



# **Statistical models for large-scale comparative metagenome analysis**

## **Dissertation**

for the award of the degree  
„Doctor rerum naturalium“  
of the Georg-August-Universität Göttingen  
within the doctoral program  
„Molecular Biology of Cells“  
of the Georg-August-University School of Science (GAUSS)

submitted by

**Kathrin Petra Abhauer**

from

**Bad Karlshafen**

Göttingen 2015



## **Members of the Thesis Committee**

**Prof. Dr. Burkhard Morgenstern**

1<sup>st</sup> Referee

Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-Universität Göttingen

**Prof. Dr. Edgar Wingender**

2<sup>nd</sup> Referee

Institute of Bioinformatics, Center for Informatics, Statistics and Epidemiology UMG, Georg-August-Universität Göttingen

**Prof. Dr. Lutz Walter**

Primate Genetics Laboratory, German Primate Center, Leibniz Institute for Primate Research

**Date of the oral examination:** February 19<sup>th</sup>, 2015



## **Affidavit**

I hereby confirm that this thesis has been written independently and with no other sources and aids than quoted.

Göttingen, January 2015

Kathrin Petra Aßhauer



# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Progress of culture-independent analysis strategies for whole microbial communities	5
1.1.1	The amplicon-based approach	5
1.1.2	The whole-metagenome sequencing approach	6
1.1.3	The impact of high-throughput next-generation sequencing (NGS) on metagenomic research	6
1.2	Analysis of 16S rRNA data	7
1.2.1	Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences	8
1.3	Analysis of whole-metagenome sequence data	9
1.3.1	Predicting the functional and metabolic capabilities of microbial communities from metagenomic shotgun data	10
1.4	Comparative analysis of microbial communities	16
1.4.1	Example: The human microbiome	17
1.4.2	Identification of closely related metagenome datasets	18
1.5	Objective and overview	19
<b>2</b>	<b>On the estimation of metabolic profiles in metagenomics</b>	<b>21</b>
<b>3</b>	<b>Exploring Neighborhoods in the Metagenome Universe</b>	<b>37</b>
<b>4</b>	<b>Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data</b>	<b>55</b>
<b>5</b>	<b>Discussion</b>	<b>63</b>
5.1	Predicting the functional and metabolic potential from the taxonomic profile	64
5.1.1	Mixture-of-Pathways: a probabilistic model for the quantification of the metabolic potential from metagenomic shotgun data	64
5.1.2	Tax4Fun: predicting the functional potential of microbial communities using 16S rRNA sequence data	65
5.2	Limitations of the prediction of functional capacities using the taxonomic profile for approximation	67
5.2.1	Measuring the reliability of taxonomic and functional predictions	67
5.2.2	The need for reference genomes: closing the gap of the unknown microbial diversity	68

5.2.3	Beyond the functional and metabolic potential . . . . .	69
5.3	Benefits of the unsupervised identification of related metagenomes . . . . .	71
5.4	CoMet-Universe - a web-server for comparative analysis of metagenomes based on protein domain signatures . . . . .	73
<b>6</b>	<b>Summary and conclusion</b>	<b>75</b>
	<b>Acronyms</b>	<b>105</b>

# Abstract

Metagenomics, as a culture-independent approach, enables the exploration of complex heterogeneous microbial communities under natural conditions by massive sequencing of community-specific DNA. Metagenomic data sets, derived from various environments, provide new insights into microbial life. Large-scale projects like the Human Microbiome Project or the Earth Microbiome Project emphasize the increasing importance of metagenomics for biomedical and ecosystem research. However, such projects are currently challenging bioinformatics due to the explosive increase in sequencing data. New computationally efficient and statistically adequate methods are required to answer the essential questions “Who is in there?” and “What are they doing?”.

In this thesis, I developed the Mixture-of-Pathways (MoP) model and Tax4Fun approach. Both methods link the taxonomic profile to a set of pre-computed reference profiles to predict the metabolic repertoire of the microbial community. Since the taxonomic profile is normally estimated to answer the question “Who is in there?”, the further use of the taxonomic profile avoids additional costs for answering the question “What are they doing?”. Tax4Fun is specifically designed for the output of 16S rRNA analysis pipelines using the SILVA database as reference, whereas the MoP model is especially conceived for metagenome sequence data and provides a robust statistical basis to describe the metabolic potential of a microbial community. The adequate metabolic modeling of metagenomes provides a concise summary of the functional variation of metagenomes across many samples, enabling the identification of relevant metabolic differences in comparative analyses.

For comparative metagenomics, the identification of similar metagenomes to a newly obtained dataset is of growing importance. For an efficient large-scale identification of closely related metagenomes within a database retrieval context, I conducted a detailed evaluation of a  $k$ -nearest-neighbor search utilizing different biological feature profiles and metrics. I demonstrated that different features and metrics can be chosen for a convenient interpretation of results in terms of the underlying features. The integration of the  $k$ -nearest-neighbor search into metagenome annotation and comparison systems is beneficial to automatically identify additional metagenomes for comparative analyses as well as to detect mislabeled or contaminated datasets by unexpected neighboring habitat labels.

The MoP approach and  $k$ -nearest-neighbor search are available to the scientific community as part of the CoMet-Universe web server application. Additionally, the MoP and Tax4Fun approach are provided as R Package.



# 1 Introduction

Microorganisms were the first form of life about 3.5 billion years ago [1]. They are genetically and metabolically highly diverse and exist across a wide range of physiological conditions where they are important members of their ecosystems [2–4]. After their discovery through a microscope by Antoni van Leeuwenhoek in the 1670s [5], microorganisms have been intensively studied and much has been learned about them. With the contribution of Robert Koch in the late 1800s [6], cultivation and isolation became the gold standard for the identification and characterization of microorganisms [7, 8]. The diversity and functions of single bacterial or archaeal species were discovered by techniques such as direct observation of the cells, biochemical testing, and differential staining. In the 20<sup>th</sup> century, genomics has evolved as an effective technology to study microorganisms and their genetic material [9, 10]. However, all these techniques require prior cultivation under laboratory conditions.

In 1985, Staley and Konopka [11] revealed that the number and diversity of cells observed microscopically far exceeded those of cells grown in culture, a phenomenon that became known as the “Great Plate Count Anomaly”. Culture-based methods are biased towards microorganisms easily grown with standard culturing techniques. The majority of microorganisms requires growth conditions that are so far unknown or difficult to obtain in the laboratory [12]. Generally, less than 1% of the microbial diversity is estimated to be cultivable [13–18]. For example, although human-associated microorganisms were among the first investigated by microscopy, most organisms are difficult to culture under convenient laboratory conditions. In a typical human gut microbiome, *Escherichia coli* accounts for at most 5% of the microorganisms [19]. The vast majority of the remaining microbial species have never been grown in the laboratory [20]. Furthermore, cultivability of microorganisms is heavily dependent on the habitat. For example, in a freshwater lake up to 10% [21] and in a marine tidal sediment 23% [22] of the microorganisms can be cultivated, whereas in natural soil and aquatic communities only a minor fraction has yet been grown in culture and characterized [23–28]. Therefore, cultivated microorganisms lead to a narrow picture of the variety of microorganisms and provide only a glimpse of the total microbial diversity. In addition, microbial communities consist of hundreds of different species. For example, one gram of soil may contain  $10^{10}$  bacteria and a ton could consist of  $4 \times 10^6$  different taxa [15, 28]. In seawater the number of bacteria is approximately  $10^6$  per milliliter from a total of about  $2 \times 10^6$  bacterial taxa [15] and the human microbiota consists of about  $10^{14}$  microbial cells that outnumber the human cells 10 to 1 [29]. This enormous number of microorganisms cannot be analyzed in appropriate time using traditional genomic and approaches in microbiology, which are therefore unsuitable to study entire microbial communities residing in a particular environment [30, 31].

A new era of microbial ecology was initiated by advances in molecular biology which allow to access the genetic material without laboratorial cultivation [32]. For studying heterogeneous microbial communities in their natural habitats, cultivation-independent approaches extract a comprehensive set of marker genes or genomes which are then sequenced and analyzed. The analysis aims in particular for a quantitative description of the microbial community under investigation. The taxonomic composition gives valuable insights into the diversity of different habitats. However, inferring which organisms are present in a sample alone does not necessarily lead to an understanding of the potential biological activity of the microbial community. The functional and metabolic capabilities of a microbial community indicate the ability to carry on functioning when conditions change. Thus, two major objectives are addressed by the investigation of microbial communities without prior cultivation: Determining the taxonomic composition (“Who is in there?”) and functional and metabolic capabilities (“What are they doing?”). Besides these two major objectives, regarding the comprehensive description of a single microbial community, the increasing number of datasets for various conditions or environments provide a wealth of opportunities for comparative studies. These studies cover the question how the findings regarding a single community compare with other communities. The exploration of similarities and dissimilarities of different samples can be summarized with the question “What are the differences that make a difference?”. In order to answer these questions, appropriate computational methods are required. The bioinformatical analysis of heterogeneous microbial communities faces a huge challenge due to the rapidly growing amount of sequencing data. Currently, the bottleneck in metagenomic studies is the analysis of the sheer amount of data and not the sequencing itself [33, 34].

In my thesis, I focus on the computationally efficient quantitative functional and metabolic description of microbial communities. Furthermore, I address the identification of appropriate samples for comparative analysis. For a better understanding of the underlying data and the resulting difficulties in the analysis, I start with a synopsis of the development of the 16S ribosomal ribonucleic acid (rRNA) gene and metagenomic sequencing approach for analyzing microbial communities (see Section 1.1). Following this, I briefly introduce existing analysis strategies for the 16S rRNA gene and metagenomic sequencing approach (see Section 1.2 and Section 1.3). Here, I review bioinformatic approaches particularly addressing the question: “What are they doing?”. In Section 1.3.1, I outline the state-of-the-art approaches for functional and metabolic profiling in metagenomics. In the first instance, I focus on the quantitative functional description of microbial communities since the resulting abundances are crucial for the subsequent metabolic characterization. Especially, I discuss the computational drawbacks using homology search for functional annotation in metagenomics (see Section 1.3.1.1), which further affects metabolic profiling. Thereafter, I summarize existing approaches for pathway reconstruction in genomics and pathway profiling in metagenomics and address the inherent methodological drawbacks (see Section 1.3.1.2). In particular, I stress the handling of the ambiguity of the function-to-pathway mapping, which can lead to an incorrect quantification and motivated me to develop a statistically adequate estimation of the metabolic potential (see Chapter 2).

Further, I address comparative metagenomic studies highlighting benefits from previous studies. Finally, I motivate an efficient large-scale identification of similar metagenomes within a database retrieval context, which represents an additional important part of my thesis.

## 1.1 Progress of culture-independent analysis strategies for whole microbial communities

In the following sections, I introduce the development of two different approaches for the investigation of microbial communities without prior cultivation. First, I treat the amplicon-based approach targeting a phylogenetic marker region for analyzing the taxonomical composition. Then, I proceed to the whole-metagenome sequencing approach for characterizing the whole collection of genes of a microbial community. Finally, I address the impact of next generation sequencing technologies on both approaches.

### 1.1.1 The amplicon-based approach

In the 1980s, the idea of studying microorganisms by the comprehensive extraction and analysis of mixed microbial nucleic acid (deoxyribonucleic acid (DNA) or ribonucleic acid (RNA)) from environmental samples was introduced by Pace and colleagues [35] and applied by Schmidt et al. [36] to a marine picoplankton community (see also [20, 37–39]). They amplified the well-conserved small subunit (SSU) 16S rRNA gene sequence selectively from total DNA extracted from an environmental sample. Subsequent to isolation and amplification by polymerase chain reaction (PCR), the sequences were cloned and sequenced. Following this, they utilized the established framework for bacterial phylogeny by Woese and Fox [40] for taxonomic classification. At that time, the analysis was restricted to certain phylogenetic markers due to time-consuming and costly sequencing. However, remarkable, unexpected insights and discoveries regarding the microbial diversity of environmental samples were gained [41–44].

This earliest high-throughput technique, known as amplicon profiling, is a powerful tool and still the most common technique for assessing and comparing the taxonomic composition of environmental microbial communities [45]. Amplicon-based analysis of community structure is focused on a targeted phylogenetic marker region. This region has to be ubiquitous in the taxonomic range of interest and variable enough to discriminate between different species. Additionally, the region must be flanked by highly conserved sequences for the purpose of amplification. Several phylogenetic markers have been defined, including ribosomal protein subunits, elongation factors, and RNA polymerase subunits [46, 47]. However, the 16S rRNA gene has been established as the gold standard for taxonomic assessment of prokaryotic microbial communities [45].

### 16S rRNA

The 16S rRNA gene is a component of the prokaryotic SSU ribosome. The 1.5 kilo base pairs (bp) multi-copy 16S rRNA gene contains both highly conserved, ubiquitous regions and nine short species-specific hypervariable regions that allow reliable reconstruction of phylogeny [48]. The conserved regions are well-suitable for PCR amplification using broad-range primers [41, 45]. Sequence variations in the 16S rRNA genes are considered to reflect the universal molecular clock of life [49]. 16S rRNA gene sequences allow to relate any organism, even uncultured and distantly related, to cultivated and characterized strains based on a single phylogenetic tree of life [50, 51]. Further, the abundances of 16S rRNA sequences are used to describe microbial

communities in terms of species richness, evenness, or related microbial diversity measurements [52].

### **1.1.2 The whole-metagenome sequencing approach**

Phylogenetic marker sequencing provides only limited functional information. At the end of the 20<sup>th</sup> century, as sequencing became cheaper, more than just phylogenetic marker genes could be studied. The improvements in sequencing techniques led to the development of the so-called whole-metagenome sequencing approach. In 1998, Jo Handelsman used the term “metagenome” to describe the collective genomes obtained from soil microflora and coined the term “metagenomics” [53]. In metagenomics, a set of genomes from an environmental community is sampled randomly by means of cultivation-independent methods. Metagenomics goes beyond 16S rRNA gene-based characterization of microbial communities by determining directly the whole collection of genes within an environmental sample. The investigation of the whole collection of genes allows the prediction of potential functions of the collective organisms in a sample.

In 2004, two research groups published results from different large-scale environmental sequencing projects. Tyson et al. [54] conducted a metagenomic analyses of a microbial community from an acid-mine drainage environment and Venter et al. [55] characterized a much more complex community of oceanic microbial assemblages from the Bermuda Atlantic Time-series Study site in the Sargasso Sea. After these two landmark pioneering studies, metagenomes from a large variety of environments such as soils [56, 57], marine sites [58, 59], human and mouse gastrointestinal tract [60–62], and extreme environments [63–65] have been sequenced. These studies gave valuable insights into microbial genomic diversity and functional complexity. Thus, culture-independent metagenomic approaches complement traditional culture-based techniques by providing a comprehensive view of the microbial communities.

### **1.1.3 The impact of high-throughput next-generation sequencing (NGS) on metagenomic research**

Since the initial study of Handelsman et al. [53], metagenome analysis has altered from time-consuming, labor-intensive sequencing of cloned DNA fragments to direct sequencing of DNA without heterologous cloning [66]. In 2005, with the advent of next-generation sequencing (NGS) [67, 68], both the amplicon- and the whole-metagenome shotgun-approach have considerably changed [69]. Using NGS, hundreds of microbial communities can be run simultaneously yielding tens of thousands of tag sequences per sample at reasonable cost [70]. But despite decreasing costs for metagenomic sequencing, they are still an order of magnitude higher compared to amplicon sequencing. Additionally, metagenomic sequencing provides lower throughput than amplicon sequencing. Since the 16S rRNA approach is relatively cheap and easy to carry out, it has been widely used to characterize microbial communities across thousands of ecosystems [71–73]. Occasionally, 16S rRNA studies are called “metagenomic”, too, as they also analyze a heterogeneous sample of community DNA [74]. However, amplicon sequencing is limited to analyze microbial community structure and does not provide insights into the functional repertoire. Therefore, the 16S rRNA approach has to be combined with whole metagenome sequencing,

which provides additional information about the collective functions and metabolism.

## **From analyzing single communities to large-scale studies**

The 16S rRNA and metagenome sequencing approach have evolved as helpful disciplines to unlock the composition of heterogeneous microbial communities in their natural habitats [45]. These analyses provide insights into how microorganisms adapt to or reshape their environment [75]. Inter-environmental metagenomic dataset comparisons can provide further insights. Here, the studies of Tringe et al. [76] and Dinsdale et al. [77] have given rise to the sequencing of multiple samples for different conditions or environments in order to determine similarities and differences. The differences in cost and explanatory potential have led to a proliferation of two-stage study designs for large scale environmental microbial community analysis. In a two-stage study, first a comprehensive collection of microbial community samples is efficiently surveyed by sequencing the 16S rRNA gene. In this step, the microbial community structure is determined with considerable sensitivity and depth of coverage in a cost-effective manner. In the second step, a subset of microbial communities or samples with particular characteristics is selected by experts for targeted in-depth profiling approaches [63, 78–80]. These approaches comprise metagenomic sequencing as well as other functional assays. Two prominent large-scale studies for this two-stage design are the Human Microbiome Project (HMP) [81] and the Earth Microbiome Project (EMP) [82, 83]. These and other microbial community studies are increasingly turning to larger sample sizes using sequencing of the 16S rRNA taxonomic marker gene [84].

Although metagenomics has great potential for new insights into the hidden world of microorganisms, many challenges still remain before this potential can be realized. DNA sequence data generation is no longer the bottleneck in microbial studies. Even laboratories with comparatively small budgets can afford metagenomic sequencing projects nowadays. In contrast, increased data volumes are posing significant challenges to the existing analysis tools [34] and even the computing of sequence similarities represents a limiting factor in metagenome analysis [85, 86].

## **1.2 Analysis of 16S rRNA data**

The microbial diversity of microbial environments can be assessed with high phylogenetic resolution by deep sequencing of particular phylogenetic marker genes. The 16S rRNA gene is the gold standard for phylogenetic studies [13, 56]. During sample preparation using primer-based amplification steps, sequences are targeted by means of taxonomically universal primers. The universal primers target conserved regions in the 16S rRNA gene. However, the choice of hyper-variable region [84, 87–89] and the amplification itself introduce bias [90–94]. Some primers do not bind in all taxa with sufficient stringency leading to a biased amplification [95, 96]. In addition, prokaryotes and eukaryotes need to be analyzed separately due to their basic difference in rDNA sequence. Further, only genomes of organisms with amplicon target genes can be captured - excluding, for example, viruses.

After amplification, the resulting amplicons are generally subjected to high-throughput sequencing by NGS platforms. Due to the inherent sequencing read length of NGS, a shorter region of the sequence must be selected to act as proxy for the full-length sequence [97, 98]. Previous

studies revealed that for many 16S rRNA microbial community studies, a read length of 250 bp is sufficient [88, 97, 99]. Commonly, studies target between one and three of the hyper-variable regions. However, there is currently no consensus on a single best region. Consequently, different groups are sequencing different regions complicating direct comparisons among studies [74].

Following the sequencing, several data processing steps have to be conducted. For analysis of 16S rRNA gene sequence data a variety of software pipelines exist (e.g Quantitative Insights Into Microbial Ecology (QIIME) [100], mothur [101], and SILVAngs [102]). The initial processing includes read quality and read length filtering [89]. Further, chimeric sequences and other PCR artifacts are identified and removed [103].

Afterwards, the 16S rRNA-based analysis of microbial communities relies on the construction of similarity clusters to infer the abundance of the composing species. For this purpose, some degree of sequence divergence is typically allowed. In practice, 95%, 97%, or 99% sequence identity cutoffs are often used [89]. The resulting sequence clusters are also referred to as operational taxonomic unit (OTU). The assignment of sequences to OTUs is referred to as binning and allows a computationally tractable representation for biological analysis. From each OTU a representative 16S rRNA gene sequence is used for taxonomic assignment. The taxonomy is assigned using either the Basic Local Alignment Search Tool (BLAST) [104] or the Ribosomal Database Project Classifier (RDP Classifier) [105]. For the taxonomic assignment, reliable reference databases and taxonomies are crucial. Several resources provide phylogenetic classification of publicly available 16S rRNA gene sequences. These resources include the greengenes [106], the Ribosomal Database Project (RDP) [107], and the SILVA [102] databases. Subsequent to the assignment of OTUs to taxonomic identities based on a reference database, further downstream ecological and diversity analyses can be carried out. Here, for example, the OTU abundances can be used to describe the microbial communities in terms of species richness, evenness, or related microbial diversity measurements [52].

However, the functional inventory of the communities can generally not be assessed when restricting analyses to the 16S rRNA gene. The genes as well as the functional and metabolic repertoire of a community are sometimes indirectly inferred [108–110]. Here, the inference is based on available genome sequences of the community members. However, for most microbial organisms the genome and thus the functional and metabolic repertoire is unknown. Recently, an approach called Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) [111] was proposed to predict functional profiles of microbial communities using 16S rRNA gene sequences and a database of reference genomes.

### **1.2.1 Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences**

The PICRUSt approach comprises two steps: the gene content and the metagenome inference step. In the gene content inference step, the gene family abundances and 16S rRNA copy number for each organism in a 16S rRNA-based phylogeny is pre-computed. PICRUSt applies an extended ancestral state reconstruction algorithm to infer the functional abundances and 16S rRNA copy numbers for organisms that have not yet been sequenced. Typically, ancestral state reconstruction algorithms are intended to recover states from ancestral organisms from measured characteristics of living organisms [112, 113]. Here, the extrapolation back in time is achieved by fitting evolu-

tionary models to the distribution of observed states using criteria such as maximum likelihood or Bayesian posterior probability [114–117]. In PICRUSt, ancestral state reconstruction algorithms are extended to infer the content of microorganisms for which no genome sequence is available by using the predicted character states from ancestral genomes and available reference genomes. To account for the distance to the next relative, a weighting scheme is implemented based on the phylogenetic distances. Per default, the greengenes phylogenetic tree of 16S rRNA gene sequences is used for phylogenetic information. Further, the Integrated Microbial Genomes (IMG) [118] database is utilized for retrieval of 16S rRNA copy number information and functional annotation in terms of KEGG Orthology (KO) [119, 120] and Clusters of Orthologous Groups (COGs) [121]. In total, 2,590 genomes with identifiers in the greengenes reference tree were used for the initial computationally intensive genome prediction step. The subsequent metagenome inference step uses the abundances of OTUs as input, e.g., as obtained from QIIME analysis. First, the OTU abundances are normalized using the pre-computed 16S rRNA copy number. Then, the normalized OTU abundances are combined with the inferred functional abundances during the gene content inference step. The resulting estimated metagenomic functional profile can be further utilized to infer the metabolic profile.

So far, PICRUSt is the only approach which addresses the limitation of 16S rRNA data to infer the taxonomic composition, showing that it is possible to predict functional profiles from 16S rRNA data in principle. But, PICRUSt predictions depend heavily on the topology of the phylogenetic tree and the inferred ancestral genomes. Especially, the lack of closely related genomes increases uncertainty in the predictions of ancestral character states. Thus, the distance to the next sequenced organism along the branches of the tree strongly influences the prediction of gene content for the tips in the tree. Since there is always a neighbor in the tree topology, PICRUSt links all OTUs, even if distances are large. This procedure can be problematic when analyzing microbial communities with a large proportion of so far not well-characterized phyla. Furthermore, different ancestral state reconstruction methods can lead to strikingly different reconstructions. Therefore, the predicted gene content can be regarded as an estimate at best.

In this work, I introduce the novel “Tax4Fun” approach which efficiently predicts the functional repertoire of a microbial community from 16S rRNA data and can directly be applied to output from the SILVAngs web server, QIIME, or any other analysis pipeline using the SILVA database as reference. Tax4Fun is entirely based on close homologies between 16S rRNA gene sequences and differs substantially from the PICRUSt approach that relies on a phylogenetic tree of the 16S sequences (see Chapter 4).

### **1.3 Analysis of whole-metagenome sequence data**

Metagenomic sequencing projects study the entire genetic material from aggregated members of the community. The sequencing results in a mixture of fragmented sequences originating from the community members. Unlike in amplicon metagenomics, whole-metagenome sequence data can provide both community composition and insights into the whole gene content of the microbial community. Thus, both fundamental questions “Who is in there?” and “What are they doing?” can be addressed [122]. However, the metagenomic samples typically contain huge collections of short sequence reads from hundreds to thousands of species, most with unknown

phylogenetic origin [123]. Though NGS methods can be applied faster, more easily, and at lower cost per base and per read than Sanger sequencing, there is a tradeoff in information quality. The cheaper sequencing comes at the cost of reduced sequence length, resulting in a sparse gene annotation. In Hamady and Knight [74], different metagenome approaches are discussed revealing the strengths and weaknesses of the different experimental approaches, sequencing methodologies, and analytical methods.

The first steps after sequencing comprise filtering for quality and control of the obtained reads [124, 125]. In case of host-associated samples, the reads have to be further screened for residual host DNA [126]. After preprocessing, the analysis comprises estimation and comparison of the taxonomic composition and the functional repertoire of the microbial community [127]. However, metagenomic analysis is heavily affected by inherent problems of the data itself including incomplete coverage, generally short read lengths, fragmented sequences from multiple species, and the sheer amount of data [125, 128, 129].

### **1.3.1 Predicting the functional and metabolic capabilities of microbial communities from metagenomic shotgun data**

Regarding the metagenome analysis, this work focuses especially on the prediction of metabolic capabilities. The quantitative functional description of a metagenome provides the basis of the subsequent metabolic characterization. In the next section, I introduce strategies for functional profiling. Further, I point out computational limitations of BLAST-based homology search for functional annotation of metagenome sequence data. In Section 1.3.1.2, I discuss existing approaches for quantitative metabolic description. Thereby, I address the computational and algorithmic drawbacks which gave rise to the development of the Mixture-of-Pathways (MoP) model (see Chapter 2).

#### **1.3.1.1 Functional profiling**

The functional repertoire of a microbial community is usually quantified by counting the number of reads assigned to functional categories. The assignment is often derived from homology search of metagenomic sequences against a database of functionally annotated genes or proteins originating from studied cultured isolates that have well-characterized genomes. The comparison against reference databases is performed either directly, using all raw metagenomic sequences, or using representatives from priorly formed sequence groups. The grouping of sequences can be performed applying a sequence similarity criterion (e.g., using CD-HIT [130, 131] with a 95% nucleotide identity) [82, 132, 133]. Further, potentially uninformative sequences can be excluded from functional annotation using gene-prediction tools, which estimate the likelihood of a DNA sequence coding for a gene. However, gene prediction for metagenomes is a non-trivial task due to frame shifts, sequencing errors, and partial genes with missing proper gene starts, stops, or even both [134, 135]. A number of tools were specifically designed to handle metagenomic prediction of coding DNA sequences (e.g., FragGeneScan [136], MetaGeneMark [137], MetaGeneAnnotator/Metagene [138, 139], and Orphelia [140]). Regardless whether raw, clustered, or predicted coding DNA sequences are used, they are classified into functional categories using reference databases such as the NCBI NR database of non-redundant protein

sequences [141], COGs, FIGfams [142], KO, or protein families such as Pfams [143].

The assignment of sequencing reads to functional categories is often performed using pairwise sequence similarity. For annotation, the best hit assignment is frequently used. Commonly, the pairwise sequence similarity is derived by utilizing an algorithm from the NCBI's BLAST suite. However, BLAST based homology search shows serious drawbacks in metagenome analysis.

### **Limitations of functional annotation of metagenome sequence data using BLAST based homology search**

By far most functional profiling approaches rely on pairwise sequence similarity in terms of a BLAST matching of metagenomic sequence reads against databases of annotated DNA and protein reference sequences. Particularly, the comparison at the protein level are challenging since shotgun sequences must be translated into polypeptides in all six reading frames and in turn, a comparison for each polypeptides must be performed against the reference database. Usually, the BLASTX tool is utilized, which has high computational demands even for a moderately sized dataset.

NCBI's BLAST suite of algorithms was introduced in the early 90s. The original BLAST program was optimized for searching individual sequence reads produced by 1990s technology, but not designed for the sheer volume of sequence data produced in nowadays metagenomic studies. Additionally, current metagenomic datasets must be compared to a much larger database of known sequences. BLAST-based analyses represent significant bottlenecks due to the quadratic runtime complexity of the all-versus-all comparison. Owing to the anticipated exponential increase in data, the problem is unlikely be addressed simply by scaling up computational resources [85]. New or improved software tools are needed to keep pace with the anticipated increase in data. Especially comparisons for large datasets, such as the HMP, could take decades when utilizing the conventional BLASTX program [78]. To cope with this bottleneck, supercomputers, accelerated BLAST programs, or both must be used [144]. Such technologies allow more than a thousandfold speed increase with only marginal loss in sensitivity compared to BLASTX [124, 145]. Specifically designed protein search tools such as RAPsearch2 [146] or protein alignment using a DNA aligner (PAUDA) [147] use reduced amino acid alphabets to decrease the overall complexity of the search. However, the reduced complexity comes at the expense of the assignment rate. For example, using PAUDA, the assignment rate is only 33% compared to BLASTX [147].

Besides the computational effort for the identification of homologs, BLAST-based analysis is hampered by the requirements of a sufficient read length and phylogenetically close species in the reference database, which are rarely met in metagenomic studies. The fragmentary nature of metagenomic data leads to partial proteins, which in turn affects the homology-based function prediction negatively. Commonly, a major part of the sequences cannot be functionally annotated due to inherent problems of similarity-based methods. In addition, reads may match genes whose function has not yet been elucidated [148]. These sequences of unknown origin or function can remain a considerable uninformative fraction of data [149].

The functional profile of a metagenome typically comprises frequencies for several thousand functional categories. In order to make functional profiles more interpretable and easier to understand, higher-level functional abstraction layers are required. A number of projects provide such functional hierarchies, e.g., SEED SubSystems [150], COGs, eggNOGs [151],

Kyoto Encyclopedia of Genes and Genomes (KEGG) [119, 120] database, and Gene Ontology (GO) [152].

Biological pathways provide key insights into the metabolic capabilities of microbial communities and have high explanatory power. The resulting pathway abundance profiles help to elucidate the relations between microbial communities and their environment and to explore significant variations of metabolic capabilities between microbial communities. Therefore, the development of methods for so-called pathway profiling based on metagenomic sequence data is of high importance for quantitative analysis.

### 1.3.1.2 Pathway profiling

This section, first the concept of pathway reconstruction in whole genome sequencing projects is introduced. Then, the inherent differences between genomics and metagenomics, which have to be taken into account, are discussed. Finally, current state-of-the-art approaches for metagenomic pathway profiling and the drawback of these approaches is reviewed.

