Freie Universität Berlin

Computational methods for the identification and quantification of microbial organisms in metagenomes

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Martin Michael Serenus Lindner

Berlin
August 2014

Betreuer:	PD Dr. Bernhard Renard
Erstgutachter:	PD Dr. Bernhard Renard
Zweitgutachter:	Jun.-Prof. Dr. Tobias Marschall
Tag der Disputation:	15. Oktober 2014

Abstract

Metagenomics allows analyzing genomic material taken directly from the environment. In contrast to classical genomics, no purification of single organisms is performed and therefore the extracted genomic material reflects the composition of the original microbial community. The possible applications of metagenomics are manifold and the field has become increasingly popular due to the recent improvements in sequencing technologies. One of the most fundamental challenges in metagenomics is the identification and quantification of organisms in a sample, called *taxonomic profiling.*

In this work, we present approaches to the following current problems in taxonomic profiling: First, differentiation between closely related organisms in metagenomic samples is still challenging. Second, the identification of novel organisms in metagenomic samples poses problems to current taxonomic profiling methods, especially when there is no suitable reference genome available.

The contribution of this thesis comprises three major projects. First, we introduce the Genome Abundance Similarity Correction (GASiC) algorithm, a method that allows differentiating between and quantifying highly similar microbial organisms in a metagenomic sample. The method first estimates the similarities between the available reference genomes with a simulation approach. Based on the similarities, GASiC corrects the observed abundances of each reference genome using a non-negative lasso approach. In several experiments we showed that the abundance estimates are highly accurate and reduce the error compared to current approaches by 5% to 60%. The approach was also successfully applied to metaproteomics.

In the second project, we developed a statistical framework to fit mixtures of discrete distribution functions to the histograms of sequencing coverage depth after mapping metagenomic reads to reference genomes. We tailored a family of distributions for this particular application and modified the expectation-maximization algorithm to also fit discrete distributions when maximum likelihood estimation of the distribution parameters is not directly possible. The most important application of our framework is the genome validity score that measures how suitable a reference genome is for a particular (metagenomic) dataset.

In the third project, we developed a taxonomic profiling tool, called MicrobeGPS. In contrast to previous approaches, MicrobeGPS identifies and characterizes organisms in a metagenome even if there are no suitable reference genomes available. Distances to existing reference genomes are measured with the genome validity score and allow the user to spot organisms for which the available reference genomes are insufficient. We demonstrated on gold standard and real metagenomic data that our approach is more accurate than other existing methods, provides more meaningful results, and handles complex microbial communities.

Taken together, these three projects enhance the current repertoire of computational methods for taxonomic profiling and enable the simultaneous quantification of highly related organisms and the identification and characterization of unknown organisms in complex metagenomic datasets.

Zusammenfassung

Die Metagenomik untersucht mit Hilfe molekularbiologischer Methoden die Gesamtheit der genetischen Information einer Biozönose. Im Gegensatz zur klassischen Genomik werden hier die einzelnenen Mikroorganismen in der Probe nicht aufgereinigt oder angezüchtet, sodass das aus einer Probe extrahierte Genmaterial die Zusammensetzung der ursprünglichen Biozönose widerspiegelt. Aufgrund der technischen Fortschritte der letzten Jahre in der Genomsequenzierung hat sich das Anwendungsspektrum der Metagenomik zunehmend verbreitert. Eine der grundlegendsten Aufgaben der Metagenomik ist jedoch weiterhin das sogenannte *Taxonomic Profiling*, die Bestimmung und Quantifizierung aller Mikroorganismen in einer Probe.

In dieser Arbeit werden Ansätze zur Lösung folgender im Zusammenhang mit Taxonomic Profiling auftretender Probleme vorgestellt: Zum einen ist die gleichzeitige Bestimmung und zahlenmäßige Erfassung – und damit auch die Unterscheidung – sehr nah verwandter Organismen in metagenomischen Proben bisher sehr ungenau. Zum anderen stellt die Bestimmung unbekannter Organismen in metagenomischen Proben die gängigen Taxonomic Profiling-Ansätze vor große Probleme, insbesondere wenn keine vergleichbaren Genome nah verwandter Organismen bekannt sind.

Der wissenschaftliche Beitrag dieser Arbeit umfasst im Wesentlichen drei Projekte. Im ersten Projekt wird der GASiC-Algorithmus (Genome Abundance Similarity Correction) vorgestellt, der es ermöglicht zwischen sehr nah verwandten Organismen in derselben Probe zu unterscheiden und deren relative Häufigkeit zu bestimmen. Im ersten Schritt berechnet die Methode die Ähnlichkeiten zwischen den Genomsequenzen bekannter, nah verwandter Organismen über einen Simulationsansatz. Mithilfe der Ähnlichkeiten korrigiert GASiC die in der Probe beobachteten Häufigkeiten der bekannten Genomsequenzen über ein nicht-negatives LASSO. In Experimenten konnte gezeigt werden, dass die korrigierten Häufigkeiten den realen Häufigkeiten sehr gut entsprechen und um 5-60% geringeren Fehler aufweisen als bisherige Ansätze. Weiterhin konnte gezeigt werden, dass sich der Ansatz auch auf Probleme der Metaproteomik übertragen lässt.

Für das zweite Projekt wurden statistische Werkzeuge entwickelt, die es erlauben, komplexe Mischungen diskreter Wahrscheinlichkeitsverteilungen an Sequenziertiefe-Histogramme anzupassen, die bei der Zuordnung im Sequenzierprozess erzeugter Genomfragmente zu bekannten Genomsequenzen entstehen. Zu diesem Zweck wurden mehrere Verteilungsfunktionen entwickelt und zusammengestellt und eine Abwandlung des Expectation-Maximization-Algorithmus vorgestellt, die es erlaubt

Verteilungen anzupassen auch wenn keine Maximum-Likelihood-Schätzung der Verteilungsparameter möglich ist. Die wichtigste Anwendung stellt das sogenannte Genome Validity-Maß dar, welches die Ähnlichkeit einer bekannten Genomsequenz zu dem in einer (metagenomischen) Probe enthaltenen Genmaterial misst.

Als dritter Beitrag wurde das Taxonomic Profiling-Programm MicrobeGPS entwickelt. Im Gegensatz zu bestehenden Ansätzen bestimmt und charakterisiert MicrobeGPS die Organismen in einer Probe, ohne die Genomsequenzen der Organismen im Voraus kennen zu müssen. Die Abstände der Organismen zu bekannten Genomsequenzen werden über das Genome Validity-Maß geschätzt und ermöglichen damit dem Benutzer Organismen zu erkennen und einzuordnen, für die es unter den bekannten Genomsequenzen keine Entsprechung gibt. Auf Daten mit Goldstandard und Realdaten konnte gezeigt werden, dass der vorgestellte Ansatz genauere Ergebnisse liefert als bestehende Methoden. Weiterhin sind die Ergebnisse von MicrobeGPS insbesondere bei sehr komplexen Biozönosen im Vergleich zu anderen Methoden aussagekräftiger und leichter zu deuten.

Zusammengenommen erweitern diese drei Beiträge den Umfang der bestehenden computergestützten Taxonomic Profiling-Methoden, indem sie es ermöglichen sehr ähnliche Organismen in einer Probe gleichzeitig zu erfassen und bisher unbekannte Organismen zu bestimmen und zu charakterisieren.

Acknowledgements

First and foremost, I want to thank my supervisor Bernhard Renard for providing the excellent scientific environment for my work that did not lack the social component. I appreciated his guidance that gave me the freedom to develop and pursue my own ideas in the projects without getting lost in dead ends. In particular, I am grateful that he always had time to give advice when I knocked on his door and took off most of the administration load that allowed me to spend my time focused on research. Furthermore, I want to thank Knut Reinert, who would have been my official supervisor if Bernhard hadn't finished his Habilitation in time. His input and feedback to my projects was always valuable and welcome and I was very pleased by his enthusiasm pushing the GASiC project forward and including it into SeqAn. I would also like to thank Tobias Marschall who willingly agreed to review this thesis.

Credit also goes to all collaborators throughout the past years. The transfer of GASiC to metaproteomics only became possible through the persistent work of Anke Penzlin, who spent hours on analyzing, converting, and processing MS data, implementing and testing Pipasic, and writing sections for the paper. My thanks also go to Maximilian Kollock for developing the fundaments of the unofficially called MaxEM framework that grew into a decent publication and still lives on in the current projects. I also want to express my gratitude to Jörg Döllinger, Wojtek Dabrowski, Andreas Nitsche and Franziska Zickmann for their valuable and indispensable contribution to the Pipasic and MaxEM publications. Furthermore, I want to thank David Weese, Jochen Singer, and Stephan Aiche for their effort implementing the GASiC method in SeqAn and Knime. Although finally not part of this thesis, I also want to acknowledge the outstanding commitment of Franziska Metge and Benjamin Strauch to the metagenomic k-mer project, for which I hope to find some time in the future to continue the work. I also want to thank all proofreaders of this thesis, who made great efforts eliminating typos and improving the content.

My colleagues of the NG4 Bioinformatics research group at the Robert Koch Institute deserve special mention for the possibility to discuss and find advice for the day-to-day research problems on the one hand, and the kind atmosphere and marvelous after-work barbecues or gaming nights on the other hand. In particular, I want to thank my fellow PhD students Franzi (for drinking approximately 1,500 cups of coffee with me), Martina (for chatting about our projects and everything else before quitting time), Mathias, Kathrin, and Vitor, the postdocs Robert and

Carlus, and the NG4 alumni for the great time during and after work. Special thanks are devoted to Wojtek, who always had a helping hand when I struggled with the computational infrastructure and who made enormous efforts fighting for our issues in the IT department.

Last but not least, I want to send the warmest thanks to my family and my friends. Whenever I needed help, I could rely on your advice and support!

