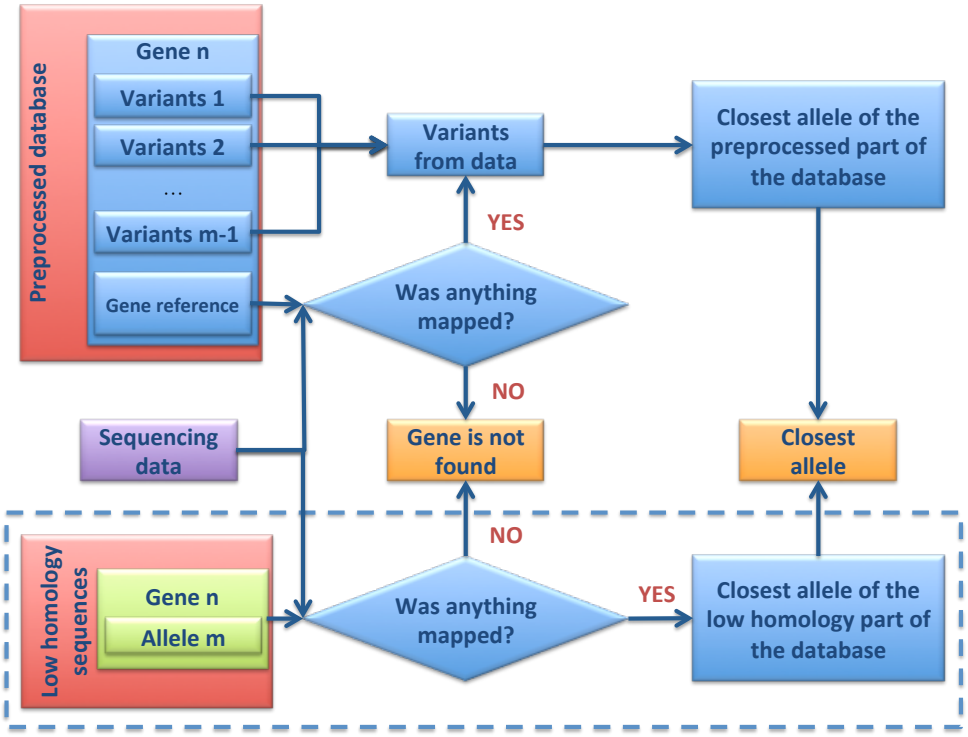

Figure 5.2: Schematic representation of the analysis part of BacTag pipeline. All of the processes are illustrated for one gene. Calculations for multiple genes are done independently in parallel. The analysis of the low homology group of sequences is highlighted by the dashed box and can be manually turned off by the user for the time efficiency.

the user provides two paired FASTQ files. After analysis initialization an output directory is created, which will serve to store the results of the analysis. The user can choose the name of the output directory, otherwise it will have the same name as the basename of the provided FASTQ files. The sequencing data analysis part of the pipeline is comprised of two steps: the main analysis and the analysis of low similarity group of sequences. If no sequences were assigned to the low similarity group during the database preprocessing, only the first step will be performed. The user can manually turn off the second step for time efficiency.

The main analysis
This part of the pipeline applies to the alleles that were not placed in the low ho-

mology group of sequences during the database preprocessing. analysed reads are mapped to the database reference, obtained after database preprocessing by concatenating all the gene reference sequences. The alignment map file is indexed and sorted and substituted to the removal of potential PCR duplicates. If there are no reads mapped to the gene reference, the gene is reported as not found in the analysed dataset. Otherwise, mapped reads are used to estimate the horizontal coverage of a gene reference at base pair resolution. The obtained BCF coverage summary is used for variant calling, the result of which is stored in VCF format. Variants are compared with variants collected for each gene allele during the preprocessing phase. Once the comparisons are done, the allele with the least difference from the sequencing data will be reported. It is also reported, if heterozygous variants were found in the sample, as that might indicate sequencing or mapping problems as well as the presence of more than one gene allele in the sequencing data. Reports for all genes are concatenated to a single result file, which is placed in the output directory.

Low homology group of sequences analysis
This part of the pipeline works with alleles that were placed in the low homology group of sequences during the database preprocessing. Sequencing reads are subjected to variant calling using each of the alleles from the low homology group as a reference (the same routine with the same parameters as for the main analysis step). If for the particular gene one of the alleles from the low homology group has fewer differences with the sequencing data in comparison to the allele reported during the main analysis, the allele from the low homology group will be reported as present in the sequencing data.

5.2.2 Pipeline testing

All the computational benchmarking was done on chimerashark Blade Server of SHARK computer cluster[3] with the maximum of 24 CPUs used at the same time.

5.2.3 Database

5.2.3.1 Genes and alleles

The database preprocessing part of the pipeline was tested using seven curated databases: *E. coli* Achtman MLST[4] (downloaded January 2018), *K. pneumoniae* Pasteur MLST[5] (downloaded October 2018), *S. pseudintermedius* MLST[6] (downloaded

[3]https://git.lumc.nl/shark/SHARK/wikis/home
[4]https://enterobase.warwick.ac.uk/species/ecoli/download_7_gene
[5]https://bigsdb.pasteur.fr/klebsiella/
[6]https://pubmlst.org/spseudintermedius/

February 2019), *P. gingivalis* MLST[7] (downloaded February 2019), *M. bovis* MLST[8] (downloaded February 2019), *Borrelia spp.* MLST[9] ([downloaded February 2019) and *Streptomyces spp.* MLST[10] ([downloaded February 2019). Each database contains sequences of variable regions of housekeeping genes: five for the *Streptomyces spp.* MLST, eight for the *Borrelia spp.* MLST and seven for all the remaining schemas.(see Table 5.1).
MLST schemas were selected for organisms from six different bacterial phyla. These organisms have a GC-content ranging between 29 and 73%. For the database preprocessing the following parameters for BWA mem and SAMtools mpileup tools were selected. Since the database consists of sequences of highly variable regions of housekeeping genes, the alignment mismatch penalty was set to 2 (4 by default) in order to provide the proper alignment for the regions where variants occur in close proximity. The minimum seed length was changed to 15 (19 by default) due to the short length of sequences in the selected database. Penalty for 5′- and 3′-end clipping was set to 100 (5 by default), forcing alignment to detect the variants located at the ends of the variable region. Single end mapped reads (anomalous read pairs, -A) were counted in order to detect variants located at the ends of the variable region. BAQ computation was disabled, as it is oversensitive to regions densely populated with variants. Bases with baseQ/BAQ lower than 13 were not skipped, since the database preprocessing is based on high quality artificial sequencing reads.

5.2.3.2 Flanking regions

The sequences of polymerase chain reaction (PCR) primers commonly applied to amplify each of the housekeeping genes (*E. coli* [340], *K. pneumoniae*[11], *S. pseudintermedius*[12], *M. bovis*[13], *P. gingivalis* [341]) for the selected MLST schemas were used to construct the flanking regions for this study. Each flanking region includes the primer sequence as well as the genomic sequence between the primer and the variable region of interest. The genomic sequence is extracted from the genome of one of the target strains for the corresponding MLST schema (see Table 5.1). In case low-sensitivity PCR primers are used (e.g., for *Borrelia spp.* MLST) or if no PCR primer sequences are available (e.g., for *Streptomyces spp.* MLST), fifty nucleotides before and after the variable regions were used as flanks. Flanking regions have the same orientation as the allele sequences in the database (see section 5.7.3, Additional file 2: Tables S2-S8).

[7] https://pubmlst.org/pgingivalis/
[8] https://pubmlst.org/mbovis/
[9] https://pubmlst.org/borrelia/
[10] https://pubmlst.org/streptomyces/
[11] Available from: http://bigsdb.pasteur.fr/klebsiella/primers_used.html. Accessed 16 Oct 2018.
[12] Available from: https://pubmlst.org/spseudintermedius/info/primers.pdf. Accessed 16 Feb 2019.
[13] Available from https://pubmlst.org/mbovis/info/M._bovis_MLST_targets_and_primers.pdf. Accessed 16 Feb 2019.

MLST database	Genes including number of alleles per gene	Number of alleles (per gene) in the low similarity group	Strain and reference sequence used for flanking region construction
E. coli	*adk* (623), *fumC* (933), *gyrB* (606), *Icd* (823), *mdh* (614), *purA* (563), *recA* (512)	*fumC* (11), *gyrB* (3), *mdh* (8)	UMN026, NC_011751.1
K. pneumoniae	*gapA* (184), *infB* (141), *mdh* (245), *pgi* (221), *phoE* (365), *rpoB* (189), *tonB* (472)	*gapA* (6), *mdh* (3), *tonB* (29)	Kp52.145, FO834906.1
S. pseudintermedius	*ack* (46), *cpn60* (96), *fdh* (26), *pta* (70), *purA* (77), *sar* (38), *tuf* (24)	-	ED99, NC_017568.1
M. bovis	*adh1* (15), *gltX* (17), *gpsA* (14), *gyrB* (25), *pta2* (23), *tdk* (15), *tkt* (26)	-	PG45, NC_014760.1
P. gingivalis	*ftsQ* (40), *gpdxJ* (37), *hagB* (37), *mcmA* (30), *pepO* (37), *pga* (27), *recA* (14)	-	ATCC 33277, NC_010729.1
Borrelia spp.	*clpA* (296), *clpX* (258), *nifS* (230), *pepX* (261), *pyrG* (269), *recG* (285), *rplB* (250), *uvrA* (261)	*clpA* (58), *clpX* (51), *nifS* (54), *pepX* (57), *pyrG* (51), *recG* (55), *rplB* (54), *uvrA* (45)	*B. hermsii* DAH, NC_010673.1
Streptomyces spp.	*atpD* (183), *gyrB* (179), *recA* (184), *rpoB* (183), *trpB* (200)	*atpD* (72), *gyrB* (147), *recA* (2), *rpoB* (6), *trpB* (69)	*S. coelicolor* A3(2), NC_003888.3

Table 5.1: Preprocessed MLST databases

5.2.3.3 Artificial test data

The sequencing data analysis part of the pipeline was validated by using artificial Illumina reads, based on the complete genomes of 30 different bacterial strains belonging to 13 different bacterial species (see Table 5.2), for which the alleles of housekeeping genes associated with the corresponding MLST schema were previously reported. Paired end FASTQ formatted reads of 100 bp were generated with an insert size of 100. For each genome, an average coverage of 80 was generated in this way.

5.2.3.4 Real test data

The analysis part of the pipeline was tested on 185 paired end Illumina WGS samples belonging to 9 different previously reported sequencing types (ST) of *E. coli* (section 5.7.3, Additional file 3: Table S9) and 98 paired end Illumina WGS samples belonging to 43 different previously reported STs of *K. pneumoniae* (section 5.7.3, Additional file 3: Table S10). Sequencing reads were downloaded from Sequence Read Archive (SRA, [342]). Prior to the analysis, the data quality check and correction (when necessary) was done for each sample using Flexiprep QC pipeline[14].

5.2.3.5 Parameters used for sequencing data analysis

The analysis of both artificial and real samples was done with the same parameters of BWA *mem* as during the database preprocessing. SAMtools *mpileup* parameters were as follow: anomalous read pairs were counted; extended BAQs were calculated for higher sensitivity but lower specificity.

[14] Available online at http://biopet-docs.readthedocs.io/en/latest/pipelines/flexiprep/

5.3 Results

5.3.1 Building the preprocessed MLST databases

We used BacTag to preprocess seven publicly available MLST databases. During this process we did not detect any duplications or partial sequences for any of the preprocessed databases. When preprocessing *E. coli* Achtman seven genes MLST database, 22 sequences (less than 0.5% of the total number of analysed sequences) belonging to three different genes were assigned to the low similarity group of sequences (see Table 5.1). The run time of the *E. coli* database preprocessing was approximately 2h. The peak memory usage was 150Mb. During the preprocessing of the *K. pneumoniae* database associated with the Pasteur seven genes MLST schema, 38 sequences (2.1% of the total number of analysed sequences) belonging to three different genes were assigned to the low similarity group of sequences. Preprocessing of databases associated with MLST schemas for *S. pseudintermedius*, *M. bovis* and *P. gingivalis* reported no sequences placed in the low similarity group of sequences. For the databases associated with the MLST schemas for Borrelia spp. and Streptomyces spp. 425 sequences (19.2% of the total number of analysed sequences) and 296 sequences (31.8% of the total number of analysed sequences) were placed in the low similarity group respectively. This large number of low similarity sequences indicates that the alleles in the analysed MLST databases are quite heterogeneous, which can be expected, considering that both aforementioned MLST schemas are genus-specific, not species-specific like other five analysed databases.
Since distance matrix is computed during the preprocessing, the expected CPU time will scale quadratically with the size of the database. We indeed found this behaviour as shown in Figure 5.3.

5.3.2 Testing BacTag on artificial data

We used the preprocessed MLST databases to reveal the presence of the corresponding housekeeping genes and to predict the allele for each of these genes in artificial sequencing data based on complete genomes of 30 different bacterial strains belonging to 15 different species. All housekeeping genes associated with the corresponding MLST schema were identified in each sample. The alleles found by the pipeline matched with the previously reported ones for each but one of the analysed genomes (see Table 5.2). The genome of *P. gingivalis* AJW4 (GenBank accession number NZ_CP011996.1) was previously reported [343] to have the allelic variants *ftsQ-16, gpdxJ-9, hagB-22, mcmA-17, pepO-22, pga-15 and recA-1*. However, BacTag analysis revealed the following set of alleles: *ftsQ-21, gpdxJ-23, hagB-1, mcmA-3, pepO-20, pga-3 and recA-7*. Manual inspection confirmed that alleles reported by BacTag are correct in case of all aforementioned genes.

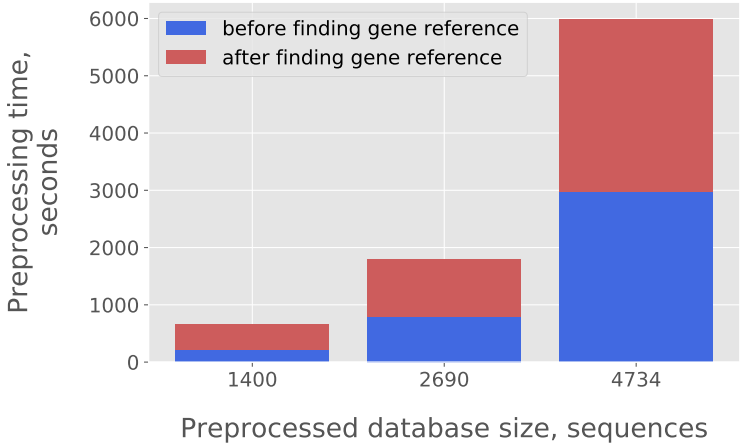

Figure 5.3: The dependence of database preprocessing time from the amount of sequences in the database.