#### Genomic pathway reconstruction

The reconstruction of biological pathways encoded by an organism is one of the first analyses in microbial whole genome sequencing projects. Pathway reconstruction is based on the assumption that experimentally identified metabolic pathways are conserved between organisms. For this purpose, the curated reference databases KEGG and MetaCyc [153] are widely used as template. Both databases contain manually curated biological pathways which represent a series of molecular interactions and reactions that concur to a biological function. For example, the curated KEGG reference database comprises the KO families, which are directly linked to KEGG Pathways.

The genomic pathway reconstruction consists of two major key steps. First, the functions of protein coding genes are predicted. This task is usually carried out using homology-based approaches, e.g., BLAST. In case of multiple BLAST hits, the highest scoring match is used to assign the sequencing reads to enzymes and pathways. Second, the pathway inference step is carried out to predict which pathways may be occurring. A complete biological pathway is inferred to be present if one or more predicted protein functions can be assigned to the pathway [154]. However, this second step is not trivial. Curated reference pathways are simplified views of the biological processes and molecular interactions. They represent unified representations which divide simultaneously occurring biological functions into subsets of functional units. A pathway can be considered as a group of functionally related genes which are usually created from literature or expert knowledge. The current knowledge about pathways is largely fragmented and far from being completely clarified. Moreover, reference pathways are unified representations which often combine the knowledge about several organisms. Therefore, not every organism necessarily performs all molecular interactions and reactions included in a reference pathway. This leads to the fact that genomes have gaps in the commonly used reference pathways [155].

The ambiguous assignments of functional categories to pathways remain another aspect which has to be addressed. Most functional categories are associated with multiple pathways. In the KEGG database, e.g., phosphoenolpyruvate carboxykinase (GTP) (KO group K01596) is associated with the tricarboxylic acid cycle (TCA cycle), glycolysis, and various signaling pathways

[156]. Both the ambiguity of the function to pathway mapping and the unified representations makes the pathway inference in genomics difficult. In Caspi et al. [153] the challenges of representing and constructing metabolic pathways are further discussed by taking the example of MetaCyc.

Pathway inference in genomics rests upon a minimal number of elements sufficient for a pathway to be considered as present. The genomic pathway reconstruction approaches follow simple rules of thumb. The naïve genomic pathway reconstruction approach assumes that a biological pathway can be considered as present if at least one element of a pathway is present in a gene set. In contrast, a conservative assumption is that a pathway is only present if all its elements exist in a gene set. These two simple rules of thumb lead to an overestimation and underestimation of the reconstructed metabolic capabilities of an organism, respectively. In the former case, a functional annotation of the phosphoenolpyruvate carboxykinase (GTP) (K01596) automatically leads to the conclusion that ten different pathways are encoded in the genome. However, this inference ignores that in certain circumstances no evidence for further molecular interactions and reactions of a particular pathway is available. In the latter case, the inference of a pathway is strongly influenced by the number of genes associated with a pathway. The smaller the number of associated genes, the more likely it is to identify all required evidences. Furthermore, both naïve genomic pathway reconstruction approaches make use of the assumption that functions are equally active in all reference pathways.

Towards a more conservative estimation, the MinPath (Minimal set of Pathways) approach was proposed. This parsimony method solved with integer programming attempts to determine the minimal set of biological pathways that have to exist in the biological system to explain the input protein sequences sampled from it [154].

A more sophisticated genomic pathway reconstruction approach is PathoLogic [157], which is implemented in the Pathway Tools software suite [158, 159]. PathoLogic implements a set of inference rules which, for example, take into account pathway variants, composition, and connectivity. In addition, manually imposed constraints inferred from completed genomes are incorporated. The considered features enable the elimination of pathways that are unlikely to occur.

However, those genomic pathway reconstruction approaches are not directly transferable to metagenomics. In metagenomics, the objective is to quantify the metabolic capabilities. Therefore, the purpose of pathway reconstruction to predict the presence of pathways does not meet the requirements in metagenomics. Accordingly, metabolic profiling requires a different modeling approach. Moreover, due to variations of the pathway-specific enzyme sets across different species the metabolic profiling approach is a highly challenging task and more complicated than in genomics. In the following section, the objective and difficulties of pathway profiling for microbial communities are discussed in detail.

## **Quantification of the metagenomic metabolic repertoire**

In a microbial community, the metabolic potential is a complex composite of the metabolic capabilities of all present organisms. Each species encodes a unique set of metabolic functions. Hence, the pathway profiling approach in metagenomics does not attempt to predict the metabolic potential of a single organism but an entire microbial community. To account for this, the entire

microbial community is often modeled as a single supra-organism [160]. To characterize the aggregated metabolic processes of microbial communities in a given environment, the entire set of functions is studied as a proxy for the microbiome's repertoire. This requires an appropriate catalog of reference pathways. However, pathways are typically modeled for single organisms. Neither KEGG nor MetaCyc [153] are currently optimized for modeling communities. More importantly, the objective is to quantify the amount of genomic material that can be explained in terms of biological pathways. Therefore, approaches which attempt to predict the presence of pathways like in genomics are unsuitable. A sufficiently deep sampling would raise evidence for almost all pathways resulting in a limited explanatory power. In contrast, the quantification of metabolic pathways enables to draw conclusions which metabolic potential is most abundant and thus potentially reveals significant metabolic differences between different conditions.

Several approaches have been proposed for a quantitative representation of the metabolic potential of a metagenome. Usually, the metabolic profile is inferred from counting the number of sequence fragments that can be assigned to a particular biological pathways. The number of sequence fragments is deduced from predicted gene functions which can be related to biological pathways. Similar to the pathway reconstruction in genomics, metabolic profiling is strongly affected by the ambiguity of function-to-pathway mapping.

In the following, three common approaches, the naïve mapping, PathoLogic, and HMP Unified Metabolic Analysis Network (HUMANN) [145], are discussed. These approaches are differing in terms of the proposed noise reduction, smoothing steps, and counting schemes. In contrast to HUMANN, the naïve mapping and the PathoLogic approach require prior gene abundance filtering since no preprocessing, e.g., outlier removal or filtering of low abundances, is performed.

### **Naïve mapping approach**

A commonly-used metabolic profiling approach is the naïve mapping approach. The naïve mapping approach increments the abundance of a pathway whenever a assigned function can be found in the metagenomic sequences. This naïve counting scheme does not consider the counts of other functions associated with the pathway. Further, the approach assumes that functions are equally active in all associated pathways. Therefore, the functional abundances are counted among all associated pathways regardless of whether a function is associated with one or more possible pathways. For example, phosphoenolpyruvate carboxykinase (GTP) (K01596) contributes to several pathways. A nonzero abundance for K01596 automatically increases the abundance of all associated pathways. Similarly to genomics, the naïve mapping strategy overestimates the metabolic potential encoded by a microbial community. In addition, both the size of a reference pathway and the distribution of functional abundances in a pathway can strongly influence the final metabolic abundances. Few high abundances can result in a highly abundant pathway despite the fact that the residual functions have either no or minor abundances.

Consequently, the naïve mapping approach is a clear simplification to compute the metabolic abundances and can give rise to a misleading description of the metabolic potential of a microbial community. However, the approach is implemented in several state-of-the-art metagenome analysis tools, e.g., in MEtaGenome ANalyzer (MEGAN) [161], MG-RAST [162] and IMG/M [163–165].

## PathoLogic

Since version 17.0 of Pathway Tools [158, 159], the genomic pathway reconstruction approach PathoLogic [157] (see Section 1.3.1.2) is extended by pathway abundance prediction for metagenomic datasets. In contrast to the presented naïve approach, the abundance counting scheme of PathoLogic takes into account the uneven distribution of gene abundances for a pathway. The abundance of a pathway is defined as the sum of gene abundances involved in the pathway divided by the number of reactions of the pathway for which gene abundances are given. However, neither the size of a reference pathway nor the ambiguity of function-to-pathway assignments is taken into account.

The PathoLogic approach is for example incorporated in the modular MetaPathways pipeline for constructing environmental pathway/genome databases from metagenome data [166]. Furthermore, the MetaPathways pipeline was applied by Hanson et al. [167] to assess the impact of different metagenomic dataset characteristics, e.g., read length, coverage, sample diversity, and taxonomic pruning on the prediction of environmental pathway/genome databases by Pathway Tools. In their study, Hanson et al. [167] pointed out different limitations of Pathway Tools for the scope of metagenomics including the multiple mapping problem and prediction hazards arising from pathway variants.

## HMP Unified Metabolic Analysis Network (HUMAnN)

The HMP Unified Metabolic Analysis Network (HUMAnN) software performs functional and metabolic profiling directly from high-throughput metagenomic short sequence reads [145]. The pipeline starts with a similarity search of cleaned short DNA reads against a functional protein sequence database. Per default a similarity search against the KO (Release 54) using the accelerated translated MBLASTX is performed. Subsequently, the output of the similarity search is used for a series of gene- and pathway-level quantification, noise reduction, and smoothing steps. First, the functional abundances of individual orthologous gene families are calculated as weighted sums of the alignments from each read, normalized by each gene family's average sequence length and alignment quality. Next, pathway reconstruction is performed using the MinPath approach [154]. In this step, the relative KO abundances are consolidated into one or more pathways. Hereby, the abundance of a KO assigned to two or more pathways is effectively duplicated. Subsequently, filtering and normalization steps based on taxonomic profiles from BLAST hits are performed to ensure that unlikely pathways are removed and normalized for genes' average copy number. After removing false positive pathway identifications, a smoothing or gap-filling step is performed to account for rare genes in abundant pathways. Finally, a coverage (presence/absence) and abundance score is assigned to each pathway resulting in the identification of present/absent pathways and modules together with their relative abundances. HUMAnN implements different filtering, normalization, and smoothing steps. So far, HUMAnN is the most sophisticated pathway profiling approach for metagenomes. However, here too, KO abundances are effectively duplicated, thus the ambiguity of the function-to-pathway mapping remains mostly unaddressed.

## Model-based estimation of pathway abundances

There are two major difficulties with these existing approaches for metabolic profiling. First, the presented metabolic profiling methods rely on a predicted functional abundance profile. However, the computational effort for the identification of homologs can become burdensome. As discussed in Section 1.3.1.1, the complexity of BLAST-based analyses represents a significant bottleneck. Accelerated translated BLAST technologies and specifically designed protein search tools such as RAPsearch2 or PAUDA reduce the overall search complexity and computational costs. The functional annotation remains computationally intense with the result that a considerable amount of central processing unit (CPU) time is spent for a moderately sized dataset. The growth in dataset size, in conjunction with computational complexity of the analysis, has left the metagenomics community in a difficult position, in terms of both financial cost and feasibility of analysis itself. Without novel algorithms for analysis, the sheer volume of sequencing data will overwhelm available resources. Besides the computational effort for the identification of homologs, difficulties arise from the inherent ambiguity in the function-to-pathway mapping. Some solutions have been proposed to adjust pathway abundance and to avoid overestimation. However, none of these approaches provides a strict probabilistic description of metagenomic sequence data and thereby may overrate the metabolic repertoire.

The objective of this work is to develop an efficient and statistically reasonable method for characterization of the metabolic potential in metagenomic samples. The method should address both, the statistically adequate modeling of the inherent ambiguous function-to-pathway mapping and the reduction of computational cost for the estimation of metabolic profiles. For estimating the fraction of sequence material that can be assigned to a particular pathway, a statistical model has to be developed. This model should be capable to provide both a sound statistical basis and a fast estimation of pathway abundances. The relative pathway abundances should provide a clear statistical meaning which permits a comparison across different studies. Further, the method should be fully operational for the growing amount and the heterogeneity of metagenomic sequence data.

## 1.4 Comparative analysis of microbial communities

Following the taxonomic and metabolic characterization of a single microbial community, the question “What are the differences that make a difference?” is addressed by comparative metagenome analyses. For comparison of microbial communities in different environments, points in time or conditions, the taxonomic, functional, and metabolic abundance profiles are used as input for techniques such as principal component analysis [168], nonmetric multidimensional scaling [169], and hierarchical clustering [170]. In addition, several metagenome annotation and comparison systems exist, e.g., MG-RAST, IMG/M, and WebMGA [171]. Furthermore, a number of dedicated comparison tools have been developed, including JCVI Metagenomics Reports (METAREP) [172], SStatistical Analysis of Metagenomic Profiles (STAMP) [173, 174], and Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline (RAMMCAP) [175].

One of the first comparative metagenomic studies were performed by Tringe et al. [76]

and Dinsdale et al. [77]. For example, Dinsdale et al. [77] conducted a broad metagenomic comparison among 45 distinct microbiomes and 42 distinct viromes. Following this, a number of other metagenomic comparisons have been realized [62, 176–179]. For instance, differences in gut microbiota composition between lean and obese individuals arouse interest in commensal microorganisms [180]. Following this, additional comparative studies have been performed to gain insight into host microbe interactions and to yield hypotheses about microbiota-based disease mechanisms [110, 181–183], which could often be confirmed by subsequent microbiota-manipulation studies [182, 184].

### 1.4.1 Example: The human microbiome

The human microbiome provides a number of biomolecular functions that are not encoded in the human genome and are necessary for human health [60]. For example, the human intestinal microbiota is involved in several nutritional, physiological, and immunological processes [185–190], and provides metabolic activity for central, carbohydrate, and amino-acid metabolism. The comparison of taxonomic, functional and pathway abundances between different conditions revealed that the microbiota composition and activity influences host metabolism and disease development such as obesity, inflammatory bowel disease, Crohn’s disease, diabetes, cancer, and allergies [183, 185, 191–195]. Conversely, the activity and composition of the microbiota is affected by age, diet, health status, and genetic background of the host [80, 183, 194, 196–203]. Therefore, understanding the interplay between the human microbiome and the human body can help to improve human health.

Several large-scale studies have been carried out to characterize microbial communities at multiple body sites and to investigate the association between the human microbiome and human development, physiology, immunity, and nutrition. For example, at the end of 2007, the US National Institutes of Health launched the Human Microbiome Project (HMP) [81] and, in early 2008, the European Commission initiated the Metagenomics of the Human Intestinal Tract (MetaHIT) project [204]. Both projects aim to characterize the microbiome of healthy individuals. First studies provide a preliminary understanding of the biological and medical significance of the human microbiome and its collective genes. In the following section, I am focusing on the HMP, which I am using as data basis for evaluating the methods proposed in Chapter 2, Chapter 3, and Chapter 4.

### The Human Microbiome Project

The HMP’s study design included extensive sampling of the human microbiome from 242 healthy adults in the United States at five clinically relevant major body sites. These are airways, skin, oral cavity, gastrointestinal tract, and vagina. Within each of these major body sites, 15 (for males) or 18 (for females) more specific body subsites were sampled, often at multiple time points, resulting in a wealth of samples [78].

The HMP study used a two-stage design. In the first instance, over 5000 samples were analyzed using 16S rRNA sequencing followed by merely about 700 samples using metagenomic sequencing. Despite the different number of data samples, both collections represent roughly equal experimental costs [79]. The efforts of the HMP have produced more than 70 million

16S rRNA gene sequences and more than 3.5 Tbp of whole-metagenome sequence data [79]. Important questions concerning the commonalities and differences among healthy individuals in both microbial taxa and functional pathways are being addressed. Using the data, it was demonstrated that body habitat accounts for much of the variation in bacterial community composition. In addition, despite considerable variation in the microbiota composition across individuals, the functional abundance profiles are quite similar [19, 205].

#### **1.4.2 Identification of closely related metagenome datasets**

With the rapid development of sequencing technologies, statistical replication of metagenomic samples became technically and financially feasible [206]. Recently, several large-scale studies with hundreds of samples were initiated [62, 78, 83, 207]. However, apart from large-scale studies, existing single or small-scale datasets in public repositories, e.g., MG-RAST or IMG/M, provide a rich resources for comparative analyses. The number of metagenome datasets is rapidly increasing and every single dataset represents a microbial community with a unique biological history, sampling location, and environmental context [208]. Thus, comparative analyses can profit from the existing wealth by comparing newly obtained datasets with existing datasets. These may serve as additional data sources or biological replicates for a statistical characterization of variations.

To derive benefit from the wealth of existing datasets, appropriate datasets have to be selected beforehand. The manual identification of datasets requires a comprehensive overview of all available datasets or at least reliable and informative metadata. Particularly, if a habitat label is rather abundant within a repository. For example, in [209] we compiled the so-called “metagenome universe”. The metagenome universe is a comprehensive metagenome collection of HMP datasets and publicly available metagenome datasets from the MG-RAST and European Bioinformatics Institute (EBI) [210] online resources. The metagenome universe comprises 1745 metagenomes each labeled with one of twelve habitat categories. The number of metagenomic datasets with habitat label “Feces + GI tract” (581) is rather abundant. Supposing that one is interested in selecting a “Feces + GI tract” sample, the question remains which of the 581 samples might be most suitable and requires in-depth investigations. Thus, sample selection solely based on metadata is difficult.

Not only the metadata categories itself, but also the reliability of metadata is another important aspect. Since we rely on the accuracy of available metadata for comparative analyses, erroneous metadata strongly influences biological interpretations. Nevertheless, mislabeling is likely during processing and pooling of samples. For example, in the study of [211] several 16S rRNA amplifications of bacterial community DNA samples collected along a time series were accidentally mislabeled [212]. Further, datasets from public repositories can also be affected by mislabeling. However, in this case, we cannot reconfirm sample labels or even resequence questionable samples. Accordingly, it is advantageous to detect and even correct erroneous metadata in some datasets. Knights et al. [212] addressed the identification of mislabeled data with a supervised classification to retrieve the correct groupings of mislabeled 16S rRNA surveys at various rates of error. However, in Knights et al. [212] only 16S rRNA data was investigated. For metagenomic datasets, it remains unclear which approach might be successful and has not been addressed so far.

In this thesis, I investigate both the identification of similar metagenomes for comparative analyses and the identification of mislabeled datasets. In contrast to [212], I do not consider supervised methods which are well-suitable in conjunction with well-defined categories and reliable labels. However, well-defined categories are commonly not available in public metagenomic repositories. Therefore, the identification of related metagenomes by means of the available sequence data alone is more promising. However, in contrast to 16S rRNA, the computational cost for all pairwise sequence comparisons between a new query dataset and  $n$  metagenomes in a repository is prohibitively expensive. Especially, on account of the growing amount of metagenome datasets, it is reasonable to compare feature profiles instead of the sequence data. However, it has not been investigated which feature profiles and which kind of metrics are most effective for identification of similar datasets to a newly obtained dataset. In this work, I utilize a  $k$ -nearest-neighbor search for identification of closely related metagenome datasets which does not depend on the quality of labels. In Chapter 3, I thoroughly investigate different feature profiles and metrics for a  $k$ -nearest-neighbor search to identify similar datasets for comparison and mislabeled datasets by means of unexpected neighboring habitat labels.

## 1.5 Objective and overview

The objective of this work is to support the development of new computationally efficient and statistically adequate methods for large-scale comparative metagenome analyses. Metagenome sequence data enable a comprehensive description of the metabolic capacities of microbial communities. However, difficulties arise from the computational effort to identify homologs and the inherent ambiguity in the function-to-pathway mapping. In this thesis, I consider the question whether the taxonomic composition in conjunction with knowledge of reference organisms may serve as a proxy for predicting the metabolic repertoire of a microbial community to reduce the total analyses costs.

The first publication (Chapter 2) introduces a novel metabolic profiling approach for metagenomics which is based on a Mixture-of-Pathways (MoP) model. With regard to the description in terms of metabolic abundances, a mixture model is introduced which provides an estimate of the fraction of sequence material that can be mapped to a particular pathway. Further, a shortcut for fast estimation is introduced. This nested model links the taxonomic profile of the metagenome to a set of pre-computed metabolic reference profiles overcoming computationally intense homology searches.

The second publication addresses the identification of similar metagenomes to a newly obtained dataset. Despite its growing importance for comparative metagenome analysis, this is a largely uninvestigated field. In Chapter 3, a  $k$ -nearest-neighbor search based on biological feature profiles is assessed for the identification of closely related metagenome datasets.

Chapter 4 contains a submitted manuscript describing the Tax4Fun approach, which is designed for the prediction of the functional repertoire of a microbial community from 16S rRNA data. So far, insights into the functional capabilities of a microbial community are not provided by restricting the analysis to 16S rRNA sequence data. In this thesis, the transformation of OTU abundances to KEGG based taxonomic profiles based on 16S rRNA sequence similarity is presented which enables the linear combination of pre-computed reference profiles to estimate

the functional capacities of a microbial community.

Both, the functional annotation approaches and the  $k$ -nearest-neighbor search approach, were evaluated with the publicly available HMP datasets [19]. For the evaluation of Tax4Fun, we used three additional datasets where both 16S rRNA and metagenome sequence data were available [110, 213–215]. To study the performance of the retrieval of similar metagenome sequencing samples in a broader more complex setting, the so-called metagenome universe was constructed from publicly available datasets.

In Chapter 5, the overall results of the publications and manuscript are discussed.

## 2 On the estimation of metabolic profiles in metagenomics

### Reference

Kathrin P. Aßhauer and Peter Meinicke. On the estimation of metabolic profiles in metagenomics. In German Conference on Bioinformatics 2013, Volume 34 of OpenAccess Series in Informatics (OASICs). Edited by Tim Beißbarth, Martin Kollmar, Andreas Leha, Burkhard Morgenstern, Anne-Kathrin Schultz, Stephan Waack, and Edgar Wingender, Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum für Informatik 2013:1-13.

### Original Contribution

KPA assembled the data basis and implemented the workflow. KPA performed the computer calculations and data analysis. KPA evaluated the new method and HUMAnN (resulting in figure 1 and tables 1-6). Both authors designed the study, wrote the manuscript, and approved the final version of the manuscript.



# On the estimation of metabolic profiles in metagenomics\*

Kathrin Petra Aßhauer and Peter Meinicke†

Department of Bioinformatics, Institute for Microbiology and Genetics  
University of Göttingen  
37077 Göttingen, Germany  
peter@gobics.de

---

## Abstract

Metagenomics enables the characterization of the specific metabolic potential of a microbial community. The common approach towards a quantitative representation of this potential is to count the number of metagenomic sequence fragments that can be assigned to metabolic pathways by means of predicted gene functions. The resulting pathway abundances make up the metabolic profile of the metagenome and several different schemes for computing these profiles have been used. So far, none of the existing approaches actually estimates the proportion of sequences that can be assigned to a particular pathway. In most publications of metagenomic studies, the utilized abundance scores lack a clear statistical meaning and usually cannot be compared across different studies. Here, we introduce a mixture model-based approach to the estimation of pathway abundances that provides a basis for statistical interpretation and fast computation of metabolic profiles. Using the KEGG database our results on a large-scale analysis of data from the Human Microbiome Project show a good representation of metabolic differences between different body sites. Further, the results indicate that our mixture model even provides a better representation than the dedicated HUMAnN tool which has been developed for metabolic analysis of human microbiome data.

**1998 ACM Subject Classification** J.3 Life and Medical Sciences – Biology and genetics

**Keywords and phrases** metagenomics, metabolic profiling, taxonomic profiling, abundance estimation, mixture modeling

**Digital Object Identifier** 10.4230/OASICS.GCB.2013.1

## 1 Introduction

In metagenomics a central task is to characterize the metabolic potential of a microbial community. The metabolic profile of a metagenome quantifies the amount of genetic material that can be attributed to metabolic pathways. The abundance of a pathway is usually estimated by the number of sequences that can be mapped to gene families with functional roles within that pathway. Several heuristics exist to compute a corresponding estimate. Using for instance the KEGG database, an abundance may be estimated by counting all BLAST best hit matches to KEGG Orthologs which are annotated for the particular pathway (see e.g. [4]). There are two major difficulties with this classical approach of metabolic profiling: First, the computational effort for the identification of homologs can become burdensome. Usually the BLASTX tool is required, which takes a considerable amount of

---

\* This work was partially supported by the Deutsche Forschungsgemeinschaft (ME3138 “Compositional descriptors for large scale comparative metagenome analysis”).

† corresponding author



© Kathrin Petra Aßhauer and Peter Meinicke;  
licensed under Creative Commons License CC-BY

German Conference on Bioinformatics 2013 (GCB'13).

Editors: T. Beißbarth, M. Kollmar, A. Leha, B. Morgenstern, A.-K. Schultz, S. Waack, E. Wingender; pp. 1–13

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

CPU-time even for a moderately sized data set. Second, the usual counting scheme lacks a probabilistic model that would provide a clear statistical interpretation of the resulting quantities. To our knowledge, none of the existing heuristics actually yields an estimate of the fraction of sequence material that can be mapped to a particular pathway. Depending on the particular method the existing tools merely provide different kinds of abundance scores [14, 12, 1, 5, 4]. Although these scores may be used for comparative analysis as well, they do not provide a strictly probabilistic description of metagenomic sequence data. Therefore, the comparison and combination with other methods or models is at least problematic. We address both problems, the algorithmic and statistical efficiency within a metabolic mixture model in terms of a mixture of pathways (MoP). This model is capable to provide both, a sound statistical basis and a fast estimation of pathway abundances. Our results on a large-scale analysis of data from the Human Microbiome Project (HMP) show the utility of our method for fast model-based estimation of pathway abundances. Further, the results for the mixture-based metabolic profiles indicate a better separation between body sites than for the profiles of the HUMAnN tool which has particularly been developed for analysis of HMP data.

## **2 Material**

### **2.1 Human Microbiome Project (HMP)**

Within the scope of the Human Microbiome Project (HMP) [3] an extensive collection of samples from healthy individuals from diverse human body sites was established allowing an insight into the functions of the healthy human microbiome. More than thousand HMP data sets are recorded in HMP's Data Acquisition and Coordination Center (DACC) Project Catalog ([http://www.hmpdacc.org/resources/data\\_browser.php](http://www.hmpdacc.org/resources/data_browser.php)) providing a comprehensive data basis for large-scale comparative studies investigating the associations of the human microbiome in healthy and diseased states.

From the HMP-DACC website we assessed the available metadata for the metagenomic samples ([http://www.hmpdacc-resources.org/hmp\\_catalog/main.cgi](http://www.hmpdacc-resources.org/hmp_catalog/main.cgi)) and the metabolic reconstruction data (<http://www.hmpdacc.org/HMMRC/>). The metabolic reconstruction data is obtained through the HMP Unified Metabolic Analysis Network (HUMAnN) pipeline [1]. HUMAnN performs functional and metabolic profiling directly from high-throughput metagenomic short sequence reads. The pipeline starts with a similarity search against a functional sequence database including the KEGG Orthologs (Release 54) using an accelerated translated BLAST implementation. Subsequently, the output is used for a series of gene- and pathway-level quantification, noise reduction, and smoothing steps resulting in the identification of present/absent pathways and modules together with their relative abundances. From the available metabolic reconstruction data, we used the "KEGG pathway abundance values – Summary file" (as of February 2013).

For our mixture modeling approach we used the reduced data samples of the HMP as describes in [10]. For comparability, the available samples and pathway abundances of HUMAnN and our mixture modeling approach were reduced to a subset of samples and pathways available in both methods. The final dataset includes 680 data samples from 14 specific body subsites, which can be grouped into five major body sites.

## 2.2 KEGG database

For the metabolic mixture modeling approach introduced here, we use the Kyoto Encyclopedia of Genes and Genomes (KEGG) database as reference knowledge base for estimating the pathway abundances of metagenomic samples [9, 8]. KEGG integrates a variety of information and provides links from gene catalogs to higher-level systematic functions of the organisms enabling biological interpretation of genomes and high-throughput datasets.

An essential part of the database with respect to metabolic profiling are the KEGG Orthologs (KO) that consist of gene groups with specific functions directly linked to known pathways in the KEGG Pathway database. Further, the KEGG Orthology is structured as a hierarchy of four flat levels: top, second, third level, and leaf nodes. While the leaf nodes represent the KEGG Orthologous groups, the third level represents the KEGG Pathways, which can be further summarized in higher level pathway classes (top and second level).

For the mixture modeling the required data reference was extracted from the KEGG database (Release 64.0).

## 2.3 MarVis

The MarVis-Suite (**Marker Visualization**) [7, 6], a toolbox originally developed for the analysis of metabolomic data, was used for filtering, clustering, and visualization of the pathway abundances. For exploration of complex pattern variation within the samples of the different body sites/subsites we used the MarVis-Cluster interface which permits high-level visualization and cluster analysis based on a one-dimensional self-organizing map (1D-SOM). The MarVis-Filter software was used for the identification of pathways overrepresented in the gastrointestinal tract samples compared to the other body subsites.

# 3 Methods

## 3.1 Taxonomic mixture modeling

The mixture model based Taxy approach provides a fast and direct estimation of taxonomic abundances in metagenomes. Taxy-Oligo [13] and Taxy-Pro [10] do not perform a taxonomic classification of sequencing reads but instead apply a mixture model to approximate the overall metagenome distribution of oligonucleotides and protein domain hits, respectively. The discrete distribution of oligonucleotides/protein domains is modeled by a mixture of organism-specific profiles as obtained from sequenced reference genomes. Because of the computational efficiency of the taxonomic mixture model approach, both methods were able to perform a large-scale analysis of sequence data from the HMP without using a computer cluster or special hardware. All reference profiles were obtained from the bacterial and archaeal genomes available in the KEGG database (Release 64.0). These genomes were also used for pre-computing the organism-specific pathway abundances for the metabolic profiling of metagenomes. For Taxy-Pro, all protein domain profiles according to the Pfam database [2] were obtained from the CoMet web server [11].