Contents

1. Introduction

1.1. Metagenomics

Metagenomics is the discipline that studies the genomics of uncultured microorganisms in their environment (Allen and Banfield, 2005; Wooley et al., 2010). In contrast to classical genomics, where the genomic material comes from a single clone, metagenomic material originates from heterogeneous communities of microbial organisms. This means, no purification of single organisms is performed and therefore the genomic material in the sample reflects the composition of the original microbial community. Further, being able to decipher the genomic sequences of uncultured organisms by means of genome sequencing provides access to the genomes of all microbial organisms (Namiki et al., 2012; Boisvert et al., 2012; Luo et al., 2012), not only the ones that can be cultivated in the laboratory – which are only the minority of all microorganisms (Torsvik et al., 1990).

The microbial communities that were analyzed in the past years cover a spectrum from very simple communities consisting of very few organisms to highly complex communities with possibly tens of thousands different bacterial and even more viral species, most of them still unknown. Among the first metagenomes that were analyzed systematically were acid mine drainage biofilms that grow under extreme conditions in highly acidic environments (Tyson et al., 2004). Due to their low complexity it was for the first time possible to assemble two almost complete genomes of organisms that could not be cultivated in the laboratory: *Leptospirillum* group II and *Ferroplasma* type II. Bioreactors, another example for low complexity microbial communities, can be used for the conversion of organic material into methane. Conversion involves only very few organisms (McInerney et al., 2009) and understanding and optimizing the interaction processes between the responsible microorganisms using metagenomics is of interest for industrial applications (Ellis et al., 2012). While the metagenomes found in other environments with extreme conditions have rather low complexity, such as in geysers or deep sea hydrothermal vents, the metagenomes of moderate habitats are populated by significantly more different organisms. For example, highly diverse metagenomes can be found in fresh water or salt water samples. Since water is an essential part of many ecosystems on earth, studying these habitats was of particular interest from the outset of metagenomics (Venter et al., 2004; DeLong et al., 2006; Rusch et al., 2007; Breitbart et al., 2009; Oh et al., 2011). Lakes and rivers are often used as drinking water reservoirs, so understanding the seasonal dynamics of microorganisms and the surveillance of the microbial composition in the reservoir can contribute to drinking water quality control (Oh et al., 2011). Since ocean microorganisms harbor a large share of the total mass of life on earth (Whitman et al., 1998), understanding their role in our ecosystem is fundamental. Today it is widely accepted that the most complex communities of microbial organisms are found in soil (Torsvik and Øvreås, 2002): current estimates range from about 2,000 to more than one million different organisms per gram of soil (Schloss and Handelsman, 2006; Gans et al., 2005). These communities have a large impact on other microbes and ecosystems; for example, soil microbes are a reservoir for antibiotic resistance genes that can also find their way to human associated microbiomes through gene transfer (Riesenfeld et al., 2004; Forsberg et al., 2014). However, soil communities are still under-explored and require new analysis methods (Xu et al., 2014; Howe et al., 2014).

By far the most effort has been put on studying the human microbiome – the entity of all microorganisms inhabiting the human body. While the human body consists of about 10^{13} cells, each person is inhabited by about one order of magnitude more (10^{14}) bacterial cells (Savage, 1977; Berg, 1996). Given the strong dependence between the human body and its microbial inhabitants (O'Hara and Shanahan, 2006), neither the analysis of the human genome is sufficient to describe the processes within the human body, nor is the analysis of a single microorganism sufficient for the understanding of a particular function or disease. Therefore, large scale projects such as the Human Micobiome Project (Peterson et al., 2009), gather huge amounts of metagenomic material from different body sites (The Human Microbiome Jumpstart Reference Strains Consortium, 2010) in order to establish a comprehensive catalog of microbial organisms inhabiting the human body. Other projects target particular microbial communities associated with the human body, e.g., the MetaHIT Consortium (Qin et al., 2010) analyzes the composition of the human gut microbiome and investigates its association to diseases such as obesity or inflammatory bowel disease.

The link between human associated metagenomics and clinical applications opened a new field: clinical metagenomics involves both understanding the dynamics and interaction between microorganisms and the human body as well as diagnostics of infectious or other microbiome related diseases. In particular the human gut microbiome is of interest for clinical applications. For example, understanding the interference of the human gut flora with antibiotic treatments can help reducing the risk of hemolytic-uremic syndrome after an antibiotic treatment (Wong et al., 2000; Sekirov et al., 2008; Hill et al., 2010). More recently, metagenomics was used to investigate an outbreak of shiga-toxigenic *Escherichia coli*, which allowed assembling a nearly complete genome of the outbreak strain purely from metagenomic data (Loman et al., 2013). In case of an outbreak, a high quality genome of the infectious agent can help researchers and clinicians developing treatments and diagnostic methods. Furthermore, the use of metagenomics for clinical applications is not restricted to prokaryotic organisms: viral metagenomics received attention in the last years due to the high genomic variability of viruses and the problem of cultivability (Mokili et al., 2012) that impedes effective response to disease outbreaks.

1.2. Taxonomic profiling

Metagenomics and the recent and ongoing development of genome sequencing technologies – such as Next Generation Sequencing (NGS), RNA-Seq, or single cell sequencing – enabled researchers studying the metagenome under various aspects. The basic question in a metagenomics experiment is typically: *Who is in the sample?* Identifying either the entirety of all organisms in the sample or checking for a particular or a small group of organisms of interest belongs to the most fundamental tasks in metagenomics. Closely related to this question is: *How many of them are in the sample?* While the absolute quantification (i.e. the number of cells of a specific organism) is not a major goal of metagenomic sequencing and is still mostly conducted with laboratory techniques such as quantitative PCR (Heid et al., 1996), the sequencing approach is used to infer relative abundances of organisms in the sample. Both the identification and quantification of organisms or higher level taxa in general was subject to extensive research and development in the past years. This part of metagenomics is summarized under the single term *taxonomic profiling*, which includes methods and strategies as diverse as the biological applications behind the sequencing data. Taxonomic profiling methods based on shotgun genome sequencing data can be roughly divided into assignment dependent and assignment independent approaches. The most relevant techniques in the context of this work are alignment based methods, a subset of the assignment dependent approaches. The following sections review the different approaches and briefly present the ideas behind established methods.

1.2.1. Assignment dependent taxonomic profiling

The class of assignment dependent taxonomic profiling approaches includes – strictly speaking – all approaches that use some kind of reference data for the identification or quantification of taxa. Possible reference data includes whole genome sequences, genes or other small parts of the genome, compositional properties of genomes such as oligo-nucleotide frequencies and codon usage, or databases of known protein sequences. Here, we focus on approaches making use of alignments to reference genomes or compositional features and discuss other approaches below.

Alignment based methods When whole reference genome sequences are available, the most intuitive approach is to map the metagenomic reads to the reference genome and use this information for the identification of taxa present in the dataset and their relative quantification. There are two possible perceptions of the same problem: from a read-centric view, the goal is to assign each read to its correct origin reference genome. From the genome-centric perspective, the goal is to identify and quantify the taxa in the sample based on the reads mapped to the reference genomes. Here, we focus on the genome-centric view. One of the first methods developed for metagenomic analyses is MEGAN (Huson et al., 2007), which analyzes BLAST read alignments (Altschul et al., 1997) to the reference genomes. Based on the BLAST quality scores and the set of genomes a read can be assigned to, MEGAN assigns the read to the lowest common ancestor (LCA) in the taxonomic tree of all involved reference genomes. The number of reads assigned to each taxon provide information about its abundance while the numbers of reads assigned to higher and lower taxa allow the user to judge the quality of the estimate. However, large numbers of reads are assigned to higher taxonomic ranks if the reference database contains many similar genomes, e.g., when multiple strains of the same organism are present. In these cases, it may not be possible to identify and quantify on the species level, but only on higher taxonomic levels. However, in particular clinical diagnostics applications require species- or even strain-accurate results. This problem was approached by the more recent tool GRAMMy (Liu et al., 2011), which estimates species level abundances for each reference genome from BLAST alignments. GRAMMy explicitly models the read assignment ambiguities in a probability matrix and thereby reflects the reference genome similarities. The problem is formulated as a finite mixture model that incorporates the read probability matrix and the genome lengths. The expectation-maximization algorithm is used to iteratively solve for the mixing parameters of the model: the relative genome abundances. A different approach has been presented more recently by Francis et al. (2013): they developed the tool Pathoscope, which constructs a Bayesian mixture model based on the likelihoods of each read alignment. The reads are then reassigned to their most likely origin by optimizing the model parameters by expectation-maximization. This allows Pathoscope to differentiate between similar strains even in cases with very low sequencing depth.

Since mapping the reads to thousands of reference genomes is very time-consuming and presents a major bottleneck in taxonomic profiling tools, different approaches have been developed to speed up the process. One approach is implemented in the tool MetaPhlAn (Segata et al., 2012), which maps the metagenomic reads to a set of previously selected marker sequences that are unique for each organism in the database. The marker sequences are carefully selected such that a read can only match to at most one marker, therefore read assignments are by construction unambiguous. This property can be exploited by the read mapper that can stop searching for matching positions if one match was found. Additionally, the small size of the marker sequence database further reduces the run time of the read mapping step.

The abundances of organisms are calculated by extrapolating the number of reads on the marker sequences to the whole genome and the accuracy is comparable to other reference based methods, such as Pathoscope. A similar approach to MetaPhlAn is implemented in the tool MetaPhyler (Liu et al., 2010).

Composition based methods Instead of reducing the size of the reference database, an alternative approach to speed up taxonomic profiling is reducing the computational effort for comparing the reads to the reference genomes. While read mappers typically search for full sequence alignments of the reads to reference genomes, in some cases it can be sufficient to compare compositional features in the reads and the references, such as the k-mer composition (i.e. the set of sub-sequences of length k) or codon usage patterns. There exist various alignment free methods for fast sequence comparison (Vinga, 2014; Vinga and Almeida, 2003; Rumble et al., 2009); for example, the recent tool Kraken (Wood and Salzberg, 2014) is a k-mer based tool specialized for fast taxonomic profiling of metagenomes. Kraken extracts k-mers from the metagenomic reads and searches the k-mers in the reference genome database. If a k-mer is present in multiple reference genomes, Kraken assigns the k-mer to the LCA in the taxonomic tree, similar to the strategy pursued in MEGAN. In this way, the set of all LCA taxa for each read is obtained and used to determine the appropriate label. Queries to the Kraken database require only very low computational effort, such that Kraken is about one order of magnitude faster than MetaPhlAn and can process about 4.1 million 100 bp reads per minute.