5.3.3 Testing BacTag on real *E. coli* and *K. pneumoniae* data

We tested BacTag on 185 *E. coli* and 97 *K. pneumoniae* clinical publicly accessible WGS datasets, with each test yielding either one of nine *E. coli* or one of 44 *K. pneumoniae* sequencing types (STs). *E. coli* samples were analysed using the preprocessed *E. coli* Achtman seven genes MLST database, while *K. pneumoniae* samples were analysed using the preprocessed *K. pneumoniae* Pasteur seven genes MLST database. Each sample was shown to contain all expected seven housekeeping genes; alleles of those genes identified using our method corresponded to the expected ones for all but one sample (see Table 5.3). This sample was checked additionally using web-based tools for the MLST [333, 334]. Results of this independent check were completely identical to the ones obtained by our pipeline and suggest that the sample belongs to *E. coli* ST95 instead of ST73. Furthermore, according to the original publication [344], MLST was never done for this and 21 other samples analysed during the same study in order to confirm their sequencing type. Thus, we conclude that in Ref. [344] one of the samples was incorrectly assigned to *E. coli* ST73.

Our pipeline reported the presence of multiple variants at the same position for eight *E. coli* samples belonging to three different STs and 55 samples of *K. pneumoniae* belonging to 24 different STs (see Table 5.3)). This might suggest the presence of contamination in the sequenced DNA samples or the existence of pseudogenes in the genome of the sampled organisms.

Species and strain	GeneBank AC	Identified alleles
E. coli 042	FN554766.1	*adk-18*, *fumC-22*, *gyrB-20*, *Icd-23*, *mdh-5*, *purA-15*, *recA-4*
E. coli E2348/69	FM180568.1	*adk-15*, *fumC-15*, *gyrB-11*, *Icd-15*, *mdh-18*, *purA-11*, *recA-11*
E. coli E24377A	CP000800.1	*adk-6*, *fumC-213*, *gyrB-33*, *Icd-1*, *mdh-24*, *purA-8*, *recA-7*
E. coli IHE3034	NC_017628.1	*adk-37*, *fumC-38*, *gyrB-19*, *Icd-37*, *mdh-17*, *purA-11*, *recA-26*
E. coli IMT5155	CP005930.1	*adk-55*, *fumC-38*, *gyrB-19*, *Icd-37*, *mdh-17*, *purA-11*, *recA-26*
E. coli RS218	NZ_CP007149.1	*adk-37*, *fumC-38*, *gyrB-19*, *Icd-37*, *mdh-17*, *purA-11*, *recA-26*
E. coli UMN026	NC_011751.1	*adk-21*, *fumC-35*, *gyrB-115*, *Icd-6*, *mdh-5*, *purA-5*, *recA-4*
S. pseudintermedius NA45	NZ_CP016072.1	*ack-2*, *cpn60-10*, *fdh-2*, *pta-1*, *purA-5*, *sar-1*, *tuf-2*
S. pseudintermedius ED99	NC_017568.1	*ack-3*, *cpn60-9*, *fdh-2*, *pta-1*, *purA-1*, *sar-1*, *tuf-1*
S. pseudintermedius HKU10-03	NC_014925.1	*ack-2*, *cpn60-55*, *fdh-3*, *pta-42*, *purA-14*, *sar-2*, *tuf-1*
M. bovis Ningxia-1	NZ_CP023663.1	*adh1-4*, *gltX-3*, *gpsA-2*, *gyr-3*, *pta2-17*, *tdk-3*, *tkt-4*
M. bovis HB0801	NC_018077.1	*adh1-4*, *gltX-3*, *gpsA-2*, *gyr-3*, *pta2-5*, *tdk-3*, *tkt-4*
M. bovis NM2012	NZ_CP011348.1	*adh1-4*, *gltX-3*, *gpsA-2*, *gyr-3*, *pta2-5*, *tdk-3*, *tkt-4*
M. bovis CQ-W70	NC_015725.1	*adh1-4*, *gltX-5*, *gpsA-2*, *gyr-3*, *pta2-5*, *tdk-3*, *tkt-4*
M. bovis PG45	NC_014760.1	*adh1-3*, *gltX-2*, *gpsA-4*, *gyr-2*, *pta2-1*, *tdk-3*, *tkt-2*
M. bovis 08M	NZ_CP019639.1	*adh1-4*, *gltX-3*, *gpsA-2*, *gyr-3*, *pta2-5*, *tdk-3*, *tkt-4*
P. gingivalis ATCC 33277	NC_010729.1	*ftsQ-5*, *gpdxJ-9*, *hagB-1*, *mcmA-1*, *pepO-1*, *pga-5*, *recA-5*
P. gingivalis AJW4	NZ_CP011996.1	*ftsQ-21*, *gpdxJ-23*, *hagB-1*, *mcmA-3*, *pepO-20*, *pga-3*, *recA-7*
P. gingivalis A7A1-28	CP013131.1	*ftsQ-1*, *gpdxJ-12*, *hagB-1*, *mcmA-1*, *pepO-1*, *pga-1*, *recA-1*

Table 5.2: To be continued on the next page

Species and strain	GeneBank AC	Identified alleles
Borrelia hermsii DAH	NC_010673.1	*clpA-68, clpX-165, nifS-149, pepX-171, pyrG-179, recG-188, rplB-157, uvrA-175*
Borrelia turicatae 91E135	NC_008710.1	*clpA-71, clpX-166, nifS-150, pepX-172, pyrG-180, recG-189, rplB-158, uvrA-176*
Borrelia anserina BA2	CP005829	*clpA-212, clpX-179, nifS-161, pepX-186, pyrG-196, recG-204, rplB-170, uvrA-188*
Borrelia recurrentis A1	NC_011244	*clpA-213, clpX-164, nifS-162, pepX-187, pyrG-197, recG-205, rplB-156, uvrA-189*
Borrelia parkeri SLO	CP005851	*clpA-214, clpX-180, nifS-163, pepX-188, pyrG-198, recG-206, rplB-171, uvrA-190*
Borrelia coriaceae Co53	CP005745	*clpA-215, clpX-181, nifS-164, pepX-189, pyrG-199, recG-207, rplB-172, uvrA-191*
Borrelia crocidurae Achema	CP003426	*clpA-216, clpX-164, nifS-165, pepX-190, pyrG-200, recG-208, rplB-173, uvrA-192*
Streptomyces coelicolor A3(2)	NC_003888.3	*atpD-127, gyrB-124, recA-131, rpoB-126, trpB-142*
Streptomyces fulvissimus DSM 40593	CP005080.1	*atpD-133, gyrB-130, recA-13, rpoB-36, trpB-147*
Streptomyces griseus NBRC 13350	NC_010572.1	*atpD-6, gyrB-8, recA-8, rpoB-8, trpB-8*
Streptomyces albidoflavus J1074	NC_020990.1	*atpD-36, gyrB-5, recA-5, rpoB-36, trpB-39*

Table 5.2: Testing the pipeline on artificial WGS data

5.3.4 Comparing BacTag with web-based tools for *E. coli* Achtman MLST

We measured the time required for the analysis, using 30 samples belonging to the ST131 with the dataset size varying from 0.2 to 3. Gb. We performed the MLST typing in two modes: with and without analysis of the low similarity sequences group. As can be seen in Figure 5.4a and b, the processing time of BacTag depended on the sequencing sample size and the analysis mode. The larger the input sequencing data is, the more time is required for typing regardless of the analysis mode. Performing the typing including the analysis of low similarity group (mode 2) increases the processing time. Including low similarity sequences into the analysis did not affect the final output, for all samples tested during this research.

SRA Run AC	Reported ST	Expected ST	Genes with multiple variants at the same position
ERR966604	95	73	-
SRR2767732	16	16	*Icd*
SRR2767734	21	21	*Icd, mdh*
SRR2970643	131	131	*fumC*
SRR2970737	131	131	*adk, fumC, gyrB, mdh, recA, purA*
SRR2970742	131	131	*fumC*
SRR2970753	131	131	*fumC*
SRR2970774	131	131	*fumC*
SRR2970775	131	131	*fumC*
SRR5973405	1164	1164	*phoE*
SRR5973308	1180	1180	*phoE*
SRR5973303	13	13	*phoE*
SRR5973253	133	133	*phoE*
SRR5973334	133	133	*phoE*
SRR5973324	1373	1373	*phoE*
SRR5973251	1426	1426	*gapA, phoE*
SRR5973269	147	147	*gapA*
SRR5973320	1876	1876	*phoE*
SRR5973351	188	188	*gapA*
SRR5973329	20	20	*phoE*
SRR5973408	2267	2276	*phoE*
SRR5973397	25	25	*phoE*
SRR5973248	258	258	*gapA*
SRR5973283	258	258	*gapA*
SRR5973279	258	258	*gapA*
SRR5973271	258	258	*gapA*
SRR5973336	258	258	*gapA*
SRR5973319	258	258	*gapA*
SRR5973317	258	258	*gapA*
SRR5973294	258	258	*gapA*
SRR5973291	258	258	*gapA*
SRR5973289	258	258	*gapA*
SRR5973400	258	258	*gapA*
SRR5973382	258	258	*gapA*
SRR5973381	258	258	*gapA*
SRR5973287	258	258	*gapA*
SRR5973240	307	307	*phoE*

Table 5.3: To be continued on the next page

SRA Run AC	Reported ST	Expected ST	Genes with multiple variants at the same position
SRR597324	307	307	*phoE*
SRR5973282	307	307	*phoE*
SRR5973280	307	307	*phoE*
SRR5973339	307	307	*phoE*
SRR5973322	307	307	*phoE*
SRR5973288	307	307	*phoE*
SRR5973396	307	307	*phoE*
SRR5973380	307	307	*phoE*
SRR5973379	307	307	*phoE*
SRR5973376	307	307	*phoE*
SRR5973373	307	307	*phoE*
SRR5973361	307	307	*phoE*
SRR5973355	307	307	*phoE*
SRR5973284	23	23	*phoE*
SRR5973332	35	35	*phoE*
SRR5973389	35	35	*phoE*
SRR5973368	35	35	*phoE*
SRR5973393	405	405	*phoE*
SRR5973311	412	412	*phoE*
SRR5973371	429	429	*tonB*
SRR5973327	466	466	*phoE*
SRR5973407	466	466	*phoE*
SRR5973239	492	492	*phoE*
SRR5973301	502	502	*phoE*
SRR5973348	753	753	*phoE*
SRR5973362	8	8	*phoE*

Table 5.3: Results of pipeline testing on real *E. coli* and *K. pneumoniae* data. Only samples with results different from expected are shown.

The same 30 samples were submitted for analysis to web-based tools for MLST typing: MLST1.8 [333] and Enterobase [334]. These methods perform the assembly of submitted WGS data and use the obtained contigs for the BLAST-based comparison with sequences in the MLST database. For both tools, the results of the WGS assembly can be downloaded after the analysis is finished, MLST 1.8 also provides information about BLAST alignments for the best matching alleles as an output. The analysis of the 30 samples with MLST 1.8 took from 299 to 569 (median 454) minutes per job, the processing time did not correlate with the input data size (Figure 5.4c). MLST 1.8 failed to perform the assembly (and thus to finish the MLST) for two samples.

Long processing time can be explained by high load of the tool server. However, that cannot be checked as it is only possible to track the time in between job submission to the server and the time when job is finished. It is unfortunately not possible to assess when the actual calculations for the particular sample started. Another tool, Enterobase, failed to perform the analysis of one sample (due to the problems with assembly) and did not define the correct ST for one other sample. However, Enterobase shows when each part of the analysing pipeline is being launched, which allowed us to determine the time required for the analysis of each sample and compare it to our tool (Figure 5.5). The processing time for Enterobase was comparable to our tool and also seems to be dependent on the size of the submitted WGS data (Figure 5.4d).

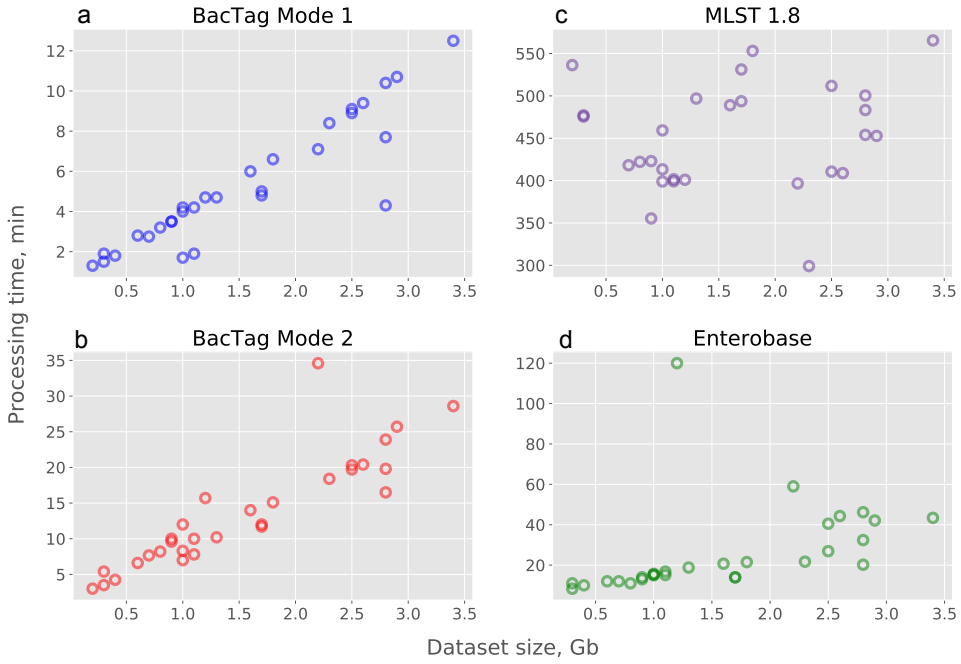

Figure 5.4: Time required for the analysis of 30 samples belonging to the ST131 by two modes of BacTag (a and b), MLST 1.8 (c) and Enterobase (d)

5.4 Discussion

In this paper we described BacTag - a new pipeline designed to perform fast and accurate gene and allele detection directly using WGS data. Our method was shown

to work faster and more accurate than most popular current bioinformatics tools due to the absence of the necessity to compare sequencing data with each sequence in the database. Instead, we preprocess the reference database once prior to the analysis in order to store all the mismatches between different alleles of the same gene. Under the assumption that all alleles of the same gene are highly similar, it is easy to check whether the gene of interest is present in the sequencing data by mapping the reads to the most "average" gene allele. Variants detected after such mapping can be compared with the information obtained during the database preprocessing in order to retrieve the allele of the detected gene. Since the database preprocessing needs to be done only once, this approach significantly reduces the time required for the analysis of multiple samples. Additionally, the possibility of parallel computation allows to speed up the database preprocessing significantly since all of the independent computations can be done in parallel.