## 3.2 Metabolic mixture modeling

For metabolic profiling, we assume that the genomic sequence material to some degree can be explained by a mixture of pathways. The mixture approach accounts for the fact that in most cases a putative gene function as observed in a sequence fragment provides evidence for more than one metabolic pathway. The statistical representation of this ambiguity of the

function-to-pathway mapping was the main motivation for the development of the following model. With  $M$  pathways  $P_i$  the probability to observe a function  $F$  encoded in sequences under this model is:

$$\tilde{p}(F) = \sum_{i=1}^M p(P_i)p(F|P_i) \quad (1)$$

The tilde indicates that  $\tilde{p}(F)$  only is an approximation of the functional profile  $p(F)$  because not every function can be explained in terms of metabolic pathways. The prior pathway probabilities  $p(P_i)$  denote the overall sequence-based abundance of functions associated with pathway  $P_i$  and correspond to the mixture weights of the model. These weights are the central model parameters, which can directly be used and interpreted in terms of the relative abundances of a metabolic profile. The conditional probability  $p(F|P_i)$  denotes the  $i$ -th pathway-specific distribution over  $N$  possible gene functions  $F_j$ . The annotation in current databases, such as KEGG, can be represented by some  $M \times N$  assignment matrix  $\mathbf{A}$  with binary entries  $A_{ij} = 1$  denoting that function  $j$  is associated with pathway  $i$ . From that assignment it follows that all functions not associated with pathway  $i$  must attain a zero conditional probability. Just from the annotation, we cannot draw any conclusions about the other probabilities. Without further knowledge the only reasonable assumption is that the  $p(F|P_i)$  are proportional to the corresponding overall function probabilities, i.e.

$$\forall i, j : p(F_j|P_i) \propto A_{ij}p(F_j). \quad (2)$$

This constraint implies that the ratio between any two non-zero function probabilities in a pathway is equal for all pathways these two functions are associated with and must equal the global ratio of the corresponding probabilities of the functional profile  $p(F)$ . With the  $N$  estimates  $\hat{p}(F_j)$  of the specific function probabilities of the profile as derived from the observed frequencies, e.g. from BLAST hit counts, we have the following estimator of the conditional probabilities:

$$\hat{p}(F_j|P_i) = \frac{A_{ij}\hat{p}(F_j)}{\sum_{k=1}^N A_{ik}\hat{p}(F_k)}. \quad (3)$$

Now let us consider the assignment probability

$$p(P|F_j) = \frac{p(P)p(F_j|P)}{\sum_{i=1}^M p(P_i)p(F_j|P_i)} \quad (4)$$

which denotes the responsibility of a pathway for a given function  $F_j$ , i.e. the contribution of a pathway to the explanation of that function. We assume that this probability is equal for all pathways the function is associated with. Without further knowledge, just with the underlying pathway annotation, there is no reason to prefer a particular pathway for the explanation of an observed function. This implies the following additional constraint:

$$\forall i, j, k : A_{kj}p(P_i|F_j) = A_{ij}p(P_k|F_j). \quad (5)$$

For a function  $F_j$  that is annotated in two pathways  $P_i$  and  $P_k$  we can obtain the ratio of the corresponding pathway abundance estimators using the former three equations (3), (4) and (5):

$$\frac{\hat{p}(P_i)}{\hat{p}(P_k)} = \frac{\hat{p}(F_j|P_k)}{\hat{p}(F_j|P_i)} = \frac{\sum_{s=1}^N A_{is}\hat{p}(F_s)}{\sum_{t=1}^N A_{kt}\hat{p}(F_t)}. \quad (6)$$

From the above proportionality, we finally obtain the estimator of the pathway probabilities:

$$\hat{p}(P_i) = \frac{\sum_{j=1}^N A_{ij} \hat{p}(F_j)}{\sum_{k=1}^M \sum_{l=1}^N A_{kl} \hat{p}(F_l)}. \quad (7)$$

Using matrix vector algebra we can compute the whole metabolic profile vector  $\mathbf{p}$  with entries  $\hat{p}(P_i)$  from the functional profile vector  $\mathbf{f}$  with entries  $\hat{p}(F_j)$  by

$$\mathbf{p} = \frac{\mathbf{A}\mathbf{f}}{\mathbf{1}^T \mathbf{A}\mathbf{f}} \quad (8)$$

where  $\mathbf{1}$  is an  $M$ -vector of ones. In an application of the above mixture model most time will be spent for the computation of the functional profile which usually requires a costly BLASTX matching of metagenomic reads against a database of functionally annotated protein sequences, such as the KEGG Orthologues. However, with our formulation in terms of a statistical model we are able to provide a shortcut that utilizes the combination with another model to obtain a hierarchical mixture of pathways. Assume that we have the functional profiles of  $K$  reference organisms as columns in an  $N \times K$  matrix  $\mathbf{F}$  and we have estimated the relative abundances of the reference organisms in a taxonomic profile vector  $\mathbf{t}$ . Then we can approximate the functional profile of the metagenome by a linear combination of reference profiles  $\mathbf{F}\mathbf{t}$ . In Taxy-Pro [10] we use this mixture model in combination with Pfam functional profiles to estimate the taxonomic abundances in a metagenome. Here, we propose a combination with  $K$  pre-computed KEGG reference profiles to predict the functional profile of a metagenome from its taxonomic profile which may be obtained by some fast method such as the oligonucleotide-based Taxy tool [13]. The estimator of the metabolic profile is then

$$\mathbf{p} = \frac{\mathbf{A}\mathbf{F}\mathbf{t}}{\mathbf{1}^T \mathbf{A}\mathbf{F}\mathbf{t}}. \quad (9)$$

Note that also the matrix product  $\mathbf{A}\mathbf{F}$  can be pre-computed to obtain  $K$  organism-specific metabolic profiles which are then just combined by the taxonomic weights  $\mathbf{t}$  of a metagenome to obtain its metabolic profile. In principle, this gives rise to a nested model where a mixture of pathways is first used for each reference organism to estimate its metabolic profile. This step has only to be performed once for each organism and therefore even a costly BLASTX analysis may be used for the “offline” training of the organism-specific models. When applied to metagenomic data a mixture of the utilized reference organisms has to be estimated by some taxonomic profiling method. In order to combine the two models the second step requires a profiling method that actually estimates the abundances in terms of the amount of sequence material that can be attributed to a particular organism. For example, this requirement is automatically fulfilled when using Taxy-Oligo [13] or Taxy-Pro [10], which we both included in the evaluation of our approach, as described above. For an application of the MoP model, it is important to check whether the metagenome composition can actually be approximated by a mixture of known reference organisms. If the reference is completely insufficient for a description of the metagenome composition, the mixture approach in general would become inadequate. Therefore, it is desirable, that the taxonomic profiling method gives us an indication of the fidelity of the abundance estimates. Both, Taxy-Oligo and Taxy-Pro provide a specific error measure to assess the adequacy of the underlying model. In this case, the fraction of oligonucleotides unexplained (FOU) and the fraction of domain-hits unexplained (FDU) should be inspected when using Taxy-Oligo and Taxy-Pro, respectively.

### 3.2.1 Workflows

For the evaluation of our model, we implemented the direct application of the metabolic mixture model as well as the nested model.

The direct application of the mixture model starts with a BLASTX analysis where the metagenomic reads are mapped against a reference database consisting of KO amino acid sequences of bacterial or archaeal origin. By default BLAST hits with E-value  $\leq 10^{-2}$  were considered to be significant. The functional profile vector  $\mathbf{f}$  is obtained by counting the KO-specific BLAST hits using a fractional increment of  $1/K$  if  $K$  different KOs simultaneously show significant hits for a particular sequence. Note that due to the computational expense of BLASTX on metagenomes, we restricted the correlation analysis (see section 4.1) to a subset of six HMP data samples from different body subsites (SRS013825, SRS016752, SRS022621, SRS024265, SRS024428, and SRS055401). For the computation of the assignment matrix  $\mathbf{A}$  the association of KOs with KEGG Pathways was extracted from the database and transformed into a binary matrix. Finally, the mixture model was applied as described above using the functional KO profile vector  $\mathbf{f}$  and matrix  $\mathbf{A}$  as input.

For the nested model, we first pre-computed the organism-specific metabolic profiles from reference genomes using all bacterial and archaeal KEGG Genomes. The KEGG Genomes were downloaded and subsequently fragmented in overlapping reads of length 400 bp with 200 bp overlap simulating a two-fold coverage of the genomes as previously described in [10]. For each reference organism, first the functional profile vector is calculated and then the metabolic profile is estimated applying the steps as described for the direct mixture approach. By combining the weights  $\mathbf{t}$  of a metagenome with the pre-computed organism-specific metabolic profiles the metabolic profile of a metagenome can be obtained in an efficient manner. Note that a BLASTX/KO analysis of the metagenome is not required in this case. For the estimation of the taxonomic profile  $\mathbf{t}$ , we were using both, Taxy-Oligo and Taxy-Pro. According to the utilized taxonomic profiling method we denote our metabolic mixture model MoP-Oligo and MoP-Pro, respectively.

## 4 Results

To validate the metabolic mixture model on a well-studied dataset, we analyzed metagenomic sequences from the Human Microbiome Project (HMP) [3]. Originally, the metabolic profiles of the HMP data have been investigated by means of the HMP Unified Metabolic Analysis Network (HUMAnN) pipeline [1]. In the following, we use the metabolic profiles of HUMAnN for comparison with the abundance estimates that we obtained from our mixture of pathways model.

### 4.1 Correlation analysis

To study the similarity of metabolic profiles across different methods we computed the Pearson and Spearman (rank) correlation coefficients of the pathway abundance estimates. First, we evaluated the fast approximation scheme using pre-computed reference profiles based on Taxy-Pro taxonomic profiles (MoP-Pro). The resulting metabolic profiles were compared with the direct application of the mixture model to KO frequencies, which were obtained from a more time consuming BLASTX analysis. For each data sample, the correlation of the pathway abundances on two different pathway hierarchy levels (second and third level) was calculated.

The means and standard deviations of all data examples of the Pearson and Spearman

correlation coefficients are shown in Table 1. The results show a very high correlation of the approximation-based and the directly obtained abundances. By reducing the number of pathways from 340 to 38 according to the third and second pathway hierarchy levels an increase of the correlation from 0.9558 to 0.9804 and 0.9491 to 0.9842 could be observed for the Taxy-Pro-based approximation. These results indicate that the approximative approach is very close to the direct approach and therefore provides a computationally attractive alternative to the BLAST-based estimation.

■ **Table 1** Correlation analysis based on the metabolic abundances obtained by applying the Taxy-Pro-based approximation and the direct mixture approach. The correlation is calculated according to Spearman and Pearson and at the third and second pathway hierarchy level.

	Pearson	Spearman
Third level	0.9557 ( $\pm$ 0.0409)	0.9491 ( $\pm$ 0.0124)
Second level	0.9803 ( $\pm$ 0.0150)	0.9842 ( $\pm$ 0.0110)

The correlations are similarly high for the even faster Taxy-Oligo variant (MoP-Oligo) with results shown in Table 2.

■ **Table 2** Correlation analysis based on the metabolic abundances obtained by applying the Taxy-Oligo-based approximation and the direct mixture approach. The correlation is calculated according to Spearman and Pearson and at the third and second pathway hierarchy level.

	Pearson	Spearman
Third level	0.9575 ( $\pm$ 0.0409)	0.9466 ( $\pm$ 0.0105)
Second level	0.9796 ( $\pm$ 0.0138)	0.9813 ( $\pm$ 0.0087)

In contrast to the high similarity of results between different variants of the mixture approach the correlation between the mixture-based pathway abundances and the HUMAnN-based profiles is comparatively low with a Pearson correlation of 0.5290 as shown in Table 3. However, the correlation is increasing when considering the second pathway level or when using the Spearman rank correlation. A maximum rank correlation of 0.9080 indicates that the coarse shape of metabolic profiles is still rather similar between different approaches. Note that the correlation with HUMAnN profiles was averaged over all 680 HMP samples.

■ **Table 3** Correlation analysis based on the metabolic abundances obtained by applying HUMAnN and the TaxyPro-based mixture model. The correlation is calculated according to Spearman and Pearson and at the third and second pathway hierarchy level.

	Pearson	Spearman
Third level	0.5290 ( $\pm$ 0.0206)	0.7588 ( $\pm$ 0.0242)
Second level	0.7884 ( $\pm$ 0.0308)	0.9080 ( $\pm$ 0.0135)

## 4.2 Nearest neighbor classification

To assess the quality of the estimated metabolic profiles we first investigated whether the body site (subsite) classification of HMP samples can be reproduced by the corresponding pathway abundances. For that purpose, we evaluated the predictive power of metabolic profiles by some nearest neighbor classification scheme using different profile distance measures. We utilized a leave-one-out cross validation, measuring the classification rate for Euclidean distance, City block metric and Shannon-Jensen divergence on profiles.

The results for body sites and subsites as shown in Table 4 reveal that the nearest neighbor classification rate is rather high and varies between 0.9735 and 0.9897 for the five body sites and between 0.8853 and 0.9235 for the 14 body subsites. For both classification problems, HUMAnN shows the highest prediction accuracy irrespective of the distance measure used. However, the two mixture variants are always very close with a maximum difference of 2.94% for the Euclidean distance on body subsite level between HUMAnN and MoP-Oligo.

■ **Table 4** Nearest neighbor classification performing a leave-one-out cross validation with the Euclidean distance, City block metric and Shannon-Jensen divergence as distance measure for the approaches HUMAnN, MoP-Pro, and MoP-Oligo.

	Body site			Body subsite		
	Euclidean	City block	Jensen-Shannon	Euclidean	City block	Jensen-Shannon
HUMAnN	0.9838	0.9897	0.9868	0.9147	0.9235	0.9132
MoP-Pro	0.9794	0.9809	0.9779	0.9103	0.9103	0.9059
MoP-Oligo	0.9735	0.9750	0.9779	0.8853	0.8956	0.9132

### 4.3 Clustering performance

For a more comprehensive analysis of profile distances, we compared the body site (subsite) classification of samples with a profile-based clustering of the data. For clustering, we used a standard hierarchical approach with average linkage, also known as UPGMA. In this context, we evaluated the same three distance measures as for the nearest neighbor classification experiment. The quality of the cluster partitioning was assessed by the Jaccard coefficient, measuring the overlap of the resulting clusters with the HMP body site (subsite) groups.

The results obtained through the application of HUMAnN, MoP-Pro, and MoP-Oligo are presented in Table 5 which shows a large variation of the clustering performance.

■ **Table 5** Cluster partitioning quality in terms of the Jaccard coefficient based on Euclidean distance, City block metric and Shannon-Jensen divergence for metabolic profiles of HUMAnN, MoP-Pro, and MoP-Oligo

	Body site			Body subsite		
	Euclidean	City block	Jensen-Shannon	Euclidean	City block	Jensen-Shannon
HUMAnN	0.4335	0.4342	0.4325	0.2361	0.3715	0.2344
MoP-Pro	0.6958	0.8817	0.6971	0.4791	0.4603	0.4801
MoP-Oligo	0.6577	0.7251	0.6382	0.3671	0.3008	0.3939

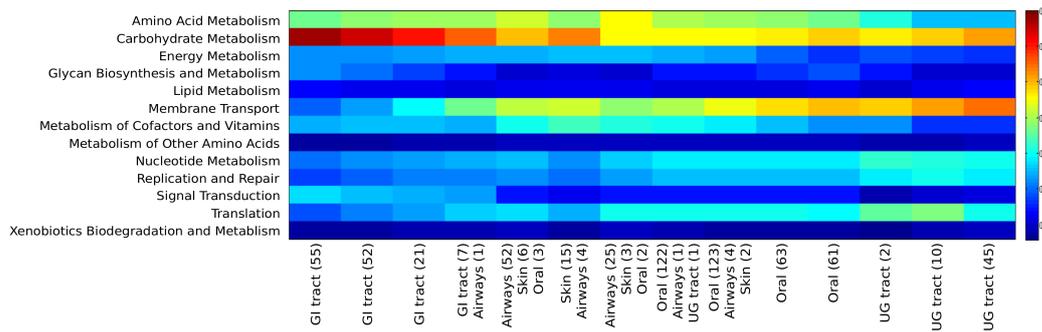
The Jaccard coefficient varied between 0.4325 and 0.8817 at body site level and between 0.2344 and 0.4801 at body subsite level. The partitioning of the MoP-Pro approach always showed the highest values on body site and subsite level. For both levels, the clustering performance of the MoP-Oligo approach is superior to HUMAnN except for the City block metric at body subsite level.

### 4.4 1D-SOM clustering and visualization

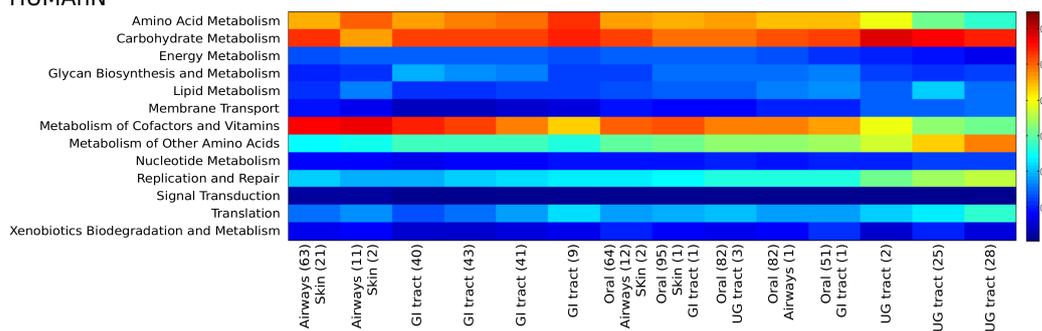
In order to study the overall variation of pathway abundance patterns over the whole range of HMP samples, we analyzed the estimated metabolic profiles with the MarVis tool. A one-dimensional self-organizing map (1D-SOM) was created using MarVis-Cluster (see Figure 1) to obtain a set of ordered prototypes well-suitable for visualization of profile variations.

Here, we utilized the second pathway level where we reduced the profiles to include just the top 10 pathways with the highest variance over all samples. Taking the union of the top 10 MoP and HUMAnN pathways we achieved a total of 13 profile dimensions that we used for 1D-SOM clustering with 14 prototypes and a unit 2-norm scaling of profile vectors. The resulting visualization indicates that most of the body sites are separated into distinct clusters (Figure 1). For the MoP profiles three major groups of clusters can be identified: gastrointestinal tract (GI tract, left side), urogenital tract (UG tract, right side), and an intermediate set of clusters from airways, oral and skin sites. Furthermore there are some interesting gradients (left to right) that show a decreasing relative abundance for *Amino Acid Metabolism*, *Carbohydrate Metabolism*, and *Signal Transduction* pathways and an increasing abundance for *Membrane Transport*, *Nucleotide Metabolism*, *Replication and Repair*, and *Translation* pathways. In contrast, the 1D-SOM based on the HUMAnN pathway profiles shows a distinct picture of the overall variation. The different body sites are not as clearly separated as for the MoP-based SOM and the overall abundance gradients of selected pathways are not as prominent as for the MoP results. The visible gradients (left to right) that show a decreasing abundance include the *Amino Acid Metabolism* and *Metabolism of Cofactors and Vitamins* pathways while an increasing abundance can be observed for *Metabolism of Other Amino Acids*, *Replication and Repair*, and *Translation* pathways.

MoP-Pro



HUMAnN



■ **Figure 1** 1D-SOM created with MarVis-Cluster at second pathway hierarchy level for MoP-Pro (upper) and HUMAnN (lower) profiles (GI tract – gastrointestinal tract; UG tract – urogenital tract). The numbers (in brackets) indicate the number of profiles (samples) assigned to the corresponding prototype (cluster) above.

## 4.5 Significant pathways

For a specific analysis of the metabolic profiles in terms of statistically significant differences in pathway abundances between different body sites we compared the gastrointestinal (GI) tract samples with all other HMP samples. To identify overrepresented pathways for the GI body site we applied an ANOVA with Holm-Bonferroni (FWER) correction on pathway abundances of the second level pathway hierarchy, filtered for pathways with an FWER below 0.05, and ranked the remaining pathways according to their fold-change in terms of the corresponding overrepresentation factor on mean abundances. In Table 6 the significant pathways of MoP-Pro and HUMAnN with a calculated fold-change larger than 1 are listed.

■ **Table 6** MarVis-Filter analysis for the identification of overrepresented pathways in the gastrointestinal tract samples in comparison to all other body subsites. All second level pathways obtained through the application of the MoP-Pro and HUMAnN approach with a fold-change larger than 1 are listed.

MoP-Pro		HUMAnN	
Pathway (second level)	Fold-Change	Pathway (second level)	Fold-Change
Transport and Catabolism	2.00	Digestive System	2.04
Signal Transduction	1.81	Endocrine System	1.12
Digestive System	1.79	Glycan Biosynthesis and Metabolism	1.07
Biosynthesis of Other Secondary Metabolites	1.53	Amino Acid Metabolism	1.06
Nervous System	1.48	Biosynthesis of Other Secondary Metabolites	1.06
Carbohydrate Metabolism	1.23	Energy Metabolism	1.01
Glycan Biosynthesis and Metabolism	1.15		
Endocrine System	1.13		
Immune System	1.12		

HUMAnN and MoP-Pro identified pathways associated with the *Digestive System*, *Endocrine System*, *Biosynthesis of Other Secondary Metabolites*, *Glycan Biosynthesis and Metabolism* to be overrepresented in GI tract samples. For all these pathways, except for the *Digestive System*, the MoP-Pro fold-change was higher than the corresponding factor of HUMAnN. Exclusively for the HUMAnN approach, pathways associated with *Amino Acid Metabolism* and *Energy Metabolism* are found to be slightly overrepresented. Furthermore, through the application of the MoP-Pro we detected five additional pathways to be overrepresented: *Transport and Catabolism*, *Signal Transduction*, *Nervous System*, *Carbohydrate Metabolism*, and *Immune System*. These additional pathways are possibly related to a mutually beneficial relationship between the gut microbiota and the host, maintaining a normal mucosal immune function and nutrient absorption. Furthermore, the overrepresentation of pathways associated with the nervous system may provide an indication for the bidirectional brain-gut interactions which have an important role in the modulation of gastrointestinal functions and possibly support the hypothesis of a communication pathway between the microbiota and the host's central nervous system [15].

## 4.6 Runtime

To get an overview of the computational cost of the different variants of the mixture modeling, we measured the approximate runtime averaged over the six selected HMP data samples (average size ~200 MB) used for the correlation analysis. For the selected data sets the

mean runtime ranges from minutes to months. The longest CPU times were required by the direct application of the mixture model due to the costly similarity searches against the KO database. On a computer with four CPUs (2.4 GHz) BLASTX searches and calculation of the metabolic profile took approximately 58 days. The fastest method was MoP-Oligo with about half a minute, followed by the MoP-Pro method with about one minute runtime in total. Once the taxonomic profile is estimated, using either MoP-Oligo or MoP-Pro, the resulting matrix vector multiplication for obtaining the metabolic profile of a metagenome can be done within a second.

## 5 Discussion

We presented a novel metabolic profiling approach for metagenomics, which is based on a mixture of pathways (MoP) model for estimation of pathway abundances. To overcome computationally intense homology searches, we implemented a shortcut to estimate the metabolic profile of a metagenome. Here, we link the taxonomic profile of the metagenome to a set of pre-computed metabolic reference profiles. The combination of the taxonomic abundance estimates, obtained through the fast methods Taxy-Oligo and Taxy-Pro, and the metabolic reference profiles, based on the KEGG database, achieves an unrivaled speed of the metabolic profiling approach.

We are aware of the difficulties in the evaluation that arise when trying to assess the quality of the resulting metabolic profiles. Therefore we restricted our evaluation to the large-scale data from the Human Microbiome Project (HMP) and to the comparison with the observations and findings for this data obtained through the HUMAnN approach. In this setup we tried to provide several views on metabolic profiles considering different aspects of quality: Our correlation analysis has shown that the pathway abundances obtained through our statistical model are slightly different when compared to the HUMAnN abundance predictions. However, we demonstrated through the nearest neighbor classification that our model based approach is at least comparable to the HUMAnN approach when considering the prediction of body sites and subsites. Considering the cluster performance analysis, our approach even outperforms the HUMAnN pipeline in most cases. Furthermore, our case study on statistically overrepresented pathways in the gastrointestinal tract provides additional insight in comparison with the results of the dedicated HUMAnN approach.

To our knowledge, the MoP approach for the first time provides a potentially unbiased estimator of the fraction of sequences that can be attributed to a particular pathway. In addition, our model-based combination with taxonomic abundance estimators also provides the fastest way to estimate the metabolic profile of a metagenome. We intend to make the method accessible via an easy-to-use interface by integration into the CoMet web server [11] (<http://comet.gobics.de>).

**Acknowledgements** We would like to thank Heiner Klingenberg, Thomas Lingner, Alexander Kaefer, and Manuel Landesfeind for fruitful discussions.

---

## References

- 1 Sahar Abubucker, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, Jeremy Zucker, Mathangi Thiagarajan, Bernard Henrissat, Owen White, Scott T. Kelley, Barbara Methé, Patrick D. Schloss, Dirk Gevers, Makedonka Mitreva, and Curtis Huttenhower. Metabolic Reconstruction

- for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Comput Biol*, 8(6):e1002358, 06 2012.
- 2 Robert D. Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E. Pollington, O. Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Research*, 38(suppl 1):D211–D222, 2010.
  - 3 The NIH HMP Working Group, Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A. Schloss, Vivien Bonazzi, Jean E. McEwen, Kris A. Wetterstrand, Carolyn Deal, Carl C. Baker, Valentina Di Francesco, T. Kevin Howcroft, Robert W. Karp, R. Dwayne Lunsford, Christopher R. Wellington, Tsegahiwot Belachew, Michael Wright, Christina Giblin, Hagit David, Melody Mills, Rachelle Salomon, Christopher Mullins, Beena Akolkar, Lisa Begg, Cindy Davis, Lindsey Grandison, Michael Humble, Jag Khalsa, A. Roger Little, Hannah Peavy, Carol Pontzer, Matthew Portnoy, Michael H. Sayre, Pamela Starke-Reed, Samir Zakhari, Jennifer Read, Bracie Watson, and Mark Guyer. The NIH Human Microbiome Project. *Genome Research*, 19(12):2317–2323, 2009.
  - 4 Daniel H. Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, 2011.
  - 5 Dazhi Jiao, Yuzhen Ye, and Haixu Tang. Probabilistic Inference of Biochemical Reactions in Microbial Communities from Metagenomic Sequences. *PLoS Comput Biol*, 9(3):e1002981, 03 2013.
  - 6 Alexander Kaefer, Manuel Landesfeind, Mareike Possienke, Kirstin Feussner, Ivo Feussner, and Peter Meinicke. MarVis-Filter: ranking, filtering, adduct and isotope correction of mass spectrometry data. *BioMed Research International*, 2012, 2012.
  - 7 Alexander Kaefer, Thomas Lingner, Kirstin Feussner, Cornelia Gobel, Ivo Feussner, and Peter Meinicke. MarVis: a tool for clustering and visualization of metabolic biomarkers. *BMC Bioinformatics*, 10(1):92, 2009.
  - 8 Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.
  - 9 Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.
  - 10 Heiner Klingenberg, Kathrin Petra Aßhauer, Thomas Lingner, and Peter Meinicke. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics*, 2013.
  - 11 Thomas Lingner, Kathrin Petra Aßhauer, Fabian Schreiber, and Peter Meinicke. CoMet - a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research*, 39(suppl 2):W518–W523, 2011.
  - 12 Victor M. Markowitz, I-Min A. Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Yuri Grechkin, Anna Ratner, Biju Jacob, Amrita Pati, Marcel Huntemann, Konstantinos Liolios, Ioanna Pagani, Iain Anderson, Konstantinos Mavromatis, Natalia N. Ivanova, and Nikos C. Kyrpides. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Research*, 40(D1):D123–D129, 2012.
  - 13 Peter Meinicke, Kathrin Petra Aßhauer, and Thomas Lingner. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics*, 27(12):1618–1624, 2011.
  - 14 Folker Meyer, Daniel Paarmann, Mark D’Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, Andreas 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.

- 15 Sang H Rhee, Charalabos Pothoulakis, and Emeran A Mayer. Principles and clinical implications of the brain–gut–enteric microbiota axis. *Nature Reviews Gastroenterology and Hepatology*, 6(5):306–314, 2009.



# 3 Exploring Neighborhoods in the Metagenome Universe

## Reference

Kathrin P. Abhauer, Heiner Klingenberg, Thomas Lingner, and Peter Meinicke: Exploring Neighborhoods in the Metagenome Universe. *International Journal of Molecular Sciences* 2014, 15(7):12364-12378.

## Original Contribution

KPA, TL, and PM designed the study. KPA and HK assembled the test data and performed the computer calculations. KPA performed the comparative  $k$ -nearest-neighbor search analysis (resulting in figure 1 and 2, supplementary table 1 and 2, and supplementary figure 1-6 and 8). HK and PM performed the dimensionality reduction analysis (resulting in figure 3 and supplementary figure 7). KPA implemented the frontend of the user-friendly CoMet-Universe web server application and integrated the  $k$ -nearest-neighbor search and 2D metagenome representation. KPA, TL, and PM wrote the manuscript. All authors read and approved the final version of the manuscript.



Article

## Exploring Neighborhoods in the Metagenome Universe

Kathrin P. Abhauer, Heiner Klingenberg, Thomas Lingner and Peter Meinicke \*

Department of Bioinformatics, Institute for Microbiology and Genetics, University of Göttingen, 37077 Göttingen, Germany; E-Mails: kathrin@gobics.de (K.P.A.); heiner@gobics.de (H.K.); thomas@gobics.de (T.L.)

\* Author to whom correspondence should be addressed; E-Mail: peter@gobics.de;  
Tel.: +49-551-39-14925.

*Received: 31 March 2014; in revised form: 23 June 2014 / Accepted: 25 June 2014 /*

*Published: 14 July 2014*

---

**Abstract:** The variety of metagenomes in current databases provides a rapidly growing source of information for comparative studies. However, the quantity and quality of supplementary metadata is still lagging behind. It is therefore important to be able to identify related metagenomes by means of the available sequence data alone. We have studied efficient sequence-based methods for large-scale identification of similar metagenomes within a database retrieval context. In a broad comparison of different profiling methods we found that vector-based distance measures are well-suitable for the detection of metagenomic neighbors. Our evaluation on more than 1700 publicly available metagenomes indicates that for a query metagenome from a particular habitat on average nine out of ten nearest neighbors represent the same habitat category independent of the utilized profiling method or distance measure. While for well-defined labels a neighborhood accuracy of 100% can be achieved, in general the neighbor detection is severely affected by a natural overlap of manually annotated categories. In addition, we present results of a novel visualization method that is able to reflect the similarity of metagenomes in a 2D scatter plot. The visualization method shows a similarly high accuracy in the reduced space as compared with the high-dimensional profile space. Our study suggests that for inspection of metagenome neighborhoods the profiling methods and distance measures can be chosen to provide a convenient interpretation of results in terms of the underlying features. Furthermore, supplementary metadata of metagenome samples in the future needs to comply with readily available ontologies for fine-grained and standardized annotation. To make profile-based  $k$ -nearest-neighbor search and the 2D-visualization of the metagenome universe available to

the research community, we included the proposed methods in our CoMet-Universe server for comparative metagenome analysis.