1.2.2. Assignment independent taxonomic profiling

A second class of taxonomic profiling methods is defined as all methods that do not require any kind of reference data, e.g., reference genomes or protein sequences (Mande et al., 2012). Since these methods almost exclusively compare features between the metagenomic reads, it is not directly possible to assign labels to the metagenomic dataset, such as the taxa that are present. Instead, these methods typically seek to create clusters of reads with similar features that can be used in further analysis steps. Binning (the process of creating read clusters) is mostly performed by comparing features between the reads, for example the GC-content or the k-mer composition (typically $k = 3, 4$, or 5). But also other features are possible: the AbundanceBin approach clusters reads into bins of similar sequencing depth (Wu and Ye, 2011). This is achieved by counting the number of occurrences of k-mers in all reads in the dataset and estimating the sequencing depth of a read via the observed abundances of the constituent k-mers. The basis for this approach is the Lander-Waterman-model (Lander and Waterman, 1988), which describes the local sequencing depth via a Poisson distribution. Under the assumption that the reads extracted by modern sequencing devices are evenly distributed over the genome –

which can only be assumed for moderate GC-contents (Dohm et al., 2008) – and the organism abundances are not evenly distributed (Angly et al., 2005; Hoffmann et al., 2007), the AbundanceBin approach is able to separate the metagenomic reads into bins, where the reads in each bin belong to mainly one organism. Although the clusters created by the binning methods neither provide information about the taxonomic identity of the organisms nor guarantee that each bin represents a single organism, it is often possible to infer quantitative traits of the metagenomic datasets from the results, such as the sample complexity (number of organisms) or the average genome length. Both the clusters itself and the secondary information can be used as preprocessing for follow-up analysis steps. One example is metagenome assembly (Dröge and McHardy, 2012; Wu and Ye, 2011), where previous clustering of the metagenomic reads into smaller bins can improve the quality of the generated contigs.

1.2.3. Other approaches

Besides the two already presented categories of taxonomic profiling methods, there exist several other approaches with high practical relevance that do not fit into one of the categories.

16S rRNA gene amplicon sequencing The analysis of the 16S small subunit ribosomal RNA (rRNA) was one of the first genotypic methods used for the identification of bacteria (Stackebrandt and Goebel, 1994; Clarridge, 2004). 16S rRNA is involved in the synthesis of all proteins in a cell and is therefore essential for all bacteria. Relationships between bacterial organisms can be inferred with phylogenetic methods from the differences on the 16S sequence; this also allows inferring the taxonomic affiliation of novel 16S fragments. Today, targeted sequencing of the 16S gene in microbial samples with NGS technologies (also called amplicon sequencing) is commonly used to assess the composition of a metagenome (Costello et al., 2009; Gilbert et al., 2012; Poretsky et al., 2014). Taken together, the advances in genome sequencing technologies and large 16S gene databases contribute to the popularity of this approach. However, due to the limited amount of sequence variability, short read length, and sequencing errors, it is often not possible to distinguish 16S sequences on the lower taxonomic levels (i.e. species, genus, or family) such that 16S profiling has a lower resolution and often lower sensitivity than whole genome based approaches (Poretsky et al., 2014).

Genome assembly A second popular approach is based on genome assembly in combination with gene prediction. Similarly to the other taxonomic profiling approaches, this approach uses whole genome sequencing reads from metagenomic samples. In a first step, these reads are assembled into larger contigs using either traditional genome assembly methods or methods specialized for metagenome assembly (Oh et al., 2011). In the next step, the assembled contigs are annotated with gene-finding methods, such as MetaGene (Noguchi et al., 2006), and identified genes are translated into proteins. These proteins can be searched in databases such as GenBank (Benson et al., 2008), Pfam (Finn et al., 2006), KEGG (Kanehisa and Goto, 2000), or COG (Tatusov et al., 2003) to retrieve taxonomic or functional annotation for the protein. The protein databases cover a much broader range of all prokaryotic organisms than the genome databases, therefore the combination of assembly and gene identification can be beneficial for under-explored metagenomic communities where only very few reference genomes are available. Alternatively, the assembled contigs can be searched against reference genomes at high error rates (e.g., using BLAST) to find more distantly related matches with higher specificity than compared to short read mapping.

This selection of taxonomic profiling methods is by far not complete and the number of published methods is constantly increasing. A broader overview of taxonomic profiling methods can be found in the review articles by Mande et al. (2012) or Teeling and Glöckner (2012).

1.3. Open problems in taxonomic profiling

The previous sections provided a brief overview over the field of metagenomics with focus on current methods for taxonomic profiling. Despite the extensive previous work in this field, there still remain unsolved problems. While the impact of some problems tends to decrease in the future due to the ongoing improvement of the technical components involved in metagenomics analyses, other problems will remain or even intensify with the technical advances. The limited number of available reference genomes, short read length and insufficient throughput of NGS devices, or limited computational resources are examples for problems of the former category. On the other hand, the complexity of reference databases and the similarity of the sequenced organisms will increase over time, making reference based identification both more precise and complicated. Other problems, such as evolving novel strains and species whose genomes differ from all previously known organisms, will also persist in the future. In this section, we describe two current challenges from the latter category in more detail.

1.3.1. Reference genome similarity

The steady, super-linear increase of the throughput of NGS devices (Stein, 2010) is one of the main promoters of the success of metagenomics: growing sequencing depths allow analyzing microbial communities in much greater detail and assembling even lower abundant genomes from environmental samples. Therefore, reference genome databases such as the NCBI RefSeq (Pruitt et al., 2007, 2014) have enormous growth rates (annual increase of NCBI RefSeq microbial genomes in 2011: 85.2%), which result both from re-sequencing of novel strains of already known organisms and from previously unknown organisms discovered in metagenomic studies. While the latter contributes to the breadth of reference databases, the former increases their depth. With high database depth it is in principle possible for taxonomic profiling to identify organisms with higher accuracy: if the reference database contains more strains of the same organism, it is more likely that the organism in the sample is similar to a reference genome in the database and thus identification down to the strain level becomes possible. On the other hand, it is more difficult to identify the correct strain in a sample if there are many highly similar reference genomes in the database. Most of the reads that were mapped to the correct reference genome will also map to the other sequenced strains. Given the differences between the sequenced organism and its reference genome due to natural variation and technical errors in the sequencing process, there can be considerable influences by incorrect mappings even on the correct reference genome, such that the identification of the correct strain is not trivial.

The presence of highly similar reference genomes in the database complicates taxonomic profiling in several ways. Taxonomic binning – (re)assigning a read to the correct reference – becomes ambiguous when the read maps to multiple reference genomes with the same mapping error and the coverage depth is similar for all reference genomes. Experiments demonstrate that there are also reads mapping uniquely to reference genomes of species that are not present in the sample (Lindner and Renard, 2015). Compensating for these effects is one of the biggest challenges for identifying organisms in the sample. Furthermore, estimating accurate relative abundances for each reference genome via the number of reads mapping to the genome is hampered by ambiguously matching reads. In particular, disentangling read assignments and relative abundances in metagenomes where two or more highly similar strains are present at different abundance levels is challenging and requires more innovation than higher throughput of genome sequencers or more computational resources.

Although the very coarse resolution of other taxonomic profiling strategies can be sufficient for some applications, such as 16S rRNA amplicon sequencing, the distinction between similar strains in a metagenome can have high practical impact. For example, the genomic difference within one species between highly pathogenic strains and strains commonly found in human associated metagenomes (skin, gut, etc.) can be very small and proper distinction based on metagenomics is only possible if sufficient strains of this organism (including the pathogenic ones) are available as reference genomes.

1.3.2. Missing reference genomes and inhomogeneous databases

The previous section described the challenges arising when many highly similar reference genomes are used for taxonomic profiling on low taxonomic levels. Another extreme case is insufficient breadth of the reference database, i.e. when no or only very few reference genomes are available for large parts of the taxonomy. For reference based taxonomic profiling methods this means that identification or quantification of organisms becomes imprecise or even impossible. For example, the NCBI bacterial genomes contain the *Akkermansia muciniphila* ATCC BAA-835 genome, which is the only representative of the class *Verrucomicrobiae* in the whole database. This means, *A. muciniphila* is the closest related taxon for all other species of the class *Verrucomicrobiae* and potentially has the highest genomic similarity. Therefore, reads from related organisms in the sample are likely to match to the only relative in the database, making the relative appear as present in the metagenome. However, when one of these isolated reference genomes is reported as present in the metagenome, it is not clear afterwards whether the organism itself is present or only a distantly related organism. This applies to approaches both based on reassigning reads to the most likely reference genome and assigning reads to their LCA in the taxonomic tree.

Furthermore, reference genome databases such as the NCBI RefSeq (Pruitt et al., 2007, 2014) are not homogeneous: while the density of reference genomes in some parts of the taxonomy is very low (as described above), model organisms such as *E. coli* or other organisms cultivable under laboratory conditions are massively over-represented. This means, while it is possible to differentiate between similar species or strains in some parts of the taxonomy, only very coarse and unreliable identifications can be expected in other cases. On the one hand, these skews in the database challenge the taxonomic profiling tools, which over-represent genomes from the crowded parts of the taxonomy and are at risk of missing organisms where only distantly related genomes are available. On the other hand, experimentalists may misinterpret the presented results and take a reported species as present although the sample only contains a distantly related organism.