Most of the existing tools for automatic gene and allele detection are based on fixed and rarely updated databases. The possibility to choose the database that will be preprocessed as well as to check the quality of that database is another essential feature of BacTag. It is important to note that the pipeline allows the user to set the parameters for the database preprocessing and sequencing data analysis. The same database, preprocessed with different parameters, allows the user to control in which case the variants for some alleles are not properly called. Thus, the user can determine the optimal parameters to detect as many of the alleles of interest as possible and apply this knowledge to the experimental design. On the other hand, preprocessing the database with the parameters of already existing sequencing data provides an estimate of the alleles that likely will not be properly detected.

While the current tools for gene allele identification require assembly of the WGS data prior to the comparison with the reference database, we chose to work directly with raw sequencing data. This was done in order to preserve the information about positions with multiple reported variants, which would be lost in case of bacterial genome assembly. That information is crucial for the detection of possible sample contaminations, presence of pseudogenes and, potentially, for extending our pipeline to metagenomic datasets. Furthermore, BacTag can work with sequencing data that for some reasons cannot be assembled.

Two main limitations of the pipeline need to be addressed. First, our approach assumes that a considerable part of the same gene alleles is highly similar. The more alleles of the same gene that do not fulfill this requirement, the slower the pipeline will work: sequences for which the pipeline will not be able to call the proper variants will be checked by direct read mapping. Second, the pipeline also does not provide proper analysis results if several alleles of the same gene are present in the sequencing data (this can be caused, among other reason, by the mixed-strain infection of the same subject, see [345, 346, 347]). More detailed evaluation of the horizontal coverage of the detected genes as well as the additional analysis of the positions with multiple variants reported could potentially help to resolve this

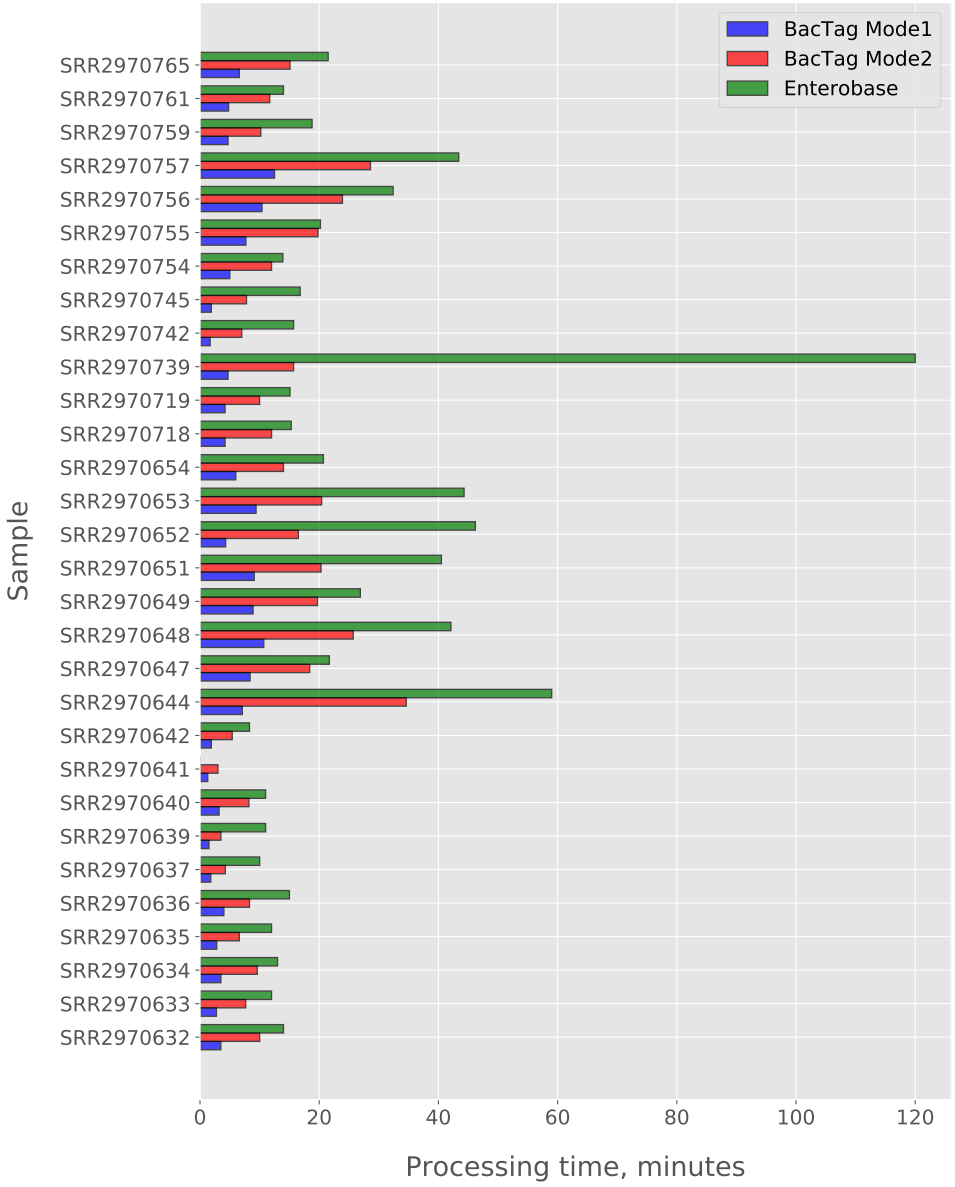

Figure 5.5: Comparing of the processing time required for the Achtman seven genes MLST analysis of 30 WGS *E. coli* samples.

problem and extend the approach in order to perform the analysis on complicated metagenomic datasets.

5.5 Conclusions

We have introduced BacTag - a new pipeline for fast and accurate gene and allele recognition based on database preprocessing and parallel computing. In contrast to the majority of already existing methods, BacTag avoids the comparison of sequencing data to each allele sequence present in the database due to the database preprocessing. While the database preprocessing provides analysis time reduction, it also provides important information about database quality. Amongst other advantages of our method are the possibility to cope with any user-provided database, and the absence of the assembly step that potentially may help extend our approach to metagenomics datasets. We believe that our approach can be useful for a wide range of projects, including bacterial subspecies classification, clinical diagnostics of bacterial infections, and epidemiological studies.

5.6 Abbreviations

- MLST - multi-locus sequence typing;
- NGS - next-generation sequencing;
- ST - sequence type;
- WGS - whole-genome shotgun sequencing.

5.7 Author Statements

5.7.1 Acknowledgements

The authors would like to thank Martijn Vermaat, Sander Bollen and Peter van 't Hof for the helpful discussions and suggestions. We also would like to thank Louk Rademaker for the feedback on this manuscript.

5.7.2 Funding information

This work is part of the research program "Forensic Science" which is funded by grant number 727.011.002 of the Netherlands Organisation for Scientific Research (NWO). The funding body had no direct influence on the design of the study, collection of samples, analysis or interpretation of the data.

5.7.3 Availability of data and materials

- All the data analysed in this study are included in this manuscript and its Additional files 1[15], 2[16] and 3[17].
- Results of the analysis done in this study are available via Figshare: https://doi.org/10.6084/m9.figshare.c.4041512.v1
- BacTag is publicly available via https://git.lumc.nl/l.khachatryan/BacTag.

[15] Available online https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-019-5723-0/MediaObjects/12864_2019_5723_MOESM1_ESM.pdf
[16] Available online https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-019-5723-0/MediaObjects/12864_2019_5723_MOESM2_ESM.pdf
[17] Available online https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-019-5723-0/MediaObjects/12864_2019_5723_MOESM3_ESM.pdf

5.7.4 Authors' contributions

LK conception, pipeline design, acquisition of data, analysis and interpretation of data, manuscript drafting; MEMK conception, acquisition of data, manuscript editing; ATB conception, general supervision; JFJL pipeline design, manuscript editing, general supervision. All authors have read and approved this manuscript.

5.7.5 Ethics approval and consent to participate

Since in this research no human material or clinical records of patients or volunteers were used, this research is out of scope for a medical ethical committee. This was verified by the Leiden University Medical Center Medical Ethical Committee.

5.7.6 Competing interests

The authors declare that they have no competing interests.

Chapter 6

General discussion and possible future improvement

As one can appreciate from this thesis, metagenomics analysis can be a relevant and vital step for the improvement of many fields including human and animal health, ecology, agriculture and forensics. This research was dedicated to a better understanding of the current situation in the field of metagenomics, and extending its present application boundaries. At first, we described, classified and evaluated popular data types, sequencing platforms and algorithms aimed to collect the information provided by microbial communities. We also improved the set of metagenomics data analysis tools by developing and testing both reference-dependent and reference-free algorithms. Below, we will summarize the most important conclusions of this thesis as answers to four important questions in the field of metagenomics.

6.1 Who is inhabiting the microbiome?

So far, the only possibility to find the answer to this question is to perform so-called reference-dependent analysis of metagenomic data, comparing the reads obtained during the microbiome sequencing with a reference database. As described in Chapter 2, we created a series of benchmark bacterial mixes with a different known distribution of species. The obtained mixes were used to estimate the resolution capacity of two different metagenomic datatypes - routine 16S and costlier WGS - and to evaluate two different approaches for the taxonomic reads classification. We have shown that the use of WGS data provides a much more accurate outcome in comparison to 16S samples. This was true for expected taxa prediction, and estimations of the abundances of the observed species. This conclusion was solid across all mixes and analysis techniques. Furthermore, we demonstrated that the same microbiome, analysed using 16S sampling by different pipelines and even using different reference databases, can produce quite distinct results. Finally, it is important to note that the constructed bacterial mixes can be utilized to evaluate future algorithms for metagenomic taxonomic profiling.
The conclusions obtained during this research finalize and supplement a series of previous reports [90, 348, 185, 349, 350, 351, 186, 352] addressing the incompetence of 16S metagenomic data in accessing the true metagenome taxonomic composition, and should be considered when planning microbiome sequencing experiments. Since the cost of producing WGS metagenomic data remains rather high, it is worth considering investigating comprehensive yet cost-effective sampling techniques for taxonomic profiling. The search of new, distinct from 16S rRNA, marker genes could be one of the possible solutions.

6.2 How complex is the investigated microbiome?

Once microbiology switched from single-genome studies to the exploration of multi-organism DNA samples, the question about the complexity of the investigated sample became the most vital one. The classical routine approaches aim to answer it by mapping the metagenome sequencing reads or assembly contigs to an annotated sequence from a reference databases. The obvious weak spot of such method is the incompleteness of current databases, as well as the discrepancy between their content and the real distribution of microbial species on our planet. Another group of techniques to estimate the metagenome complexity use the sequencing of multiple samples of the same metagenome cultivated under different conditions, and analyse the reads or contigs co-occurrences. The main weakness of such methods is their technical and computational difficulty.
In Chapter 3 we proposed a reference-free method to estimate the complexity of a metagenome. Our approach was designed to classify reads within a single long read metagenomic dataset using only the sequencing information, particularly *k*-mers. This so far unique approach featured an unsupervised machine learning tSNE algorithm for non-linear dimensionality reduction, as well as a subsequent density-based clustering technique. We have shown that *k*-mer profiles can reveal relationships between reads within a single metagenome using a series of simulated long read metagenomic datasets as well as the real PacBio RSII bioreactor microbiome sequencing data.
The obtained results are highly important, as they prove the concept of substructures detection within a single metagenome operating only with the information purely found in the sequencing reads alone. The possibility of reference-free deconvolution of metagenomic data benefits the field of metagenomics greatly, as it contributes not only to the estimation of metagenome complexity, but also improves the metagenomic data assembly and enables the investigation of new bacterial species. The main limitations of the described approach - restricted number of reads that can be analysed - is caused by memory issues when calculating the dissimilarity matrix between *k*-mer profiles. We believe that in the future, this issue can be solved by calculating the distances between *k*-mer profiles "on the go", and storing only the most informative ones. The constant improvement in quality and accessibility of long-reads sequencing techniques provides a great perspective for this approach in the future.

6.3 How to compare different metagenomes?

As was mentioned in the introduction to this thesis, comparative metagenomics strictly speaking does not necessarily require reference-based metagenome profiling. However, most of the scientific research uses reference-based methods to address the difference between two distinct metagenomes. In Chapter 4 we demonstrated that the comparison of metagenomic data performed using a reference-free approach provides much better resolution and allows to fetch the patterns lost during the standard reference-dependent techniques. In this thesis we presented kPal - a *k*-mer based method, that was used to resolve the level of relatedness between microbiomes. We tested kPal on a series of simulated metagenomes with different copy number of closely related bacterial genomes. Our method was sensitive to temporal changes in microbiome composition. To check whether our reference-free approach could distinguish between different human metagenomes, we tested it on a set of gut and palm 16S metagenomes, collected from different people in a period of 6 months. kPal could distinguish the datasets not only by the metagenome origin (gut or skin), but also by person! This result was better than the one demonstrated by the homology-based approach, which failed to cluster metagenomes per person in case of skin samples. The obtained results are highly significant as they allow to look at the comparative metagenomics under a different angle.
While the existing tools are following the "first annotate, then compare" model, we proposed a contrasting "first compare, then annotate" algorithm, when the comparison of the annotation-free profiles (in our case *k*-mer profiles) is followed by the investigation of the *k*-mers that contribute the most to the observed dissimilarities. The further investigation of the most informative *k*-mers and reads from which these *k*-mers belong, could allow to fetch the DNA sequences that might possibly be lost during the routine reference-based techniques. This idea can be developed further as a base for many different projects, for example metagenomics-based disease diagnostics. Another possible application is the search for species specific to a particular environment, body habitat, diet, or a person. This opens a set of new possibilities for fields like forensics, where the resolution of reference-dependent techniques was not enough to use metagenomic data in routine experiments.

6.4 What is the possible pathogenic impact of the metagenome?

Many different strategies can be implemented to find the functional profile of a metagenome. Among them are using a mapping to existing reference databases, and predicting possible functional genes with supervised machine learning techniques. Recently separated branch of metagenomics - meta-transcriptomics - provides researchers with community-wide gene expression (RNA-seq) data, which can be further utilized for metagenome functionality annotation. However, standard approaches for functional profiling fail to annotate the metagenomic data on the "sub-gene" level, when the information about allele of the particular gene is desired. In the meantime, it is known that different alleles are often responsible for distinct types of virulence. Therefore, it is important to rapidly detect not only the gene of interest, but also the relevant allele. Consequently, an approach that allows a "super-zoom" to a gene sequence, as well as a database providing the user with sequences of different alleles of the same gene, were required. Current methods are limited to mapping reads to each of the known allele reference, which is a time-consuming procedure. The other strategy is the assembly of sequencing reads with the subsequent mapping of the obtained contigs to the known allele references. The last algorithm provides fast and accurate results, but cannot be extended to metagenomic samples, since the assembly dismantles the possible variations in case of two different alleles of the same gene in the sample.

