Metagenomics : beyond the horizon of current implementations and methods

Khachatryan, L.

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.