The problem of incomplete and inhomogeneous reference databases is one of the main obstacles for the broad applicability of reference based taxonomic profiling methods. Since most current metagenomic research projects have only limited prior knowledge about the microbial composition and the genomes are not known yet, it is more important to obtain a broad overview of the composition than to have high resolution identification for few taxa. Due to their incompleteness and inhomogeneity, current reference genome collections compete with the breadth of 16S based studies (Poretsky et al., 2014). Although reference genome based taxonomic profiling will become more practical in the next years given the current technological growth rates (Stein, 2010), newly evolving microbial species, novel discoveries, and mutations in known organisms will also be a challenge for reference based methods in the future. Here, the development of a new class of methods that handles these problems is required.

1.3.3. Practical relevance

The background of this thesis and the main motivation behind the choice of the topic is my work at the Robert Koch Institute, the German central federal institute for public health and disease control and prevention. Making metagenomics accessible for diagnostics of infectious diseases could potentially enhance disease control by reducing the amount of time between a disease outbreak and identification of the disease agent. The unbiased approach of metagenomics has the benefit over traditional diagnostic tests that it is in principle possible to detect a novel, possibly unknown agent where no established test methods are available. Also mixed infections by multiple agents where the symptoms of the infections are superimposed can be detected with metagenomics, even in cases when the symptoms would suggest only one of infectious agents or a completely different one. However, it is not possible to use metagenomics for reliable diagnostics given the current status of the technological development. While reference free or 16S based approaches are technically not accurate enough to clearly identify an infection given a metagenomic sample, the reference based approaches struggle with the problems described above. Therefore, the approaches presented in this thesis are meant to improve reference based taxonomic profiling of metagenomic data and thus contribute to the development of metagenomics based diagnostics of infectious diseases.

1.4. Terminology and abbreviations

The term *coverage* is used frequently in the context of read mapping and metagenomics. However, its meaning is ambiguous and we will therefore follow the proposal by Rodriguez-R and Konstantinidis (2014) and strictly differentiate between the terms coverage and coverage depth. Genome *coverage depth* is the number of sequencing reads mapped to a specific position on a reference genome and is closely related with the term *sequencing depth.* While the sequencing depth describes the number of reads that were generated for each position in a genome in the sequencing process, the coverage depth describes the number of sequencing reads that were actually mapped to a position on a reference genome. Under ideal conditions sequencing depth and coverage depth are equal; however, considering errors in the reference genomes, mutations, repeats, and technical errors in the mapping process, the coverage depth typically differs from the sequencing depth. The *coverage* of a genome is of particular interest for metagenomics: it is defined as the fraction of the genome with non-zero coverage depth, i.e. the fraction of the positions in the genome that was actually covered by reads.

1.4.1. List of abbreviations

Abbreviation	Explanation
16S	16S ribosomal RNA, a component of prokaryotic ribosomes
AMD	Acid Mine Drainage
ANI	Average Nucleotide Identity
AUC	Area under the ROC curve
AVGRE	Average Relative Error
bp	Base pairs
BR	Brighton Red strain of the Cowpox virus
DWV	Deformed Wing Virus
EBI	European Bioinformatics Institute
EM	Expectation-Maximization algorithm
F	F-measure, the harmonic mean of precision and recall
FAMeS	An *in silico* metagenomic dataset
GCP	Genome Coverage depth Profile
HMP	Human Microbiome Project
Kre	Krefeld strain of the Cowpox Virus
LASSO	Least Absolute Shrinkage and Selection Operator
LCA	Lowest Common Ancestor
MC	Mock Community, an *in vitro* metagenomic community
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
NB	Negative Binomial distribution
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PSM	Peptide Spectrum Matches
qRT-PCR	Quantitative Real-Time Polymerase Chain Reaction
RDP	Read Distance Profile
RefSeq	Reference Sequence database of the NCBI
ROC	Receiver Operating Characteristic
RRMSE	Relative Root Mean Squared Error
SAM	Sequence Alignment/Map format
SRA	Sequence Read Archive of the NCBI.
STEC	Shiga-Toxigenic Escherichia Coli, a pathogenic bacterium
USR	Unique Source Reads
VDV	Varroa Destructor Virus
z,p,n,t	zero, Poisson, negative binomial, tail distribution
ZI	Zero-Inflated distribution

1.5. Thesis outline

This thesis describes computational approaches in metagenomics to improve reference based taxonomic profiling of microbial communities. In particular, the previously described problems with similar genomes and skewed and incomplete reference genome databases are addressed and possible solutions are presented. The three main contributions are described in chapters 2, 3, and 4. Chapter 5 discusses and summarizes the impact of the three contributions and concludes the results. All contributions were developed under the guidance of Bernhard Renard.

Chapter 2 addresses the problem of genome abundance estimation when highly similar reference genomes are available or the sample contains highly similar organisms. A novel method, *GASiC*, is presented and described in detail. Experiments demonstrate that GASiC is able to identify and quantify highly related organisms in metagenomic datasets and provides more accurate results than related tools. The chapter is based on the publication:

> *Metagenomic abundance estimation and diagnostic testing on species level.* M. S. Lindner and B. Y. Renard. *Nucleic Acids Research*, 41 (1): e10, 2013.

The concept behind GASiC was successfully transferred to metaproteomics and implemented in the method *Pipasic*. Pipasic was developed together with Anke Penzlin, the biological data for the evaluation was provided by Joerg Doellinger and Wojtek Dabrowski (both ZBS1, Robert Koch Institute). Pipasic was submitted as a conference contribution to ISMB 2014 and published as a journal article. Bernhard Renard and Anke Penzlin contributed to drafting the manuscript.

> *Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics.* A. Penzlin[1], M. S. Lindner[1], J. Doellinger, P. W. Dabrowski, A. Nitsche, and B. Y. Renard. *Bioinformatics*, 30 (12): i149–i156, 2014.

Chapter 3 describes a framework for fitting distribution functions to coverage depth profiles of genomes. The observed patterns in the profile can provide information both about the reference genome and the dataset. For the case of metagenomics, it is for example possible to calculate the *genome validity*, an estimate for the divergence between a reference genome and its closest relative in the metagenomic dataset. This project was developed together with Maximilian Kollock and Franziska Zickmann and published in the following article:

> *Analyzing genome coverage profiles with applications to quality control in metagenomics.* M. S. Lindner, M. Kollock, F. Zickmann, and B. Y. Renard. *Bioinformatics*, 29 (10): 1260–1267, 2013.

[1] Joint first authors

A novel taxonomic profiling approach is presented in chapter 4. The developed method *MicrobeGPS* finds candidate organisms in a metagenomic sample that are described by the available reference genomes. If only very few genomes are available, MicrobeGPS reports the closest related genome together with its distance to the candidate organism. The applicability and performance of this approach is demonstrated on metagenomic datasets from different habitats. This chapter is based on:

> *Metagenomic Profiling of Known and Unknown Microbes with MicrobeGPS.* M. S. Lindner and B. Y. Renard. *PLOS ONE*, 10 (2): e0117711, 2015.

2. Detection and quantification of highly similar organisms with GASiC

Detecting an organism in a metagenomic sample and determining its abundance are elementary tasks for taxonomic profiling tools. In the era of high throughput NGS technologies and large databases of reference genomes, the most accurate taxonomic profiling tools rely on aligning the metagenomic reads to the reference genomes. However, similar reference genomes of closely related organisms or genes shared over multiple species induce ambiguity in the read mapping process and impede simple interpretation of the data. Therefore, relative quantification of reference genomes with high similarity is often problematic for current methods, such that quantification on the species level is not always possible and multiple strategies were developed to approach this problem.

One way is to align reads against a comprehensive reference sequence database using BLAST (Altschul et al., 1997) and to subsequently analyze the results with tools such as MEGAN (Huson et al., 2007). As reads – especially short NGS reads – often match to multiple genomes, MEGAN assigns these ambiguous reads to nodes in the phylogenetic tree by finding the lowest common ancestor (LCA) node of all matching sequences. Assigning the reads to the LCA reduces the risk of a too optimistic assignment and thus of obtaining false positive matches; with the disadvantage that quantification may only be possible at a low resolution. Furthermore, MEGAN discards nodes with insufficient support, i.e. when the number of reads assigned to a node does not exceed a user defined threshold. The graphical user interface makes MEGAN highly suitable for the visual inspection of metagenomic data. Yet, MEGANs read counts are influenced by several factors such as genome sizes or the presence of similar genomes in the phylogenetic tree, which makes MEGAN less suitable for quantitative metagenomic analyses.

Another tool based on read alignment, GAAS (Angly et al., 2009), uses an iterative procedure to estimate improved relative genome abundances and an average genome length. To this end, GAAS calculates genome length corrected alignment qualities (*E-values*) for all matching reads and uses this information to iteratively calculate weights for each reference genome. Yet, ambiguities of read matches are only considered indirectly via the corrected E-values, which is only suitable if the reference genomes have low similarity.

GRAMMy (Xia et al., 2011) successively improves on GAAS as it explicitly models read assignment ambiguities in a probability matrix. The problem is formulated as a finite mixture model, which incorporates the read probability matrix and the genome lengths. The expectation-maximization (Dempster et al., 1977) algorithm is used to iteratively solve for the mixing parameters of the model: the relative genome abundances. In contrast to the previous methods, GRAMMy seeks to reflect the reference genome similarities in the mixture model. Yet, the similarity parameters are estimated from the alignment qualities of the reads to the reference genomes rather than from the reference genomes directly and are thus not accurate enough to allow robust abundance estimation in the case of highly similar reference genomes.

We observed that high similarity of reference sequences challenges all described methods. This can be problematic, for instance in diagnostic settings, when the distinction between presence and absence of single species or relative abundance levels are of eminent importance. To overcome this limitation, we present *Genome Abundance Similarity Correction* (GASiC), a versatile algorithm to estimate corrected abundances on the species level by directly accounting for the reference genome similarities. We demonstrate that GASiC is able to provide accurate abundance estimates for reference genomes with high sequence similarity and for complex metagenomic communities. Its simulation-based approach makes GASiC more independent from biases introduced by the sequencing technology, differences in genome sizes, or composition and structure of the reference sequences. Furthermore, GASiC provides statistical tests for the presence of a species in the sample.