We developed BacTag (see Chapter 5), a distributed bioinformatics pipeline for fast and accurate bacterial gene and allele typing using clinical WGS sequencing data. The major advantage of this approach is a preprocessing procedure in which signatures of candidate alleles are identified and stored in a database. The subsequent identification of alleles in clinical samples is done using these signatures instead of using a traditional exhaustive search. This tool can be successfully used for diagnostic purposes. Also, and because this particular approach can be applied to uncultured samples, we expect to implement this method for cases in which time is of the essence. BacTag currently is not designed to work with samples where more than one allele of the same gene is present. However, unlike in case with other similar tools, this issue can be fixed in the future by detailed evaluation of the coverage depth, as well as the additional analysis of heterozygous variants sites. The development of reference databases, containing the allelic sequences of virulent genes, is another direction that still can be improved. Some progress in this direction is done for antibiotic-resistance genes, however, the great number of possibly virulent genes and their alleles is still not included in such databases. In the era of rising antimicrobial resistance and the existence of so-called "super-bacteria", a fast and accurate bioinformatic analysis providing the possible pathogenic impact of a microbial sample can be crucial for human health.

Bibliography

[1] WB Whitman, DC Coleman, and WJ Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583, 1998.

[2] AA Juwarkar, SK Yadav, PR Thawale, P Kumar, SK Singh, and T Chakrabarti. Developmental strategies for sustainable ecosystem on mine spoil dumps: a case of study. *Environmental monitoring and assessment*, 157(1-4):471–481, 2009.

[3] C Pedros-Alio. Dipping into the rare biosphere. *Sciense*, 5809:192, 2007.

[4] C Pedros-Alio. The rare bacterial biosphere. *Annual review of marine science*, 4:449–466, 2012.

[5] GT Pecl, MB Araujo, JD Bell, J Blanchard, TC Bonebrake, I-C Chen, TD Clark, RK Colwell, F Danielsen, B Evengaard, et al. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science*, 355(6332):eaai9214, 2017.

[6] K Todar. Bacteria and archaea and the cycles of elements in the environment. *Retrieved June*, 7:2014, 2012.

[7] M Paumann, G Regelsberger, C Obinger, and GA Peschek. The bioenergetic role of dioxygen and the terminal oxidase(s) in cyanobacteria. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1707(2-3):231–253, 2005.

[8] PG Falkowski and LV Godfrey. Electrons, life and the evolution of earth's oxygen cycle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1504):2705–2716, 2008.

[9] C Gougoulias, JM Clark, and LJ Shaw. The role of soil microbes in the global carbon cycle: tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems. *Journal of the Science of Food and Agriculture*, 94(12):2362–2371, 2014.

[10] MG Klotz, DA Bryant, and TE Hanson. The microbial sulfur cycle. *Frontiers in microbiology*, 2:241, 2011.

[11] J Chen, Y Li, Y Tian, C Huang, D Li, Q Zhong, and X Ma. Interaction between microbes and host intestinal health: modulation by dietary nutrients and gut-brain-endocrine-immune axis. *Current Protein and Peptide Science*, 16(7):592–603, 2015.

[12] SE Erdman and T Poutahidis. Microbes and oxytocin: benefits for host physiology and behavior. In *International review of neurobiology*, volume 131, pages 91–126. Elsevier, 2016.

[13] E Patterson, JF Cryan, GF Fitzgerald, RP Ross, TG Dinan, and C Stanton. Gut

microbiota, the pharmabiotics they produce and host health. *Proceedings of the Nutrition Society*, 73(4):477–489, 2014.

[14] LK Ursell, W van Treuren, JL Metcalf, M Pirrung, A Gewirtz, and R Knight. Replenishing our defensive microbes. *Bioessays*, 35(9):810–817, 2013.

[15] TA Van der Meulen, HJM Harmsen, H Bootsma, FKL Spijkervet, FGM Kroese, and A Vissink. The microbiome–systemic diseases connection. *Oral diseases*, 22(8):719–734, 2016.

[16] GJM Christensen and H Bruggemann. Bacterial skin commensals and their role as host guardians. *Beneficial microbes*, 5(2):201–215, 2013.

[17] Y Belkaid and S Tamoutounour. The influence of skin microorganisms on cutaneous immunity. *Nature Reviews Immunology*, 16(6):353, 2016.

[18] PC Calder. Feeding the immune system. *Proceedings of the Nutrition Society*, 72(3):299–309, 2013.

[19] DA Cowan, J-B Ramond, TP Makhalanyane, and P de Maayer. Metagenomics of extreme environments. *Current opinion in microbiology*, 25:97–102, 2015.

[20] P Blum. *Archaea: Ancient Microbes, Extreme Environments, and the Origin of Life*, volume 50. Gulf Professional Publishing, 2001.

[21] L Tazi, DP Breakwell, AR Harker, and KA Crandall. Life in extreme environments: microbial diversity in great salt lake, utah. *Extremophiles*, 18(3):525–535, 2014.

[22] C Bang, T Dagan, P Deines, N Dubilier, WJ Duschl, S Fraune, U Hentschel, H Hirt, N Hulter, T Lachnit, et al. Metaorganisms in extreme environments: do microbes play a role in organismal adaptation? *Zoology*, 2018.

[23] J Rifkin. *The biotech century*. Sonoma County Earth First/Biotech Last, 1998.

[24] S Farmer. Probiotic, lactic acid-producing bacteria and uses thereof, October 8 2002. US Patent 6,461,607.

[25] JR Postgate. Economic importance of sulphur bacteria. *Phil. Trans. R. Soc. Lond. B*, 298(1093):583–600, 1982.

[26] M Fernandez, B Del Rio, DM Linares, MC Martin, and MA Alvarez. Real-time polymerase chain reaction for quantitative detection of histamine-producing bacteria: use in cheese production. *Journal of dairy science*, 89(10):3763–3769, 2006.

[27] A Mayra-Makinen and M Bigret. Industrial use and production of lactic acid bacteria. *Food sciense and Technology*, 139:175–198, 2004.

[28] BS Dien, MA Cotta, and TW Jeffries. Bacteria engineered for fuel ethanol production: current status. *Applied microbiology and biotechnology*, 63(3):258–266, 2003.

[29] SF Bender, C Wagg, and MGA van der Heijden. An underground revolution: biodiversity and soil ecological engineering for agricultural sustainability. *Trends in Ecology & Evolution*, 31(6):440–452, 2016.

[30] M-N Xing, X-Z Zhang, and H Huang. Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. *Biotechnology advances*, 30(4):920–929, 2012.

[31] M Wainwright, J Lederberg, and J Lederberg. History of microbiology. *Encyclopedia of microbiology*, 2:419–437, 1992.

[32] Y Stanier, M Doudoroff, EA Adelberg, et al. General microbiology. *General microbiology.*, 1958.

[33] N Hall. Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, 210(9):1518–1525, 2007.

[34] SE Hasnain. Impact of human genome sequencing on microbiology. *Indian journal of medical microbiology*, 19(3):114, 2001.

[35] JT Staley and A Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology*, 39(1):321–346, 1985.

[36] A Escobar-Zepeda, A Vera-Ponce de Leon, and A Sanchez-Flores. The road to metagenomics: from microbiology to dna sequencing technologies and bioinformatics. *Frontiers in genetics*, 6:348, 2015.

[37] National Research Council et al. *The new science of metagenomics: revealing the secrets of our microbial planet*. National Academies Press, 2007.

[38] C Simon and R Daniel. Achievements and new knowledge unraveled by metagenomic approaches. *Applied microbiology and biotechnology*, 85(2):265–276, 2009.

[39] R Knight, A Vrbanac, B C Taylor, A Aksenov, C Callewaert, J Debelius, A Gonzalez, T Kosciolek, L-I McCall, D McDonald, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, page 1, 2018.

[40] S Junemann, N Kleinbolting, S Jaenicke, C Henke, J Hassa, J Nelkner, Y Stolze, S P Albaum, A Schluter, A Goesmann, et al. Bioinformatics for ngs-based metagenomics and the application to biogas research. *Journal of biotechnology*, 261:10–23, 2017.

[41] DJ Lane, B Pace, GJ Olsen, DA Stahl, ML Sogin, and NR Pace. Rapid determination of 16s ribosomal rna sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, 82(20):6955–6959, 1985.

[42] TM Schmidt, EF DeLong, and NR Pace. Analysis of a marine picoplankton community by 16s rrna gene cloning and sequencing. *Journal of bacteriology*, 173(14):4371–4378, 1991.

[43] PJ Turnbaugh, RE Ley, M Hamady, CM Fraser-Liggett, R Knight, and JI Gordon. The human microbiome project. *Nature*, 449(7164):804, 2007.

[44] HMP Integrative. The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, 16(3):276, 2014.

[45] V Robles-Alonso and F Guarner. From basic to applied research: lessons from the human microbiome projects. *Journal of clinical gastroenterology*, 48:S3–S4, 2014.

[46] SM Bakhtiar, JG LeBlanc, E Salvucci, A Ali, R Martin, P Langella, J-M Chatel, A Miyoshi, LG Bermudez-Humaran, and V Azevedo. Implications of the human microbiome in inflammatory bowel diseases. *FEMS microbiology letters*, 342(1):10–17, 2013.

[47] J Lloyd-Price, G Abu-Ali, and C Huttenhower. The healthy human microbiome. *Genome medicine*, 8(1):51, 2016.

[48] XC Morgan, N Segata, and C Huttenhower. Biodiversity and functional genomics in the human microbiome. *Trends in genetics*, 29(1):51–58, 2013.

[49] X C Morgan, T L Tickle, H Sokol, D Gevers, K L Devaney, D V Ward, J A Reyes, S A Shah, N LeLeiko, S B Snapper, et al. Dysfunction of the intestinal

microbiome in inflammatory bowel disease and treatment. *Genome biology*, 13(9):R79, 2012.

[50] D Gevers, S Kugathasan, L A Denson, Y Vazquez-Baeza, W Van Treuren, B Ren, E Schwager, D Knights, S J Song, M Yassour, et al. The treatment-naive microbiome in new-onset crohn's disease. *Cell host & microbe*, 15(3):382–392, 2014.

[51] C Pedros-Alio. Genomics and marine microbial ecology. *International Microbiology*, 9(3):191–197, 2006.

[52] PG Falkowski, RT Barber, and V Smetacek. Biogeochemical controls and feedbacks on ocean primary production. *Science*, 281(5374):200–206, 1998.

[53] CA Suttle. Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801, 2007.

[54] NA Bokulich, ZT Lewis, K Boundy-Mills, and DA Mills. A new perspective on microbial landscapes within food production. *Current opinion in biotechnology*, 37:182–189, 2016.

[55] M Trindade, LJ van Zyl, J Navarro-Fernández, and A Abd Elrazak. Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Frontiers in microbiology*, 6:890, 2015.

[56] MM Zhang, Y Qiao, EL Ang, and H Zhao. Using natural products for drug discovery: the impact of the genomics era. *Expert opinion on drug discovery*, 12(5):475–487, 2017.

[57] SM Techtmann and TC Hazen. Metagenomic applications in environmental monitoring and bioremediation. *Journal of industrial microbiology & biotechnology*, 43(10):1345–1354, 2016.

[58] JL Metcalf, ZZ Xu, A Bouslimani, P Dorrestein, DO Carter, and R Knight. Microbiome tools for forensic science. *Trends in biotechnology*, 35(9):814–823, 2017.

[59] SE Schmedes, AE Woerner, NMM Novroski, FR Wendt, JL King, KM Stephens, and B Budowle. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. *Forensic Science International: Genetics*, 32:50–61, 2018.

[60] EN Hanssen, E Avershina, K Rudi, P Gill, and L Snipen. Body fluid prediction from microbial patterns for forensic application. *Forensic Science International: Genetics*, 30:10–17, 2017.

[61] L Brinkac, TH Clarke, H Singh, C Greco, A Gomez, MG Torralba, B Frank, and KE Nelson. Spatial and environmental variation of the human hair microbiota. *Scientific reports*, 8(1):9017, 2018.

[62] N Fierer, CL Lauber, N Zhou, D McDonald, EK Costello, and R Knight. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences*, 107(14):6477–6481, 2010.

[63] S Lax, JT Hampton-Marcell, SM Gibbons, GB Colares, D Smith, JA Eisen, and JA Gilbert. Forensic analysis of the microbiome of phones and shoes. *Microbiome*, 3(1):21, 2015.

[64] RR Dunn, N Fierer, JB Henley, JW Leff, and HL Menninger. Home life: factors structuring the bacterial diversity found within and between homes. *PloS one*, 8(5):e64133, 2013.

[65] GE Flores, ST Bates, JG Caporaso, CL Lauber, JW Leff, R Knight, and N Fierer. Diversity, distribution and sources of bacteria in residential

kitchens. *Environmental microbiology*, 15(2):588–596, 2013.

[66] GE Flores, ST Bates, D Knights, CL Lauber, J Stombaugh, R Knight, and N Fierer. Microbial biogeography of public restroom surfaces. *PloS one*, 6(11):e28132, 2011.

[67] S Lax, DP Smith, J Hampton-Marcell, SM Owens, KM Handley, NM Scott, SM Gibbons, P Larsen, BD Shogan, S Weiss, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, 345(6200):1048–1052, 2014.

[68] SE Schmedes, AE Woerner, and B Budowle. Forensic human identification using skin microbiomes. *Applied and environmental microbiology*, pages AEM–01672, 2017.

[69] F Schluenzen, A Tocilj, R Zarivach, J Harms, M Gluehmann, D Janell, A Bashan, H Bartels, I Agmon, F Franceschi, et al. Structure of functionally activated small ribosomal subunit at 3.3 å resolution. *Cell*, 102(5):615–623, 2000.

[70] CR Woese, O Kandler, and ML Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

[71] CR Woese and GE Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[72] WG Weisburg, SM Barns, DA Pelletier, and DJ Lane. 16s ribosomal dna amplification for phylogenetic study. *Journal of bacteriology*, 173(2):697–703, 1991.

[73] J Wagner, P Coupland, H P Browne, T D Lawley, S C Francis, and J Parkhill. Evaluation of pacbio sequencing for full-length bacterial 16s rrna gene classification. *BMC microbiology*, 16(1):274, 2016.