**Keywords:** metagenomics; functional profile; taxonomic profile; metagenome comparison

---

## 1. Introduction

With the rapidly increasing number of sequenced metagenomes in current databases it has become important to be able to compare novel metagenomic data with the existing data on a large scale [1,2]. In particular, the identification of closely related metagenome datasets (“neighbors”) to a newly obtained dataset is of growing importance for downstream analysis. Firstly, inspection of the neighbors and their associated annotations can be used as a final quality control of the dataset and may reveal unexpected flaws of the sampling, sequencing or data processing procedures. For instance, neighbors with unexpected habitat labels may indicate some contamination of the sample [3]. Secondly, related metagenome datasets in the neighborhood can be used as additional data sources for comparative analyses. Similar to biological replicates in gene expression analysis or homology extension in sequence analysis, the neighbors may be used for a statistical characterization of variations. However, manually identifying neighboring datasets on the basis of metadata can be misleading with the currently available coarse-grained and non-standardized annotation categories. If, for instance, the existing habitat annotations are used for sample selection, it is unclear which metagenomes are good neighbors for a data-driven comparative analysis, in particular, if a habitat label is rather abundant or rather sparse within the database.

Because metagenomic data usually consists of huge collections of short anonymous sequences, the comparison of two metagenomes is notoriously difficult. In analogy to comparative genomics a comparison may be conducted on a sequence-by-sequence basis to identify all pairwise similarities between two metagenomic data sets [4]. However, the computational cost for all pairwise sequence comparisons between a new query data set and  $n$  metagenomes in a database is prohibitively expensive due to the average size of a single file that may comprise several millions of sequences. Therefore, instead of the sequences it is reasonable to compare feature profiles that can represent relevant aspects of the functional and taxonomic composition of metagenomic sequence data [5–11]. But so far, it is unclear what kind of features and which metrics are most suitable for the comparison of metagenomes.

We present here a study of profile-based methods for nearest neighbor identification according to metagenome habitat annotation, using a broad spectrum of profile representations and distance metrics. Our results indicate that taxonomic as well as functional profiles can be used to retrieve related metagenomes in a database with a high confidence. Furthermore, we found that several standard metrics such as the City block or Euclidean distance are well-suitable for the identification of biologically meaningful nearest neighbors. In this context, we also investigated the performance of dimensionality reduction methods for visualization of the “metagenome universe”, where unsupervised kernel regression [12] showed the best representation in terms of neighborhood conservation.

## 2. Results and Discussion

The rapidly growing number of publicly available metagenomes nowadays requires efficient tools to compare and relate a novel metagenome to those in databases. In this study, we investigate the possibility to detect and visually explore metagenomic neighbors based on taxonomic, functional and metabolic profiles. In the following, we will first present the results of our evaluation of neighborhood accuracy and then discuss the opportunities and difficulties of a dimensionality-reduced representation of metagenome profiles for visual inspection.

### 2.1. Neighborhood Accuracy

The neighborhood accuracy measures the fraction of metagenomes with the same habitat label among the  $k$  nearest neighbors as obtained from a leave-one-out cross-validation. It is an estimator of the posterior probability to find related metagenomes within a local neighborhood of the profile space. For profile-based approaches the achievable accuracy depends on the particular feature space and the distance metrics that is used for comparison.

#### 2.1.1. HMP Collection

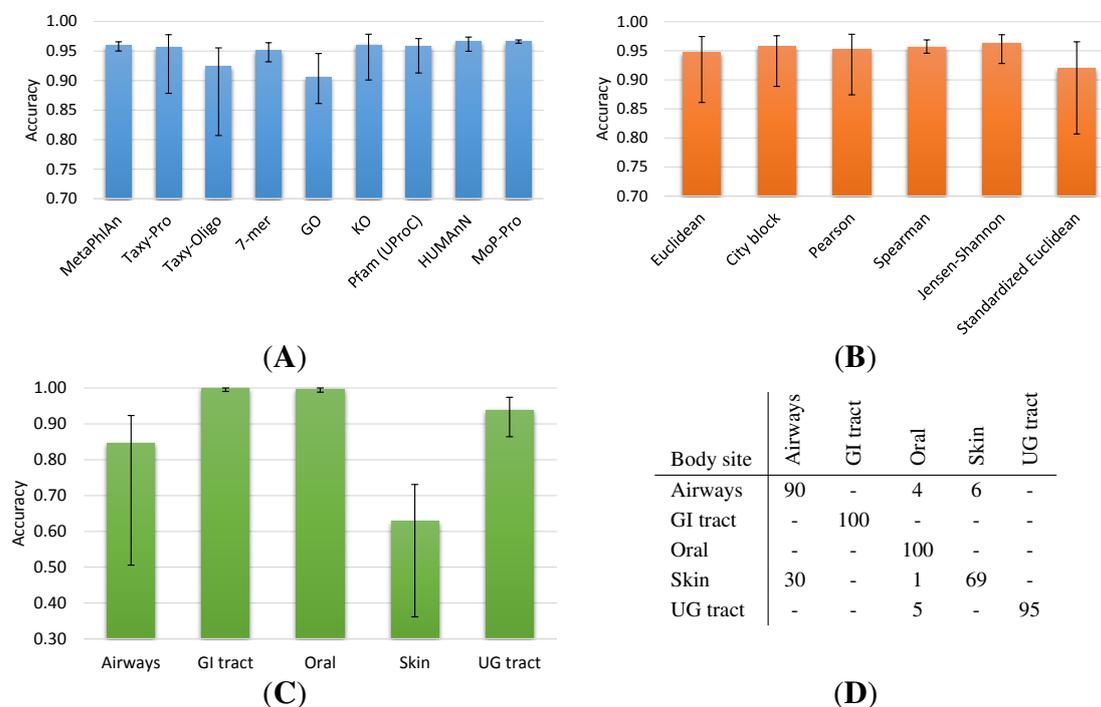
The Human Microbiome Project (HMP [13], see also Section 3.1.1.) provides high-quality sequencing data and a consistent habitat annotation of metagenomes in terms of distinct body sites. Therefore, we expect only a small overlap of HMP samples from different body sites, indicating a suitable benchmark dataset for the evaluation of metagenome profiling methods. Originally, the phylogenetic, functional, and metabolic profile of the HMP data have been investigated by means of the HMP Unified Metabolic Analysis Network (HUMANn) pipeline [15], the Metagenomic Phylogenetic Analysis (MetaPhlAn) tool [14] and a Gene Ontology (GO) Slim analysis. Besides these annotations we also used different taxonomical, functional and metabolic profiling methods as described in Section 3.2.1. and evaluated the  $k$  nearest neighbors according to Section 3.3.

Figure 1 shows the neighborhood accuracy on the HMP dataset for different profiling methods, metrics and body sites. Figure 1A indicates that in general a high fraction ( $\approx 90\%$  to  $97\%$  on average) of equally-labelled neighbors can be detected by all methods. Here, the MetaPhlAn and MoP-Pro methods show very little variation of the accuracy with respect to the underlying profile distance measure. On the other side, Taxy-Oligo and GO show a relatively low accuracy on average and are much more susceptible with respect to the distance metric. The GO Slim profile space has the lowest dimensionality and it seems to require a nonlinear metric or a more suitable normalization, while the relatively low accuracy of Taxy-Oligo is mainly caused by the standardized Euclidean metric (see Figure S1) that seems to be unsuitable for the corresponding profiles. This distance measure showed the lowest average accuracy for most of the methods (see Figure 1B), but as an exceptional case it did improve the performance of the 7-mer approach (see Figure S1).

Figure 1B also indicates the Spearman metric as the most robust distance measure with respect to the choice of the profiling method, however, the conversion of category counts to ranks for the calculation of this metric is problematic when only a few counts are present for many categories. Except for the

GO profile space, the City block metric generally showed a high accuracy and allows a fast calculation of distances as well as an intuitive interpretation. Further inspecting the City block results, we found that three HMP body sites (“GI tract”, “UG tract”, “Oral”) allow a high neighborhood accuracy for all methods, while the “Skin” and “Airways” categories show a low average accuracy and a large variation with respect to the utilized method (Figures 1C and S2). The low accuracy cannot be attributed to particular profiling methods or metrics (see Figures S3 and S4) and thus indicates a systematic overlap of categories. Indeed, the “Skin” body site comprises only a few datasets (26 samples) and the confusion matrix of the neighborhood evaluation (Figure 1D) indicates a large fraction of neighbor misassignments to the “Airways” category. Because the Airways samples have been taken from nose regions there might be a natural overlap with skin-associated microbial communities.

**Figure 1.** Neighborhood accuracy on Human Microbiome Project (HMP) data for different profiling methods, metrics and body sites. (A) Accuracy of profiling methods with average/minimum/maximum over six different metrics; (B) Accuracy of distance metrics with average/minimum/maximum over all nine profiling methods; (C) Body site-specific accuracy for City block metric averaged over nine profiling methods; (D) Confusion matrix of neighborhood evaluation for different body sites according to UProC protein domain profiles and City block metric. Values represent rounded percentages and entries lower than 0.5 are omitted.



Further characterization of the overlap in terms of the profile space distances turned out to be difficult because the corresponding neighborhood patterns can vary considerably. To illustrate this variation for the Airways and Skin body sites, we represented the metagenome neighborhood of two query metagenomes from the Airways category in terms of multidimensional scaling (MDS) plots and a hierarchical clustering analysis (HCA) of the neighboring functional profiles (see Figure S5). Based

on the evaluation of  $k = 10$  neighbors in the UProC domain profile space using the City block metric, the two Airways query metagenomes are in one case assigned to the right habitat (6 correct labels) and in the other case misclassified (4 correct labels). In the first case (Figure S5A,B) the five nearest neighbors of the query are grouped into a cluster consisting of four correctly (Airways) and one incorrectly (Skin) assigned neighbor(s). Another cluster shows a mixed composition of metagenomes from the Skin and Airways categories. In the second example (Figure S5C,D) the Airways query metagenome is grouped within a cluster of four Skin samples. However, another cluster consisting of four Airways samples and one Skin sample is located nearby. Although these examples indicate the difficulties of overlapping habitats, they do not allow inferences about the reasons of possible misclassifications. Here, further statistical analysis based on taxonomic, functional or metabolic features of the metagenomic neighbors would be necessary.

### 2.1.2. Metagenome Universe Collection

To investigate whether the findings on the HMP dataset collection could be reproduced with a more diverse range of biomes, we analyzed a set of 1745 publicly available metagenomes associated with twelve different habitat categories (“metagenome universe”, see Section 3.1.2. for details). Here, we expect that the overlap of categories is larger than in the HMP collection, since not all labels actually describe distinct environments. We excluded MetaPhlAn and the HUMAnN pipeline from the analysis for computational reasons and the Taxy-Oligo method because of its shortcomings regarding the profiling of viral metagenomes [9].

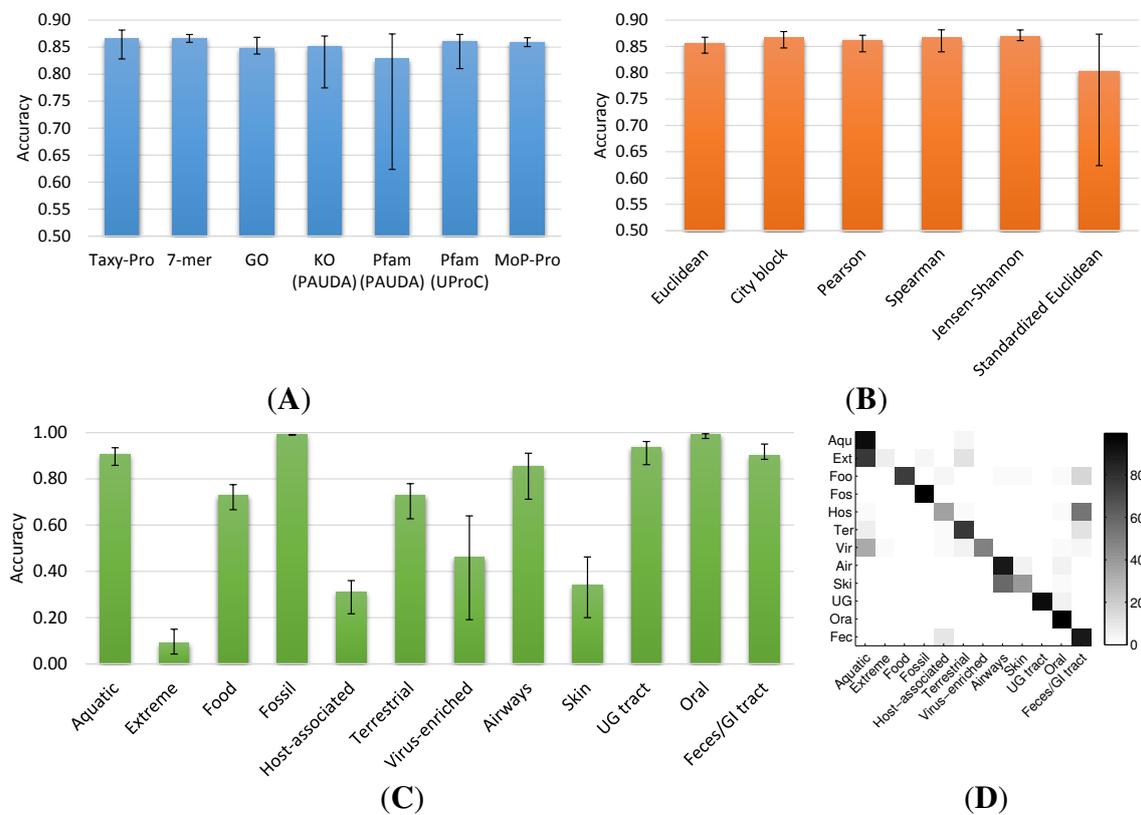
Figure 2 shows the neighbor detection performance on the metagenome universe collection for different profiling methods, metrics and habitats. In general, the average neighborhood accuracy of all methods is slightly lower (~83% to 87%) than on the HMP dataset (see Figure 2A). In particular, the protein alignment using a DNA aligner (PAUDA) method for detection of significant Pfam protein domains and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs shows a substantially lower accuracy and higher variation. This is mainly caused by use of the standardized Euclidean metric (see Figure 2B), which seems to be susceptible to small Pfam/KO counts resulting from the low sensitivity of the PAUDA similarity detection.

Concentrating on more robust metrics such as the City block and Spearman distance, we observe large differences in the ability to correctly detect neighbors for different habitat categories (see Figures 2C, S6 and S7). In particular, the categories “Extreme”, “Virus-enriched”, “Host-associated” and “Skin” indicate low accuracies and/or large variations with respect to the utilized profiling method. While the performance of the oligonucleotide-based 7-mer method noticeably decreases for virus-enriched and skin metagenomes, the GO method shows particularly low accuracy for the host-associated category.

Considering the habitat annotation of the metagenomes, the difficulties of the evaluation of neighborhood detection become apparent. For instance, the “Extreme”, “Virus-enriched” and “Host-associated” categories just provide a rather unspecific labelling of datasets. A closer look at the confusion matrix associated with our neighborhood evaluation using UProC indicates a systematic overlap of the “Extreme” and “Virus-enriched” categories with the “Aquatic” habitat and of “Host-associated” environments with the “Feces/GI tract” category (see Figure 2D and Table S1).

This can be well explained by the natural overlap of the annotation, which does not define mutually exclusive habitat categories in this case.

**Figure 2.** Neighborhood accuracy on metagenome universe collection for different methods and habitats. **(A)** Accuracy of profiling methods with average/minimum/maximum over six different metrics; **(B)** Accuracy of distance metrics with average/minimum/maximum over all seven profiling methods; **(C)** Habitat-specific accuracy for City block metric averaged over seven profiling methods; **(D)** Heatmap of confusion matrix for different habitats according to UProC protein domain profiles and City block metric. Habitat labels on y-axis abbreviated to three letters.

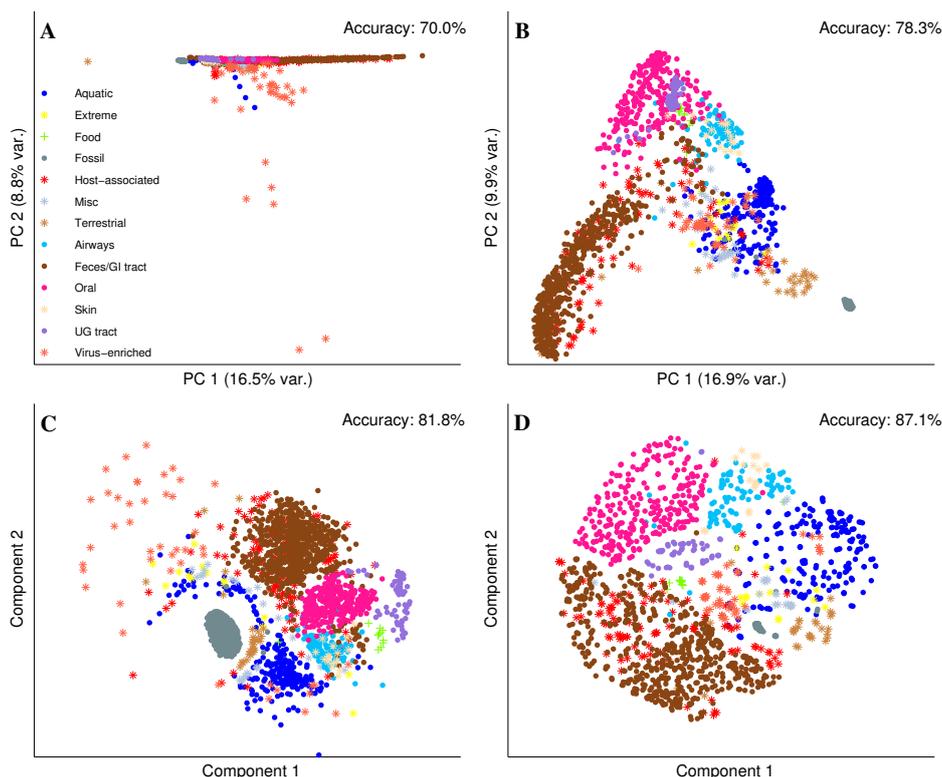


## 2.2. Visual Exploration of the Metagenome Universe

The objective of the dimensionality reduction was to obtain a two-dimensional representation of the comprehensive metagenome collection (“metagenome universe”) for scatter plot visualization. In a suitable scatter plot, data points appear closer to each other on the plot when they reflect similar properties. Therefore, the adjacent data points should correspond to related metagenomes with the 2D neighborhoods reflecting the habitat labeling. To obtain the scatter plots, we applied different dimension reduction methods to the UProC protein domain profiles of metagenomes. First, we applied classical principal component analysis (PCA) which showed the well-known susceptibility to outliers ([25], see Figure 3). In this case a few virus-enriched metagenomes are spanning the whole scatter plot and one has to zoom into the main part of the distribution to see meaningful neighborhoods. This is also reflected by the 2D Euclidean neighborhood accuracy which is only 72%. Plotting

subsequent principal components (e.g., PC 2 and PC 3) against each other or removing a few apparent (viral) outliers did not enhance the overview given by the PCA plot (data not shown). Only the complete removal of viral metagenomes from the database yields a visualization of the metagenome universe with distinguishable clusters according to habitats (see Figure S8). Using a City block distance matrix, classical multidimensional scaling (MDS) shows a more suitable sketch of the distribution with a considerably reduced influence of the virus-enriched metagenomes. This also resulted in an increased neighborhood accuracy of 78.3% for the MDS coordinates which show an interesting distribution. The shape corresponds to the so-called horseshoe effect which is well-known for MDS and occurs when only the distances between nearby points are representative [26]. Thus, we speculate that for unrelated habitats the distance between protein domain profile vectors does not reflect biologically meaningful differences. We also used the City block distances as an input for the Sammon mapping which also shows a good clustering of metagenomes according to their habitat and a slightly increased neighborhood accuracy of 81.8%. The most convincing result we achieved with unsupervised kernel regression which showed the best utilization of the image area and the highest neighborhood accuracy. In this case, the 87.1% accuracy in 2D was nearly as good as for the original space of the high-dimensional Pfam profiles (87.4%).

**Figure 3.** 2D representation of metagenome universe for different dimension reduction methods using UProC protein domain profile space. Markers represent metagenome datasets with colors corresponding to habitat labels as provided in legend in subfigure (A) Principal component analysis (PCA) using Euclidean metric with dimension-specific variance in parantheses; (B) Multidimensional scaling (MDS) using City block metric with dimension-specific variance in parantheses; (C) Sammon mapping using City block metric; (D) Unsupervised kernel regression (UKR) using City block metric.



### 3. Materials and Methods

In the following, we will describe the datasets and the experimental setup used in this study. First, we will give an overview of the two different collections of metagenome datasets and the methods used to compute the taxonomic, functional and metabolic profiles. Finally, we present the profile distance measures for neighbor detection and the dimensionality reduction methods used for visualization.

#### 3.1. Metagenome Dataset Collections

##### 3.1.1. HMP Collection

The Human Microbiome Project (HMP, [13]) provides an extensive collection of samples from human body sites of healthy individuals for large-scale comparative studies. More than a thousand HMP data sets have been recorded and are publicly available in HMP's Data Acquisition and Coordination Center (DACC) Project Catalog [29]. From the HMP-DACC website we obtained the available metadata for the metagenomic samples including taxonomic and functional annotations [30]. The taxonomic annotation comprises the results of the Metagenomic Phylogenetic Analysis (MetaPhlan) tool [14,31]. Further, we used the summary matrix of the Gene Ontology (GO) Slim analysis [32] ("GO Slim Summary File") and the functional and metabolic reconstruction data as precomputed through the HMP Unified Metabolic Analysis Network (HUMANN) pipeline [15,33] ("KEGG pathway abundance values—Summary file" and "Enzyme Abundance Data").

For our evaluation, we used 750 clinical study-related samples of HMP data as described in [9] (see also Supplementary Information). Furthermore, we restricted our evaluation to those body sites for which at least ten samples were available. The final dataset includes 640 data samples from five major body sites (see Table S2A).

##### 3.1.2. Metagenome Universe

In addition to the HMP datasets, we used a large collection of publicly available metagenome datasets from the MG-RAST [16] and European Bioinformatics Institute (EBI) online resources [17] to compile a "metagenome universe". For this purpose, all publicly accessible dataset files from the MG-RAST website [34] were downloaded in December 2012. We selected the FASTA files that passed the MG-RAST quality control filters and removed datasets with less than 1000 hits to Pfam protein domains. We used the metadata annotation of MG-RAST to assign each of the resulting 664 metagenomes to one of twelve habitat categories (see Table S2B). Furthermore, we downloaded the "project.csv" file from the EBI metagenomic projects website [35] and used it to obtain all associated FASTA files that also passed the EBI quality control. After filtering datasets with less than 1000 hits to Pfam domains, 821 of the 1307 samples were used for our reference database.

To reduce redundancy in the database, we computed Pfam domain profiles of all metagenomes and selected one representative file from datasets with a high profile correlation ( $>0.995$ ). Furthermore, we assessed taxonomic coverage quality values in terms of the "fraction of domains unexplained" (FDU, see [9]) for all metagenomes and removed those with an FDU value above 0.6. Finally, we removed datasets associated with profiles that had hits to less than 400 different Pfam families from our database.

As an exception, we did not apply this procedure to virus-enriched metagenomes, *i.e.*, datasets with a high fraction of viral DNA (>20% as measured by Taxy-Pro). The total number of datasets according to habitat categories can be found in Table S2B. A CSV-formatted list containing the metagenome identifiers and habitat labels as used in our evaluation can be found in Supplementary Dataset.

### 3.2. Profiling Methods

We used a variety of different profiling methods with largely varying dimensionality ranging from 61 (GO) to 16,384 (7-mer oligonucleotide frequencies). The theoretical dimensionality of the different profile spaces and the actual number of non-zero dimensions can be found in Table S3.

#### 3.2.1. Pfam Protein Domain Annotation

The ultrafast protein classification (UProC) that is part of the CoMet web server [8] was used for computation of the functional profiles according to the Pfam 27 database. The Pfam profiles also served for estimation of taxonomic and metabolic abundances with the protein-based mixture models (Taxy-Pro, MoP-Pro). For metagenome universe datasets we used the Pfam profiles to calculate GO functional profiles according to the HMP GO Slim ontology scheme. For this purpose, we downloaded the Pfam to GO mapping from the GO website [36] and counted all associations of GO Slim terms with Pfam domains detected in a metagenome.

#### 3.2.2. Taxonomic Profiling

The mixture model-based Taxy approach provides a computationally efficient and direct estimation of taxonomic abundances in metagenomes. Taxy-Oligo [18] and Taxy-Pro [9] apply a mixture model to approximate the overall metagenome distribution of oligonucleotides and protein domain hits, respectively. For the evaluation on HMP data, all reference profiles were obtained from 1912 bacterial and 133 archaeal genomes available in the KEGG database (release 64.0). These genomes were also used for precomputing the organism-specific pathway abundances for the metabolic profiling of metagenomes. For each reference genome we computed oligonucleotide (7-mers) and protein domain signatures. To measure the influence of the taxonomic model, the raw 7-mer oligonucleotide frequencies were used as an additional profile space.

For the evaluation of the metagenome universe, all archaeal, bacterial and viral genomes were downloaded from the National Center for Biotechnology Information (NCBI) FTP server [37,38]. They were complemented by 53 Eukaryotic genomes, 33 from diArk [19] and 20 from NCBI. As described in [9], we also included virus-enriched metagenomes to manage the underrepresentation of viral diversity in genome databases. For each reference, the Pfam profile was calculated and profiles with low coverage (<1000 Pfam hits) were excluded from downstream analysis. In addition, we removed similar profiles (correlation of >99% on phylum level) to reduce reference profile redundancy. This process reduced the number of reference genomes to 2199, including 157 Archaea, 1617 Bacteria, 50 Eukaryota, 273 Viruses and 102 viral metagenomes.

### 3.2.3. Mixture-of-Pathways

The Mixture-of-Pathways (MoP) model extends the taxonomic mixture model to a statistically adequate modeling of the metabolic potential of metagenomes [20]. MoP is based on a mixture model of pathways for the estimation of relative KEGG pathway abundances. To overcome computationally intense homology searches, we used the MoP-Pro approach introduced in [20]. MoP-Pro implements a shortcut to estimate the metabolic profile of a metagenome by linking the taxonomic profile of the metagenome to a set of pre-computed metabolic reference profiles. Here, organism-specific metabolic profiles in terms of KEGG Ortholog groups are computed for all bacterial and archaeal genomes in the KEGG database. Combining these organism-specific profiles according to the taxonomic profile of the metagenome, we estimate the relative pathway abundances by the posterior probabilities of the metabolic mixture model described in [20].

### 3.2.4. Protein Alignment Using a DNA Aligner (PAUDA) Annotation

The protein alignment using a DNA aligner (PAUDA) approach performs a protein database search [21]. PAUDA converts all protein sequences into pseudo DNA by mapping the amino acid alphabet onto a four-lettered alphabet. Then the read aligner Bowtie2 is used to compare the pseudo DNA reads with a pseudo DNA database. The statistical significance of matches is calculated based on protein alignments of the backtranslated protein sequences. PAUDA runs  $\sim 10,000$  times faster than BLASTX, while achieving about one-third of the assignment rate of reads to KEGG orthology groups. In this study, PAUDA was used to perform a search against the functional sequence database including all KEGG Orthologs of bacterial and archaeal origin available in the KEGG database (Release 64.0) and protein domain families in the Pfam database (Release 27). Here we extracted all full length sequences labeled according to their Pfam ID from the 'Pfam-A.full' multiple alignment file. The homology search was executed in `--fast` mode with default parameters. In the case of multiple matches, only the best hit is considered.

## 3.3. Nearest Neighbor Analysis

In our study, we introduce the concept of identifying neighbors of a query metagenome within a database of annotated reference metagenomes based on their taxonomic or functional profiles. For evaluation of the neighbor detection we performed a leave-one-out cross-validation on all metagenome profiles using a  $k$ -nearest-neighbor search with  $k = 10$ . As an accuracy measure we counted the fraction of profiles in the neighborhood with the same habitat label as the query profile. Here, we used the habitat assignments of publicly available metagenomes (see above) as obtained from their annotation. We utilized different linear and nonlinear metrics in the profile space to calculate the distances between pairs of metagenomes.

Let  $\mathbf{x}$  and  $\mathbf{y}$  be taxonomic or functional profile vectors of two metagenomes, then the City block (or  $L_1$ ) distance between  $\mathbf{x}$  and  $\mathbf{y}$  can be calculated according to

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i| \quad (1)$$

Note that in case of relative abundances, *i.e.*,  $\sum_i x_i = \sum_i y_i = 1$ , the City block distance corresponds to the Bray–Curtis dissimilarity, which is widely used in ecology for comparison of two assemblages [22]. Analogously to the City block metric, the Euclidean (or  $L_2$ ) distance can be computed according to

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2} \quad (2)$$

A standardized version of the Euclidean distance can be obtained by normalizing each profile dimension with respect to its standard deviation, *i.e.*,

$$d_{2s}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i \left( \frac{x_i - y_i}{\sigma_i} \right)^2} \quad (3)$$

The Pearson correlation coefficient

$$\varrho(\mathbf{x}, \mathbf{y}) = \frac{\sum_i (x_i - \mu^{\mathbf{x}})(y_i - \mu^{\mathbf{y}})}{\sqrt{\sum_i (x_i - \mu^{\mathbf{x}})^2 \cdot \sum_i (y_i - \mu^{\mathbf{y}})^2}} \quad (4)$$

between two metagenome profiles can be utilized as a distance according to  $d_P(\mathbf{x}, \mathbf{y}) = 1 - \varrho(\mathbf{x}, \mathbf{y})$ . Similarly, Spearman's rank correlation coefficient defines a distance metric  $d_S(\mathbf{x}, \mathbf{y}) = 1 - \varrho(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ , whereby  $\hat{\mathbf{x}}$  corresponds to a representation of the profile  $\mathbf{x}$  with values converted to ranks.

Finally, we used the Jensen–Shannon divergence, a symmetrized version of the Kullback–Leibler divergence  $d_{KL}$ , to measure the distance between metagenomes. The Jensen–Shannon divergence is defined by

$$d_{JS}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}d_{KL}(\mathbf{x}, \mathbf{m}) + \frac{1}{2}d_{KL}(\mathbf{y}, \mathbf{m}) \quad (5)$$

whereby  $d_{KL}(\mathbf{x}, \mathbf{y}) = \sum_i x_i \ln \left( \frac{x_i}{y_i} \right)$  and  $\mathbf{m} = \frac{1}{2}(\mathbf{x} + \mathbf{y})$ . To prevent numerical problems we excluded profile dimensions that did not contribute any counts from the computation.