The GASiC workflow is depicted in Figure 2.1. As in most reference based methods, the reads are first aligned to every genome in a set of references and the number of reads matching to each genome is counted. We call these counts the *observed abundances*, as opposed to the *abundance estimates* that we want to obtain in the end. In the next step, GASiC constructs a similarity matrix encoding the alignment similarities between the reference sequences. The similarity matrix and the observed abundances are then used together in a linear system of equations, where GASiC solves for the corrected abundances using a constrained optimization routine to obtain the estimates. The whole procedure can be iterated using bootstrap (Efron, 1979) samples from the original dataset. This yields more stable abundance estimates and provides an intuitive non-parametric statistical test for the presence of a species.

We first introduce some notation that will be used in the following. Starting from the experiment side, the sequencing dataset is denoted as D, containing N reads in total. The reads may originate from a set of M Species $S = \{S_i,\ i = 1..M\}$ with known reference sequences or possibly from other sources (noise, contaminants) with no relation to any species in S. S_i is synonymously used for both the species itself as well as its reference sequence. For quantification of species we use the term *abundance*, which is the number of reads belonging to the species divided by the total number of reads N. Due to amplification biases, this abundance may not represent

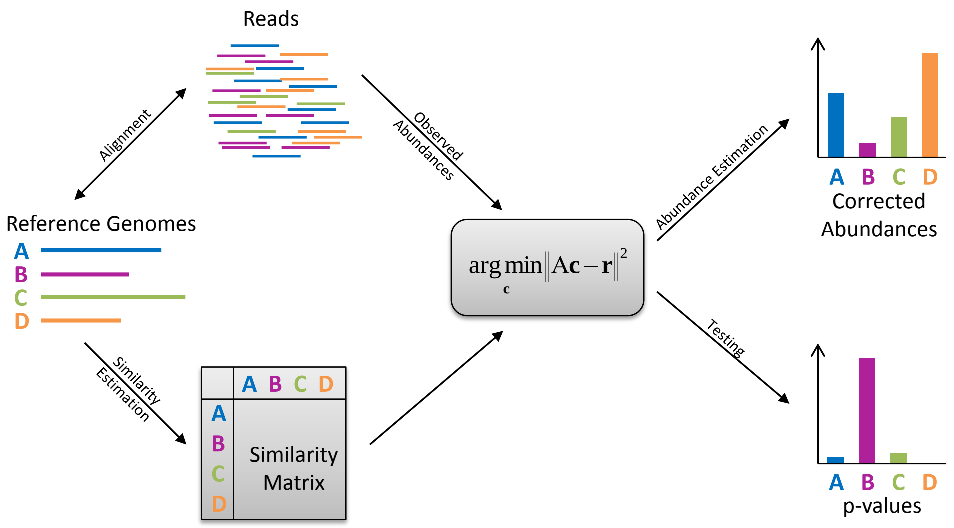


Figure 2.1.: GASiC workflow. Metagenomic reads are first aligned to the reference genomes and matching reads are counted for each genome (observed abundances). GASiC then uses the reference genomes to construct a similarity matrix encoding the genome similarities while considering influences of the applied sequencing technology. The similarity matrix and the observed abundances are used in a linear system of equations to model the influence of reference genome similarities on read alignment. GASiC solves the system of equations using a constrained optimization routine to calculate the estimated true abundances of the reference genomes in the dataset. Bootstrapping from the reads delivers stable abundance estimates and allows GASiC to test for the presence of each species in the dataset.

the true absolute abundance of the species in the data, but may be valuable when comparing abundances of multiple (in particular similar) species.

2.1. Genome similarity estimation

2.1.1. Alignment

The reads in D are aligned to all species S with an alignment method suitable for the characteristics of D. Then, we count the number of reads r_i from D that were successfully aligned to S_i, irrespective of the number of matching positions in S_i or matches to other species. In particular, we neither restrict ourselves to unique matches only, nor assume any phylogenetic structure within the S_i, as is done for example in MEGAN. If the dataset only contains very dissimilar species, the read counts r_i may already be suitable estimates for the true abundances. Otherwise, the

r_i are in general highly disturbed and dominated by shared matches, such that the r_i can not directly be used as abundance estimates.

2.1.2. Similarity estimation

A proper similarity estimation of the reference sequences is required to achieve accurate similarity correction of the r_i. The similarities between sequences are encoded in a similarity matrix $A = (a_{ij})$, $i, j = 1..M$, where a_{ij} denotes the probability that a read drawn from S_i can be aligned to S_j. In practice, we simulate a set of reads from every reference S_i with a read simulator that is able to imitate the sequencing technology and error characteristics of D. For example, Mason (Holtgrewe, 2010) and Grinder (Angly et al., 2012) simulate Illumina, 454, and Sanger reads, and dwgsim (`http://sourceforge.net/projects/dnaa/`) simulates Illumina, ABI SOLiD, and IonTorrent reads. Then, we align the simulated reads of S_i to S_j using the very same settings as for aligning the reads in dataset D and count the number of matching reads $\tilde{r}_{ij}$. The matrix entries are then estimated as $a_{ij} = \frac{\tilde{r}_{ij}}{\tilde{r}_{ii}}$.

The key element of similarity estimation is a proper read simulation since we use the simulated reads to estimate the reference genome similarities, the source of ambiguous alignments. Thus, the simulated reads should have the read characteristics and the error characteristics of the instrument (read length, paired/single end, etc.) and should cover the reference genome at least once.

For very complex metagenomic communities with a high number of species M, the calculation of the complete similarity matrix may become infeasible because of its computational complexity $O(M^2)$. We recommend to first estimate similarities using for example fast k-mer based methods (Reinert et al., 2009) and refine the estimates via the simulation approach only for genomes with sufficiently high (e.g., $a_{ij} > 0.01$) similarity.

2.2. Abundance estimation

2.2.1. Similarity correction

We introduce a linear model to correct the observed number of reads r_i for the influence of the genome similarity using the similarity matrix A. Let c_i denote the true, but unknown, abundance of species S_i. We then assume that the observed abundance r_i is a mixture of the true abundances c_j of all species S_j, weighted with the estimated probability a_{ij} that a read from j can be aligned to i:

$$\sum_j a_{ij} c_j = r_i.$$

To simplify notation, we use a matrix representation of the true and the observed abundances, i.e. $\mathbf{c} = (c_1, c_2, ..., c_M)^T$ and $\mathbf{r} = (r_1, r_2, ..., r_M)^T$. In matrix notation, this can be written as:

$$A\mathbf{c} = \mathbf{r}.$$

Since direct inversion of the matrix A may result in instable abundance estimates, we formulate the solution for $\mathbf{c}$ as a non-negative LASSO (Efron et al., 2004; Renard et al., 2008) problem:

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmin}} ||A\mathbf{c} - \mathbf{r}||_2$$
$$\text{s.t. } \hat{c}_i \geq 0 \ \forall i \text{ and } \sum_i |\hat{c}_i| \leq 1$$

The constraints enforce the result to be meaningful, i.e. each estimated relative abundance $\hat{c}_i$ must be equal to or greater than zero and the sum of all relative abundances must be less than or equal to one. The first conditions also ensure that the correction produces abundances lower than or equal to the measured abundances. The last condition allows the presence of reads from a totally unrelated species, since the abundances are allowed to sum up to less than or equal to one. It also enforces the sparsity of results such that only meaningful contributions have abundances larger than 0. We solve the constraint optimization problem with the COBYLA method implemented in Python SciPy (`www.scipy.org/`).

2.2.2. Error estimation and testing

We apply a bootstrapping procedure on the steps described before, first, to estimate how errors in the input data propagate through the correction algorithm and, second, to calculate p-values to test for the presence of a species in the sample. To this end, we generate B bootstrap samples from the dataset D and perform similarity correction for each sample separately, yielding a distribution $\hat{c}_{i,b}$ $b = 1..B$ of abundances for each species i. We calculate the average abundance $\overline{c_i}$ and estimate the standard error $\sigma_i = \sqrt{\text{VAR}(\hat{c}_{i,b})}$. To test whether a species is present in the sample, we count how many bootstrap samples yielded a higher abundance than an a priori defined detection threshold t:

$$p(c_i > t) = \frac{\#(c_{i,b} > t, b = 1..B)}{B}.$$

2.2.3. Quality check

As the composition of the reference genome set is critical for the complete method, GASiC offers an additional quality check after the alignment to reference genomes.

The quality check step analyzes the outputted SAM files of the read alignment tool and provides helpful statistics to the user to judge the appropriateness of the results. Besides reporting statistical measures, such as the number of mapped reads or the average genome coverage depth, GASiC generates a coverage depth histogram that often allows the user to exclude certain genomes from the reference set or to detect possibly important missing reference genomes (see also chapter 3 and Lindner et al. (2013)). For example, a high number of uncovered bases in combination with a typical Poisson distribution at higher coverage depth may indicate that the considered species is not contained in the dataset, but a closely related species. In addition to the statistics and the histogram, GASiC produces warning messages in critical setups, e.g., when the dataset may be too small for abundance estimation or large parts of the genome are not covered although there is evidence for the genome in the dataset.

2.2.4. Implementation

We implemented GASiC in Python (Van Rossum and Drake Jr, 1995), making extensive use of the high performance scientific computing libraries SciPy and NumPy (`www.scipy.org/`). Since GASiC is independent from the choice of the alignment algorithm and read simulator, we already integrated interfaces to a set of tools. The user can add custom interfaces easily, a brief manual is provided within the code. We set value on comprehensible and well documented code, such that GASiC can easily be adapted to the users needs without deeper knowledge of Python.

GASiC requires the widespread SAM alignment format (Li et al., 2009) as output from the alignment tool to analyze the results, since most alignment tools either directly support SAM output or alignment results can be readily converted into SAM files.