[74] GJ Olsen, R Overbeek, N Larsen, TL Marsh, MJ McCaughey, MA Maciukenas, W-M Kuan, TJ Macke, Y Xing, and CR Woese. The ribosomal database project. *Nucleic Acids Research*, 20(suppl):2199–2200, 1992.

[75] JR Cole, Q Wang, JA Fish, B Chai, DM McGarrell, Y Sun, CT Brown, A Porras-Alfaro, CR Kuske, and JM Tiedje. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic acids research*, 42(D1):D633–D642, 2013.

[76] P Yilmaz, LW Parfrey, P Yarza, J Gerken, E Pruesse, C Quast, T Schweer, J Peplies, W Ludwig, and FO Glockner. The SILVA and "all-species living tree project (ltp)" taxonomic frameworks. *Nucleic acids research*, 42(D1):D643–D648, 2013.

[77] D McDonald, MN Price, J Goodrich, EP Nawrocki, TZ DeSantis, A Probst, GL Andersen, R Knight, and P Hugenholtz. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610, 2012.

[78] B Yang, Y Wang, and P-Y Qian. Sensitivity and correlation of hypervariable regions in 16s rrna genes in phylogenetic analysis. *BMC bioinformatics*, 17(1):135, 2016.

[79] CM Burke and AE Darling. A method for high precision sequencing of near full-length 16s rrna genes on an illumina miseq. *PeerJ*, 4:e2492, 2016.

[80] J Shin, S Lee, M-J Go, SY Lee, SC Kim, C-H Lee, and B-K Cho. Analysis of

the mouse gut microbiome using full-length 16s rrna amplicon sequencing. *Scientific reports*, 6:29681, 2016.

[81] S Ceuppens, D De Coninck, N Bottledoorn, F Van Nieuwerburgh, and M Uyttendaele. Microbial community profiling of fresh basil and pitfalls in taxonomic assignment of enterobacterial pathogenic species based upon 16S rRNA amplicon sequencing. *International journal of food microbiology*, 257:148–156, 2017.

[82] FE Dewhirst, Z Shen, MS Scimeca, LN Stokes, T Boumenna, T Chen, BJ Paster, and JG Fox. Discordant 16S and 23S rRNA gene phylogenies for the genus helicobacter: implications for phylogenetic inference and systematics. *Journal of bacteriology*, 187(17):6106–6118, 2005.

[83] S Hong, J Bunge, C Leslin, S Jeon, and SS Epstein. Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal*, 3(12):1365, 2009.

[84] PHA Timmers, HCA Widjaja-Greefkes, CM Plugge, and AJM Stams. Evaluation and optimization of PCR primers for selective and quantitative detection of marine anme subclusters involved in sulfate-dependent anaerobic methane oxidation. *Applied microbiology and biotechnology*, 101(14):5847–5859, 2017.

[85] RJ Case, Y Boucher, I Dahllof, C Holmstrom, WF Doolittle, and S Kjelleberg. Use of 16s rrna and rpob genes as molecular markers for microbial ecology studies. *Applied and environmental microbiology*, 73(1):278–288, 2007.

[86] FE Angly, PG Dennis, A Skarshewski, I Vanwonterghem, P Hugenholtz, and GW Tyson. Copyrighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, 2(1):11, 2014.

[87] M Perisin, M Vetter, JA Gilbert, and J Bergelson. 16stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *The ISME journal*, 10(4):1020, 2016.

[88] S Louca, M Doebeli, and L W Parfrey. Correcting for 16s rrna gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, 6(1):41, 2018.

[89] MGI Langille, J Zaneveld, JG Caporaso, D McDonald, D Knights, JA Reyes, JC Clemente, DE Burkepile, RLV Thurber, R Knight, et al. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature biotechnology*, 31(9):814, 2013.

[90] R Ranjan, A Rani, A Metwally, HS McGee, and DL Perkins. Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications*, 469(4):967–977, 2016.

[91] M Tessler, J S Neumann, E Afshinnekoo, M Pineda, R Hersch, L F M Velho, B T Segovia, F A Lansac-Toha, M Lemke, R DeSalle, et al. Large-scale differences in microbial biodiversity discovery between 16s amplicon and shotgun sequencing. *Scientific reports*, 7(1):6589, 2017.

[92] CB Driscoll, TG Otten, NM Brown, and TW Dreher. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in genomic sciences*, 12(1):9, 2017.

[93] BL Brown, M Watson, SS Minot, MC Rivera, and RB Franklin. Minion nanopore sequencing of environmental metagenomes: a synthetic approach. *GigaScience*, 6(3):1–10, 2017.

[94] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data. 2010.

[95] M Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[96] AM Bolger, M Lohse, and B Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[97] S Lindgreen, KL Adair, and PP Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, 6:19233, 2016.

[98] PHA Sneath, RR Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[99] PD Schloss, SL Westcott, T Ryabin, JR Hall, M Hartmann, EB Hollister, RA Lesniewski, BB Oakley, DH Parks, CJ Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

[100] JG Caporaso, J Kuczynski, J Stombaugh, K Bittinger, FD Bushman, EK Costello, N Fierer, AG Pena, JK Goodrich, JI Gordon, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335, 2010.

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

[102] Stephen F Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[103] B Buchfink, C Xie, and DH Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59, 2014.

[104] M Hamada, Y Ono, K Asai, and MC Frith. Training alignment parameters for arbitrary sequencers with last-train. *Bioinformatics*, 33(6):926–928, 2016.

[105] H Li and R Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[106] B Langmead and SL Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357, 2012.

[107] WJ Kent. BLAT - the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[108] AV Aho, JE Hopcroft, and JD Ullman. On finding lowest common ancestors in trees. *SIAM Journal on computing*, 5(1):115–132, 1976.

[109] DH Huson, AF Auch, J Qi, and SC Schuster. Megan analysis of metagenomic data. *Genome research*, 17(3):000–000, 2007.

[110] DE Wood and SL Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.

[111] D Kim, L Song, FP Breitwieser, and SL Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, 2016.

[112] M Burrows and DJ Wheeler. A block-sorting lossless data compression algorithm. 1994.

[113] P Ferragina and G Manzini. Indexing compressed text. *Journal of the ACM (JACM)*, 52(4):552–581, 2005.

[114] AL Delcher, S Kasif, RD Fleischmann, J Peterson, O White, and SL Salzberg. Alignment of whole genomes. *Nucleic acids research*, 27(11):2369–2376, 1999.

[115] R Ounit, S Wanamaker, TJ Close, and S Lonardi. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, 16(1):236, 2015.

[116] TAK Freitas, P-E Li, MB Scholz, and PSG Chain. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic acids research*, 43(10):e69–e69, 2015.

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

[118] F Meyer, D Paarmann, M D'Souza, R Olson, EM Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008.

[119] M Rho, H Tang, and Y Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research*, 38(20):e191–e191, 2010.

[120] J Droge, I Gregor, and AC McHardy. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*, 31(6):817–824, 2014.

[121] B Liu, T Gibbons, M Ghodsi, T Treangen, and M Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *Genome biology*, 12(1):P11, 2011.

[122] S Sunagawa, DR Mende, G Zeller, F Izquierdo-Carrasco, SA Berger, JR Kultima, LP Coelho, M Arumugam, J Tap, HB Nielsen, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, 10(12):1196, 2013.

[123] D Koslicki, S Foucart, and G Rosen. Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics*, 29(17):2096–2102, 2013.

[124] D Koslicki, S Foucart, and G Rosen. Wgsquikr: fast whole-genome shotgun metagenomic classification. *PloS one*, 9(3):e91784, 2014.

[125] P Menzel, KL Ng, and A Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications*, 7:11257, 2016.

[126] N-P Nguyen, S Mirarab, B Liu, M Pop, and T Warnow. Tipp: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.

[127] GGZ Silva, DA Cuevas, BE Dutilh, and RA Edwards. Focus: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ*, 2:e425, 2014.

[128] S Hunter, M Corbett, H Denise, M Fraser, A Gonzalez-Beltran, C Hunter, P Jones, R Leinonen, C McAnulla, E Maguire, et al. Ebi metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606, 2013.

[129] E Quevillon, V Silventoinen, S Pillai, N Harte, N Mulder, R Apweiler, and R Lopez. Interproscan: protein domains identifier. *Nucleic acids research*, 33(suppl_2):W116–W120, 2005.

[130] J-H Lee, H Yi, and J Chun. rrnaselector: a computer program for selecting ribosomal rna encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology*, 49(4):689, 2011.

[131] A Bateman, L Coin, R Durbin, RD Finn, V Hollich, S Griffiths-Jones, A Khanna, M Marshall, S Moxon, ELL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.

[132] DH Haft, JD Selengut, and O White. The tigrfams database of protein families. *Nucleic acids research*, 31(1):371–373, 2003.

[133] TK Attwood and ME Beck. Prints–a protein motif fingerprint database. *Protein Engineering, Design and Selection*, 7(7):841–848, 1994.

[134] N Hulo, A Bairoch, V Bulliard, L Cerutti, E De Castro, PS Langendijk-Genevaux, M Pagni, and CJA Sigrist. The prosite database. *Nucleic acids research*, 34(suppl_1):D227–D230, 2006.

[135] DWA Buchan, SCG Rison, JE Bray, D Lee, F Pearl, JM Thornton, and CA Orengo. Gene3d: structural assignments for the biologist and bioinformaticist alike. *Nucleic acids research*, 31(1):469–473, 2003.

[136] IF Spellerberg and PJ Fedor. A tribute to claude shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'shannon–wiener'index. *Global ecology and biogeography*, 12(3):177–179, 2003.

[137] EH Simpson. Measurement of diversity. *nature*, 1949.

[138] JR Bray and JT Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349, 1957.

[139] C Lozupone, M Hamady, and R Knight. Unifrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics*, 7(1):371, 2006.

[140] A Kislyuk, S Bhatnagar, J Dushoff, and JS Weitz. Unsupervised statistical clustering of environmental shotgun sequences. *BMC bioinformatics*, 10(1):316, 2009.

[141] DD Roumpeka, RJ Wallace, F Escalettes, I Fotheringham, and M Watson. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in genetics*, 8:23, 2017.

[142] Y-W Wu and Y Ye. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, 18(3):523–534, 2011.

[143] Y Wang, HCM Leung, S-M Yiu, and FYL Chin. Metacluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *Journal of Computational Biology*, 19(2):241–249, 2012.

[144] T Van Lang, T Van Hoai, et al. A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. *Algorithms for Molecular Biology*, 10(1):2, 2015.

[145] K Song, J Ren, G Reinert, M Deng, MS Waterman, and F Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in bioinformatics*, 15(3):343–353, 2013.

[146] S Girotto, C Pizzi, and M Comin. Metaprob: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics*, 32(17):i567–i575, 2016.

[147] X Ding, F Cheng, C Cao, and X Sun. Dectico: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. *BMC bioinformatics*, 16(1):323, 2015.

[148] H Cui and X Zhang. Alignment-free supervised classification of metagenomes by recursive svm. *BMC genomics*, 14(1):641, 2013.

[149] W Liao, J Ren, K Wang, S Wang, F Zeng, Y Wang, and F Sun. Alignment-free transcriptomic and metatranscriptomic comparison using sequencing signatures with variable length markov chains. *Scientific reports*, 6:37243, 2016.

[150] CC Laczny, C Kiefer, V Galata, T Fehlmann, C Backes, and A Keller. Busybee web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic acids research*, 45(W1):W171–W179, 2017.

[151] Y Wang, H Hu, and X Li. Mbmc: An effective markov chain approach for binning metagenomic reads from environmental shotgun sequencing projects. *Omics: a journal of integrative biology*, 20(8):470–479, 2016.

[152] RM Kotamarti, M Hahsler, D Raiford, M McGee, and MaH Dunham. Analyzing taxonomic classification using extensible markov models. *Bioinformatics*, 26(18):2235–2241, 2010.

[153] H-S Seok, W Hong, and J Kim. Estimating the composition of species in metagenomes by clustering of next-generation read sequences. *Methods*, 69(3):213–219, 2014.

[154] VI Ulyantsev, SV Kazakov, VB Dubinkina, AV Tyakht, and DG Alexeev. Metafast: fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics*, 32(18):2760–2767, 2016.

[155] VB Dubinkina, DS Ischenko, VI Ulyantsev, AV Tyakht, and DG Alexeev. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC bioinformatics*, 17(1):38, 2016.

[156] S Chatterji, I Yamazaki, Z Bai, and JA Eisen. Compostbin: A dna composition-based algorithm for binning environmental shotgun reads. In *Annual International Conference on Research in Computational Molecular Biology*, pages 17–28. Springer, 2008.

[157] M Comin and M Schimd. Fast comparison of genomic and meta-genomic reads with alignment-free measures based on quality values. *BMC medical genomics*, 9(1):36, 2016.

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

[159] SV Thankachan, SP Chockalingam, Y Liu, A Apostolico, and S Aluru. Alfred: a practical method for alignment-free distance computation. *Journal of Computational Biology*, 23(6):452–460, 2016.

[160] Y Wang, X Lei, S Wang, Z Wang, N Song, F Zeng, and T Chen. Effect of k-tuple length on sample-comparison with high-throughput sequencing data. *Biochemical and biophysical research communications*, 469(4):1021–1027, 2016.

[161] C Rinke, P Schwientek, A Sczyrba, NN Ivanova, IJ Anderson, J-F Cheng, A Darling, S Malfatti, BK Swan, EA Gies, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431, 2013.

[162] CE Mason and S Tighe. Focus on metagenomics. *Journal of Biomolecular Techniques: JBT*, 28(1):1, 2017.

[163] S Kumar, KK Krishnani, B Bhushan, and MP Brahmane. Metagenomics: retrospect and prospects in high throughput age. *Biotechnology research international*, 2015, 2015.

[164] DB Roszak and RR Colwell. Survival strategies of bacteria in the natural environment. *Microbiological reviews*, 51(3):365, 1987.

[165] EJ Stewart. Growing unculturable bacteria. *Journal of bacteriology*, pages JB–00345, 2012.

[166] KI Mohr. Diversity of myxobacteria-we only see the tip of the iceberg. *Microorganisms*, 6(3), 2018.

[167] M Hamady, C Fraser-Liggett, R Knight, et al. The human microbiome project: Exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804–10, 2007.

[168] W-L Wang, S-Y Xu, Z-G Ren, L Tao, J-W Jiang, and S-S Zheng. Application of metagenomics in the human gut microbiome. *World journal of gastroenterology: WJG*, 21(3):803, 2015.