### 3.4. Dimensionality Reduction

For visualization of the high-dimensional metagenome profile data we compared different dimensionality reduction methods: principal component analysis (PCA, [23]), classical multidimensional scaling (MDS, [23]), Sammon mapping [24] and unsupervised kernel regression (UKR, [12]). Sammon mapping and MDS were both based on City block ( $L_1$ ) distances and UKR was used with the  $L_1$ -kernel. The iterative optimization schemes of the Sammon and UKR methods were initialized with the MDS and  $L_1$ -kernel PCA, respectively. For computation we used the dimensionality reduction [39] and UKR [40] toolboxes in MATLAB. No additional parameters (hyperparameters) were required by any of the chosen methods. The resulting 2D coordinates of the dimensionality-reduced representation of all metagenomes were used for the neighborhood evaluation based on an Euclidean distance.

#### 4. Conclusions

The focus of our study has been on the comparison of unsupervised methods for metagenome similarity search. The aim was not to introduce a particular method that has been tuned to provide the best classification performance for a given labeling of the data. If the prediction of certain categories is the main objective, then supervised methods can be used that explicitly utilize the label information for parameter optimization [27]. However, our results indicate that the labeling of metagenomic data may also give rise to uncertain categories that are not well represented in terms of profile similarity. Therefore, a supervised approach may be adequate for a rather specific task if well-defined categories and reliable labels are available, for instance to predict a certain disease in a medical context. In contrast, an unsupervised approach to metagenome similarity computation can be more general and may even provide the potential for the discovery of novel or unexpected relationships. Furthermore, the performance of unsupervised methods does not depend on the quality of labels and mislabeled data may even be identified by inconsistent neighborhoods in profile space. On the other hand, metagenomic database retrieval would largely benefit from high-quality metadata and therefore the increasing acceptance of the “Minimum Information about a Metagenome Sequence” (MIMS) specification [28] will multiply the utility of profile-based metagenome comparison. We are aware of the fact that the coarse habitat-oriented labeling that we used in our comparison can only give a first impression of what is actually possible with profile similarity detection. However, the results indicate that a sequence feature-based identification of meaningful metagenomic neighbors is possible and computationally efficient for a wide range of profiles and distance metrics. Although we identified certain combinations that should not be used, in general no single metric or profiling method systematically outperformed the other methods in terms of the neighborhood accuracy. This implies that the profile space and the distance measure can in principle be chosen to allow a convenient interpretation of results in terms of the underlying features. In this context, protein families and metabolic pathways can provide a biologically more powerful representation than oligonucleotide-based features. With biologically meaningful profile features at hand our approach for neighbor identification allows subsequent in-depth analysis such as the identification and interpretation of features which contribute most to the distance between two metagenomes. Therefore, we have started to integrate a  $k$ -nearest-neighbor search based on protein domain frequency features in the CoMet-Universe server [41], which already implements some of the techniques that we have evaluated in our study.

#### Acknowledgments

We would like to thank two anonymous reviewers for their comments. This work was partially funded by a DFG grant (ME 3138) to P.M.

#### Author Contributions

K.P.A., T.L., and P.M. designed the study. K.P.A. and H.K. assembled the test data and performed the computer calculations. K.P.A. performed the comparative  $k$ -nearest-neighbor search analysis. H.K. and P.M. performed the dimensionality reduction analysis. K.P.A. implemented the frontend of the user-friendly CoMet-Universe web server application and integrated the  $k$ -nearest-neighbor search and

2D metagenome representation. K.P.A., T.L., and P.M. wrote the manuscript. All authors read and approved the final version of the manuscript.

### Conflicts of Interest

The authors declare no conflict of interest.

### References

1. Delmont, T.O.; Malandain, C.; Prestat, E.; Larose, C.; Monier, J.M.; Simonet, P.; Vogel, T.M. Metagenomic mining for microbiologists. *ISME J.* **2011**, *5*, 1837–1843.
2. Teeling, H.; Glöckner, F.O. Current opportunities and challenges in microbial metagenome analysis—A bioinformatic perspective. *Brief. Bioinform.* **2012**, *13*, 728–742.
3. Knights, D.; Kuczynski, J.; Charlson, E.S.; Zaneveld, J.; Mozer, M.C.; Collman, R.G.; Bushman, F.D.; Knight, R.; Kelley, S.T. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **2011**, *8*, 761–763.
4. Mailliet, N.; Lemaitre, C.; Chikhi, R.; Lavenier, D.; Peterlongo, P. Compareads: Comparing huge metagenomic experiments. *BMC Bioinform.* **2012**, *13*, S10.
5. Li, W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinform.* **2009**, *10*, 359.
6. Mitra, S.; Klar, B.; Huson, D.H. Visual and statistical comparison of metagenomes. *Bioinformatics* **2009**, *25*, 1849–1855.
7. Mitra, S.; Rupek, P.; Richter, D.C.; Urich, T.; Gilbert, J.A.; Meyer, F.; Wilke, A.; Huson, D.H. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinform.* **2011**, *12*, S21.
8. Lingner, T.; Abhauer, K.P.; Schreiber, F.; Meinicke, P. CoMet—A web server for comparative functional profiling of metagenomes. *Nucleic Acids Res.* **2011**, *39*, W518–W523.
9. Klingenberg, H.; Abhauer, K.P.; Lingner, T.; Meinicke, P. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics* **2013**, *29*, 973–980.
10. Sanli, K.; Karlsson, F.H.; Nookaew, I.; Nielsen, J. FANTOM: Functional and taxonomic analysis of metagenomes. *BMC Bioinform.* **2013**, *14*, 38.
11. Su, X.; Xu, J.; Ning, K. Meta-Storms: Efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. *Bioinformatics* **2012**, *28*, 2493–2501.
12. Meinicke, P.; Klanke, S.; Memisevic, R.; Ritter, H. Principal surfaces from unsupervised kernel regression. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1379–1391.
13. Group, T.N.H.W.; Peterson, J.; Garges, S.; Giovanni, M.; McInnes, P.; Wang, L.; Schloss, J.A.; Bonazzi, V.; McEwen, J.E.; Wetterstrand, K.A.; *et al.* The NIH human microbiome project. *Genome Res.* **2009**, *19*, 2317–2323.

14. Segata, N.; Waldron, L.; Ballarini, A.; Narasimhan, V.; Jousson, O.; Huttenhower, C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **2012**, *9*, 811–814.
15. Abubucker, S.; Segata, N.; Goll, J.; Schubert, A.M.; Izard, J.; Cantarel, B.L.; Rodriguez-Mueller, B.; Zucker, J.; Thiagarajan, M.; Henrissat, B.; *et al.* Metabolic reconstruction for Metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **2012**, *8*, e1002358.
16. Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.M.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; *et al.* The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* **2008**, *9*, 386.
17. Brooksbank, C.; Bergman, M.T.; Apweiler, R.; Birney, E.; Thornton, J. The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res.* **2014**, *42*, 18–25.
18. Meinicke, P.; Aßhauer, K.P.; Lingner, T. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics* **2011**, *27*, 1618–1624.
19. Hammesfahr, B.; Odronitz, F.; Hellkamp, M.; Kollmar, M. diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. *BMC Res. Notes* **2011**, *4*, 338.
20. Aßhauer, K.P.; Meinicke, P. On the estimation of metabolic profiles in metagenomics. In *German Conference on Bioinformatics 2013*; Beißbarth, T., Kollmar, M., Leha, A., Morgenstern, B., Schultz, A.K., Waack, S., Wingender, E., Eds.; OpenAccess Series in Informatics (OASISs); Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik: Dagstuhl, Germany, 2013; Volume 34, pp. 1–13.
21. Huson, D.H.; Xie, C. A poor man's BLASTX—high-throughput metagenomic protein database search using PAUDA. *Bioinformatics* **2014**, *30*, 38–39.
22. Bray, J.R.; Curtis, J.T. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **1957**, *27*, 325–349.
23. Ripley, B.D.; Hjort, N.L. *Pattern Recognition and Neural Networks*, 1st ed.; Cambridge University Press: New York, NY, USA, 1995.
24. Sammon, J.W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **1969**, *18*, 401–409.
25. Hubert, M.; Engelen, S. Robust PCA and classification in biosciences. *Bioinformatics* **2004**, *20*, 1728–1736.
26. Diaconis, P.; Goel, S.; Holmes, S. Horseshoes in multidimensional scaling and local kernel methods. *Ann. Appl. Stat.* **2008**, *2*, 777–807.
27. Liu, Z.; Hsiao, W.; Cantarel, B.L.; Drabek, E.F.; Fraser-Liggett, C. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics* **2011**, *27*, 3242–3249.
28. Yilmaz, P.; Kottmann, R.; Field, D.; Knight, R.; Cole, J.R.; Amaral-Zettler, L.; Gilbert, J.A.; Karsch-Mizrachi, I.; Johnston, A.; Cochrane, G.; *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **2011**, *29*, 415–420.

29. Human Microbiome Project DACC - HMPDACC Data Browser. Available online: [http://www.hmpdacc.org/resources/data\\_browser.php](http://www.hmpdacc.org/resources/data_browser.php) (accessed on 6 February 2013).
30. Human Microbiome Project DACC - HMP Project Catalog - View Dataset. Available online: <http://www.hmpdacc.org/catalog/grid.php?dataset=metagenomic> (accessed on 8 July 2014).
31. Human Microbiome Project DACC - HMSMCP. Available online: <http://hmpdacc.org/HMSMCP> (accessed on 11 February 2013).
32. Human Microbiome Project DACC - HMGS. Available online: <http://hmpdacc.org/HMGS> (accessed on 7 February 2013).
33. Human Microbiome Project DACC - HMMRC. Available online: <http://www.hmpdacc.org/HMMRC> (accessed on 5 April 2013).
34. MG-RAST - Home. Available online: <http://metagenomics.anl.gov/> (accessed on 6 November 2012).
35. EBI Metagenomics: Archiving, Analysis and Integration of Metagenomics Data < EBI metagenomics < EMBL-EBI. Available online: <https://www.ebi.ac.uk/metagenomics/> (accessed on 24 January 2014).
36. Gene Ontology Consortium — Gene Ontology Consortium. Available online: <http://www.geneontology.org/external2go/pfam2go> (accessed on 4 October 2013).
37. Index von <ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/>. Available online: <ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/> (accessed on 5 November 2013).
38. Index von <ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/>. Available online: <ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/> (accessed on 5 November 2013).
39. Matlab Toolbox for Dimensionality Reduction. Available online: [http://homepage.tudelft.nl/19j49/Matlab\\_Toolbox\\_for\\_Dimensionality\\_Reduction.html](http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html) (accessed on 30 September 2011).
40. Stefan Klanke: UKR Toolbox. Available online: [http://www.sklanke.de/ukr\\_toolbox.zip](http://www.sklanke.de/ukr_toolbox.zip) (accessed on 27 March 2014).
41. CoMet-Universe: Home. Available online: <http://comet2.gobics.de> (accessed on 31 March 2014).

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).



# 4 Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data

## Manuscript

The following manuscript was submitted to Bioinformatics as Application Note.

Supplementary material is available on <http://tax4fun.gobics.de>.

Kathrin P. Aßhauer<sup>1</sup>, Bernd Wemheuer<sup>2</sup>, Rolf Daniel<sup>2</sup> and Peter Meinicke<sup>1</sup>

<sup>1</sup> Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-University Göttingen, 37077 Göttingen, Germany

<sup>2</sup> Department of Genomic and Applied Microbiology and Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University Göttingen, 37077 Göttingen, Germany

## Original Contribution

KPA and PM designed the project. KPA and BW performed the computer calculations. KPA evaluated all approaches (resulting in figure 1). KPA and BW set the requirements of the tool. KPA wrote the source code and set up the Tax4Fun webpage. RD and BW assisted with biological expertise and extensively tested the software. KPA and PM wrote the manuscript. All authors read and approved the final version of the manuscript.



# Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data

Kathrin P. Abhauer<sup>1,\*</sup>, Bernd Wemheuer<sup>2</sup>, Rolf Daniel<sup>2</sup>  
and Peter Meinicke<sup>1</sup>

<sup>1</sup>Department of Bioinformatics, Institute of Microbiology and Genetics, Georg-August-University Göttingen, 37077 Göttingen, Germany

<sup>2</sup>Department of Genomic and Applied Microbiology and Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University Göttingen, 37077 Göttingen, Germany

**Motivation:** The characterization of phylogenetic and functional diversity are key elements in the analysis of microbial communities. Amplicon-based sequencing of marker genes, such as 16S rRNA, is a powerful tool for assessing and comparing the structure of microbial communities at a high phylogenetic resolution. Because 16S rRNA sequencing is more cost-effective than whole metagenome shotgun sequencing, marker gene analysis is frequently used for broad studies that involve a large number of different samples. However, in comparison to shotgun sequencing approaches, insights into the functional capabilities of the community get lost when restricting the analysis to taxonomic assignment of 16S rRNA data.

**Results:** Tax4Fun is a software package that predicts the functional capabilities of microbial communities based on 16S rRNA datasets. We evaluated Tax4Fun on a range of paired metagenome/16S rRNA datasets to assess its performance. Our results indicate that Tax4Fun provides a good approximation to functional profiles obtained from metagenomic shotgun sequencing approaches.

**Availability:** Tax4Fun is an open-source R package and applicable to output as obtained from the SILVAngs web server or the application of QIIME with a SILVA database extension.

Tax4Fun is freely available for download at <http://tax4fun.gobics.de/>.

---

\*to whom correspondence should be addressed

# 1 Introduction

Amplicon-based sequencing of marker genes is widely used for large-scale studies that involve many different sampling sites or time series. The common 16S rRNA gene-based analysis is a powerful tool for assessing the phylogenetic distribution of a metagenome but does not provide insights into the communities metabolic potential. Therefore, the prediction of the functional capabilities of a microbial community based on marker gene data would be highly beneficial. As a particular difficulty of such a predictive approach for most organisms in marker gene databases the genome and therefore the functional repertoire is not known. For instance, the SILVA SSU rRNA database [1] (SILVA 115 – full release) contains 3,808,884 rRNA sequences whereas KEGG (Release 71.1) [2] only comprises 2982 complete prokaryotic genomes. Recently, the PICRUSt approach was proposed to predict functional profiles of microbial communities using 16S rRNA gene sequences [3]. PICRUSt infers unknown gene content by an extended ancestral state reconstruction algorithm. The algorithm uses a phylogenetic tree of 16S rRNA gene sequences to link operational taxonomic units (OTUs) with gene content. Thus, PICRUSt predictions depend on the topology of the tree and the distance to the next sequenced organism. Because a nearest neighbor within the tree topology always exists, PICRUSt links all OTUs, even if distances are large. This procedure can be problematic when analyzing microbial communities with a large proportion of so far not well-characterized phyla.

Here, we present Tax4Fun, a novel tool for functional community profiling based on 16S rRNA data. In Tax4Fun the linking of 16S rRNA gene sequences with the functional annotation of sequenced prokaryotic genomes is realized with a nearest neighbor identification based on a minimum 16S rRNA sequence similarity. Tax4Fun can be applied to the output of 16S rRNA analysis pipelines that can perform a mapping of 16S rRNA gene reads to SILVA. The results of Tax4Fun indicate that the correlation of functional predictions with the metagenome profile is higher as compared to the PICRUSt tool.

# 2 Implementation

Our method provides a prediction of functional profiles on the basis of SILVA-labeled OTU abundances. After preprocessing and clustering of the 16S rRNA sequencing reads the resulting OTUs have to be assigned to reference sequences in the SILVA database. The SILVA assignment counts are then transformed to functional profiles using Tax4Fun, which proceeds in three steps.

First, the SILVA-based 16S rRNA profile is transformed to a taxonomic profile of the prokaryotic KEGG organisms. The linear transformation is realized by a pre-computed association matrix. The matrix was built from a BLASTN analysis where we extracted 16S rRNA gene sequences of all prokaryotic KEGG organisms and searched them against the SILVA SSU Ref NR database (Release 115). For the assignment, we require a sufficient sequence similarity according to a threshold on the BLAST bitscore ( $>1500$ ). A non-zero entry in this sparse matrix represents a valid assignment of a SILVA sequence identifier to one of the KEGG organisms. In case that  $K$  different KEGG organisms simultaneously show significant hits for a SILVA 16S rRNA gene sequence the corresponding entries in the association matrix were set to  $1/K$ .

Then, the estimated abundances of KEGG organisms are normalized by the 16S rRNA copy

number obtained from the NCBI genome annotations. Finally, the normalized taxonomic abundances are used to linearly combine the pre-computed functional profiles of the KEGG organisms for the prediction of the functional profile of the microbial community. The organism-specific reference profiles are estimated with the same method as used for the Taxy-Pro reference profiles [4]. For a fast computation of the organism-specific and metagenomic functional KEGG Ortholog (KO) profiles, we utilized UProC [5] and PAUDA [6], respectively.

### 3 Results

We applied Tax4Fun and PICRUSt to a collection of paired metagenome/16S rRNA datasets that have also been used in the original PICRUSt study [7–11]. Before applying Tax4Fun, the SILVA-based 16S rRNA profile was computed using the QIIME tool [12] or the SILVAngs web server [1], respectively. For each paired dataset, the Spearman correlation of the whole metagenome and the 16S rRNA-predicted KO profile was calculated. The resulting correlation coefficients are shown in Figure 1 for the UProC-based functional profiles.

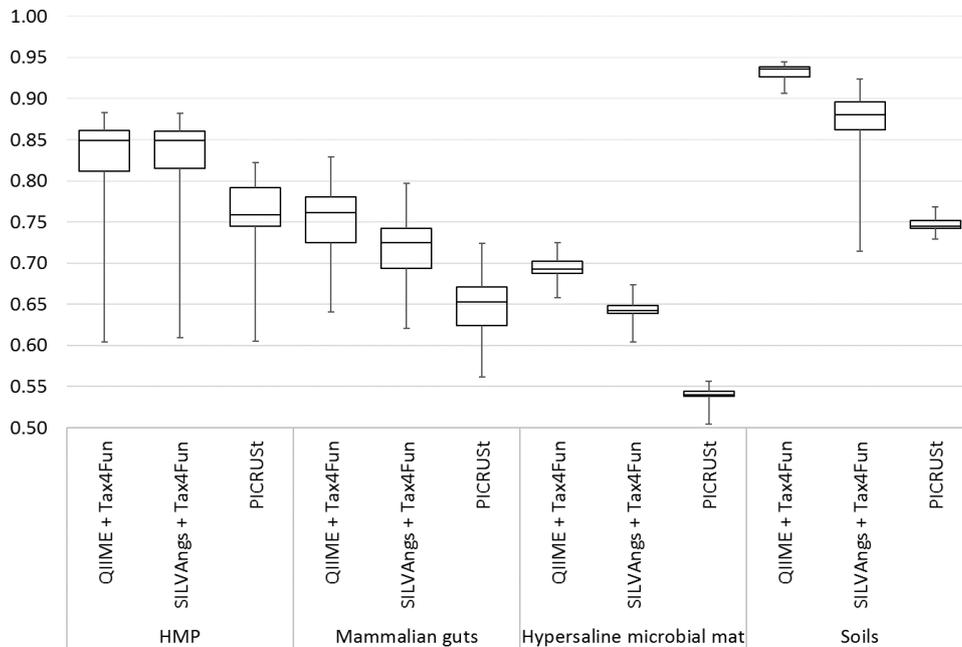


Figure 1: Spearman correlations between metagenomic and 16S-predicted functional profiles for comparison of Tax4Fun and PICRUSt on paired datasets from the human microbiome (HMP), mammalian guts, Guerrero Negro hypersaline microbial mat and soils.

Using Tax4Fun, the median of the correlation coefficient varies between 0.8706 (soils) and 0.6427 (Guerrero Negro hypersaline microbial mat). In comparison with PICRUSt the correlation of Tax4Fun is significantly higher for all four datasets according to a nonparametric sign test ( $p$ -value  $<0.001$ ). Similar results are obtained using the PAUDA tool for estimation of the functional profiles (see Tax4Fun web site).

Further, we compared the coverage of the analysis pipelines in terms of the fraction of reads that were classified by QIIME/SILVAngs and the percentage of OTUs that were mapped to KEGG organisms using Tax4Fun. Especially for the soil samples, we observed rather low fractions of 16S rRNA sequences that were finally used to predict the functional profiles (SILVAngs: 0.02%, Tax4Fun: 4.78%; QIIME: 95.21%, Tax4Fun: 55.36%). Contrary, the coverage for the human microbiome and mammalian guts datasets is rather high for both QIIME/SILVAngs and Tax4Fun (SILVAngs: 95%, Tax4Fun 95%). For all datasets, the coverage values are shown on the Tax4Fun website. Thus, the coverage of taxonomic assignments should always be inspected to check the reliability of predictions, in particular when using SILVAngs.

## 4 Conclusion

Tax4Fun predicts the functional profile of a microbial community just from 16S rRNA sequence data. Our approach cannot replace whole metagenome profiling but is useful to supplement 16S rRNA analyses in metagenome pre-studies or in situations where shotgun sequencing is prohibitively expensive, e.g. for broad surveys in microbial ecology applications. We evaluated our method on four paired data collections from different habitats and compared it to the PICRUSt tool. The results indicate a high correlation of the predicted Tax4Fun profiles with the corresponding functional profiles obtained from whole metagenome sequence data. Moreover, the results show that Tax4Fun outperforms PICRUSt on all test datasets. Tax4Fun allows easy processing of the output from SILVAngs, QIIME or any other analysis pipeline using the SILVA database as reference. The implementation in R facilitates further statistical analyses of the Tax4Fun predictions, which can be processed within the same R environment.

## Acknowledgement

**Funding:** Grants from the Deutsche Forschungsgemeinschaft (ME 3138, to P.M. in part, TRR51, to R.D. in part).

## References

- [1] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic Acids Research* 2013, **41**(D1):D590–D596, [<http://dx.doi.org/10.1093/nar/gks1219>].
- [2] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic Acids Research* 2014, **42**(D1):D199–D205, [<http://dx.doi.org/10.1093/nar/gkt1076>].
- [3] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C: **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.** *Nature Biotechnology* 2013, **31**(9):814–821, [<http://dx.doi.org/10.1038/nbt.2676>].
- [4] Klingenberg H, Aßhauer KP, Lingner T, Meinicke P: **Protein signature-based estimation of metagenomic abundances including all domains of life and viruses.** *Bioinformatics* 2013, **29**(8):973–980, [<http://dx.doi.org/10.1093/bioinformatics/btt077>].
- [5] Meinicke P: **UProC: tools for ultra-fast protein domain classification.** *Bioinformatics* in press.
- [6] Huson DH, Xie C: **A poor man’s BLASTX–high-throughput metagenomic protein database search using PAUDA.** *Bioinformatics* 2014, **30**:38–39, [<http://dx.doi.org/10.1093/bioinformatics/btt254>].
- [7] Human Microbiome Project Consortium: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486**(7402):207–214, [<http://dx.doi.org/10.1038/nature11234>].
- [8] Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrisat B, Knight R, Gordon JI: **Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans.** *Science* 2011, **332**(6032):970–974, [<http://dx.doi.org/10.1126/science.1198719>].
- [9] Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG: **Cross-biome metagenomic analyses of soil microbial communities and their functional attributes.** *Proceedings of the National Academy of Sciences* 2012, **109**(52):21390–21395, [<http://dx.doi.org/10.1073/pnas.1215210110>].
- [10] Harris JK, Caporaso JG, Walker JJ, Spear JR, Gold NJ, Robertson CE, Hugenholtz P, Goodrich J, McDonald D, Knights D, Marshall P, Tufo H, Knight R, Pace NR: **Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat.** *The ISME Journal* 2013, **7**:50–60, [<http://dx.doi.org/10.1038/ismej.2012.79>].

- [11] Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P: **A Bioinformatician's Guide to Metagenomics**. *Microbiology and Molecular Biology Reviews* 2008, **72**(4):557–78, [<http://dx.doi.org/10.1128/MMBR.00009-08>].
- [12] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data**. *Nature Methods* 2010, **7**(5):335–336, [<http://dx.doi.org/10.1038/nmeth.f.303>].

## 5 Discussion

In this work, I introduced algorithmically efficient statistical models for comparative metagenome analysis. The Mixture-of-Pathways (MoP) model provides a strictly probabilistic description of metagenomic sequence data in terms of KEGG Pathways (see Chapter 2). To avoid the expense to compute the functional profile in terms of KEGG Ortholog groups via homology search, I introduced the nested metabolic mixture model (see Chapter 2). In the nested model, the metabolic profile of the metagenome is predicted from its taxonomic profile by a linear combination of pre-computed metabolic reference profiles. The organism-specific profiles are estimated by applying the MoP model to all prokaryotic genomes in the KEGG database.

Further, I transferred the idea of predicting the metabolic profile from the taxonomic profile to 16S rRNA sequence data. In Chapter 4, I introduced Tax4Fun, a novel tool for functional community profiling based on 16S rRNA data. For Tax4Fun, I realized a linking of 16S rRNA sequences with the functional annotation of sequenced prokaryotic genomes with a nearest neighbor identification based on a minimum 16S rRNA sequence similarity. The linking, based on close homologies between 16S rRNA gene sequences, enables the transformation of SILVA-labeled OTU abundances to KEGG based taxonomic profiles. Like in the nested metabolic mixture model, the functional profile is computed from the transformed KEGG based taxonomic profile by a linear combination of pre-computed functional reference profiles. Tax4Fun utilizes the SILVA database as reference and can be applied to the output of 16S rRNA analysis pipelines performing a mapping of 16S rRNA reads to SILVA, e.g., by using the QIIME or the SILVAngs web server [102].

The results of the nested metabolic mixture model and Tax4Fun approach reveal that the taxonomic composition can be used to approximate the metabolic profile of a microbial community. In metagenomics, the application of the nested mixture model obviates the need for the computationally demanding estimation of functional abundances by homology search. Moreover, the Tax4Fun approach broadens the perspective of microbial communities based on the more cost-effective 16S rRNA sequencing. The application of Tax4Fun extends the explanatory potential of 16S rRNA studies from assessment of the phylogenetic distribution to valuable insights into the metabolic capabilities of microbial communities.

Further, I investigated the important problem how to automatically identify metagenomes that are closely related to a newly obtained dataset. First and foremost, similar metagenomes can serve as additional data sources in comparative analyses. Second, unexpected habitat labels of the related metagenomes provide an indication of potentially mislabeled datasets. However, the unsupervised identification of closely related metagenomes is a mostly uninvestigated field for comparative metagenome analysis, I evaluated different feature profiles and metrics for a

$k$ -nearest-neighbor search in the study “Exploring Neighborhoods in the Metagenome Universe” (see Chapter 3).

I implemented the nested metabolic mixture model and  $k$ -nearest-neighbor search approach in the CoMet-Universe webserver (<http://comet2.gobics.de/>) which enables exploratory analysis and comparison of metagenome data. Further, I implemented the Tax4Fun and the metabolic mixture model approach in R. The R Package can be easily integrated into existing analysis pipelines and facilitates further statistical analyses. Both approaches can process large datasets in reasonable time without using a computer cluster or special hardware. Due to the computational efficiency, both tools are well suitable for large-scale analysis on the growing amount of metagenomic sequence data.

## 5.1 Predicting the functional and metabolic potential from the taxonomic profile

The study of microbial communities using either 16S rRNA or whole-metagenomic sequencing has become a routinely used tool. Metagenomics provides a comprehensive view of the taxonomic composition as well as the functional and metabolic potential of heterogeneous microbial communities under natural conditions. In the following two sections, I separately discuss the MoP model and the Tax4Fun approach for a quantitative metabolic characterization.

### 5.1.1 Mixture-of-Pathways: a probabilistic model for the quantification of the metabolic potential from metagenomic shotgun data

For analyzing microbial communities in terms of their metabolic potential, a central task is the assignment of sequencing reads to functionally characterized categories and the subsequent transformation into pathway abundances. Several approaches using different counting schemes have been proposed for computing the pathway abundances. However, existing approaches do not properly address the ambiguity of the function-to-pathway mapping (see Section 1.3.1.2).

In this work, I introduced a metabolic mixture model in terms of a mixture of pathways which addresses the ambiguity of the function-to-pathway mapping (see Chapter 2). The MoP model provides an estimate of the fraction of sequence material that can be mapped to a particular pathway. The strict probabilistic description of metagenomic sequence data in terms of pathways enables a clear statistical interpretation of the resulting quantities. However, the direct application of the metabolic mixture model requires a functional profile as input to estimate the fraction of sequence material that can be mapped to a particular pathway. The computational effort for the identification of homologs can become burdensome and most time will be spent on the computation of the functional profile (see Section 1.3.1.1). To illustrate, running BLASTX requires thousands of CPU hours per million reads. In a metagenomic study of 12 permafrost samples [63] the functional annotation against the KEGG database of approximately 246 million reads reportedly took 800,000 CPU hours [147]. Both, the MoP model and existing approaches are impaired by the effort to compute the functional abundances.

In Chapter 2, I further addressed the computational challenges related to the prediction of the functional capacities. The deep impact of the taxonomic distribution on the functional properties

have been demonstrated in several microbial community studies [108–110]. Since the taxonomic composition of a metagenomic sample can be efficiently estimated, I investigated whether the functional capacities of a microbial community can be predicted from the taxonomic distribution. In Chapter 2, I introduced a shortcut to estimate the metabolic profile of a metagenome avoiding the expense to compute the functional profile. The metabolic profile is predicted by linking the taxonomic profile of the metagenome to a set of pre-computed reference profiles. For the taxonomic abundance estimation, I used the fast mixture model based methods Taxy [216] and Taxy-Pro [217]. When using Taxy-Pro, the nested model is further denoted MoP-Pro. By using Taxy or Taxy-Pro the taxonomic profile can be computed on a local standard PC in a few minutes. The subsequent linear combination of reference profiles is achieved in an unrivaled speed without the need of a computationally intense homology search for the functional profile of a metagenomic sample. The reference profiles are pre-computed by applying the MoP model to BLAST-derived functional abundances from 2045 prokaryotic reference organisms available in the KEGG database (Release 64.0). However, this computationally intense homology search has to be performed once.

The evaluation of the runtime of the different metabolic mixture modeling variants pointed out that the total CPU time can be considerably decreased using the nested metabolic mixture model. For HMP data samples with average size of 200 MB, the computational cost on a computer with four CPUs (2.4 GHz) can be reduced from 58 days to about one minute runtime in total. Due to the computational efficiency, even a large-scale comparison of many metagenomic datasets can be realized in reasonable time.

Regardless of the differences in computational cost, the evaluation on datasets from the large-scale HMP study indicates a high correlation between the predicted metabolic profiles obtained through the direct and nested metabolic mixture model. Moreover, our analysis of the clustering performance shows that the nested mixture model outperforms the HUMAnN approach in most cases. The case study on statistically overrepresented pathways further demonstrates that the nested model provides valuable insights into the metabolic capacities of microbial communities. Further, additional insights in comparison to the HUMAnN approach could be obtained. Thus, the MoP approach provides both a fast and potentially unbiased estimator of the metabolic profile of a metagenome.