The GASiC tool and source code is available for download at `http://sourceforge.net/projects/gasic/`. An implementation of the GASiC method in SeqAn (Döring et al., 2008) and Knime (Berthold et al., 2009) was created by David Weese, Stephan Aiche and Jochen Singer from the SeqAn team at FU Berlin and is available from `https://github.com/seqan/knime_seqan_workflows/tree/master/metagenomics_gasic_workflow/`.

2.3. Experiments and results

We sought to corroborate the key features of GASiC with corresponding experiments. First, we compared GASiC to previous methods on a common reference dataset. Second, we demonstrate GASiCs power to disambiguate abundances of highly similar bacteria and to test for the presence of species. Third, we present a potential application besides metagenomics: we analyzed a published viral dataset and compared GASiCs results to abundance levels obtained by a quantitative PCR method.

2.3.1. FAMeS dataset

The established metagenomic FAMeS (Mavromatis et al., 2007) reference datasets contain shotgun sequencing reads of 113 microbial species mixed into three datasets with low, medium, and high complexity. The low complexity dataset `simLC` simulates a bioreactor community with one dominant and many low abundant genomes. The `simMC` dataset mimics a moderately complex community, as for example found in acid mine drainage biofilms, with few dominating species flanked by low abundant ones. A typical metagenomic dataset with high complexity and no dominant species is simulated in `simHC`. Each dataset consists of approximately 100,000 Sanger reads with approximate read length of 1,000 bp randomly selected from the 113 sequenced microbial genomes, thus the exact number of reads per species, the origin of every read, and the reference genomes are available. The sequence read data including detailed information for all datasets was downloaded from `http://fames.jgi-psf.org/`. The reference sequences were downloaded from NCBI using the provided Taxon IDs.

Xia *et al.* compared the performance of the tools MEGAN, GAAS, and GRAMMy on the FAMeS dataset, see Xia et al. (2011) for details. To extend this comparison, we repeated the experiment with GASiC under the same conditions. We estimated the corrected abundances with GASiC and measured the *Relative Root Mean Squared Error* and *Averarage Relative Error* (RRMSE and AVGRE) on all datasets. RRMSE measures the sum of *squared* relative errors, while AVGRE is the sum of *absolute* relative errors, thus, RRMSE is more sensitive to outliers. Given the true abundances t_i and the corrected abundances c_i, $i = 1..M$, the error measures are defined as follows:

$$RRMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\frac{|c_j - t_j|}{t_j}\right)^2}$$

$$AVGRE = \frac{1}{M}\sum_{j=1}^{M}\frac{|c_j - t_j|}{t_j}.$$

For the construction of the similarity matrix, we simulated 20,000 Sanger reads for each species using the Mason (Holtgrewe, 2010) read simulator. Mason comes with a built-in error model for Sanger reads and thus needs only very few tuning parameters. The exact command was `mason sanger -N 20000 -nm 961 -ne 145 -sq -o [reads] [reference]`, creating reads of length 961 bp with standard error 145 bp. These numbers were estimated from the FAMeS datasets. Then, the distance matrix was constructed as described; explicit calculation of the distance matrix has complexity $O(M^2)$ in the number of reference sequences M and required almost 500 GB of hard disk memory and two days computation on 4 processors. Since most entries in the similarity matrix were exactly or close to 0, we recommend to use, e.g., fast k-mer comparison (Reinert et al., 2009) to select candidate genomes with a minimum similarity and only measure their distance by simulating and aligning reads and to set all other similarities to exactly 0.

Alignments to the reference genomes were performed with Bowtie 2 (beta4) (Langmead and Salzberg, 2012), which can align long reads and handle indels. We used Bowtie 2 with default parameters in the `−−local` mode, which allows mismatches on the ends of the reads. Bowtie 2 only reported the first found matching position for every read. The complete command was `bowtie2 -U [reads] -x [reference] -S [output] -p 4 −−local -M 0`.

The error measures of MEGAN, GAAS, and GRAMMy, as reported in Xia et al. (2011), and GASiC are compared in Table 2.1. GASiC strongly reduces the estimated errors on all three datasets compared to the competing methods; the strongest error reduction compared to GRAMMy is achieved on the high complexity simHC dataset, where the error rates are reduced by 51.9% and 60.5% for RRMSE and AVGRE, respectively. It is notable that the relative difference between RRMSE and AVGRE is smaller for GRAMMy than for GASiC: For GASiC, RRMSE is about twice as high as AVGRE. Since RRMSE is more sensitive to outliers, this indicates that the GASiC error is mainly dominated by few outliers, while the majority of abundance estimates has very low error. The overall high increase of accuracy demonstrates GASiCs ability to quantify low abundances correctly, even when a large number of reference genomes is used. Also the differing genome lengths (ranging from 1.0 Mbp to 9.7 Mbp) did not pose an obstacle for GASiC.

2.3.2. Mixed E. coli / STEC dataset

In the second experiment, we combined two real datasets, *E. coli* DH10B and *E. coli* TY-2482, in selected fractions. Both datasets were acquired with an IonTorrent PGM device. The *E. coli* DH10B dataset was recorded by Lifetech as a challenge dataset and contains 522,099 reads with an average length of 93.9 bp and is freely available upon registration on the IonTorrent Community homepage. The *E. coli* TY-2482 dataset contains 977,971 reads with average length 181.7 bp. Dataset sources are provided in Table A.1. *E. coli* TY-2482, which is also known as shigatoxigenic *E. coli* (STEC), is highly similar to *E. coli* DH10B and received attention in the so called *German 2011 STEC outbreak* and we therefore, respectively, term the datasets *E. coli* and *STEC* for a better differentiation. To eliminate the read length differences in the datasets, we trimmed all reads to 80 bp and discarded shorter reads. These reads were used to create 11 datasets with varying *E. coli*

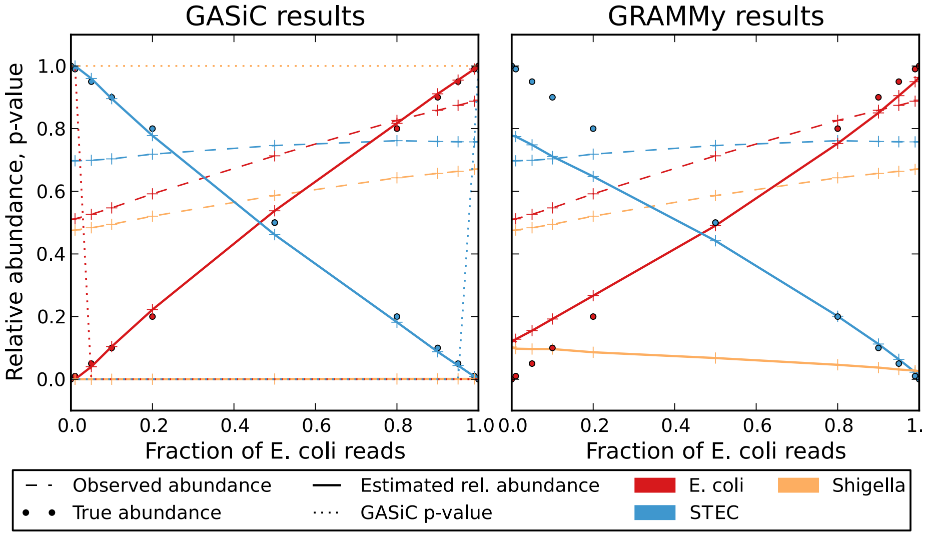


Figure 2.2.: Comparison of GASiC and GRAMMy on synthetic datasets with varying concentrations of real *E. coli* and STEC reads. Both algorithms estimated the relative abundances of the highly similar bacteria *E. coli*, STEC, and *Shigella flexneri* in all datasets and GASiC tested (p-value) for the absence of each bacterium. GRAMMy was challenged by the similarity of the bacteria and deviated strongly from the expected relative concentrations. For *S. flexneri*, which was not present in the sample, GRAMMy incorrectly estimates abundances up to 10%. GASiC provided more stable abundance estimates at all concentrations and also correctly identified *S. flexneri* as not present in the dataset and accordingly assigned high p-values.

Table 2.1.: Benchmark comparison on FAMeS datasets. In addition to MEGAN-based, GAAS, and GRAMMy abundance estimates (Xia et al., 2011), we calculated abundance estimates with GASiC for all reference genomes in the FAMeS datasets simLC, simMC, and simHC. The four tools are compared by their relative error (Relative Root Mean Square Error and Average Relative Error). GASiC reduces the relative error on all datasets and improves on GRAMMy, the best existing tool, by up to 60%. Best results are achieved on the high complexity dataset simHC, indicating that GASiC provides a particularly large benefit for complex mixtures where more corrections are necessary and low concentrations exist, which are more difficult to estimate.

	simLC		**simMC**		**simHC**	
Tool	**RRMSE**	**AVGRE**	**RRMSE**	**AVGRE**	**RRMSE**	**AVGRE**
MEGAN	48.6%	39.3%	50.0%	40.6%	50.2%	40.8%
GAAS	433.8%	152.5%	171.4%	11.6%	507.9%	165.8%
GRAMMy	20.0%	14.0%	25.6%	19.7%	21.6%	14.7%
GASiC	18.7%	9.1%	17.5%	10.9%	10.4%	5.8%

and STEC concentrations. Each dataset consisted of 400,000 reads, the fractions of *E. coli* reads were 0.0, 0.01, 0.05, 0.10, 0.20, 0.50, 0.80, 0.90, 0.95, 0.99, 1.0, the remaining reads were filled from the STEC dataset.