[169] S Al Khodor, B Reichert, and IF Shatat. The microbiome and blood pressure: can microbes regulate our blood pressure? *Frontiers in pediatrics*, 5:138, 2017.

[170] R Kolde, EA Franzosa, G Rahnavard, AB Hall, H Vlamakis, C Stevens, MJ Daly, RJ Xavier, and C Huttenhower. Host genetic variation and its microbiome interactions within the human microbiome project. *Genome medicine*, 10(1):6, 2018.

[171] M Hattori and T D Taylor. The human intestinal microbiome: a new frontier of human biology. *DNA research*, 16(1):1–12, 2009.

[172] Atsushi Kouzuma, Shunichi Ishii, and Kazuya Watanabe. Metagenomic insights into the ecology and physiology of microbes in bioelectrochemical systems. *Bioresource technology*, 2018.

[173] FH Coutinho, GB Gregoracci, JM Walter, CC Thompson, and FL Thompson. Metagenomics sheds light on the ecology of marine microbes and their viruses. *Trends in Microbiology*, 2018.

[174] RJ Boissy, DJ Romberger, WA Roughead, L Weissenburger-Moser, JA Poole, and TD LeVan. Shotgun pyrosequencing metagenomic analyses of dusts from swine confinement and grain facilities. *PloS one*, 9(4):e95578, 2014.

[175] B Carbonetto, N Rascovan, R Alvarez, A Mentaberry, and MP Vazquez. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage systems in argentine pampas. *PloS one*, 9(6):e99949, 2014.

[176] JQ Su, B Wei, CY Xu, M Qiao, and YG Zhu. Functional metagenomic characterization of antibiotic resistance genes in agricultural soils from China. *Environment international*, 65:9–15, 2014.

[177] SJ Finley, ME Benbow, and GT Javan. Microbial communities associated with human decomposition and their potential use as postmortem clocks. *International journal of legal medicine*, 129(3):623–632, 2015.

[178] A Fornaciari. Environmental microbial forensics and archaeology of past pandemics. *Microbiology spectrum*, 5(1), 2017.

[179] Q Peng, X Wang, M Shang, J Huang, G Guan, Y Li, and B Shi. Isolation of a novel alkaline-stable lipase from a

metagenomic library and its specific application for milkfat flavor production. *Microbial cell factories*, 13(1):1, 2014.

[180] M Drancourt, C Bollet, A Carlioz, R Martelin, J-P Gayral, and D Raoult. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *Journal of clinical microbiology*, 38(10):3623–3630, 2000.

[181] G Muyzer, EC De Waal, and AG Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and environmental microbiology*, 59(3):695–700, 1993.

[182] M Balvociute and DH Huson. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC genomics*, 18(2):114, 2017.

[183] JM Janda and SL Abbott. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9):2761–2764, 2007.

[184] J-H Ahn, B-Y Kim, J Song, and H-Y Weon. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *Journal of Microbiology*, 50(6):1071–1074, 2012.

[185] JP Brooks, DJ Edwards, MD Harwich, MC Rivera, JM Fettweis, MG Serrano, RA Reris, NU Sheth, B Huang, P Girerd, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC microbiology*, 15(1):66, 2015.

[186] AJ Pinto and L Raskin. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PloS one*, 7(8):e43093, 2012.

[187] SW Kembel, M Wu, JA Eisen, and JL Green. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS computational biology*, 8(10):e1002743, 2012.

[188] EN Hanssen, KH Liland, P Gill, and L Snipen. Optimizing body fluid recognition from microbial taxonomic profiles. *Forensic Science International: Genetics*, 37:13–20, 2018.

[189] GJ Olsen and CR Woese. Ribosomal rna: a key to phylogeny. *The FASEB journal*, 7(1):113–123, 1993.

[190] P Yilmaz, R Kottmann, E Pruesse, C Quast, and FO Glockner. Analysis of 23s rrna genes in metagenomes–a case study from the global ocean sampling expedition. *Systematic and applied microbiology*, 34(6):462–469, 2011.

[191] L Yang, Z Tan, D Wang, L Xue, M-X Guan, T Huang, and R Li. Species identification through mitochondrial rrna genetic analysis. *Scientific reports*, 4:4089, 2014.

[192] AE Budding, M Hoogewerf, CMJE Vandenbroucke-Grauls, and PHM Savelkoul. Automated broad-range molecular detection of bacteria in clinical samples. *Journal of clinical microbiology*, 54(4):934–943, 2016.

[193] FCA Quaak, T van Duijn, J Hoogenboom, AD Kloosterman, and I Kuiper. Human-associated microbial populations as evidence in forensic casework. *Forensic Science International: Genetics*, 36:176–185, 2018.

[194] AE Woerner, NMM Novroski, FR Wendt, A Ambers, R Wiley, SE Schmedes, and B Budowle. Forensic human identification with targeted microbiome markers using nearest neighbor classification. *Forensic Science International: Genetics*, 38:130–139, 2019.

[195] J Jovel, J Patterson, W Wang, N Hotte, S O'Keefe, T Mitchel, T Perry, D Kao, AL Mason, KL Madsen, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology*, 7:459, 2016.

[196] N Shah, H Tang, TG Doak, and Y Ye. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. In *Biocomputing 2011*, pages 165–176. World Scientific, 2011.

[197] B Steven, LV Gallegos-Graves, SR Starkenburg, PS Chain, and CR Kuske. Targeted and shotgun metagenomic approaches provide different descriptions of dryland soil microbial communities in a manipulated field study. *Environmental microbiology reports*, 4(2):248–256, 2012.

[198] A Almeida, AL Mitchell, A Tarkowska, and RD Finn. Benchmarking taxonomic assignments based on 16s rrna gene profiling of the microbiota from commonly sampled environments. *BMC Genomics*, 17(1), 2016.

[199] T Sijen. Molecular approaches for forensic cell type identification: on mrna, mirna, dna methylation and microbial markers. *Forensic Science International: Genetics*, 18:21–32, 2015.

[200] TH Clarke, A Gomez, H Singh, KE Nelson, and LM Brinkac. Integrating the microbiome as a resource in the forensics toolkit. *Forensic Science International: Genetics*, 30:141–147, 2017.

[201] R D'Amore, U Z Ijaz, M Schirmer, JG Kenny, R Gregory, AC Darby, M Shakya, M Podar, C Quince, and N Hall. A comprehensive benchmarking study of protocols and sequencing platforms for 16s rrna community profiling. *BMC genomics*, 17(1):55, 2016.

[202] A Sczyrba, P Hofmann, P Belmann, D Koslicki, S Janssen, J Droge, I Gregor, S Majda, J Fiedler, E Dahms, and et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063, 2017.

[203] MA Peabody, T Van Rossum, R Lo, and FSL Brinkman. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC bioinformatics*, 16(1):362, 2015.

[204] K Mavromatis, N Ivanova, K Barry, H Shapiro, E Goltsman, AC McHardy, I Rigoutsos, A Salamov, F Korzeniewski, M Land, and et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature methods*, 4(6):495, 2007.

[205] ABR McIntyre, R Ounit, E Afshinnekoo, RJ Prill, E Henaff, N Alexander, SS Minot, D Danko, J Foox, S Ahsanuddin, and et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome biology*, 18(1):182, 2017.

[206] AM Walsh, F Crispie, O O'Sullivan, L Finnegan, MJ Claesson, and PD Cotter. Species classifier choice is a key consideration when analysing low-complexity food microbiome data. *Microbiome*, 6(1):50, 2018.

[207] VC Piro, M Matschkowski, and BY Renard. Metameta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, 5(1):101, 2017.

[208] A Escobar-Zepeda, EE Godoy-Lozano, L Raggi, L Segovia, E Merino, RM Gutierrez-Rios, K Juarez, AF Licea-Navarro, L Pardo-Lopez, and A Sanchez-Flores. Analysis of sequencing strategies and tools for

taxonomic annotation: Defining standards for progressive metagenomics. *Scientific reports*, 8(1):12034, 2018.

[209] E Plummer, J Twin, D M Bulach, SM Garland, and SN Tabrizi. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *Journal of Proteomics & Bioinformatics*, 8(12):283, 2015.

[210] H Hasman, D Saputra, T Sicheritz-Ponten, O Lund, CA Svendsen, N Frimodt-Moller, and FM Aarestrup. Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples. *Journal of clinical microbiology*, pages JCM–02452, 2013.

[211] A Klindworth, E Pruesse, T Schweer, J Peplies, C Quast, M Horn, and FO Glockner. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*, 41(1):e1–e1, 2013.

[212] A Bankevich, S Nurk, D Antipov, AA Gurevich, M Dvorkin, AS Kulikov, VM Lesin, SI Nikolenko, S Pham, AD Prjibelski, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[213] SY Anvar, L Khachatryan, M Vermaat, M van Galen, I Pulyakhina, Y Ariyurek, K Kraaijeveld, JT den Dunnen, P de Knijff, P Ac't Hoen, et al. Determining the quality and complexity of next-generation sequencing data without a reference genome. *Genome biology*, 15(12):555, 2014.

[214] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[215] T Magoc and SL Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011.

[216] NA O'Leary, MW Wright, JR Brister, S Ciufo, D Haddad, R McVeigh, B Rajput, B Robbertse, B Smith-White, D Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2015.

[217] FP Breitwieser and SL Salzberg. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *BioRxiv*, page 084715, 2016.

[218] A Wilke, T Harrison, J Wilkening, D Field, EM Glass, N Kyrpides, K Mavrommatis, and F Meyer. The m5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC bioinformatics*, 13(1):141, 2012.

[219] HB Mann and DR Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[220] Y Benjamini, Yand Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[221] C Van Rijsbergen. Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle*

upon Tyne Seminar on Data Base Systems, pages 1–14, 1979.

[222] N Nagarajan, C Cook, MP Di Bonaventura, H Ge, A Richards, KA Bishop-Lilly, R DeSalle, TD Read, and M Pop. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC genomics*, 11(1):242, 2010.

[223] DJ Edwards and KE Holt. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(1):2, 2013.

[224] EA Grice and JA Segre. The skin microbiome. *Nature Reviews Microbiology*, 9(4):244, 2011.

[225] S Bikel, A Valdez-Lara, F Cornejo-Granados, K Rico, S Canizales-Quinteros, X Soberon, L Del Pozo-Yauner, and A Ochoa-Leyva. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Computational and structural biotechnology journal*, 13:390–401, 2015.

[226] MJ Gosalbes, JJ Abellan, A Durban, AE Perez-Cobas, A Latorre, and A Moya. Metagenomics of human microbiome: beyond 16S rDNA. *Clinical Microbiology and Infection*, 18:47–49, 2012.

[227] S Maccaferri, E Biagi, and P Brigidi. Metagenomics: key to human gut microbiota. *Digestive diseases*, 29(6):525–530, 2011.

[228] R Martin, S Miquel, P Langella, and LG Bermudez-Humaran. The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, 5(3):413–423, 2014.

[229] SL Edmonds-Wilson, NI Nurinova, CA Zapka, N Fierer, and M Wilson. Review of human hand microbiome research. *Journal of dermatological science*, 80(1):3–12, 2015.

[230] HE Blum. The human microbiome. *Advances in medical sciences*, 62(2):414–420, 2017.

[231] E Holmes, JV Li, JR Marchesi, and JK Nicholson. Gut microbiota composition and activity in relation to host metabolic phenotype and disease risk. *Cell metabolism*, 16(5):559–564, 2012.

[232] AP Bhatt, MR Redinbo, and SJ Bultman. The role of the microbiome in cancer development and therapy. *CA: a cancer journal for clinicians*, 67(4):326–344, 2017.

[233] I Cho and MJ Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260, 2012.

[234] JL Sonnenburg and F Backhed. Diet–microbiota interactions as moderators of human metabolism. *Nature*, 535(7610):56, 2016.

[235] BH Mullish, JR Marchesi, MR Thursz, and HRT Williams. Microbiome manipulation with faecal microbiome transplantation as a therapeutic strategy in clostridium difficile infection. *QJM: An International Journal of Medicine*, 108(5):355–359, 2014.

[236] RD Moloney, L Desbonnet, G Clarke, TG Dinan, and JF Cryan. The microbiome: stress, health and disease. *Mammalian Genome*, 25(1-2):49–74, 2014.

[237] AV Contreras, B Cocom-Chan, G Hernandez-Montes, T Portillo-Bobadilla, and O Resendis-Antonio. Host-microbiome interaction and cancer: Potential application in precision medicine. *Frontiers in physiology*, 7:606, 2016.

[238] C He, Y Shan, and W Song. Targeting gut microbiota as a possible therapy for diabetes. *Nutrition Research*, 35(5):361–367, 2015.

[239] CJ Marx. Can you sequence ecology? metagenomics of adaptive diversification. *PLoS biology*, 11(2):e1001487, 2013.

[240] S Hiraoka, C-C Yang, and W Iwasaki. Metagenomics and bioinformatics in microbial ecology: current status and beyond. *Microbes and environments*, 31(3):204–212, 2016.

[241] R Tiwari, L Nain, N E Labrou, and P Shukla. Bioprospecting of functional cellulases from metagenome for second generation biofuel production: a review. *Critical reviews in microbiology*, 44(2):244–257, 2018.

[242] MOA Sommer, GM Church, and G Dantas. A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion. *Molecular systems biology*, 6(1):360, 2010.

[243] V Kunin, A Copeland, A Lapidus, K Mavromatis, and P Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews*, 72(4):557–578, 2008.

[244] SS Mande, MH Mohammed, and TS Ghosh. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, 13(6):669–681, 2012.

[245] Leandro N Lemos, Roberta R Fulthorpe, Eric W Triplett, and Luiz FW Roesch. Rethinking microbial diversity analysis in the high throughput sequencing era. *Journal of microbiological methods*, 86(1):42–51, 2011.

[246] P Janssen, L Goldovsky, V Kunin, N Darzentas, and CA Ouzounis. Genome coverage, literally speaking: The challenge of annotating 200 genomes with 4 million publications. *EMBO reports*, 6(5):397–399, 2005.

[247] KB Akondi and VV Lakshmi. Emerging trends in genomic approaches for microbial bioprospecting. *Omics: a journal of integrative biology*, 17(2):61–70, 2013.

[248] JC Hunter-Cevera. The value of microbial diversity. *Current Opinion in Microbiology*, 1(3):278–285, 1998.

[249] NR Pace. Mapping the tree of life: progress and prospects. *Microbiology and molecular biology reviews*, 73(4):565–576, 2009.

[250] J-D Grattepanche, LF Santoferrara, GB McManus, and LA Katz. Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends in microbiology*, 22(8):432–437, 2014.

[251] L Zinger, A Gobet, and T Pommier. Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, 21(8):1878–1896, 2012.