### **5.1.2 Tax4Fun: predicting the functional potential of microbial communities using 16S rRNA sequence data**

In Chapter 2, I examined whether the functional properties of microbial communities can be estimated based on taxonomic profiles predicted from metagenome sequence data. However, about 90% of publicly available microbial community datasets have been taxonomically investigated using the amplicon based 16S rRNA approach [162, 208]. The analysis of 16S rRNA data is a very common tool for large sample collections of microbial communities in ecological and human microbiome research, but inherently limited to the characterization of the taxonomic composition and phylogenetic diversity (see Section 1.1.1). With the restriction on 16S rRNA marker genes, insights into the functional inventory of the communities generally get lost. So far, the PICRUSt tool provides the only approach which addresses this drawback, showing that in principle it is possible to predict functional profiles from 16S rRNA data.

In Chapter 4, I introduced the Tax4Fun approach which efficiently predicts the functional repertoire of a microbial community from 16S rRNA data. Tax4Fun transfers the nested mixture model approach to 16S rRNA studies (see Chapter 4). The core of the Tax4Fun approach is a transformation of the SILVA-labeled OTU abundances to a taxonomic profile based on KEGG reference organisms and a subsequent linking with pre-computed functional or metabolic abundances. For the transformation, I implemented a conservative mapping of 16S rRNA gene sequences to sequenced reference genomes. The association between 16S rRNA gene sequences from the SILVA database and sequenced reference organisms is established based on a homology search. Here, only associations for relationships to closely related sequenced organisms are allowed where a reasonable conclusion can be drawn with certainty. The transformation is entirely based on close homologies between 16S rRNA gene sequences and substantially differs from the PICRUSt approach that relies on a phylogenetic tree of the 16S rRNA sequences. The PICRUSt approach infers unknown gene content from the closest sequenced organism based on extended ancestral state reconstruction algorithm using a phylogenetic tree of 16S rRNA sequences as reference. However, the distance to the next closely related sequenced organism is not restricted. Consequently, PICRUSt considers all taxonomic units regardless of the distance to the next closely related sequenced organism Section 1.2.1. This procedure can be problematic when analyzing microbial communities with a large proportion of so far not well-characterized phyla.

I thoroughly assessed the performance of Tax4Fun and compared the predictions of both Tax4Fun and PICRUSt with functional profiles obtained from whole metagenome sequence data on several test datasets [78, 110, 213–215]. In this work, I used the well-established QIIME software package and the web-based NGS analysis pipeline of the SILVA rRNA gene database project (SILVAngs 1.0) [102] for classification of 16S rRNA reads. But, any other 16S rRNA pipeline is conceivable. The application of Tax4Fun to a range of paired data samples, where both 16S rRNA and metagenome data is available, revealed a strong correspondence between the predicted functional profiles for both data types. The evaluation indicates that Tax4Fun provides a good approximation to the functional profiles and consistently outperforms the PICRUSt approach. In addition, I investigated the application of KEGG Pathway reference profiles. As pathway reference profiles, I used the metabolic reference profiles computed for the MoP-Pro approach (see Chapter 2). The metabolic abundances predicted by the Tax4Fun approach and the direct annotation of metagenome data by MoP-Pro have a high correspondence. Detailed investigations have shown that the Tax4Fun predictions produce very similar biological findings as reported in the original publication for the functional annotation of the metagenome data samples.

Tax4Fun provides valuable insights into the functional and metabolic repertoire of microbial communities for which only 16S rRNA surveys are available. Furthermore, microbial communities can be efficiently sampled with the cost-effective 16S rRNA strategy and conclusions about both the taxonomic and functional properties can be drawn. In two-stage study designs (see Section 1.1.3), the functional properties can be used for targeted in-depth second-stage profiling sample selection. All in all, the explanatory potential of 16S rRNA studies is increased with the new Tax4Fun approach.

## 5.2 Limitations of the prediction of functional capacities using the taxonomic profile for approximation

The nested MoP model and the Tax4Fun approach provide a fast estimation of functional and pathway abundances from metagenome and 16S rRNA data. Both studies demonstrated that knowing “Who is in there?” may serve as a proxy for the substantial question “What are they doing?”. However, both the MoP-Pro and Tax4Fun approach use a KEGG based taxonomic profile restricted to prokaryotes for linear combination with pre-computed functional or metabolic reference profiles. The reference profiles again are limited to KEGG Orthologs and Pathways for which an annotation in at least one prokaryotic organism exists. Since the reference data is restricted to prokaryotic annotations, the models exclude the annotation of the eukaryotic or viral subset of a microbial community. Thus, both approaches do not infer capabilities restricted to viral or eukaryotic members of a microbial community.

In addition, the approach is limited to known capacities of the KEGG reference organisms. Therefore, potentially exclusive functions encoded in unknown organisms cannot be revealed. This is reflected by the number of sequences that lack a functional annotation or cannot be assigned to a reference organism. Typically, one-third of the metagenome sequences lacks a reference genome and about half of the genes miss a known function [62, 149, 204, 218, 219]. The majority of the sequences are hypothetical or conserved hypothetical genes. Therefore, the analysis of microbial communities is affected by the restriction of functional prediction to genes with known function and the large number of organisms that have not been cultured, or even sequenced. In the following, I discuss both points individually.

### 5.2.1 Measuring the reliability of taxonomic and functional predictions

Taxonomic and functional predictions of microbial communities are impaired by the number of sequences that lack a functional annotation or cannot be assigned to a reference organism. Different measurements can be considered for quantifying the proportion that is not included in the taxonomic or functional characterization. For the nested MoP model, we proposed the estimation of the taxonomic abundances by applying Taxy or Taxy-Pro. For Taxy and Taxy-Pro, the fraction of oligonucleotides unexplained (FOU), the fraction of sequences unassigned (FSU), and the fraction of domain hits unexplained (FDU), respectively, were introduced which are important indicators of the reliability of the functional and taxonomic abundance estimates. Generally, the lower the value of one of these indicators, the more we can rely on the taxonomic profile. In Taxy-Pro, the FSU corresponds to the amount of sequences without Pfam domain hits. A corresponding measure can be applied to the direct MoP model approach. Here, the proportion of sequences without significant hits to the KEGG Orthologs can be used as measure of uncertainty for the underlying functional profile. The smaller the number of sequences used to estimate the functional abundance profile, the higher the uncertainty of the subsequent metabolic estimates.

For Tax4Fun, both the taxonomic classification rate of 16S rRNA sequencing reads and the amount of classified sequences mapping to the prokaryotic KEGG reference organisms have to be taken into account. The quality of a 16S rRNA data analysis pipeline, e.g. QIIME or SILVAngs, can be determined in terms of the fraction of sequences classified. Further, the fraction

of sequences that contribute to the functional predictions must be determined. For Tax4Fun, the corresponding measure is termed fraction of taxonomic units unexplained (FTU) which reflects the amount of sequences assigned to a taxonomic unit and not transferable to KEGG reference organisms.

If the quality of one of these measures is completely insufficient for a description of a microbial community in terms of taxonomic or functional abundances, the direct and nested metabolic mixture model as well as the Tax4Fun approach become inadequate. For example, the highly diverse soil data samples show a very poor taxonomic description using SILVAngs. Only 0.02 percent of the reads can be taxonomically assigned. From these only 4.78 percent can be used for functional annotation using the Tax4Fun approach. Likewise, the functional annotation of the soil metagenome sequence data show a low fraction of annotated sequences. For both datasets, a relatively small amount of sequences is used for functional inference. The poor coverage may indicate a large proportion of unknown functions and organisms in the soil datasets which is consistent with the commonly described complexity and heterogeneity of soil ecosystems [220]. In contrast, the HMP data samples exhibit a very high taxonomic and functional annotation coverage for both the 16S rRNA and metagenome datasets. In the case of HMP, the measurements indicate a good coverage with the existing reference organisms and thus a reliable prediction. Consequently, the measurements described above have to be taken into consideration in order to assess the confidence of the predictions. In case one of the measurements indicates an unreliable prediction, the derived profiles should be treated with caution.

### **5.2.2 The need for reference genomes: closing the gap of the unknown microbial diversity**

The coverage of sequencing reads in terms of functional and taxonomical annotation is limiting the analysis of microbial communities. Often a substantial fraction of metagenome sequence reads cannot be assigned to a function or organism. In addition, a number of organisms are only known by their 16S rRNA gene sequence. Both issues re-emphasize the narrow picture of the microbial diversity through culture-based methods. Reference genome sequences are needed for both the taxonomic and functional annotation of metagenome data as well as for the inference of functional properties from 16S rRNA data. Especially, MoP-Pro and Tax4Fun will highly benefit from a higher coverage of the microbial diversity since the functional prediction is derived from the annotation of available reference organisms. In particular, the transformation of 16S rRNA sequence identifiers to KEGG reference organisms is highly affected by the limited number of available reference organisms. For example, the Tax4Fun association matrix attempts to link 3,808,884 SSU Parc aligned rRNA sequences (SILVA Release 115) to 2045 sequenced prokaryotic KEGG genomes (KEGG Release 64.0) with functional annotation of high quality. However, one must consider that not every 16S rRNA sequences correspond to a species since the copy number per genome vary from 1 to 15 or more copies [221] which commonly differ by more than 1% [222]. However, it is estimated that the amount of 16S rRNA variants is 2.5 fold greater than the number of bacterial species [223]. Despite these facts, a substantial dimensional difference of sequenced organisms and those only known by their 16S rRNA gene sequence is currently existing.

Systematic investigations are indispensable in order to increase the coverage of the microbial

diversity and to eliminate the cultivation bias towards easy grown microorganisms. Initiatives such as the HMP [224] and the Genomic Encyclopedia of Bacteria and Archaea (GEBA) [225] produce reference genomes by the thousands. These and other initiatives make use of the inherent advantages of the new sequencing technologies. The high throughput and low cost led to an increase in the production of "complete" genome sequences. More than 500 new species are now described annually [226, 227]. However, a general change from a few finished high-quality genomes to a high number of draft genomes can be found [228]. Nevertheless, this procedure helps to close the gap of under-represented organisms along with the contribution to the improvement of the annotation of microbial communities.

Additionally, it is often not possible to differentiate between closely related species due to the restricted resolution of the 16S rRNA gene. This limits both the diversity estimates obtained by OTU construction and the calculation of the functional or metabolic abundances by the taxonomic composition [229, 230]. For example, if the taxonomic resolution at species level is poor and a 16S rRNA gene sequence is assigned to the genus *Pseudomonas* this can indicate the presence of either a beneficial or pathogenic bacterium in a sample with distinct functional properties [231]. Another striking example of a limited resolution of the 16S rRNA sequence occurs in the two strains of *Escherichia coli*, O157:H7 and K-12. Although both *Escherichia coli* strains have extremely closely related 16S rRNA sequences, there is considerable variation in the genomes. For these two strains, it is reported that they differ in hundreds of genes and have significant differences in their major functions [232, 233].

### 5.2.3 Beyond the functional and metabolic potential

Although a wealth of knowledge can be gained from metagenomic and amplicon-based community characterizations, these methods only provide information about the functional potential. The inherent limitations of the techniques itself constitute one of the biggest drawbacks. Using metagenomics or the taxonomic composition of a microbial community, all genes or members serve as basis for the functional and metabolic abundance estimation. However, a large proportion of the microorganisms in a given environment is inactive at a particular time. In addition, not only living but also dead cells are considered which can make up a considerable proportion of a sample. For instance, in fecal samples more than half of the cells are non-viable or heavily damaged [234]. Further, the techniques cannot distinguish between expressed and non-expressed genes in a sample [235]. Because both active and inactive microorganisms and expressed and non-expressed genes are used for the calculations, the approaches are incapable of displaying the actual metabolic activity. Therefore, metagenomics in general and especially the MoP(-Pro) and Tax4Fun approach give insights into the functional and metabolic potential of organisms in a selected habitat. Whether identified metabolic pathways are functional, to what extent, and under which conditions remain to be quantified. Therefore, targeted methods that are able to detect the expressed genetic information, proteins, and metabolites of a microbial community are necessary to go beyond the potential capacities.

## **Outlook: metatranscriptomics, metaproteomics, and metametabolomics**

Recently, more complementary meta'omics technologies became available to describe entire ecosystems. These complementary techniques, e.g. metatranscriptomics, metaproteomics [236], and metametabolomics, are more likely to consider the active constituents of the microbial consortia, thus able to reveal the dynamics of genes, proteins, and metabolites in an environment.

Transcriptional level control of gene expression enables microorganisms to rapidly adapt to changing environmental conditions. In metatranscriptomics, the whole transcriptome of a microbial community is analyzed [237]. By investigating the actively transcribed ribosomal and messenger RNA (mRNA), metatranscriptomics can reveal which microbial organisms are active and which genes are actually being expressed in different environments and to what extent. However, metatranscriptomics faces several methodological challenges which are absent in metagenome-based studies. In contrast to DNA, the RNA is highly unstable with a short half-life and a rapid turnover rate [238, 239] which makes recovery of high-quality RNA from environmental samples challenging. Only a small fraction of the transcripts represent mRNA derived sequences. The majority of the extracted RNA comprises ribosomal RNA (rRNA) molecules [240]. However, only protein coding sequences matter to reveal which genes and pathways are expressed under a given condition. Therefore, an enrichment or separation of mRNA from the total RNA pool is necessary [241]. Further, metatranscriptomics as much is affected as metagenomics by the large fraction of reads without significant hits to any known gene sequence in the databases [149].

Given that proteins are much more stable than mRNAs [242], a proteome-based analysis is expected to provide a more accurate view of the functionality of a given environment [243]. In metaproteomics, the complete proteome of an environmental sample under a given set of conditions at a specific point in time is studied [244]. In metaproteomics, proteins are extracted from a mixed microbial community sample and quantified using two-dimensional polyacrylamide gel electrophoresis and mass spectrometry [236]. Here, too, the drawback of metaproteomics is the low extraction yield and the lack of reference sequences in databases for functional assignments of protein fragments. Due to the complexity of protein extraction, separation and identification, metaproteomics is still in its infancy and has less widely been used than metagenomics and metatranscriptomics [74, 245, 246].

Further, the quantification of community metabolites by metametabolomics is a promising approach [247, 248]. Metametabolomics could provide a qualitative and quantitative measure of all low molecular-weight molecules involved in metabolic reactions and required for the maintenance, growth, and function of a microbial community. Changes at the metabolome are expected to be amplified relative to changes at the transcriptome or proteome level, and thus the metabolome can provide informative details about key metabolic pathways.

In conclusion, metagenomic studies provide a snapshot of the genetic composition of a microbial community. Combining metagenomic studies with metatranscriptomic, metaproteomic, or metametabolomic studies offers additional insights into the dynamics of microbial communities, e.g., [240]. However, integration of these vast and diverse meta'omic datasets will be challenging.

## 5.3 Benefits of the unsupervised identification of related metagenomes

A wealth of knowledge can be gained from the metagenomic and amplicon-based community characterizations. Beyond the functional capacities and taxonomic composition of a single community, the similarities as well as dissimilarities of microbial communities are subsequently investigated. Replicated experimental design and large-scale projects with multiple samples in different conditions or environments allow for robust statistical analysis. However, not only these specifically designed projects, but also existing data repositories provide a rapidly growing source of information for comparative studies. For example, the popular web-based MG-RAST server [162], which serves as data repository and analysis pipeline, contains 140250 metagenome datasets of which 20289 are publicly available (as of October 16th, 2014). Finding samples similar to a given query sample is becoming a central operation. In particular, the identification of closely related metagenome datasets to a newly obtained dataset is of growing importance for downstream comparative analysis.

A conceivable starting point for selecting metagenome datasets is the associated metadata describing the environmental context and the experimental methods of a sample. However, supplementary high-quality metadata is often not available or not in a standardized fine-grained format. Metagenomic database retrieval and comparisons between datasets would largely benefit from standards for metadata collection and high-quality metadata. In this regard, the Minimum Information about a Metagenome Sequence (MIMS) [249] specification is increasingly accepted. However, a large number of currently available metagenome data lacks this information. Although the existing habitat annotations are used for sample selection, it is unclear which metagenomes are good neighbors for a comparative analysis, in particular, if a habitat label is rather abundant within the database (see Section 1.4.2).

Thus, unsupervised methods are desirable for the identification of related metagenomes. Although, the computation of similar datasets needs to be reasonably fast. The identification of similar metagenomes by means of all pairwise sequence comparisons between a new query dataset and all metagenomes in a repository is prohibitively expensive. In contrast, feature profile-based comparison avoids computationally expensive pairwise sequence comparisons. However, the feasibility of profile-based comparison starts from the premise that the feature profiles itself are efficiently computable and captures biologically meaningful information gathered during the preceding non-comparative analysis. For example, the taxonomic, functional, and pathway profile is normally already calculated to answer the questions “Who is in there?” and “What are they doing?”, thus the further use avoids additional computational costs. However, so far, it was unclear what kind of features and which metrics are most suitable for the identification of closely related metagenomes. In this work, I addressed this topic in the context of a nearest neighbor identification.

In Chapter 3, I performed a leave-one-out cross-validation using a  $k$ -nearest-neighbor search. The accuracy of the different feature profiles was assessed in terms of the fraction of profiles in the neighborhood with the same habitat label as the query profile. Here, I used two metagenome collections in order to evaluate the  $k$ -nearest-neighbor search. First, I made use of the HMP dataset collection which provides high-quality sequencing data and a consistent habitat annotation

of metagenomes in terms of five distinct body sites. Second, I utilized the “metagenome universe” dataset collection of 1745 publicly available metagenomes and their associated assignments in 12 habitats [209]. The neighborhood accuracy varied on average from 90% to 97% for the HMP and from 83% to 87% for the metagenome universe collection. Presumably, both the high-quality sequencing data and the consistent habitat annotation of the HMP collection have a beneficial effect on neighborhood detection. It is easily conceivable that methodological differences and biases along the analysis steps lead to the poorer neighborhood accuracy. In the metagenome universe collection different DNA extraction, sequencing technologies, etc. were applied. Related to the DNA extraction protocol and read length, it has previously been reported that both factors influence the quantitative description of a metagenome [250], thus presumably affecting the subsequent  $k$ -nearest-neighbor search. Nevertheless, as reported in [250] and revealed in Chapter 3, the methodological variations do not appear to limit global comparisons.

Notably, the differences in accuracy cannot be attributed to a particular profiling method or metric, but rather to the quality of the currently used habitat labels. For example, for well-defined labels a neighborhood accuracy of 100% can be achieved. In contrast, a large fraction of samples from the “Skin” and “Host-associated” category were misassigned to the “Airways” and “Feces/GI tract” category, respectively. Presumably, there is a natural overlap of airways- and skin-associated microbial communities due to the proximity of the sampling location of both body regions. Also, the non-mutually exclusive habitat categories “Host-associated” and “Feces/GI tract” lead to a systematical confusion. The results indicate that the categories “Host-associated” and “Feces/GI tract” are not mutually exclusive or at least are not well represented in terms of profile similarity. To exemplify, depending on the study objective, the investigator can assign to an animal fecal samples the label “Feces/GI tract” focusing on the type of sample or “Host-associated” focusing on the sample source.

Moreover, the results reveal another application area of the  $k$ -nearest-neighbor search. Apart from the initial intention to identify similar datasets for comparison and mislabeled datasets, the  $k$ -nearest-neighbor search also appears well-suited for the definition and verification of non-overlapping habitat categories. Identified non-mutually exclusive categories, as in the case of “Host-associated” and “Feces/GI tract”, should be reconsidered. However, as long as this issue is not addressed and categories are overlapping, metadata should not be considered for the identification of similar metagenomes. In contrast, the  $k$ -nearest-neighbor search considers all samples irrespective of the habitat labels. Thus, for the currently existing habitat labels, the  $k$ -nearest-neighbor search is more general and may even provide the potential for the discovery of novel or unexpected relationships.

The accuracy differences are mainly attributed to the quality of the currently used habitat labels, thus a wide range of profiles and distance metrics can be applied for the  $k$ -nearest-neighbor search. Since the functional, taxonomic, and metabolic profiles are already calculated to answer the questions “Who is in there?” and “What are they doing?”, the further utilization of the computed profiles reduces the additionally required computational costs considerably. However, one should select a feature representation and distance measures that allow a convenient interpretation in terms of the underlying features and subsequent in-depth analyses. Given the fact that taxonomic, functional, and metabolic features provide a biologically more meaningful representation than oligonucleotide-based features, those features should be preferred. Albeit, the MoP-Pro approach shows very little variation of the accuracy with respect to the underlying profile distance measure.

Hence, the MoP-Pro approach not only provides a reasonably rapidly calculable pathway profile but also a suitable feature profile for the neighborhood detection. Further, the profiles can be utilized in downstream analysis to identify similarities and differences between samples at the pathway level. In conclusion, the  $k$ -nearest-neighbor search utilizing a wide range of biological feature profiles and metrics enables an efficient, automatic identification of additional metagenomes for comparative analyses. In addition, the neighboring habitat labels can provide an indication of mislabeled or contaminated datasets.

## 5.4 CoMet-Universe - a web-server for comparative analysis of metagenomes based on protein domain signatures

In order to make the MoP-Pro approach and the  $k$ -nearest-neighbor search accessible for the scientific community, I integrated them into the CoMet-Universe web server (<http://comet2.gobics.de>). The CoMet-Universe web server is especially developed for high-performance comparative metagenomics and integrates several tools for the analysis and comparison of user-supplied sequence data. Beyond the analysis of uploaded metagenome data, the user has the possibility to compare a particular metagenome with more than thousand pre-computed profiles from a broad variety of publicly available datasets or with previously uploaded data from the same user.

The ultrafast protein classification (UProC) tool (<http://uproc.gobics.de/>) provides the basis for all analyses implemented in CoMet-Universe. UProC enables a computationally efficient protein domain classification according to the Pfam database. Using UProC, the processing of large amounts of unassembled short read data is orders of magnitude faster than with a conventional BLAST-based approach. Further, the analysis pipeline implements the mixture modeling approaches Taxy-Pro and MoP-Pro for taxonomic and metabolic profiling of metagenomic sequences. Due to the computational efficiency of all prediction methods used, the user-supplied multi-FASTA sequence files can be processed by the CoMet-Universe engine within a few minutes. On that account, CoMet-Universe is well-suitable to cope with large collections. Subsequent to the fast calculation of the functional, taxonomic, and metabolic profile, the user can analyze the annotated metagenomics datasets by the supplied graphical and tabular summaries including interactive Krona charts [251] for profile visualization.

Further, I included the visualization method proposed in [209]. With the help of the visualization of the metagenome universe in a 2D scatter plot the user has the possibility of a final quality control of the dataset by inspecting the neighbors and their associated annotations. As discussed in Chapter 3, the inspection may reveal unexpected flaws of the sampling, sequencing or data processing procedures. For instance, neighbors with unexpected habitat labels may indicate some contamination of the sample.

In addition to the analysis of user-supplied metagenome shotgun sequence data, CoMet-Universe supports the comparative analysis. The incorporation of precomputed profiles from publicly available metagenome datasets allows to compare a selected metagenome directly with other metagenomes. In order to facilitate the selection of interesting related metagenome datasets, the results of the  $k$ -nearest-neighbor search based on Pfam protein domain frequencies can be used as starting point for detailed comparisons. After selection of appropriate and

interesting metagenomes for comparison, the statistical differences between the selected samples are computed. For analysis of statistical differences between the selected metagenome samples, the number of different Pfam domain families between two samples is used to determine highly variable Pfam families across all samples and statistically enriched GO terms associated with these families. Furthermore, the significantly differing domain families are used to perform multi-dimensional scaling (MDS) and a hierarchical clustering analysis (UPGMA).

At the moment, CoMet-Universe is designed to perform the whole annotation workflow after uploading the metagenome sequence data. Especially with the increasing size of data the upload of the data becomes increasingly difficult. The reduction of the upload size of the data is a promising approach. The direct upload of the Pfam protein domain abundances suggests itself since they are the basis for all further analyses and UProC is available as command line tool. This would significantly reduce the data transmission and storage. Therefore, it would be necessary to extend the internal CoMet-Universe engine in order to be able to handle both data types, multi-FASTA sequence files as well as files containing the domain frequencies. The prediction of KEGG Orthologs by CoMet-Universe is another striking point since UProC can also provide KEGG Orthologs classification, in principle. Thus, the integration would enable the user to get another very comprehensive functional overview about the functional properties of a metagenome. Further, the KEGG Ortholog profiles could be used to apply the direct MoP model approach instead of the currently implemented MoP-Pro approach. Following completion of the evaluation of UProC regarding the classification performance for KEGG Orthologs this additional feature would further increase the value of CoMet-Universe.

CoMet-Universe is well-suitable for scientists without extensive programming skills. However, for analyzing a collection of newly sequenced metagenome datasets the one by one data upload and subsequent download of all computed results become a laborious task. For large-scale datasets and scientists with programming skills, it is preferable to offer stand-alone tools which facilitate the integration of the corresponding methods in specialized local workflows. In addition, locally installed pipelines allow a much more flexible exploratory data analysis and comparison. Therefore, the functionality of MoP-Pro and Tax4Fun is additionally available as an open-source R Package. Tax4Fun is applicable to output as obtained through the SILVAngs web server (<https://www.arb-silva.de/ngs/>) or the application of QIIME against the SILVA database [102]. The corresponding package allows the combination with an adapted R-based workflow, featuring more sophisticated tools for graphical representation and statistical testing.

## 6 Summary and conclusion

Metagenomics, as a culture-independent approach, enables the exploration of complex microbial communities by massive sequencing of community-specific DNA. The analysis of metagenomic sequence data involves the functional classification of sequencing reads. The resulting functional profile typically comprises the frequencies of several thousand functional categories which are often summarized in terms of easily understandable biological pathways. For the functional characterization, a homology search is usually utilized which requires a considerable amount of CPU-time even for a moderately sized dataset. This computational drawback is unlikely to be addressed simply by scaling up computational resources and affects both the functional and subsequent metabolic characterization. Furthermore, the ambiguous assignment of functions to pathways impairs the quantification in terms of biological pathways.

Several approaches have been proposed to describe the metabolic potential of microbial communities. However, none of these approaches provides a strict probabilistic description of metagenomic sequence data and thereby may overrate the metabolic capabilities. Furthermore, all existing methods assume as input a functional profile. So far, the HUMAnN tool has been the most sophisticated approach implementing several filtering, normalization, and smoothing steps to adjust pathway abundances and to avoid overestimation.

In this work, I developed an efficient and statistically reasonable method for characterizing the metabolic potential in metagenomic samples. In comparison to other metagenomic pathway profiling methods, the expensive computation of the functional profile is avoidable. The mixture model-based approach provides a shortcut to estimate the pathway profile of the metagenome by a linear combination of reference profiles. The nested MoP model approach provides a concise summary of the functional variation across many samples in reasonable time. This enables the fast identification of relevant metabolic differences.

Further, I investigated the computation of the functional profile from 16S rRNA sequence data. By applying Tax4Fun, the previously restricted analysis of the taxonomic composition is extended by the functional capabilities. In contrast to the PICRUSt approach, Tax4Fun does not depend on the topology of a phylogenetic tree but close homologies between 16S rRNA gene sequences. Further, PICURST applies the greengenes database, whereas Tax4Fun utilizes the high quality SILVA database. Further, Tax4Fun narrows down the inference to reliable linkages and does no attempt to links all OTUs to the next sequenced organism. This moderate inference is especially beneficial for analyzing microbial communities with a high proportion of so far not well-characterized phyla where phylogenetic distances to the next sequenced organism can be large. Altogether, the Tax4Fun evaluation shows a good approximation of the functional profiles and outperforms PICRUSt on all test datasets.

Both the Tax4Fun and nested MoP model approach are provided as R Package in order to facilitate the integration into existing analysis pipelines and further statistical analyses. Further, CoMet-Universe provides an interactive user interface for exploratory analysis and comparison of metagenome data. Here, the comparison with more than thousand precomputed profiles in the CoMet database is facilitated by the implemented  $k$ -nearest-neighbor search approach. The integration of the  $k$ -nearest-neighbor search into metagenome annotation and comparison systems is beneficial to automatically identify additional metagenomes for comparative analyses as well as to detect mislabeled or contaminated datasets by unexpected neighboring habitat labels. Overall, CoMet-Universe and the Tax4Fun R Package are valuable tools for the analysis of microbial communities. Both tools can process large-scale datasets in reasonable time and are well-suitable for answering the fundamental questions “What are they doing?” and “What are the differences that make a difference?” for the growing amount of sequence data from microbial communities.

# Bibliography

- [1] Altermann W, Kazmierczak J: **Archean microfossils: a reappraisal of early life on Earth.** *Research in Microbiology* 2003, **154**(9):611–617, [<http://dx.doi.org/10.1016/j.resmic.2003.08.006>].
- [2] Arrigo KR: **Marine microorganisms and global nutrient cycles.** *Nature* 2005, **437**(7057):349–355, [<http://dx.doi.org/10.1038/nature04159>].
- [3] van der Heijden MGA, Bardgett RD, van Straalen NM: **The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems.** *Ecology Letters* 2008, **11**(3):296–310, [<http://dx.doi.org/10.1111/j.1461-0248.2007.01139.x>].
- [4] Whitman WB, Coleman DC, Wiebe WJ: **Prokaryotes: The unseen majority.** *Proceedings of the National Academy of Sciences* 1998, **95**(12):6578–6583, [<http://www.pnas.org/content/95/12/6578.abstract>].
- [5] Gest H: **The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society.** *Notes and Records of the Royal Society of London* 2004, **58**(2):187–201, [<http://rsnr.royalsocietypublishing.org/content/58/2/187.abstract>].
- [6] Koch R, Gaffky G, Pfuhl E: *Gesammelte Werke von ROBERT KOCH, Volume 1.* Leipzig: Thieme 1912.
- [7] Palleroni NJ: **Prokaryotic diversity and the importance of culturing.** *Antonie van Leeuwenhoek* 1997, **72**:3–19, [<http://dx.doi.org/10.1023/A%3A1000394109961>].
- [8] Janssen PH: **New cultivation strategies for terrestrial microorganisms.** In *Accessing uncultivated microorganisms: from the environment to organisms and genomes and back.* Edited by Zengler K, Washington: ASM Press 2008:173–192.
- [9] Joyce EA, Chan K, Salama NR, Falkow S: **Redefining bacterial populations: a post-genomic reformation.** *Nature Reviews Genetics* 2002, **3**(6):462–473, [<http://dx.doi.org/10.1038/nrg820>].