All combined datasets were analyzed with GASiC and GRAMMy, the two best performing tools from the previous experiment. In addition to the *E. coli* and the STEC references, we included *S. flexneri* as phantom reference. Herewith, we challenged the tools, first, to distinguish highly similar reference genomes over a wide range of abundances and, second, to exclude reference genomes not present in the data. Both GASiC and GRAMMy were applied to all 11 datasets using the *E. coli*, STEC, and *S. flexneri* reference genomes. For GRAMMy, reads were aligned as described in the original paper using BLAT with default settings and the results were then passed to the GRAMMy pipeline to run the EM estimation of the abundances. For GASiC, we used Bowtie (Langmead et al., 2009) to align the reads to the reference genomes and analyzed the output SAM (Li et al., 2009) files. We used the following command to invoke the alignment: `bowtie -S -p 2 -q -3 30 -v 2 [index] [reads] > [samfile]`. Note that we allowed up to 2 mismatches in total and discarded the last 30 bp from the read.

Figure 2.2 shows the estimated relative abundances of both tools for *E. coli*, STEC, and *S. flexneri*. Detailed results are reported in Table 2.2. The raw abundance estimates based on the read counts do not allow proper abundance estimation for all datasets, the raw abundances of *E. coli* and STEC are always very close to each other, irrespective of the true concentration in the dataset. E.g., the raw abundance of STEC in the pure *E. coli* dataset would be more than 70% and the abundance of

S. flexneri, which is not present in any dataset, varies between 40% and 70%. Here, both approaches improve significantly on the raw abundance estimates. Overall, GASiC abundance estimates are closer to the ground truth than the GRAMMy abundance estimates, especially in the case of low abundances. It persistently rules out the presence of all phantom references correctly, where the diagnostic detection threshold t in GASiC was set to disregard abundances below 1%. The statistical test for the presence of a genome assigns high p-values to *S. flexneri* in all datasets, to STEC and *E. coli* only at concentrations of 1% or below, proving GASiC suitable for detecting the presence of low abundant genomes.

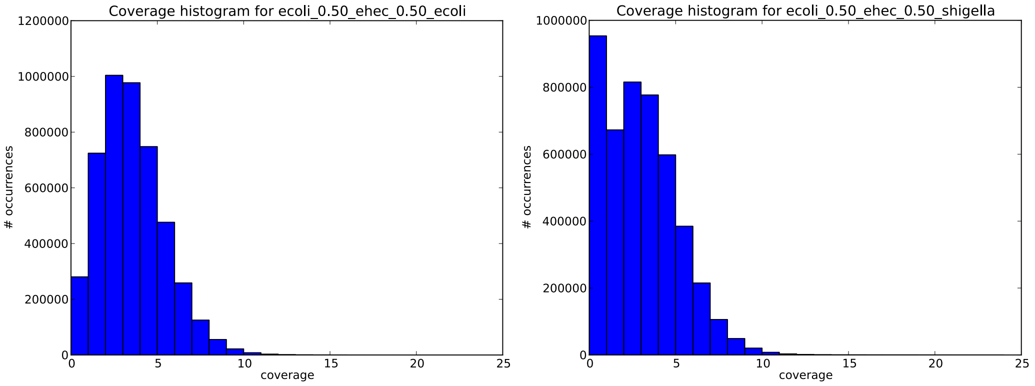
In follow-up experiments, we challenged GASiC under complicated conditions. First, we sought to test how robust the results are with respect to the size of the reference sequence database. Therefore, we increased the number of phantom references and used six reference genomes that are similar to *E. coli* and STEC: *Escherichia fergusonii*, *Klebsiella pneumoniae*, *Pantoea ananatis*, and again *S. flexneri*. We report the results in Table A.2. GASiC consequently estimates zero abundance and high p-values for all additional genomes, while the estimates for *E. coli* and STEC are consistent with the previous experiment (maximum absolute difference < 0.004). We conclude that additional genomes in the reference set seem not to affect the accuracy of GASiCs estimates, as long as the correct reference genomes are in the set.

Next, we simulated a very distant unknown species in the metagenome. Therefore, we enlarged the mixed datasets by adding randomly generated reads that simulated the unknown species. Here, we call the set of available reference sequences a *closed subset* of all genomes present in the dataset, as we expect no reads belonging to the missing genomes to be ambiguously aligned to the genomes in the reference set. The numbers reported in Table A.3 show that the additional reads have no influence on GASiCs estimates. Therefore, GASiC should be able to provide reliable estimates in cases when not all reference genomes are available, as long as the missing genomes are not similar to the reference genomes used in the reference set. As the reads did not match to any of the reference sequences, GASiCs estimates were not affected by the noise reads.

Furthermore, we removed the STEC genome from the reference set to simulate the effect of having a novel species in the dataset with high similarity to existing ones. Abundances were estimated by both GASiC and GRAMMy in order to see how the methods handle this difficult situation. We report the results in Tables A.4 and A.5. We observe that both methods have severe problems estimating the true abundance of the reference sequences and respond to the additional STEC reads by overestimating the abundances of genomes similar to STEC, where GASiC produced overall better estimates than GRAMMy. Yet, genomes with very small genomic distance (here: *P. ananatis*) are not affected by the missing reference sequence, corroborating the findings of the previous experiment on closed subsets. In this case, GASiCs

Table 2.2.: Comparison on mixed *E. coli* / STEC datasets. The abundances of *E. coli*, STEC and *S. flexneri* in the 11 mixed datasets were estimated with GASiC and GRAMMy. *Frac* denotes the true fraction of a species in the dataset, *Aln* is the abundance estimated by counting the number of mapped reads. *GASiC* is the relative species abundance estimated by GASiC, *P* it the p-value assigned by GASiC. *GRAMMy* denotes the relative species abundance by GRAMMy.

	E. coli					**STEC**					***S. flexneri***				
Dataset	**Frac**	**Aln**	**GASiC**	**P**	**GRAMMy**	**Frac**	**Aln**	**GASiC**	**P**	**GRAMMy**	**Frac**	**Aln**	**GASiC**	**P**	**GRAMMy**
1	0.00	0.509	0.000	1.00	0.120	1.00	0.698	1.000	0.00	0.778	0.00	0.475	0.000	1.00	0.103
2	0.01	0.512	0.000	1.00	0.128	0.99	0.697	1.000	0.00	0.775	0.00	0.476	0.000	1.00	0.098
3	0.05	0.526	0.041	0.00	0.155	0.95	0.699	0.959	0.00	0.748	0.00	0.484	0.000	1.00	0.097
4	0.10	0.547	0.105	0.00	0.192	0.90	0.703	0.895	0.00	0.711	0.00	0.494	0.000	1.00	0.096
5	0.20	0.592	0.222	0.00	0.267	0.80	0.719	0.777	0.00	0.647	0.00	0.521	0.000	1.00	0.086
6	0.50	0.713	0.538	0.00	0.490	0.50	0.746	0.461	0.00	0.442	0.00	0.586	0.001	1.00	0.068
7	0.80	0.826	0.817	0.00	0.753	0.20	0.761	0.182	0.00	0.201	0.00	0.643	0.001	1.00	0.046
8	0.90	0.859	0.911	0.00	0.850	0.10	0.759	0.088	0.00	0.113	0.00	0.657	0.001	1.00	0.037
9	0.95	0.874	0.955	0.00	0.905	0.05	0.758	0.044	0.00	0.064	0.00	0.664	0.001	1.00	0.031
10	0.99	0.888	0.991	0.00	0.950	0.01	0.758	0.008	0.80	0.023	0.00	0.670	0.001	1.00	0.028
11	1.00	0.891	0.998	0.00	0.964	0.00	0.758	0.001	0.99	0.011	0.00	0.672	0.001	1.00	0.026



(a)

(b)

Figure 2.3.: Coverage depth histogram produced by the quality check step in GASiC after mapping 400,000 reads (200,000 *E. coli*, 200,000 STEC) to the *E. coli* (a) and *S. flexneri* (b) reference genome. (a) The coverage depth histogram has a Poisson-like shape, as one would expect it for a genome present in the dataset. (b) The coverage depth histogram has a Poisson-like shape for higher coverages, but shows an unnaturally high amount of uncovered bases. This indicates that the genome is not present in the dataset.

quality check (see Figure 2.3) provides useful information to the experimentator, as it suggests that *S. flexneri* may not be present in the dataset. This contradicts GASiCs estimates and should encourage the experimentator to check manually whether a reference genome is missing.

As a last experiment, we tested GASiCs performance when low quality reference genomes are presented to the method. Therefore, we repeated the original experiment with *E. coli*, STEC and *S. flexneri* and replaced the *E. coli* reference genome by a set of contigs assembled from the original *E. coli* reads. For the assembly, we used Mira (Chevreux et al., 1999) using default settings for IonTorrent reads. The assembly yielded 711 contigs in total, 154 of which were longer than 1,000 bp, summing up to 4.4 Mbp (compare: *E. coli* has 4.6Mbp). We used the 154 contig sequences longer than 1,000 bp, which covered about 95% of the *E. coli* genome. Since the Mason read simulator is able to simulate reads from a set of contigs and Bowtie can map reads to multiple sequences at once, the GASiC method is able to deal with loose collections of reference sequences. The GASiC estimates for the *E. coli* "draft genome" and the STEC and *S. flexneri* genomes are shown in Table A.6. The results show that the GASiC estimates for this setup are comparable to the case with complete reference genomes.

2.3.3. Viral RNA quantification

To demonstrate a possible application of GASiC beyond metagenomics, we analyzed RNA data from a study on viral recombination in *Apis mellifera*, the honey bee. Moore et al. (2011) analyzed viral RNA of 40 honeybee pupae, many of them infested by varroa destructor mites. The viral RNA was purified and the corresponding cDNA was sequenced on an Illumina GAII. The raw data contains 16.8 million paired-end reads with length 2×72 bp. The authors identified novel recombinations of the two *Picornavirales*, Deformed Wing Virus (DWV) and Varroa Destructor Virus-1 (VDV-1): VDV-1_{DVD} and VDV-1_{VVD}. All genomes are available from NCBI, accessions are provided in Table A.1. We estimated the similarity of the original sequences and the recombinants via whole genome alignment with Geneious v. 5.5.0 (beta) and found that the four viral genomes show a high sequence similarity, ranging from 84% to 96% identical bases (see Table 2.3).