[252] B Szalkai, I Scheer, K Nagy, BG Vertessy, and V Grolmusz. The metagenomic telescope. *PloS one*, 9(7):e101605, 2014.

[253] GL Rosen, R Polikar, DA Caseiro, SD Essinger, and BA Sokhansanj. Discovering the unknown: improving detection of novel species and genera from short reads. *BioMed Research International*, 2011, 2011.

[254] Y Ono, K Asai, and M Hamada. Pbsim: Pacbio reads simulator—toward accurate genome assembly. *Bioinformatics*, 29(1):119–121, 2012.

[255] E Van Eijk, SY Anvar, HP Browne, WY Leung, J Frank, AM Schmitz, AP Roberts, and WK Smits. Complete

genome sequence of the *Clostridium difficile* laboratory strain 630δ erm reveals differences from strain 630, including translocation of the mobile element ctn 5. *BMC genomics*, 16(1):31, 2015.

[256] MF Haroon, S Hu, Y Shi, M Imelfort, J Keller, P Hugenholtz, Z Yuan, and GW Tyson. Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature*, 500(7464):567, 2013.

[257] MJ Chaisson and G Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *BMC bioinformatics*, 13(1):238, 2012.

[258] C-S Chin, P Peluso, FJ Sedlazeck, M Nattestad, GT Concepcion, A Clum, C Dunn, R O'Malley, R Figueroa-Balderas, A Morales-Cruz, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050, 2016.

[259] L van der Maaten and G Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[260] M Ester, H-P Kriegel, J Sander, X Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[261] J Frank, S Lucker, RHAM Vossen, MSM Jetten, RJ Hall, HJM Op den Camp, and SY Anvar. Resolving the complete genome of kuenenia stuttgartiensis from a membrane bioreactor enrichment using single-molecule real-time sequencing. *Scientific reports*, 8(1):4580, 2018.

[262] DB Goldstein, A Allen, J Keebler, EH Margulies, S Petrou, S Petrovski, and S Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics*, 14(7):460, 2013.

[263] A Nekrutenko and J Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667, 2012.

[264] M Costello, TJ Pugh, TJ Fennell, C Stewart, L Lichtenstein, JC Meldrim, JL Fostel, DC Friedrich, D Perrin, D Dionne, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research*, 41(6):e67–e67, 2013.

[265] C Alkan, BP Coe, and EE Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363, 2011.

[266] JM Kidd, N Sampas, F Antonacci, T Graves, R Fulton, HS Hayden, C Alkan, M Malig, M Ventura, G Giannuzzi, et al. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature methods*, 7(5):365, 2010.

[267] H Li and N Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 2010.

[268] J Kuczynski, CL Lauber, WA Walters, LW Parfrey, JC Clemente, D Gevers, and R Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47, 2012.

[269] S Subramanian and S Kumar. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome research*, 13(5):838–844, 2003.

[270] J Sved and A Bird. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences*, 87(12):4692–4696, 1990.

[271] Ms Csuros, L Noe, and G Kucherov. Reconsidering the significance of genomic word frequencies. *Trends in Genetics*, 23(11):543–546, 2007.

[272] C Acquisti, G Poste, D Curtiss, and S Kumar. Nullomers: really a matter of natural selection? *PloS one*, 2(10):e1022, 2007.

[273] J Josse, AD Kaiser, and A Kornberg. Enzymatic synthesis of deoxyribonucleic acid VIII. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *Journal of Biological Chemistry*, 236(3):864–875, 1961.

[274] B Chor, D Horn, N Goldman, Y Levy, and T Massingham. Genomic DNA k-mer spectra: models and modalities. *Genome biology*, 10(10):R108, 2009.

[275] R Hariharan, R Simon, MR Pillai, and TD Taylor. Comparative analysis of DNA word abundances in four yeast genomes using a novel statistical background model. *PloS one*, 8(3):e58038, 2013.

[276] B Jiang, JS Liu, and ML Bulyk. Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers. *Bioinformatics*, 29(11):1390–1398, 2013.

[277] Y Liu, J Schroder, and B Schmidt. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics*, 29(3):308–315, 2012.

[278] H Chae, J Park, S-W Lee, KP Nephew, and S Kim. Comparative analysis using k-mer and k-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes. *Nucleic acids research*, 41(9):4783–4791, 2013.

[279] DR Kelley, MC Schatz, and SL Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11(11):R116, 2010.

[280] R Chikhi and P Medvedev. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37, 2013.

[281] A Brazma, I Jonassen, J Vilo, and E Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome research*, 8(11):1202–1215, 1998.

[282] GE Sims, S-R Jun, GA Wu, and S-H Kim. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of the National Academy of Sciences*, pages pnas–0813249106, 2009.

[283] JT Simpson. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, 30(9):1228–1235, 2014.

[284] T Lappalainen, M Sammeth, MR Friedlander, P AC't Hoen, J Monlong, MA Rivas, M Gonzalez-Porta, N Kurbatova, T Griebel, PG Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506, 2013.

[285] P AC't Hoen, MR Friedlander, J Almlof, M Sammeth, I Pulyakhina, SY Anvar, JFJ Laros, HPJ Buermans, O Karlberg, M Brannvall, et al. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature biotechnology*, 31(11):1015, 2013.

[286] WA Kosters and JFJ Laros. Metrics for mining multisets. In *Research and Development in Intelligent systems XXIV*, pages 293–303. Springer, 2008.

[287] PJ Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[288] J Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[289] G Lunter and M Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, 21(6):936–939, 2011.

[290] AR Quinlan and IM Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[291] HH Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, and R Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[292] JG Caporaso, CL Lauber, EK Costello, D Berg-Lyons, A Gonzalez, J Stombaugh, D Knights, P Gajer, J Ravel, N Fierer, et al. Moving pictures of the human microbiome. *Genome biology*, 12(5):R50, 2011.

[293] KJ Stacey, GR Young, F Clark, DP Sester, TL Roberts, S Naik, MJ Sweet, and DA Hume. The molecular basis for the lack of immunostimulatory activity of vertebrate DNA. *The Journal of Immunology*, 170(7):3614–3620, 2003.

[294] P Kaufmann, A Pfefferkorn, M Teuber, and L Meile. Identification and quantification of bifidobacterium species isolated from food with genus-specific 16S rRNA-targeted probes by colony hybridization and PCR. *Applied and Environmental Microbiology*, 63(4):1268–1273, 1997.

[295] KJV Nordstrom, MC Albani, GVe James, C Gutjahr, B Hartwig, F Turck, U Paszkowski, G Coupland, and K Schneeberger. Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature biotechnology*, 31(4):325, 2013.

[296] SN Gardner and BG Hall. When whole-genome alignments just won't work: kSNP v2 software for alignment-free snp discovery and phylogenetics of hundreds of microbial genomes. *PloS one*, 8(12):e81760, 2013.

[297] G Marcais and C Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

[298] M Crusoe, G Edvenson, J Fish, A Howe, E McDonald, J Nahum, K Nanlohy, H Ortiz-Zuazaga, J Pell, J Simpson, et al. The khmer software package: enabling efficient sequence analysis. *URL http://dx. doi. org/10.6084/m9. figshare*, 979190, 2014.

[299] DS DeLuca, JZ Levin, A Sivachenko, T Fennell, M-D Nazaire, C Williams, M Reich, W Winckler, and G Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, 2012.

[300] N Segata, D Boernigen, TL Tickle, XC Morgan, WS Garrett, and C Huttenhower. Computational metaomics for microbial community studies. *Molecular systems biology*, 9(1):666, 2013.

[301] KT Konstantinidis, A Ramette, and JM Tiedje. The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 361(1475):1929–1940, 2006.

[302] M Schloter, M Lebuhn, T Heulin, and A Hartmann. Ecology and evolution of bacterial microdiversity. *FEMS microbiology reviews*, 24(5):647–660, 2000.

[303] DL Hartl and DE Dykhuizen. The population genetics of *Escherichia coli*. *Annual review of genetics*, 18(1):31–68, 1984.

[304] PA Cotter and VJ DiRita. Bacterial virulence gene regulation: an evolutionary perspective. *Annual Reviews in Microbiology*, 54(1):519–565, 2000.

[305] RW Jackson, E Athanassopoulos, G Tsiamis, JW Mansfield, A Sesma, DL Arnold, MJ Gibbon, J Murillo, JD Taylor, and A Vivian. Identification of a pathogenicity island, which contains genes for virulence and avirulence, on a large native plasmid in the bean pathogen pseudomonas syringae pathovar phaseolicola. *Proceedings of the National Academy of Sciences*, 96(19):10875–10880, 1999.

[306] Antimicrobial Resistance WHO. Global report on surveillance. *Antimicrobial Resistance, Global Report on Surveillance*, 2014.

[307] PM Bennett. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology*, 153(S1):S347–S357, 2008.

[308] T Foster. Staphylococcus. 1996.

[309] S Jarraud, C Mougel, J Thioulouse, G Lina, H Meugnier, F Forey, X Nesme, J Etienne, and F Vandenesch. Relationships between *Staphylococcus aureus* genetic background, virulence factors, agr groups (alleles), and human disease. *Infection and immunity*, 70(2):631–641, 2002.

[310] R Urwin and MCJ Maiden. Multilocus sequence typing: a tool for global epidemiology. *Trends in microbiology*, 11(10):479–487, 2003.

[311] MCJ Maiden, JA Bygraves, E Feil, G Morelli, J E Russell, R Urwin, Q Zhang, J Zhou, K Zurth, DA Caugant, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145, 1998.

[312] M Dreyer, L Aguilar-Bultet, S Rupp, C Guldimann, R Stephan, Al Schock, A Otter, Ge Schupbach, S Brisse, M Lecuit, et al. *Listeria monocytogenes* sequence type 1 is predominant in ruminant rhombencephalitis. *Scientific reports*, 6:36419, 2016.

[313] MD Ismail, I Ali, S Hatt, EA Salzman, AW Cronenwett, CF Marrs, AH Rickard, and B Foxman. Association of *Escherichia coli* ST131 lineage with risk of urinary tract infection recurrence among young women. *Journal of global antimicrobial resistance*, 13:81–84, 2018.

[314] S Jena, S Panda, KC Nayak, and DV Singh. Identification of major sequence types among multidrug-resistant *Staphylococcus epidermidis* strains isolated from infected eyes and healthy conjunctiva. *Frontiers in Microbiology*, 8:1430, 2017.

[315] C-R Usein, AS Ciontea, CM Militaru, M Condei, S Dinu, M Oprea, D Cristea, V Michelacci, G Scavia, LC Zota, et al. Molecular characterisation of human shiga toxin-producing *Escherichia coli* OMLST26 strains: results of an outbreak investigation, romania, february to august 2016. *Eurosurveillance*, 22(47), 2017.

[316] MH Antwerpen, K Prior, A Mellmann, S Höppner, WD Splettstoesser, and D Harmsen. Rapid high resolution genotyping of *Francisella tularensis* by

whole genome sequence comparison of annotated genes ("MLST+"). *PLoS One*, 10(4):e0123298, 2015.

[317] VI Siarkou, F Vorimore, N Vicari, S Magnino, A Rodolakis, Y Pannekoek, K Sachse, D Longbottom, and K Laroucau. Diversification and distribution of ruminant *Chlamydia abortus* clones assessed by MLST and mlva. *PLoS One*, 10(5):e0126433, 2015.

[318] B Heym, M Le Moal, L Armand-Lefevre, and M-H Nicolas-Chanoine. Multilocus sequence typing (MLST) shows that the 'Iberian'clone of methicillin-resistant *Staphylococcus aureus* has spread to france and acquired reduced susceptibility to teicoplanin. *Journal of Antimicrobial Chemotherapy*, 50(3):323–329, 2002.

[319] Y Yamaoka. *Helicobacter pylori* typing as a tool for tracking human migration. *Clinical Microbiology and Infection*, 15(9):829–834, 2009.

[320] Z Qi, Y Cui, Q Zhang, and R Yang. Taxonomy of *Yersinia pestis*. In *Yersinia pestis: Retrospective and Perspective*, pages 35–78. Springer, 2016.

[321] W Wade. Unculturable bacteria—the uncharacterized organisms that cause oral infections. *Journal of the Royal Society of Medicine*, 95(2):81–83, 2002.

[322] S Bhattacharya, N Vijayalakshmi, and SC Parija. Uncultivable bacteria: Implications and recent trends towards identification. *Indian journal of medical microbiology*, 20(4):174, 2002.

[323] M Pinto, V Borges, M Antelo, M Pinheiro, A Nunes, J Azevedo, MJ Borrego, J Mendonca, D Carpinteiro, L Vieira, and et al. Genome-scale analysis of the non-cultivable treponema pallidum reveals extensive within-patient genetic variation. *Nature microbiology*, 2(1):16190, 2017.

[324] D Smajs, M Strouhal, and S Knauf. Genetics of human and animal uncultivable treponemal pathogens. *Infection, Genetics and Evolution*, 61:92–107, 2018.

[325] KW Larssen, A Nor, and K Bergh. Rapid discrimination of *Staphylococcus epidermidis* genotypes in a routine clinical microbiological laboratory using single nucleotide polymorphisms in housekeeping genes. *Journal of medical microbiology*, 67(2):169–182, 2018.

[326] SA Nachappa, SM Neelambike, C Amruthavalli, and NB Ramachandra. Detection of first-line drug resistance mutations and drug–protein interaction dynamics from tuberculosis patients in south india. *Microbial Drug Resistance*, 24(4):377–385, 2018.

[327] SS Khoramrooz, SA Dolatabad, FM Dolatabad, M Marashifard, M Mirzaii, H Dabiri, A Haddadi, SM Rabani, HRG Shirazi, and D Darban-Sarokhalil. Detection of tetracycline resistance genes, aminoglycoside modifying enzymes, and coagulase gene typing of clinical isolates of *Staphylococcus aureus* in the southwest of iran. *Iranian journal of basic medical sciences*, 20(8):912, 2017.

[328] J Sarvari, A Bazargani, MR Kandekar-Ghahraman, A Nazari-Alam, M Motamedifar, et al. Molecular typing of methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* isolates from shiraz teaching hospitals by PCR-rflp of coagulase gene. *Iranian journal of microbiology*, 6(4):246, 2014.

[329] RA Viau, LM Kiedrowski, BN Kreiswirth, M Adams, F Perez, D Marchaim, DM Guerrero, KS Kaye, LK Logan, MV Villegas, et al. A comparison of molecular typing methods applied to *Enterobacter cloacae* complex:

hsp60 sequencing, rep-PCR, and MLST. *Pathogens & immunity*, 2(1):23, 2017.

[330] KA Jolley and MCJ Maiden. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11(1):595, 2010.