- [10] Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R: **Microbiology in the post-genomic era.** *Nature Reviews Microbiology* 2008, **6**(6):419–430, [<http://dx.doi.org/10.1038/nrmicro1901>].
- [11] Staley JT, Konopka A: **Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats.** *Annual Review of Microbiology* 1985, **39**:321–346, [<http://dx.doi.org/10.1146/annurev.mi.39.100185.001541>].
- [12] Zengler K: **Central Role of the Cell in Microbial Ecology.** *Microbiology and Molecular Biology Reviews* 2009, **73**(4):712–729, [<http://dx.doi.org/10.1128/MMBR.00027-09>].
- [13] Amann RI, Ludwig W, Schleifer KH: **Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation.** *Microbiological Reviews* 1995, **59**:143–169, [<http://mmbbr.asm.org/content/59/1/143.abstract>].
- [14] Kellenberger E: **Exploring the unknown. The silent revolution of microbiology.** *EMBO reports* 2001, **2**:5–7, [<http://dx.doi.org/10.1093/embo-reports/kve014>].
- [15] Curtis TP, Sloan WT, Scannell JW: **Estimating prokaryotic diversity and its limits.** *Proceedings of the National Academy of Sciences* 2002, **99**(16):10494–10499, [<http://dx.doi.org/10.1073/pnas.142680199>].
- [16] Rappé MS, Giovannoni SJ: **THE UNCULTURED MICROBIAL MAJORITY.** *Annual Review of Microbiology* 2003, **57**:369–394, [<http://dx.doi.org/10.1146/annurev.micro.57.030502.090759>].
- [17] Giovannoni SJ, Britschgi TB, Moyer CL, Field KG: **Genetic diversity in Sargasso Sea bacterioplankton.** *Nature* 1990, **345**(6270):60–63, [<http://dx.doi.org/10.1038/345060a0>].
- [18] Giovannoni SJ, DeLong EF, Schmidt TM, Pace NR: **Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton.** *Applied and Environmental Microbiology* 1990, **56**(8):2572–2575, [<http://aem.asm.org/content/56/8/2572.abstract>].
- [19] Human Microbiome Project Consortium: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486**(7402):207–214, [<http://dx.doi.org/10.1038/nature11234>].
- [20] Pace NR, Stahl DA, Lane DJ, Olsen GJ: **The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences.** In *Advances in Microbial Ecology, Volume 9 of Advances in Microbial Ecology*. Edited by Marshall K, Springer 1986:1–55, [[http://dx.doi.org/10.1007/978-1-4757-0611-6\\_1](http://dx.doi.org/10.1007/978-1-4757-0611-6_1)].

- [21] Bruns A, Nübel U, Cypionka H, Overmann J: **Effect of Signal Compounds and Incubation Conditions on the Culturability of Freshwater Bacterioplankton.** *Applied and Environmental Microbiology* 2003, **69**(4):1980–1989, [<http://dx.doi.org/10.1128/AEM.69.4.1980-1989.2003>].
- [22] Köpke B, Wilms R, Engelen B, Cypionka H, Sass H: **Microbial Diversity in Coastal Subsurface Sediments: a Cultivation Approach Using Various Electron Acceptors and Substrate Gradients.** *Applied and Environmental Microbiology* 2005, **71**(12):7819–7830, [<http://dx.doi.org/10.1128/AEM.71.12.7819-7830.2005>].
- [23] Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ: **Microbial diversity in the deep sea and the underexplored “rare biosphere”.** *Proceedings of the National Academy of Sciences* 2006, **103**(32):12115–12120, [<http://dx.doi.org/10.1073/pnas.0605127103>].
- [24] Kemp PF, Aller JY: **Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us.** *FEMS Microbiology Ecology* 2004, **47**(2):161–177, [[http://dx.doi.org/10.1016/S0168-6496\(03\)00257-5](http://dx.doi.org/10.1016/S0168-6496(03)00257-5)].
- [25] Nalin R, Ranjard L, Nazaret S, Simonet P: **La biologie moléculaire en écologie microbienne du sol: application à l’analyse de la structure des communautés bactériennes.** *Bulletin de la Société Française de Microbiologie* 1998, **13**:21–26.
- [26] Sait M, Hugenholtz P, Janssen PH: **Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys.** *Environmental Microbiology* 2002, **4**(11):654–666, [<http://dx.doi.org/10.1046/j.1462-2920.2002.00352.x>].
- [27] Torsvik V, Øvreås L: **Microbial diversity and function in soil: from genes to ecosystems.** *Current Opinion in Microbiology* 2002, **5**(3):240–245, [[http://dx.doi.org/10.1016/S1369-5274\(02\)00324-7](http://dx.doi.org/10.1016/S1369-5274(02)00324-7)].
- [28] Torsvik V, Goksøyr J, Daae FL: **High diversity in DNA of soil bacteria.** *Applied and Environmental Microbiology* 1990, **56**(3):782–787, [<http://aem.asm.org/content/56/3/782.abstract>].
- [29] Savage DC: **Microbial ecology of the gastrointestinal tract.** *Annual Review of Microbiology* 1977, **31**:107–133, [<http://dx.doi.org/10.1146/annurev.mi.31.100177.000543>].
- [30] Embley TM, Stackebrandt E: **Species in practice: exploring uncultured prokaryote diversity in natural samples.** In *Species: the units of biodiversity, Volume 54*. Edited by Claridge MF, Dawah HA, Wilson MR, London: Chapman and Hall 1997:61–81.
- [31] Torsvik V, Øvreås L, Thingstad TF: **Prokaryotic Diversity–Magnitude, Dynamics, and Controlling Factors.** *Science* 2002, **296**(5570):1064–1066, [<http://dx.doi.org/10.1126/science.1071698>].

- [32] Woese CR: **There must be a prokaryote somewhere: microbiology's search for itself.** *Microbiological Reviews* 1994, **58**:1–9, [<http://mbr.asm.org/content/58/1/1.abstract>].
- [33] Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F: **From genomics to metagenomics.** *Current Opinion in Biotechnology* 2012, **23**:72–76, [<http://dx.doi.org/10.1016/j.copbio.2011.12.017>].
- [34] Scholz MB, Lo CC, Chain PSG: **Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis.** *Current Opinion in Biotechnology* 2012, **23**:9–15, [<http://dx.doi.org/10.1016/j.copbio.2011.11.013>].
- [35] Pace NR, Stahl DA, Lane DJ, Olsen GJ: **Analyzing natural microbial populations by rRNA sequences.** *American Society for Microbiology News* 1985, **51**:4–12.
- [36] Schmidt TM, DeLong EF, Pace NR: **Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing.** *Journal of Bacteriology* 1991, **173**(14):4371–4378, [<http://jb.asm.org/content/173/14/4371.abstract>].
- [37] Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR: **Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses.** *Proceedings of the National Academy of Sciences* 1985, **82**(20):6955–6959, [<http://www.pnas.org/content/82/20/6955.abstract>].
- [38] Pace NR: **A Molecular View of Microbial Diversity and the Biosphere.** *Science* 1997, **276**(5313):734–740, [<http://dx.doi.org/10.1126/science.276.5313.734>].
- [39] Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA: **Microbial ecology and evolution: a ribosomal RNA approach.** *Annual Review of Microbiology* 1986, **40**:337–365, [<http://dx.doi.org/10.1146/annurev.mi.40.100186.002005>].
- [40] Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proceedings of the National Academy of Sciences* 1977, **74**(11):5088–5090, [<http://dx.doi.org/10.1073/pnas.74.11.5088>].
- [41] Hugenholtz P, Pace NR: **Identifying microbial diversity in the natural environment: a molecular phylogenetic approach.** *Trends in Biotechnology* 1996, **14**(6):190–197, [[http://dx.doi.org/10.1016/0167-7799\(96\)10025-1](http://dx.doi.org/10.1016/0167-7799(96)10025-1)].
- [42] Hugenholtz P, Goebel BM, Pace NR: **Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity.** *Journal of Bacteriology* 1998, **180**(18):4765–4774, [<http://jb.asm.org/content/180/18/4765.short>].
- [43] Lozupone CA, Knight R: **Global patterns in bacterial diversity.** *Proceedings of the National Academy of Sciences* 2007, **104**(27):11436–11440, [<http://dx.doi.org/10.1073/pnas.0611525104>].

- [44] Ward DM, Weller R, Bateson MM: **16S rRNA sequences reveal numerous uncultured microorganisms in a natural community.** *Nature* 1990, **345**(6270):63–65, [<http://dx.doi.org/10.1038/345063a0>].
- [45] Tringe SG, Hugenholtz P: **A renaissance for the pioneering 16S rRNA gene.** *Current Opinion in Microbiology* 2008, **11**(5):442–446, [<http://dx.doi.org/10.1016/j.mib.2008.09.011>].
- [46] Bocchetta M, Ceccarelli E, Creti R, Sanangelantoni AM, Tiboni O, Cammarano P: **Arrangement and nucleotide sequence of the gene (*fus*) encoding elongation factor G (EF-G) from the hyperthermophilic bacterium *Aquifex pyrophilus*: phylogenetic depth of hyperthermophilic bacteria inferred from analysis of the EF-G/*fus* sequences.** *Journal of Molecular Evolution* 1995, **41**(6):803–812, [<http://dx.doi.org/10.1007/BF00173160>].
- [47] Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S: **Use of 16S rRNA and *rpoB* Genes as Molecular Markers for Microbial Ecology Studies.** *Applied and Environmental Microbiology* 2007, **73**:278–288, [<http://dx.doi.org/10.1128/AEM.01177-06>].
- [48] Neefs JM, Van de Peer Y, Hendriks L, De Wachter R: **Compilation of small ribosomal subunit RNA sequences.** *Nucleic Acids Research* 1990, **18**(Suppl):2237–2317, [<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC331875/>].
- [49] Woese CR: **Bacterial evolution.** *Microbiological Reviews* 1987, **51**(2):221–271.
- [50] Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proceedings of the National Academy of Sciences* 1990, **87**(12):4576–4579, [<http://dx.doi.org/10.1073/pnas.87.12.4576>].
- [51] McInerney JO, Cotton JA, Pisani D: **The prokaryotic tree of life: past, present... and future?** *Trends in Ecology & Evolution* 2008, **23**(5):276–281, [<http://dx.doi.org/10.1016/j.tree.2008.01.008>].
- [52] Lozupone CA, Knight R: **Species divergence and the measurement of microbial diversity.** *FEMS Microbiology Reviews* 2008, **32**(4):557–578, [<http://dx.doi.org/10.1111/j.1574-6976.2008.00111.x>].
- [53] Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM: **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.** *Chemistry & Biology* 1998, **5**(10):R245–R249, [[http://dx.doi.org/10.1016/S1074-5521\(98\)90108-9](http://dx.doi.org/10.1016/S1074-5521(98)90108-9)].
- [54] Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428**(6978):37–43, [<http://dx.doi.org/10.1038/nature02340>].

- [55] Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental Genome Shotgun Sequencing of the Sargasso Sea.** *Science* 2004, **304**(5667):66–74, [<http://dx.doi.org/10.1126/science.1093857>].
- [56] Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM: **Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms.** *Applied and Environmental Microbiology* 2000, **66**(6):2541–2547, [<http://dx.doi.org/10.1128/AEM.66.6.2541-2547.2000>].
- [57] Lee SW, Won K, Lim HK, Kim JC, Choi GJ, Cho KY: **Screening for novel lipolytic enzymes from uncultured soil microorganisms.** *Applied Microbiology and Biotechnology* 2004, **65**(6):720–726, [<http://dx.doi.org/10.1007/s00253-004-1722-3>].
- [58] Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC: **The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.** *PLoS Biology* 2007, **5**(3):e77, [<http://dx.doi.org/10.1371/journal.pbio.0050077>].
- [59] Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, Somerfield PJ, Huse S, Joint I: **The seasonal structure of microbial communities in the Western English Channel.** *Environmental Microbiology* 2009, **11**(12):3132–3139, [<http://dx.doi.org/10.1111/j.1462-2920.2009.02017.x>].
- [60] Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic Analysis of the Human Distal Gut Microbiome.** *Science* 2006, **312**(5778):1355–1359, [<http://dx.doi.org/10.1126/science.1124234>].
- [61] Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI: **The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice.** *Science Translational Medicine* 2009, **1**(6):6ra14, [<http://dx.doi.org/10.1126/scitranslmed.3000322>].
- [62] Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F,

- Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J, MITC, Antolín M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariáz G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Mérieux A, Melo Minardi R, M'rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P: **Enterotypes of the human gut microbiome.** *Nature* 2011, **473**(7346):174–180, [<http://dx.doi.org/10.1038/nature09944>].
- [63] Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK: **Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw.** *Nature* 2011, **480**(7377):368–371, [<http://dx.doi.org/10.1038/nature10576>].
- [64] Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE: **De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities.** *The ISME Journal* 2012, **6**:81–93, [<http://dx.doi.org/10.1038/ismej.2011.78>].
- [65] Yergeau E, Hogues H, Whyte LG, Greer CW: **The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses.** *The ISME Journal* 2010, **4**(9):1206–1214, [<http://dx.doi.org/10.1038/ismej.2010.41>].
- [66] Snyder LAS, Loman N, Pallen MJ, Penn CW: **Next-Generation Sequencing—the Promise and Perils of Charting the Great Microbial Unknown.** *Microbial Ecology* 2009, **57**:1–3, [<http://dx.doi.org/10.1007/s00248-008-9465-9>].
- [67] Shendure J, Ji H: **Next-generation DNA sequencing.** *Nature Biotechnology* 2008, **26**(10):1135–1145, [<http://dx.doi.org/10.1038/nbt1486>].
- [68] Metzker ML: **Sequencing technologies - the next generation.** *Nature Reviews Genetics* 2010, **11**:31–46, [<http://dx.doi.org/10.1038/nrg2626>].
- [69] Shokralla S, Spall JL, Gibson JF, Hajibabaei M: **Next-generation sequencing technologies for environmental DNA research.** *Molecular Ecology* 2012, **21**(8):1794–1805, [<http://dx.doi.org/10.1111/j.1365-294X.2012.05538.x>].
- [70] Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R: **Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex.** *Nature Methods* 2008, **5**(3):235–237, [<http://dx.doi.org/10.1038/nmeth.1184>].
- [71] Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R: **Moving pictures of the human microbiome.** *Genome Biology* 2011, **12**(5):R50, [<http://dx.doi.org/10.1186/gb-2011-12-5-r50>].

- [72] Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, Somerfield P, Fuhrman JA, Field D: **Defining seasonal marine microbial community dynamics.** *The ISME Journal* 2012, **6**(2):298–308, [<http://dx.doi.org/10.1038/ismej.2011.107>].
- [73] Gilbert JA, Jansson JK, Knight R: **The Earth Microbiome project: successes and aspirations.** *BMC Biology* 2014, **12**:69, [<http://dx.doi.org/10.1186/s12915-014-0069-1>].
- [74] Hamady M, Knight R: **Microbial community profiling for human microbiome projects: Tools, techniques, and challenges.** *Genome Research* 2009, **19**(7):1141–1152, [<http://dx.doi.org/10.1101/gr.085464.108>].
- [75] National Research Council: *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Washington, DC: The National Academies Press 2007, [[http://www.nap.edu/catalog.php?record\\_id=11902](http://www.nap.edu/catalog.php?record_id=11902)].
- [76] Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative Metagenomics of Microbial Communities.** *Science* 2005, **308**(5721):554–557, [<http://dx.doi.org/10.1126/science.1107851>].
- [77] Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F: **Functional metagenomic profiling of nine biomes.** *Nature* 2008, **452**(7187):629–632, [<http://dx.doi.org/10.1038/nature06810>].
- [78] Human Microbiome Project Consortium: **A framework for human microbiome research.** *Nature* 2012, **486**(7402):215–221, [<http://dx.doi.org/10.1038/nature11209>].
- [79] Tickle TL, Segata N, Waldron L, Weingart U, Huttenhower C: **Two-stage microbial community experimental design.** *The ISME Journal* 2013, **7**(12):2330–2339, [<http://dx.doi.org/10.1038/ismej.2013.139>].
- [80] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI: **Human gut microbiome viewed across age and geography.** *Nature* 2012, **486**(7402):222–227, [<http://dx.doi.org/10.1038/nature11053>].
- [81] NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S,

- Read J, Watson B, Guyer M: **The NIH Human Microbiome Project.** *Genome Research* 2009, **19**(12):2317–2323, [<http://dx.doi.org/10.1101/gr.096651.109>].
- [82] Gilbert JA, Dupont CL: **Microbial Metagenomics: Beyond the Genome.** *Annual Review of Marine Science* 2011, **3**:347–371, [<http://dx.doi.org/10.1146/annurev-marine-120709-142811>].
- [83] Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, Desai N, Eisen JA, Evers D, Field D, Feng W, Huson D, Jansson J, Knight R, Knight J, Kolker E, Konstantindis K, Kostka J, Kyrpides N, Mackelprang R, McHardy A, Quince C, Raes J, Sczyrba A, Shade A, Stevens R: **Meeting Report: The Terabase Metagenomics Workshop and the Vision of an Earth Microbiome Project.** *Standards in Genomic Sciences* 2010, **3**(3):243–248, [<http://dx.doi.org/10.4056/sigs.1433550>].
- [84] Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R: **Experimental and analytical tools for studying the human microbiome.** *Nature Reviews Genetics* 2012, **13**:47–58, [<http://dx.doi.org/10.1038/nrg3129>].
- [85] Angiuoli SV, White JR, Matalka M, White O, Fricke WF: **Resources and Costs for Microbial Sequence Analysis Evaluated Using Virtual Machines and Cloud Computing.** *PLoS ONE* 2011, **6**(10):e26624, [<http://dx.doi.org/10.1371/journal.pone.0026624>].
- [86] Chen K, Pachter L: **Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities.** *PLoS Computational Biology* 2005, **1**(2):106–112, [<http://dx.doi.org/10.1371/journal.pcbi.0010024>].
- [87] Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H, Hugenholtz P: **Experimental factors affecting PCR-based estimates of microbial species richness and evenness.** *The ISME Journal* 2010, **4**(5):642–647, [<http://dx.doi.org/10.1038/ismej.2009.153>].
- [88] Liu Z, DeSantis TZ, Andersen GL, Knight R: **Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers.** *Nucleic Acids Research* 2008, **36**(18):e120, [<http://dx.doi.org/10.1093/nar/gkn491>].
- [89] Schloss PD: **The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies.** *PLoS Computational Biology* 2010, **6**(7):e1000844, [<http://dx.doi.org/10.1371/journal.pcbi.1000844>].
- [90] Suzuki MT, Giovannoni SJ: **Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR.** *Applied and Environmental Microbiology* 1996, **62**(2):625–630, [<http://aem.asm.org/content/62/2/625.abstract>].
- [91] von Wintzingerode F, Göbel UB, Stackebrandt E: **Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis.** *FEMS Microbiology*

- Reviews* 1997, **21**(3):213–229, [<http://dx.doi.org/10.1111/j.1574-6976.1997.tb00351.x>].
- [92] Kanagawa T: **Bias and artifacts in multitemplate polymerase chain reactions (PCR).** *Journal of Bioscience and Bioengineering* 2003, **96**(4):317–323, [[http://dx.doi.org/10.1016/S1389-1723\(03\)90130-7](http://dx.doi.org/10.1016/S1389-1723(03)90130-7)].
- [93] Hong S, Bunge J, Leslin C, Jeon S, Epstein SS: **Polymerase chain reaction primers miss half of rRNA microbial diversity.** *The ISME Journal* 2009, **3**(12):1365–1373, [<http://dx.doi.org/10.1038/ismej.2009.89>].
- [94] Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmiento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG: **Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities.** *Environmental Microbiology* 2014, **16**(9):2659–2671, [<http://dx.doi.org/10.1111/1462-2920.12250>].
- [95] Baker GC, Smith JJ, Cowan DA: **Review and re-analysis of domain-specific 16S primers.** *Journal of Microbiological Methods* 2003, **55**(3):541–555, [<http://dx.doi.org/10.1016/j.mimet.2003.08.009>].
- [96] Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M: **Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA genotyping bacterial community analysis.** *FEMS Microbiology Ecology* 2007, **60**(2):341–350, [<http://dx.doi.org/10.1111/j.1574-6941.2007.00283.x>].
- [97] Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R: **Short pyrosequencing reads suffice for accurate microbial community analysis.** *Nucleic Acids Research* 2007, **35**(18):e120, [<http://dx.doi.org/10.1093/nar/gkm541>].
- [98] Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Welch DM, Relman DA, Sogin ML: **Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing.** *PLoS Genetics* 2008, **4**(11):e1000255, [<http://dx.doi.org/10.1371/journal.pgen.1000255>].
- [99] Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L: **Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing.** *PLoS ONE* 2008, **3**(7):e2836, [<http://dx.doi.org/10.1371/journal.pone.0002836>].
- [100] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data.** *Nature Methods* 2010, **7**(5):335–336, [<http://dx.doi.org/10.1038/nmeth.f.303>].

- [101] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF: **Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.** *Applied and Environmental Microbiology* 2009, **75**(23):7537–7541, [<http://dx.doi.org/10.1128/AEM.01541-09>].
- [102] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic Acids Research* 2013, **41**(D1):D590–D596, [<http://dx.doi.org/10.1093/nar/gks1219>].
- [103] Schloss PD, Gevers D, Westcott SL: **Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies.** *PLoS ONE* 2011, **6**(12):e27310, [<http://dx.doi.org/10.1371/journal.pone.0027310>].
- [104] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215**(3):403–410, [[http://dx.doi.org/10.1016/S0022-2836\(05\)80360-2](http://dx.doi.org/10.1016/S0022-2836(05)80360-2)].
- [105] Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.** *Applied and Environmental Microbiology* 2007, **73**(16):5261–5267, [<http://dx.doi.org/10.1128/AEM.00062-07>].
- [106] DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB.** *Applied and Environmental Microbiology* 2006, **72**(7):5069–5072, [<http://dx.doi.org/10.1128/AEM.03006-05>].
- [107] Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM: **Ribosomal Database Project: data and tools for high throughput rRNA analysis.** *Nucleic Acids Research* 2014, **42**(D1):D633–D642, [<http://dx.doi.org/10.1093/nar/gkt1244>].
- [108] Barott KL, Rodriguez-Mueller B, Youle M, Marhaver KL, Vermeij MJA, Smith JE, Rohwer FL: **Microbial to reef scale interactions between the reef-building coral *Montastraea annularis* and benthic algae.** *Proceedings of the Royal Society of London B: Biological Sciences* 2011, **279**(1733):1655–1664, [<http://dx.doi.org/10.1098/rspb.2011.2155>].
- [109] Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C: **Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.** *Genome Biology* 2012, **13**(9):R79, [<http://dx.doi.org/10.1186/gb-2012-13-9-r79>].

- [110] Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrisat B, Knight R, Gordon JI: **Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans.** *Science* 2011, **332**(6032):970–974, [<http://dx.doi.org/10.1126/science.1198719>].
- [111] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepille DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C: **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.** *Nature Biotechnology* 2013, **31**(9):814–821, [<http://dx.doi.org/10.1038/nbt.2676>].
- [112] Pauling L, Zuckerkandl E: **Chemical Paleogenetics. Molecular “Restoration Studies” of Extinct Forms of Life.** *Acta Chemica Scandinavica* 1963, **17**:9–16, [<http://dx.doi.org/10.3891/acta.chem.scand.17s-0009>].
- [113] Omland KE: **The Assumptions and Challenges of Ancestral State Reconstructions.** *Systematic Biology* 1999, **48**(3):604–611.
- [114] Koshi JM, Goldstein RA: **Probabilistic reconstruction of ancestral protein sequences.** *Journal of Molecular Evolution* 1996, **42**(2):313–320.
- [115] Yang Z, Kumar S, Nei M: **A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences.** *Genetics* 1995, **141**(4):1641–50, [<http://www.genetics.org/content/141/4/1641.abstract>].
- [116] Huelsenbeck JP, Bollback JP: **Empirical and Hierarchical Bayesian Estimation of Ancestral States.** *Systematic Biology* 2001, **50**(3):351–366, [<http://dx.doi.org/10.1080/10635150119871>].
- [117] Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**(8):754–755, [<http://dx.doi.org/10.1093/bioinformatics/17.8.754>].
- [118] Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC: **IMG 4 version of the integrated microbial genomes comparative analysis system.** *Nucleic Acids Research* 2014, **42**(D1):D560–D567, [<http://dx.doi.org/10.1093/nar/gkt963>].
- [119] Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28**:27–30, [<http://dx.doi.org/10.1093/nar/28.1.27>].
- [120] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic Acids Research* 2014, **42**(D1):D199–D205, [<http://dx.doi.org/10.1093/nar/gkt1076>].

- [121] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41, [<http://dx.doi.org/10.1186/1471-2105-4-41>].
- [122] Davenport CF, Tümmler B: **Advances in computational analysis of metagenome sequences.** *Environmental Microbiology* 2013, **15**:1–5, [<http://dx.doi.org/10.1111/j.1462-2920.2012.02843.x>].
- [123] Tringe SG, Rubin EM: **Metagenomics: DNA sequencing of environmental samples.** *Nature Reviews Genetics* 2005, **6**(11):805–814, [<http://dx.doi.org/10.1038/nrg1709>].
- [124] Gevers D, Pop M, Schloss PD, Huttenhower C: **Bioinformatics for the Human Microbiome Project.** *PLoS Computational Biology* 2012, **8**(11):e1002779, [<http://dx.doi.org/10.1371/journal.pcbi.1002779>].
- [125] Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P: **A Bioinformatician's Guide to Metagenomics.** *Microbiology and Molecular Biology Reviews* 2008, **72**(4):557–78, [<http://dx.doi.org/10.1128/MMBR.00009-08>].
- [126] Schmieder R, Edwards R: **Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets.** *PLoS ONE* 2011, **6**(3):e17288, [<http://dx.doi.org/10.1371/journal.pone.0017288>].
- [127] Thomas T, Gilbert J, Meyer F: **Metagenomics - a guide from sampling to data analysis.** *Microbial Informatics and Experimentation* 2012, **2**:3, [<http://dx.doi.org/10.1186/2042-5783-2-3>].
- [128] Temperton B, Giovannoni SJ: **Metagenomics: microbial diversity through a scratched lens.** *Current Opinion in Microbiology* 2012, **15**(5):605–612, [<http://dx.doi.org/10.1016/j.mib.2012.07.001>].
- [129] Wooley JC, Godzik A, Friedberg I: **A primer on Metagenomics.** *PLoS Computational Biology* 2010, **6**(2):e1000667, [<http://dx.doi.org/10.1371/journal.pcbi.1000667>].
- [130] Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**(13):1658–1659, [<http://dx.doi.org/10.1093/bioinformatics/btl1158>].
- [131] Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-generation sequencing data.** *Bioinformatics* 2012, **28**(23):3150–3152, [<http://dx.doi.org/10.1093/bioinformatics/bts565>].
- [132] Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010, **26**(19):2460–2461, [<http://dx.doi.org/10.1093/bioinformatics/btq461>].

- [133] Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F: **Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes.** *Cold Spring Harbor Protocols* 2010, **2010**:pdb.prot5368, [<http://dx.doi.org/10.1101/pdb.prot5368>].
- [134] Hoff KJ: **The effect of sequencing errors on metagenomic gene prediction.** *BMC Genomics* 2009, **10**:520, [<http://dx.doi.org/10.1186/1471-2164-10-520>].
- [135] Teeling H, Glöckner FO: **Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective.** *Briefings in Bioinformatics* 2012, **13**(6):728–742, [<http://dx.doi.org/10.1093/bib/bbs039>].
- [136] Rho M, Tang H, Ye Y: **FragGeneScan: predicting genes in short and error-prone reads.** *Nucleic Acids Research* 2010, **38**(20):e191, [<http://dx.doi.org/10.1093/nar/gkq747>].
- [137] Zhu W, Lomsadze A, Borodovsky M: **Ab initio gene identification in metagenomic sequences.** *Nucleic Acids Research* 2010, **38**(12):e132, [<http://dx.doi.org/10.1093/nar/gkq275>].
- [138] Noguchi H, Park J, Takagi T: **MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.** *Nucleic Acids Research* 2006, **34**(19):5623–5630, [<http://dx.doi.org/10.1093/nar/gkl723>].
- [139] Noguchi H, Taniguchi T, Itoh T: **MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes.** *DNA Research* 2008, **15**(6):387–396, [<http://dx.doi.org/10.1093/dnares/dsn027>].
- [140] Hoff KJ, Lingner T, Meinicke P, Tech M: **Orphelia: predicting genes in metagenomic sequencing reads.** *Nucleic Acids Research* 2009, **37**(suppl 2):W101–W105, [<http://dx.doi.org/10.1093/nar/gkp327>].
- [141] NCBI Resource Coordinators: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Research* 2014, **42**(D1):D7–D17, [<http://dx.doi.org/10.1093/nar/gkt1146>].
- [142] Meyer F, Overbeek R, Rodriguez A: **FIGfams: yet another set of protein families.** *Nucleic Acids Research* 2009, **37**(20):6643–6654, [<http://dx.doi.org/10.1093/nar/gkp698>].
- [143] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M: **Pfam: the protein families database.** *Nucleic Acids Research* 2014, **42**(D1):D222–D230, [<http://dx.doi.org/10.1093/nar/gkt1223>].
- [144] Weinstock GM: **Genomic approaches to studying the human microbiota.** *Nature* 2012, **489**(7415):250–256, [<http://dx.doi.org/10.1038/nature11553>].