[331] M Inouye, H Dashnow, L-A Raven, MB Schultz, BJ Pope, T Tomita, J Zobel, and KE Holt. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11):90, 2014.

[332] R Tewolde, T Dallman, U Schaefer, CL Sheppard, P Ashton, B Pichon, M Ellington, C Swift, J Green, and A Underwood. MOST: a modified MLST typing tool based on short read sequencing. *PeerJ*, 4:e2308, 2016.

[333] MV Larsen, S Cosentino, S Rasmussen, C Friis, H Hasman, R Marvig, L Jelsbak, TS Pontén, DW Ussery, FM Aarestrup, et al. Multilocus sequence typing of total genome sequenced bacteria. *Journal of clinical microbiology*, pages JCM–06094, 2012.

[334] N-F Alikhan, Z Zhou, MJ Sergeant, and M Achtman. A genomic overview of the population structure of salmonella. *PLoS genetics*, 14(4):e1007261, 2018.

[335] A Gupta, IK Jordan, and L Rishishwar. stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics*, 33(1):119–121, 2016.

[336] AJ Page, N-F Alikhan, HA Carleton, T Seemann, JA Keane, and LS Katz. Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microbial genomics*, 3(8), 2017.

[337] H Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[338] P Danecek, A Auton, G Abecasis, CA Albers, E Banks, MA DePristo, RE Handsaker, G Lunter, GT Marth, ST Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

[339] VI Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.

[340] T Wirth, D Falush, R Lan, F Colles, P Mensa, LH Wieler, H Karch, PR Reeves, MCJ Maiden, H Ochman, and et al. Sex and virulence in escherichia coli: an evolutionary perspective. *Molecular microbiology*, 60(5):1136–1151, 2006.

[341] A Koehler, H Karch, T Beikler, TF Flemmig, S Suerbaum, and H Schmidt. Multilocus sequence analysis of porphyromonas gingivalis indicates frequent recombination. *Microbiology*, 149(9):2407–2415, 2003.

[342] R Leinonen, H Sugawara, M Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[343] M Enersen, I Olsen, AJ van Winkelhoff, and DA Caugant. Multilocus sequence typing of porphyromonas gingivalis strains from different geographic origins. *Journal of clinical microbiology*, 44(1):35–41, 2006.

[344] F Alhashash, X Wang, K Paszkiewicz, M Diggle, Z Zong, and A McNally. Increase in bacteraemia cases in the east midlands region of the uk due to mdr escherichia coli ST73: high levels of genomic and plasmid diversity in causative isolates. *Journal of Antimicrobial Chemotherapy*, 71(2):339–343, 2015.

[345] T Cohen, PD van Helden, D Wilson, C Colijn, MM McLaughlin, I Abubakar, and RM Warren. Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clinical microbiology reviews*, 25(4):708–719, 2012.

[346] M Dzunkova, A Moya, X Chen, C Kelly, and G D'Auria. Detection of mixed-strain infections by facs and ultra-low input genome sequencing. *Gut microbes*, pages 1–5, 2018.

[347] KE Raven, T Gouliouris, J Parkhill, and SJ Peacock. Genome-based analysis of enterococcus faecium bacteremia associated with recurrent and mixed-strain infection. *Journal of clinical microbiology*, 56(3):e01520–17, 2018.

[348] G Yu, D Fadrosh, JJ Goedert, J Ravel, and AM Goldstein. Nested pcr biases in interpreting microbial community structure in 16s rrna gene sequence datasets. *PLoS One*, 10(7):e0132253, 2015.

[349] M Schirmer, UZ Ijaz, R D'Amore, N Hall, WT Sloan, and C Quince. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. *Nucleic acids research*, 43(6):e37–e37, 2015.

[350] D-L Sun, X Jiang, QL Wu, and N-Y Zhou. Intragenomic heterogeneity in 16s rrna genes causes overestimation of prokaryotic diversity. *Applied and environmental microbiology*, pages AEM–01282, 2013.

[351] K Kennedy, MW Hall, MDJ Lynch, G Moreno-Hagelsieb, and JD Neufeld. Evaluating bias of illumina-based bacterial 16s rrna gene profiles. *Applied and environmental microbiology*, pages AEM–01451, 2014.

[352] R Poretsky, LM Rodriguez, C Luo, D Tsementzi, and KT Konstantinidis. Strengths and limitations of 16s rrna gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One*, 9(4):e93827, 2014.

Samenvatting

Dankzij de ontwikkelingen in sequentietechnieken zijn metagenomen een rijke bron van informatie geworden voor vele wetenschappelijke disciplines zoals menselijke en dierlijke gezondheidszorg, ecologie, forensisch onderzoek, landbouw en voedselproductie. Een gedetailleerde analyse van metagenomische data is daarom van groot belang om alle aanwezige informatie te onthullen. Hierbij proberen wetenschappers meestal het antwoord te vinden op drie hoofdvragen:

- Welke organismen zijn aanwezig in het metagenoom?
- Wat doen ze daar?
- Wat is het verschil tussen metagenomen?

Traditioneel worden de antwoorden op de eerste twee vragen verkregen door middel van zogeheten "referentie-gebaseerde methoden", waarbij metagenomische data eerst vergeleken wordt met bekende genomen, genen of reactieketens. Een duidelijk nadeel van deze technieken is de onvolledigheid van bestaande databases: microbiële gemeenschappen bestaan veelal uit honderd tot duizenden onbekende bacteriën, omdat informatie over deze bactieriën ontbreekt is de nauwkeurigheid van referentie-afhankelijke methoden beperkt. Daarom worden referentie-vrije methoden populairder in de vergelijkende metagenomica. In mijn onderzoek tracht ik de metagenomische analyse te verbeteren in twee richtingen: mét en zonder referentie-databases (zie hoofdstuk 3 en 4).

Voor de referentie-vrije analyse van verscheidene Next Generation Sequencing datasets ontwikkelden wij een methode gebaseed op *k*-meren (kPal). We laten zien dat onze aanpak gebruikt kan worden voor twee soorten metagenomische analyse: om het niveau van verwantschap tussen twee microbiomen te kwantificeren (hoofdstuk 3), en om de genetische informatie binnen één metagenoom te classificeren (hoofdstuk 4). We hebben kPal getest op een reeks gesimuleerde metagenomen met verschillende aantallen van nauw verwante bacteriële genomen. Onze methode bleek in staat tijdelijke verandering in microbiotische compositie te detecteren. Om

te controleren of deze referentie-vrije methode het verschil tussen menselijke metagenomen kan blootleggen, hebben we onze methode ook getest op 16S metagenomen van ingewanden en de huid van verschillende testpersonen over een periode van 6 maanden. kPal kan niet alleen het verschil zien tussen de afkomst (ingewanden of huid) van het metagenoom, het kan ook het onderscheid zien tussen de verschillende testpersonen! Dit resultaat is beter dan referentie-afhankelijke methoden laten zien, die namelijk niet de huid-monsters van verschillende personen kunnen onderscheiden.
We hebben onze op *k*-meren gebaseerde methode ook toegepast om genetische sequenties te classificeren in één metagenomische dataset. Naast een aantal gesimuleerde metagenomische datasets hebben we ook data verkregen van een bioreactor microbioom met behulp van het PacBio RSII platform. We laten zien dat de *k*-mer profielen relaties kunnen onthullen tussen genetische sequenties in een enkel metagenoom, waarmee we de sequenties kunnen clusteren per soort. Deze resultaten zijn zeer belangrijk, omdat ze bewijzen dat het mogelijk is om structuren te detecteren binnen een enkel metagenoom met slechts de informatie die in het metagenoom zelf beschikbaar is. Onze referentie-vrije methode kan dus gebruikt worden voor vergelijkende metagenomica. Bovendien kunnen we sequenties in een enkel metagenoom classificeren, waardoor we de in een monster aanwezige genomen kunnen ontwaren.

Daarnaast hebben we de grenzen van referentie-afhankelijke technieken onderzocht in enkele studies (hoofdstuk 2 en 5).
Ons eerste doel was om de twee meest populaire datasoorten voor referentie-afhankelijke taxonomische profilering te vergelijken: de amplicon-gebaseerde 16S data versus de Whole Genome Sequencing (WGS; volledige genoom-sequentie) data (hoofdstuk 2). Voor dit onderzoek creërden wij een reeks kunstmatige bacteriële mengsels, elk met een andere verdeling van soorten. Deze mengsels werden gebruikt om de nauwkeurigheid van de twee datasoorten te bepalen, en om verscheidene methoden voor taxonomische classificatie te evalueren. Onze resultaten laten zien dat WGS-data veel nauwkeurigere resultaten oplevert dan 16S data. Daarmee verwerpen we dat wijdverbreide mening dat 16S data toereikend is voor de analyse van metagenomische monsters.
Tot slot hebben we de toepasbaarheid van referentie-afhankelijke methoden vergroot door een pipeline te maken die klinische monsters kan analyseren met mogelijk meer dan één pathogeen (hoofdstuk 5). Hiervoor ontwikkelden we BacTag, een gedistribueerde bioinformatica pipeline voor een snelle en accurate typering van bacteriële genen en allelen in klinische WGS-data. Het grote voordeel van onze methode bestaat uit een voorbewerkingsprocedure waarin de signatuur van elk mogelijk allel wordt geïdentificeerd en opgeslagen in een database. De daaropvolgende identificatie van allelen in een klinisch monster wordt gedaan aan de hand van deze signaturen in plaats van een traditionele uitputtende zoektocht. Omdat deze

methode ook toegepast kan worden op ongecultiveerde monsters, kan de methode goed gebruikt worden voor gevallen waar een snelle analyse van belang is.

Publications

- S. Y. Anvar, L. Khachatryan, M. Vermaat, M. van Galen, I. Pulyakhina, Y. Ariyurek, K. Kraaijeveld, J. T. den Dunnen, P. de Knijff, P. A. C. 't Hoen, and J. F. J. Laros
  Determining the quality and complexity of next-generation sequencing data without a reference genome
  Genome Biology, 2014 15:555 doi 10.1186/s13059-014-0555-3

- L. Khachatryan, M. E. M. Kraakman, A. T. Bernards, and J. F. J. Laros
  BacTag - a pipeline for fast and accurate gene and allele typing in bacterial sequencing data
  BMC Genomics, 2019 20:338 doi 10.1186/s12864-019-5723-0

- L. Khachatryan, R. H. de Leeuw, M. E. M. Kraakman, N. Pappas, M. te Raa, H. Mei, P. de Knijff, and J. F. J. Laros
  Taxonomic classification and abundance estimation using 16S and WGS - a comparison using controlled reference samples
  Forensic Science International: Genetics, 2020 46:102257
  doi 10.1016/j.fsigen.2020.102257

- L. Khachatryan, S. Y. Anvar, R. H. A. M. Vossen, and J. F. J. Laros
  Reference-free resolving of long-read metagenomic data
  bioRxiv 2019 https://doi.org/10.1101/811760

Acknowledgements

I wish to thank my dear husband Louk Rademaker, who was my landmark on the dark path I took to finally finish this thesis. This journey would be sad and boring without you.
My great appreciation goes to my mother Ludmila Khachatryan and my departed father Artavazd Khachatryan, for their unconditional believe in me and constant reminders how miserable my life without PhD title would be. My brothers, Arsen Khachatryan and Levon Khachatryan, thank you for being such great supporters and for providing me with countless dark humour jokes regarding processes slowing down the defence procedure. My large Russian-Armenian family deserves a great appreciation for their genuine trust in my intellectual abilities.
I would like to thank my parents-in-law, Ruth Noorduyn and Jan Rademaker, who would always lend me a hand in case of trouble. My family-in-law (both Rademaker and Noorduyn sides), thank you all for making me feel a true part of such a great and interesting folk.
I would like to thank Chana for being my first and best Dutch friend. Special appreciation goes to Irina for setting a great example and inspiring me both scientifically and personally. I am eternally grateful to Ivo who was covering my back myriad of times (mostly due to my absolutely brilliant knack for finding trouble), and, of course, for the extended number of useful advices and hints. My dear friends Svetlana and Lena (Beletkaia), thank you for all the good time we've spent together during my time as a PhD student. Nadya and Lena (Cherniogloglo), my friends from the time in University, thank you for sharing my good and bad moments and accommodating me in your home and in your hearts.
I wish to acknowledge the support provided by the fellow colleagues from LUMC. I am particularly grateful to Igor Sidorov and Alexander Gorbalenya, my MS work supervisors, who kept being interested in my scientific career and provided me with an advice and support even after I left their group.
I would like to offer my special thanks to Alessandra Sequeira who kindly allowed me to use her artwork for this thesis cover. It was truly a pleasure to browse through many of her works to find one matching the spirit of my research.

Curriculum vitae

Lusine Khachatryan was born on February 15th 1990 in Jermuk, USSR (currently Armenia). She moved together with her family to Svoboda, Kursk district, Russia in early 1994. She graduated with honours from Svoboda secondary general education school in 2007. Same year she was admitted to Lomonosov Moscow State University (the School of Bioengineering and Bioinformatics). During first three years in University she was working as a bioengineer intern in the Belozersky Institute of Physiochemical Biology (Moscow, Russia) investigating the fragmentation of Potyvirus A coat protein by plant caspase-like protein and learning to create genetically modified plants using *Agrobacterium*-mediated transformation. In a year of 2010 Lusine was one of the 10 students selected for one-month bioinformatics internship in Leiden University Medical Center (Leiden, The Netherlands), after which she decided to continue her scientific career as bioinformatician. She spent one year as an bioinformatics intern in the Institute for Genetics and Selection of Industrial Microorganisms (Moscow, Russia) studying the binding sites of transcriptional factors specific for bidirectional promoters with different tissue expression pattern. Her MS project was dedicated to design and $in-silico$ validation of serotype-specific polymerase chain reaction for human rhino- and enteroviruses and was performed as a collaboration between Lomonosov Moscow State University (Moscow, Russia) and Leiden University Medical Center (Leiden, The Netherlands). She graduated from Lomonosov Moscow State University with honours in 2012, her MS thesis work was specifically acknowledged by the University defence committee. In August 2012 Lusine continued her academic career as a PhD student in the department of Human Genetics in Leiden University Medical Center (Leiden, The Netherlands). Her PhD research was dedicated metagenomics - new and rapidly developing branch of molecular microbiology. Particularly, she developed several approaches and investigated the limits of various already existing methods for metagenomics analysis regarding different types of sequencing data. This work resulted a number of publications and was presented at many national and international conferences. From September 2018 Lusine is hired as a Scientist in R&D facility of Philip Morris International (Neuchatel, Switzerland) where she is working on improving the anal-

ysis pipelines for metagenomic and metatranscriptomic data as well as investigating the changes of microflora in health and disease.