- [145] Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C: **Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome.** *PLoS Computational Biology* 2012, **8**(6):e1002358, [<http://dx.doi.org/10.1371/journal.pcbi.1002358>].
- [146] Zhao Y, Tang H, Ye Y: **RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data.** *Bioinformatics* 2012, **28**:125–126, [<http://dx.doi.org/10.1093/bioinformatics/btr595>].
- [147] Huson DH, Xie C: **A poor man's BLASTX–high-throughput metagenomic protein database search using PAUDA.** *Bioinformatics* 2014, **30**:38–39, [<http://dx.doi.org/10.1093/bioinformatics/btt254>].
- [148] Godzik A: **Metagenomics and the protein universe.** *Current Opinion in Structural Biology* 2011, **21**(3):398–403, [<http://dx.doi.org/10.1016/j.sbi.2011.03.010>].
- [149] Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I, Somerfield PJ, Mühling M: **The Taxonomic and Functional Diversity of Microbes at a Temperate Coastal Site: A 'Multi-Omic' Study of Seasonal and Diel Temporal Variation.** *PLoS ONE* 2010, **5**(11):e15545, [<http://dx.doi.org/10.1371/journal.pone.0015545>].
- [150] Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V: **The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes.** *Nucleic Acids Research* 2005, **33**(17):5691–5702, [<http://dx.doi.org/10.1093/nar/gki866>].
- [151] Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P: **eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.** *Nucleic Acids Research* 2010, **38**(suppl 1):D190–D195, [<http://dx.doi.org/10.1093/nar/gkp951>].
- [152] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**:25–29, [<http://dx.doi.org/10.1038/75556>].
- [153] Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA,

- Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.** *Nucleic Acids Research* 2014, **42**(D1):D459–D471, [<http://dx.doi.org/10.1093/nar/gkt1103>].
- [154] Ye Y, Doak TG: **A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes.** *PLoS Computational Biology* 2009, **5**(8):e1000465, [<http://dx.doi.org/10.1371/journal.pcbi.1000465>].
- [155] Osterman A, Overbeek R: **Missing genes in metabolic pathways: a comparative genomics approach.** *Current Opinion in Chemical Biology* 2003, **7**(2):238–251, [[http://dx.doi.org/10.1016/S1367-5931\(03\)00027-9](http://dx.doi.org/10.1016/S1367-5931(03)00027-9)].
- [156] Izui K, Matsumura H, Furumoto T, Kai Y: **PHOSPHOENOLPYRUVATE CARBOXYLASE: A New Era of Structural Biology.** *Annual Review of Plant Biology* 2004, **55**:69–84, [<http://dx.doi.org/10.1146/annurev.arplant.55.031903.141619>].
- [157] Karp PD, Latendresse M, Caspi R: **The Pathway Tools Pathway Prediction Algorithm.** *Standards in Genomic Sciences* 2011, **5**(3):424–429, [<http://dx.doi.org/10.4056/sigs.1794338>].
- [158] Karp PD, Paley S, Romero P: **The Pathway Tools software.** *Bioinformatics* 2002, **18**(suppl 1):S225–S232, [[http://dx.doi.org/10.1093/bioinformatics/18.suppl\\_1.S225](http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S225)].
- [159] Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R: **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.** *Briefings in Bioinformatics* 2010, **11**:40–79, [<http://dx.doi.org/10.1093/bib/bbp043>].
- [160] Borenstein E: **Computational systems biology and *in silico* modeling of the human microbiome.** *Briefings in Bioinformatics* 2012, **13**(6):769–780, [<http://dx.doi.org/10.1093/bib/bbs022>].
- [161] Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4.** *Genome Research* 2011, **21**(9):1552–1560, [<http://dx.doi.org/10.1101/gr.120618.111>].
- [162] Meyer F, Paarmann D, D’Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA: **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.** *BMC Bioinformatics* 2008, **9**:386, [<http://dx.doi.org/10.1186/1471-2105-9-386>].

- [163] Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, Liolios K, Pagani I, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC: **IMG/M: the integrated metagenome data management and comparative analysis system.** *Nucleic Acids Research* 2012, **40**(D1):D123–D129, [<http://dx.doi.org/10.1093/nar/gkr975>].
- [164] Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC: **IMG/M 4 version of the integrated metagenome comparative analysis system.** *Nucleic Acids Research* 2014, **42**(D1):D568–D573, [<http://dx.doi.org/10.1093/nar/gkt919>].
- [165] Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Jacob B, Ratner A, Liolios K, Pagani I, Huntemann M, Mavromatis K, Ivanova NN, Kyrpides NC: **IMG/M-HMP: A Metagenome Comparative Analysis System for the Human Microbiome Project.** *PLoS ONE* 2012, **7**(7):e40151, [<http://dx.doi.org/10.1371/journal.pone.0040151>].
- [166] Konwar KM, Hanson NW, Pagé AP, Hallam SJ: **MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information.** *BMC Bioinformatics* 2013, **14**:202, [<http://dx.doi.org/10.1186/1471-2105-14-202>].
- [167] Hanson NW, Konwar KM, Hawley AK, Altman T, Karp PD, Hallam SJ: **Metabolic pathways for the whole community.** *BMC Genomics* 2014, **15**:619, [<http://dx.doi.org/10.1186/1471-2164-15-619>].
- [168] Jolliffe I: *Principal Component Analysis*, John Wiley & Sons, Ltd 2005 [<http://dx.doi.org/10.1002/0470013192.bsa501>].
- [169] Kruskal JB: **Nonmetric multidimensional scaling: A numerical method.** *Psychometrika* 1964, **29**:115–129.
- [170] Everitt BS, Landau S, Leese M, Stahl D: *Cluster Analysis*. John Wiley & Sons, Ltd 2011, [<http://dx.doi.org/10.1002/9780470977811.index>].
- [171] Wu S, Zhu Z, Fu L, Niu B, Li W: **WebMGA: a customizable web server for fast metagenomic sequence analysis.** *BMC Genomics* 2011, **12**:444, [<http://dx.doi.org/10.1186/1471-2164-12-444>].
- [172] Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methé BA, Yooseph S: **METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics.** *Bioinformatics* 2010, **26**(20):2631–2632, [<http://dx.doi.org/10.1093/bioinformatics/btq455>].
- [173] Parks DH, Beiko RG: **Identifying biologically relevant differences between metagenomic communities.** *Bioinformatics* 2010, **26**(6):715–721, [<http://dx.doi.org/10.1093/bioinformatics/btq041>].

- [174] Parks DH, Tyson GW, Hugenholtz P, Beiko RG: **STAMP: statistical analysis of taxonomic and functional profiles.** *Bioinformatics* 2014, [<http://dx.doi.org/10.1093/bioinformatics/btu494>].
- [175] Li W: **Analysis and comparison of very large metagenomes with fast clustering and functional annotation.** *BMC Bioinformatics* 2009, **10**:359, [<http://dx.doi.org/10.1186/1471-2105-10-359>].
- [176] Xie W, Wang F, Guo L, Chen Z, Sievert SM, Meng J, Huang G, Li Y, Yan Q, Wu S, Wang X, Chen S, He G, Xiao X, Xu A: **Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries.** *The ISME Journal* 2011, **5**(3):414–426, [<http://dx.doi.org/10.1038/ismej.2010.144>].
- [177] Lamendella R, Domingo JWS, Ghosh S, Martinson J, Oerther DB: **Comparative fecal metagenomics unveils unique functional capacity of the swine gut.** *BMC Microbiology* 2011, **11**:103, [<http://dx.doi.org/10.1186/1471-2180-11-103>].
- [178] Quaiser A, Zivanovic Y, Moreira D, López-García P: **Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara.** *The ISME Journal* 2011, **5**(2):285–304, [<http://dx.doi.org/10.1038/ismej.2010.113>].
- [179] Steffen MM, Li Z, Effler TC, Hauser LJ, Boyer GL, Wilhelm SW: **Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents.** *PLoS ONE* 2012, **7**(8):e44002, [<http://dx.doi.org/10.1371/journal.pone.0044002>].
- [180] Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: **Obesity alters gut microbial ecology.** *Proceedings of the National Academy of Sciences* 2005, **102**(31):11070–11075, [<http://dx.doi.org/10.1073/pnas.0504978102>].
- [181] Morgan XC, Segata N, Huttenhower C: **Biodiversity and functional genomics in the human microbiome.** *Trends in Genetics* 2013, **29**:51–58, [<http://dx.doi.org/10.1016/j.tig.2012.09.005>].
- [182] Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, Kau AL, Rich SS, Concannon P, Mychaleckyj JC, Liu J, Houghton E, Li JV, Holmes E, Nicholson J, Knights D, Ursell LK, Knight R, Gordon JI: **Gut Microbiomes of Malawian Twin Pairs Discordant for Kwashiorkor.** *Science* 2013, **339**(6119):548–554, [<http://dx.doi.org/10.1126/science.1229000>].
- [183] Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI: **A core gut microbiome in obese and lean twins.** *Nature* 2009, **457**(7228):480–484, [<http://dx.doi.org/10.1038/nature07540>].

- [184] David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ: **Diet rapidly and reproducibly alters the human gut microbiome.** *Nature* 2014, **505**(7484):559–563, [<http://dx.doi.org/10.1038/nature12820>].
- [185] Bäckhed F, Ding H, Wang T, Hooper LV, Koh GY, Nagy A, Semenkovich CF, Gordon JI: **The gut microbiota as an environmental factor that regulates fat storage.** *Proceedings of the National Academy of Sciences* 2004, **101**(44):15718–15723, [<http://dx.doi.org/10.1073/pnas.0407076101>].
- [186] Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI: **Host-Bacterial Mutualism in the Human Intestine.** *Science* 2005, **307**(5717):1915–1920, [<http://dx.doi.org/10.1126/science.1104816>].
- [187] Hooper LV, Littman DR, Macpherson AJ: **Interactions Between the Microbiota and the Immune System.** *Science* 2012, **336**(6086):1268–1273, [<http://dx.doi.org/10.1126/science.1223490>].
- [188] Kelly D, King T, Aminov R: **Importance of microbial colonization of the gut in early life to the development of immunity.** *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 2007, **622**(1-2):58–69, [<http://dx.doi.org/10.1016/j.mrfmmm.2007.03.011>].
- [189] Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL: **An Immunomodulatory Molecule of Symbiotic Bacteria Directs Maturation of the Host Immune System.** *Cell* 2005, **122**:107–118, [<http://dx.doi.org/10.1016/j.cell.2005.05.007>].
- [190] Nicholson JK, Holmes E, Wilson ID: **Gut microorganisms, mammalian metabolism and personalized health care.** *Nature Reviews Microbiology* 2005, **3**(5):431–438, [<http://dx.doi.org/10.1038/nrmicro1152>].
- [191] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F: **Gut metagenome in European women with normal, impaired and diabetic glucose control.** *Nature* 2013, **498**(7452):99–103, [<http://dx.doi.org/10.1038/nature12198>].
- [192] Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, Roca J, Dore J: **Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.** *Gut* 2006, **55**(2):205–211, [<http://dx.doi.org/10.1136/gut.2005.073817>].
- [193] Sekirov I, Russell SL, Antunes LCM, Finlay BB: **Gut Microbiota in Health and Disease.** *Physiological Reviews* 2010, **90**(3):859–904, [<http://dx.doi.org/10.1152/physrev.00045.2009>].
- [194] Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest.** *Nature* 2006, **444**(7122):1027–1031, [<http://dx.doi.org/10.1038/nature05414>].

- [195] Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z, Järnerot G, Tysk C, Jansson JK, Engstrand L: **A Pyrosequencing Study in Twins Shows That Gastrointestinal Microbial Profiles Vary With Inflammatory Bowel Disease Phenotypes.** *Gastroenterology* 2010, **139**(6):1844–1854.e1, [<http://dx.doi.org/10.1053/j.gastro.2010.08.049>].
- [196] Claesson MJ, Cusack S, O’Sullivan O, Greene-Diniz R, de Weerd H, Flannery E, Marchesi JR, Falush D, Dinan T, Fitzgerald G, Stanton C, van Sinderen D, O’Connor M, Harnedy N, O’Connor K, Henry C, O’Mahony D, Fitzgerald AP, Shanahan F, Twomey C, Hill C, Ross RP, O’Toole PW: **Composition, variability, and temporal stability of the intestinal microbiota of the elderly.** *Proceedings of the National Academy of Sciences* 2011, **108 Suppl 1**:4586–4591, [<http://dx.doi.org/10.1073/pnas.1000097107>].
- [197] Claesson MJ, Jeffery IB, Conde S, Power SE, O’Connor EM, Cusack S, Harris HMB, Coakley M, Lakshminarayanan B, O’Sullivan O, Fitzgerald GF, Deane J, O’Connor M, Harnedy N, O’Connor K, O’Mahony D, van Sinderen D, Wallace M, Brennan L, Stanton C, Marchesi JR, Fitzgerald AP, Shanahan F, Hill C, Ross RP, O’Toole PW: **Gut microbiota composition correlates with diet and health in the elderly.** *Nature* 2012, **488**(7410):178–184, [<http://dx.doi.org/10.1038/nature11319>].
- [198] De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P: **Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa.** *Proceedings of the National Academy of Sciences* 2010, **107**(33):14691–14696, [<http://dx.doi.org/10.1073/pnas.1005963107>].
- [199] Kinross J, Nicholson JK: **Gut microbiota: Dietary and social modulation of gut microbiota in the elderly.** *Nature Reviews Gastroenterology and Hepatology* 2012, **9**(10):563–564, [<http://dx.doi.org/10.1038/nrgastro.2012.169>].
- [200] Mai V, McCrary QM, Sinha R, Glei M: **Associations between dietary habits and body mass index with gut microbiota composition and fecal water genotoxicity: an observational study in African American and Caucasian American volunteers.** *Nutrition Journal* 2009, **8**:49, [<http://dx.doi.org/10.1186/1475-2891-8-49>].
- [201] Schokker D, Zhang J, Zhang LL, Vastenhouw SA, Heilig HGJ, Smidt H, Rebel MJM, Smits MA: **Early-Life Environmental Variation Affects Intestinal Microbiota and Immune Development in New-Born Piglets.** *PLoS ONE* 2014, **9**(6):e100040, [<http://dx.doi.org/10.1371/journal.pone.0100040>].
- [202] Voreades N, Kozil A, Weir TL: **Diet and the development of the human intestinal microbiome.** *Frontiers in Microbiology* 2014, **5**:494, [<http://dx.doi.org/10.3389/fmicb.2014.00494>].
- [203] Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD: **Linking Long-Term Dietary Patterns with Gut Microbial**

**Enterotypes.** *Science* 2011, **334**(6052):105–108, [<http://dx.doi.org/10.1126/science.1208344>].

- [204] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MITC, Bork P, Ehrlich SD, Wang J: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464**(7285):59–65, [<http://dx.doi.org/10.1038/nature08821>].
- [205] Gosalbes MJ, Abellan JJ, Durbán A, Pérez-Cobas AE, Latorre A, Moya A: **Metagenomics of human microbiome: beyond 16s rDNA.** *Clinical Microbiology and Infection* 2012, **18**(Suppl. 4):47–49, [<http://dx.doi.org/10.1111/j.1469-0691.2012.03865.x>].
- [206] Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, Hugenholtz P, van der Lelie D, Meyer F, Stevens R, Bailey MJ, Gordon JI, Kowalchuk GA, Gilbert JA: **Unlocking the potential of metagenomics through replicated experimental design.** *Nature Biotechnology* 2012, **30**(6):513–520, [<http://dx.doi.org/10.1038/nbt.2235>].
- [207] Gilbert JA, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J, Ley R, Fierer N, Field D, Kyrpides N, Glöckner FO, Klenk HP, Wommack KE, Glass E, Docherty K, Gallery R, Stevens R, Knight R: **The Earth Microbiome Project: Meeting report of the “1 EMP meeting on sample selection and acquisition” at Argonne National Laboratory October 6 2010.** *Standards in Genomic Sciences* 2010, **3**(3):249–253, [<http://dx.doi.org/10.4056/aigs.1443528>].
- [208] Pagani I, Liolios K, Jansson J, Chen IMA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Research* 2012, **40**(D1):D571–D579, [<http://dx.doi.org/10.1093/nar/gkr1100>].
- [209] Aßhauer KP, Klingenberg H, Lingner T, Meinicke P: **Exploring neighborhoods in the metagenome universe.** *International Journal of Molecular Science* 2014, **15**(7):12364–12378, [<http://dx.doi.org/10.3390/ijms150712364>].
- [210] Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-Beltran A, Hunter C, Jones P, Leinonen R, McAnulla C, Maguire E, Maslen J, Mitchell A, Nuka G, Oisel A, Pesseat S, Radhakrishnan R, Rocca-Serra P, Scheremetjew M, Sterk P, Vaughan D, Cochrane G, Field D, Sansone SA: **EBI metagenomics—a new resource for the analysis and archiving of metagenomic data.** *Nucleic Acids Research* 2014, **42**(D1):D600–D606, [<http://dx.doi.org/10.1093/nar/gkt961>].
- [211] Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE: **Succession of microbial consortia in the developing infant gut microbiome.**

- Proceedings of the National Academy of Sciences* 2011, **108**(Supplement 1):4578–4585, [<http://dx.doi.org/10.1073/pnas.1000081107>].
- [212] Knights D, Kuczynski J, Koren O, Ley RE, Field D, Knight R, DeSantis TZ, Kelley ST: **Supervised classification of microbiota mitigates mislabeling errors.** *The ISME Journal* 2011, **5**(4):570–573, [<http://dx.doi.org/10.1038/ismej.2010.148>].
- [213] Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG: **Cross-biome metagenomic analyses of soil microbial communities and their functional attributes.** *Proceedings of the National Academy of Sciences* 2012, **109**(52):21390–21395, [<http://dx.doi.org/10.1073/pnas.1215210110>].
- [214] Harris JK, Caporaso JG, Walker JJ, Spear JR, Gold NJ, Robertson CE, Hugenholtz P, Goodrich J, McDonald D, Knights D, Marshall P, Tufo H, Knight R, Pace NR: **Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat.** *The ISME Journal* 2013, **7**:50–60, [<http://dx.doi.org/10.1038/ismej.2012.79>].
- [215] Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, von Mering C, Bebout BM, Pace NR, Bork P, Hugenholtz P: **Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat.** *Molecular Systems Biology* 2008, **4**:198, [<http://dx.doi.org/10.1038/msb.2008.35>].
- [216] Meinicke P, Abhauer KP, Lingner T: **Mixture models for analysis of the taxonomic composition of metagenomes.** *Bioinformatics* 2011, **27**(12):1618–1624, [<http://dx.doi.org/10.1093/bioinformatics/btr266>].
- [217] Klingenberg H, Abhauer KP, Lingner T, Meinicke P: **Protein signature-based estimation of metagenomic abundances including all domains of life and viruses.** *Bioinformatics* 2013, **29**(8):973–980, [<http://dx.doi.org/10.1093/bioinformatics/btt077>].
- [218] Garcia-Etxebarria K, Garcia-Garcerà M, Calafell F: **Consistency of metagenomic assignment programs in simulated and real data.** *BMC Bioinformatics* 2014, **15**:90, [<http://dx.doi.org/10.1186/1471-2105-15-90>].
- [219] Prakash T, Taylor TD: **Functional assignment of metagenomic data: challenges and applications.** *Briefings in Bioinformatics* 2012, **13**(6):711–727, [<http://dx.doi.org/10.1093/bib/bbs033>].
- [220] Daniel R: **The metagenomics of soil.** *Nature Reviews Microbiology* 2005, **3**(6):470–478, [<http://dx.doi.org/10.1038/nrmicro1160>].
- [221] Klappenbach JA, Saxman PR, Cole JR, Schmidt TM: **rrndb: the Ribosomal RNA Operon Copy Number Database.** *Nucleic Acids Research* 2001, **29**:181–184, [<http://dx.doi.org/10.1093/nar/29.1.181>].
- [222] Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P, Gerz EA, Jin Z, Lee P, Yang L, Poles M, Brown SM, Sotero S, Desantis T, Brodie E, Nelson K, Pei Z: **Diversity of**

**16S rRNA Genes within Individual Prokaryotic Genomes.** *Applied and Environmental Microbiology* 2010, **76**(12):3886–3897, [<http://dx.doi.org/10.1128/AEM.02953-09>].

- [223] Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF: **Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons.** *Journal of Bacteriology* 2004, **186**(9):2629–2635, [<http://dx.doi.org/10.1128/JB.186.9.2629-2635.2004>].
- [224] Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCarrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, Young S, Warren WC, Badger J, Crabtree J, Markowitz VM, Orvis J, Cree A, Ferreira S, Fulton LL, Fulton RS, Gillis M, Hemphill LD, Joshi V, Kovar C, Torralba M, Wetterstrand KA, Abouelleil A, Wollam AM, Buhay CJ, Ding Y, Dugan S, FitzGerald MG, Holder M, Hostetler J, Clifton SW, Allen-Vercoe E, Earl AM, Farmer CN, Liolios K, Surette MG, Xu Q, Pohl C, Wilczek-Boney K, Zhu D: **A Catalog of Reference Genomes from the Human Microbiome.** *Science* 2010, **328**(5981):994–999, [<http://dx.doi.org/10.1126/science.1183605>].
- [225] Kyrpides NC, Hugenholtz P, Eisen JA, Woyke T, Göker M, Parker CT, Amann R, Beck BJ, Chain PSG, Chun J, Colwell RR, Danchin A, Dawyndt P, Dedeurwaerdere T, DeLong EF, Detter JC, De Vos P, Donohue TJ, Dong XZ, Ehrlich DS, Fraser C, Gibbs R, Gilbert J, Gilna P, Glöckner FO, Jansson JK, Keasling JD, Knight R, Labeda D, Lapidus A, Lee JS, Li WJ, Ma J, Markowitz V, Moore ERB, Morrison M, Meyer F, Nelson KE, Ohkuma M, Ouzounis CA, Pace N, Parkhill J, Qin N, Rossello-Mora R, Sikorski J, Smith D, Sogin M, Stevens R, Stingl U, Suzuki KI, Taylor D, Tiedje JM, Tindall B, Wagner M, Weinstock G, Weissenbach J, White O, Wang J, Zhang L, Zhou YG, Field D, Whitman WB, Garrity GM, Klenk HP: **Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains.** *PLoS Biology* 2014, **12**(8):e1001920, [<http://dx.doi.org/10.1371/journal.pbio.1001920>].
- [226] Fournier PE, Drancourt M, Colson P, Rolain JM, La Scola B, Raoult D: **Modern clinical microbiology: new challenges and solutions.** *Nature Reviews Microbiology* 2013, **11**(8):574–585, [<http://dx.doi.org/10.1038/nrmicro3068>].
- [227] Janda JM, Abbott SL: **16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls.** *Journal of Clinical Microbiology* 2007, **45**(9):2761–2764, [<http://dx.doi.org/10.1128/JCM.01228-07>].
- [228] Klassen JL, Currie CR: **Gene fragmentation in bacterial draft genomes: extent, con-**

- sequences and mitigation.** *BMC Genomics* 2012, **13**:14, [<http://dx.doi.org/10.1186/1471-2164-13-14>].
- [229] Fox GE, Wisotzkey JD, Jurtshuk P: **How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity.** *International Journal of Systematic Bacteriology* 1992, **42**:166–170, [<http://dx.doi.org/10.1099/00207713-42-1-166>].
- [230] Stackebrandt E, Goebel B: **Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology.** *International Journal of Systematic Bacteriology* 1994, **44**(4):846–849, [<http://dx.doi.org/10.1099/00207713-44-4-846>].
- [231] Schloss PD, Westcott SL: **Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis.** *Applied and Environmental Microbiology* 2011, **77**(10):3219–3226, [<http://dx.doi.org/10.1128/AEM.02810-10>].
- [232] Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.** *DNA Research* 2001, **8**:11–22, [<http://dx.doi.org/10.1093/dnares/8.1.11>].
- [233] Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamouisis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR: **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7.** *Nature* 2001, **409**(6819):529–533, [<http://dx.doi.org/10.1038/35054089>].
- [234] Ben-Amor K, Heilig H, Smidt H, Vaughan EE, Abee T, de Vos WM: **Genetic Diversity of Viable, Injured, and Dead Fecal Bacteria Assessed by Fluorescence-Activated Cell Sorting and 16S rRNA Gene Analysis.** *Applied and Environmental Microbiology* 2005, **71**(8):4679–4689, [<http://dx.doi.org/10.1128/AEM.71.8.4679-4689.2005>].
- [235] Sorek R, Cossart P: **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity.** *Nature Reviews Genetics* 2010, **11**:9–16, [<http://dx.doi.org/10.1038/nrg2695>].
- [236] Wilmes P, Bond PL: **The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms.** *Environmental Microbiology* 2004, **6**(9):911–920, [<http://dx.doi.org/10.1111/j.1462-2920.2004.00687.x>].
-

- [237] Gilbert JA, Hughes M: **Gene Expression Profiling: Metatranscriptomics**. In *High-Throughput Next Generation Sequencing, Volume 733 of Methods in Molecular Biology*. Edited by Kwon YM, Ricke SC, Humana Press 2011:195–205, [[http://dx.doi.org/10.1007/978-1-61779-089-8\\_14](http://dx.doi.org/10.1007/978-1-61779-089-8_14)].
- [238] Poretsky RS, Bano N, Buchan A, LeClerc G, Kleikemper J, Pickering M, Pate WM, Moran MA, Hollibaugh JT: **Analysis of Microbial Gene Transcripts in Environmental Samples**. *Applied and Environmental Microbiology* 2005, **71**(7):4121–4126, [<http://dx.doi.org/10.1128/AEM.71.7.4121-4126.2005>].
- [239] Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA: **Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre**. *Environmental Microbiology* 2009, **11**(6):1358–1375, [<http://dx.doi.org/10.1111/j.1462-2920.2008.01863.x>].
- [240] Urich T, Lanzén A, Stokke R, Pedersen RB, Bayer C, Thorseth IH, Schleper C, Steen IH, Ovreas L: **Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics**. *Environmental Microbiology* 2014, **16**(9):2699–2710, [<http://dx.doi.org/10.1111/1462-2920.12283>].
- [241] He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R, Hugenholtz P: **Validation of two ribosomal RNA removal methods for microbial metatranscriptomics**. *Nature Methods* 2010, **7**(10):807–812, [<http://dx.doi.org/10.1038/nmeth.1507>].
- [242] Taverna DM, Goldstein RA: **Why are proteins marginally stable?** *Proteins: Structure, Function, and Bioinformatics* 2002, **46**:105–109, [<http://dx.doi.org/10.1002/prot.10016>].
- [243] Hettich RL, Sharma R, Chourey K, Giannone RJ: **Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities**. *Current Opinion in Microbiology* 2012, **15**(3):373–380, [<http://dx.doi.org/10.1016/j.mib.2012.04.008>].
- [244] Simon C, Daniel R: **Metagenomic Analyses: Past and Future Trends**. *Applied and Environmental Microbiology* 2011, **77**(4):1153–1161, [<http://dx.doi.org/10.1128/AEM.02345-10>].
- [245] Bastida F, Moreno J, Nicolás C, Hernandez T, Garcia C: **Soil metaproteomics: a review of an emerging environmental science. Significance, methodology and perspectives**. *European Journal of Soil Science* 2009, **60**(6):845–859, [<http://dx.doi.org/10.1111/j.1365-2389.2009.01184.x>].
- [246] Lacerda CMR, Reardon KF: **Environmental proteomics: applications of proteome profiling in environmental microbiology and biotechnology**. *Briefings in Functional Genomics and Proteomics* 2009, **8**:75–87, [<http://dx.doi.org/10.1093/bfpgp/elp005>].

- [247] Turnbaugh PJ, Gordon JI: **An Invitation to the Marriage of Metagenomics and Metabolomics.** *Cell* 2008, **134**(5):708–713, [<http://dx.doi.org/10.1016/j.cell.2008.08.025>].
- [248] Wikoff WR, Anfora AT, Liu J, Schultz PG, Lesley SA, Peters EC, Siuzdak G: **Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites.** *Proceedings of the National Academy of Sciences* 2009, **106**(10):3698–3703, [<http://dx.doi.org/10.1073/pnas.0812874106>].
- [249] Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PSG, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glöckner FO: **Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications.** *Nature Biotechnology* 2011, **29**(5):415–420, [<http://dx.doi.org/10.1038/nbt.1823>].
- [250] Delmont TO, Simonet P, Vogel TM: **Mastering methodological pitfalls for surviving the metagenomic jungle.** *BioEssays* 2013, **35**(8):744–754, [<http://dx.doi.org/10.1002/bies.201200155>].
- [251] Ondov BD, Bergman NH, Phillippy AM: **Interactive metagenomic visualization in a Web browser.** *BMC Bioinformatics* 2011, **12**:385, [<http://dx.doi.org/10.1186/1471-2105-12-385>].

# Acknowledgements

I would like to thank all the people who have supported and accompanied me throughout the last years. First, I want to thank my supervisor Professor Dr. Burkhard Morgenstern for leading me to the field of bioinformatics, accommodating me in his group, and supporting me in my work. I am much obliged to my advisor Dr. Peter Meinicke for his support throughout the last years. Without his comments and suggestions, this thesis would not have been possible. I want to thank Professor Dr. Wingender and Professor Dr. Walter for support and fruitful discussions in the thesis committee meetings. I would also like to thank Dr. Bernd Wemheuer for very helpful discussions about metagenomics.

My colleagues always provided me an enjoyable and scientifically inspiring environment. I am especially grateful for the effort that Anne-Kathrin Schultz put into the proof reading of this thesis. In particular, I would like to thank Manuel Landesfeind and Alexander Kaever for fruitful discussions and for always giving me motivating feedback, it was a pleasure to work with them.

Many thanks to all my friends and my family for their incomparable and inexhaustible support and patience. They always helped me stay on the right track and always believed in me. I would not have made it without them.



# Acronyms

<b>rRNA</b>	ribosomal ribonucleic acid	4
<b>DNA</b>	deoxyribonucleic acid	5
<b>RNA</b>	ribonucleic acid	5
<b>PCR</b>	polymerase chain reaction	5
<b>bp</b>	base pairs	5
<b>NGS</b>	next-generation sequencing	6
<b>HMP</b>	Human Microbiome Project [81]	7
<b>EMP</b>	Earth Microbiome Project [82, 83]	7
<b>QIIME</b>	Quantitative Insights Into Microbial Ecology [100]	8
<b>OTU</b>	operational taxonomic unit	8
<b>BLAST</b>	Basic Local Alignment Search Tool [104]	8
<b>RDP Classifier</b>	Ribosomal Database Project Classifier [105]	8
<b>RDP</b>	Ribosomal Database Project [107]	8
<b>PICRUSt</b>	Phylogenetic Investigation of Communities by Reconstruction of Un- observed States [111]	8
<b>IMG</b>	Integrated Microbial Genomes [118]	9
<b>KO</b>	KEGG Orthology	9
<b>COGs</b>	Clusters of Orthologous Groups [121]	9
<b>PAUDA</b>	protein alignment using a DNA aligner [147]	11
<b>KEGG</b>	Kyoto Encyclopedia of Genes and Genomes [119, 120]	12
<b>GO</b>	Gene Ontology [152]	12
<b>TCA cycle</b>	tricarboxylic acid cycle	12
<b>HUMAnN</b>	HMP Unified Metabolic Analysis Network [145]	14
<b>MEGAN</b>	MEtaGenome ANalyzer [161]	14

<b>CPU</b>	central processing unit	16
<b>EBI</b>	European Bioinformatics Institute [210]	18
<b>FOU</b>	fraction of oligonucleotides unexplained	67
<b>FSU</b>	fraction of sequences unassigned	67
<b>FDU</b>	fraction of domain hits unexplained	67
<b>FTU</b>	fraction of taxonomic units unexplained	68
<b>GEBA</b>	Genomic Encyclopedia of Bacteria and Archaea [225]	69