We estimated the viral abundances with GASiC for both the original and the recombinant genomes in the published NGS dataset used for identifying the recombinant genomes. For similarity estimation, we simulated reads with Mason. Due to the short length of the viral genomes, 10,000 simulated reads per virus were enough to cover the whole sequence. The exact command for the simulation was `mason illumina -N 10000 -hi 0 -hs 0 -n 72 -sq -o [reads] [reference]`. As for the *E. coli* dataset, we used Bowtie to align the reads to the reference genomes. To reduce the computational effort, we only used the first read of every read pair and discarded the second one. In the original dataset, both reads are concatenated as one contiguous sequence; to only align the first read, we configured Bowtie to ignore the last 72 bp of each read via `-3 72`. The complete command was `bowtie -S -p 4 -q -3 72 [index] [reads] > [samfile]`. To align the simulated reads for the calculation of the distance matrix, we simply omitted the `-3 72` parameter. The total runtime of GASiC (incl. alignment) was 41 minutes on one CPU. The peak RAM consumption was 1.3 GB.

This data posed a particularly difficult problem, since the reference sequences showed up to 96% sequence identity (see Table 2.3). Furthermore, since the considered species are RNA viruses, the reference sequences are only representatives for *quasispecies clouds* of highly similar sequences (Fishman and Branch, 2009). As the divergence of a quasispecies cloud is lower than the distance between the considered reference sequences ($< 4\%$), GASiC should be able to correct for the given similarities, although we expect the results to be not as precise as in other experiments.

GASiCs estimates are shown in Figure 2.4 and Table 2.4, demonstrating that the high sequence similarities caused strong corrections to the number of matching reads. After correction, VDV-1_{DVD} was estimated as the most abundant virus while very low abundances were estimated for VDV-1. The high p-value (p=0.53) suggests that VDV-1 is not present in the dataset. Furthermore, we see that recruiting only

Table 2.3.: Bee virus similarities. The pairwise sequence similarities were obtained by pairwise sequence alignment of the reference genome sequences with Geneious and represent the fraction of the genomic positions shared by both organisms. Similarities above 0.9 mean almost identical genomic sequences and are therefore hard to distinguish.

	VDV-1	**DWV**	**VDV-1$_{\mathrm{DVD}}$**	**VDV-1$_{\mathrm{VVD}}$**
VDV-1	1	0.841	0.916	0.924
DWV	0.841	1	0.911	0.906
VDV-1$_{\mathrm{DVD}}$	0.916	0.911	1	0.964
VDV-1$_{\mathrm{VVD}}$	0.924	0.906	0.964	1

unique matches to estimate abundances would be misleading in this case, suggesting DWV as most abundant virus. We compared our estimates to the qRT-PCR results reported by Moore *et al.*, although they used different bee pupae for qRT-PCR than for sequencing. Yet, the results should be comparable since all pupae were collected from the same apiary. For comparison, we used the data reported in Table 1 in Moore et al. (2011). Under the assumption that the virus levels are comparable for each bee, we calculated the relative virus levels for each bee individually and then averaged over all 25 bees. Similar to our findings, Moore *et al.* also found no evidence for VDV-1 and measured significant levels of VDV-1$_{\mathrm{DVD}}$ in all examined 25 bee pupae. DWV was found in 23 of 25 pupae, but at lower levels than VDV-1$_{\mathrm{DVD}}$, and VDV-1$_{\mathrm{VVD}}$ was found in 15 of 25 pupae. A direct quantitative comparison with our estimates is not possible due to the differing biological samples and due to our estimates possibly being distorted by the quasispecies cloud nature of the viral RNA. Nevertheless, the virus levels obtained by Moore et al. coincide with the abundance estimates calculated by GASiC.

Table 2.4.: Results on viral metagenome. We estimated the abundances of the highly similar bee viruses DWV, VDV-1, VDV-1$_{\mathrm{DVD}}$, and VDV-1$_{\mathrm{VVD}}$ in the viral RNA dataset acquired by Moore et al. (2011) with GASiC and GRAMMy. The results were compared to the qRT-PCR levels reported in the original paper. Here, we averaged the qRT-PCR levels of multiple bees to make the levels comparable to the abundance estimates derived from the sequencing data.

	All reads	**Unique**	**GASiC**	**p-value**	**GRAMMy**	**qRT-PCR**
DWV	6,294,759	1,344,388	0.217	0	0.338	0.090
VDV-1	5,511,043	2,302	0.011	0.533	0.106	0
VDV-1$_{\mathrm{DVD}}$	12,564,082	988,001	0.636	0	0.279	0.751
VDV-1$_{\mathrm{VVD}}$	10,479,047	241,748	0.137	0	0.277	0.159

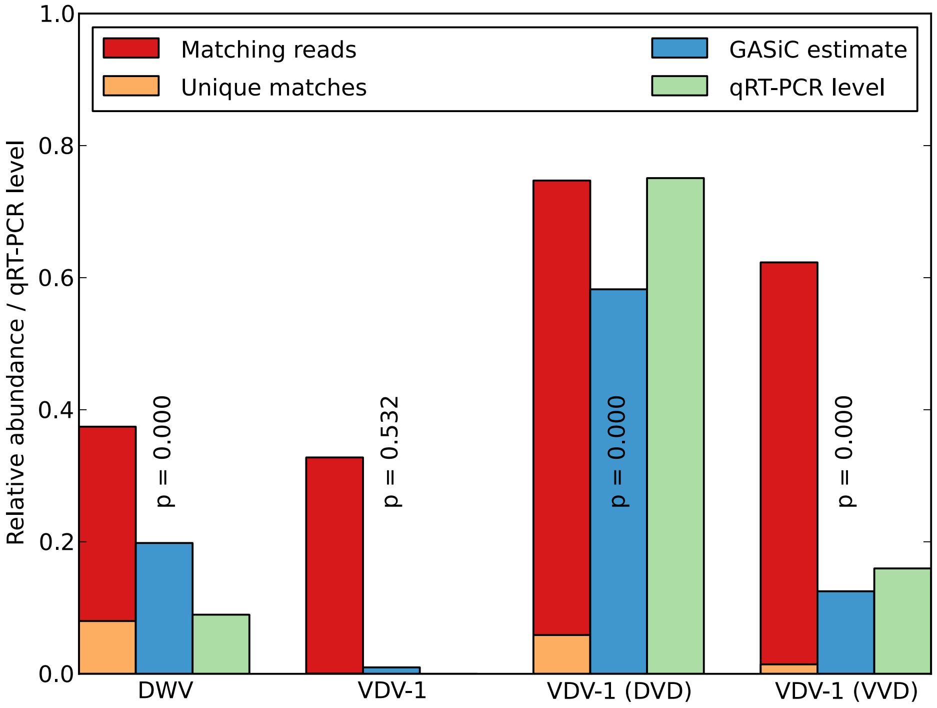


Figure 2.4.: Estimation of viral abundances based on NGS and qRT-PCR. GASiC estimated the abundances of the highly similar bee viruses DWV, VDV-1, VDV-1_{DVD}, and VDV-1_{VVD} in the viral RNA dataset acquired by Moore et al. (2011). The abundances are displayed in relation to the total number of reads. GASiCs estimates coincide with the qRT-PCR quantification in the original paper: VDV-1_{DVD} was estimated as the most abundant virus and VDV-1 was correctly identified as not present in the dataset. The displayed relative qRT-PCR levels were calculated as described in this chapter. Interestingly, only considering the unique reads would have yielded misleading estimates (DWV as most abundant) in this experiment due to the high reference similarities.

Furthermore, we estimated the viral abundances with GRAMMy to compare both tools on data with highly similar reference genomes. To this end, we aligned the first read of each pair (as for GASiC) to all four reference genomes using BLAT with default settings. The results were then passed to the GRAMMy pipeline to run the EM estimation of the abundances. The total runtime of the GRAMMy pipeline was 133 minutes on one CPU, the peak RAM consumption was 7.5 GB. We report GRAMMy's and GASiCs estimates jointly in Table 2.4. We observe substantial differences between the GRAMMy estimates and the qRT-PCR/GASiC estimates. VDV-1 could not be found in the PCR experiment and was estimated to insignificantly low abundances by GASiC, yet, GRAMMy estimates $(10.6 \pm 0.3)\%$ abundance for VDV-1. GRAMMy estimates DWV as most abundant virus, whereas the other methods identify VDV-1_{DVD} as most abundant and only observe relatively low abundances for DWV. The both recombinants, having very high similarity, were estimated by GRAMMy to about equal abundances of 27%.

2.4. Discussion of results

Our experiments demonstrate GASiCs wide range of applicability in species quantification tasks. The FAMeS benchmark dataset consists of very few but long reads, thus only a very small number of reads is available for each reference genome. While the long reads are ideal for metagenomic assembly and are thus frequently used for metagenomic analyses, the low number of reads encumbers quantification and thus challenges the algorithms. We demonstrated that GASiC greatly outperforms all current competing algorithms on the FAMeS benchmark dataset. On the other hand, we demonstrated in the *E. coli*/STEC experiment that GASiC handles mixtures of short read (80 bp) datasets of highly similar species better than GRAMMy, the best competing algorithm, and provides reliable tests for the presence of a species in the dataset. Also the different data sources did not challenge GASiC: while the aforementioned two datasets are bacterial DNA sequences, the bees dataset from the last experiment contains viral RNA reads. Also the extremely high sequence similarity (up to 96% nucleotide identity) of the viral reference sequences did not challenge GASiC.

This generality is mainly due to the fact that GASiC is independent from the underlying alignment algorithm: genomic similarities are estimated by aligning simulated reads to the reference genomes using the very same alignment tool and settings as for aligning the metagenomic reads. Thereby, tool characteristics are automatically canceled out.

Furthermore, GASiC is independent from any taxonomic information, genome annotation, or marker genes. Thus, GASiC is not restricted to the bacterial or viral domain only, but can be applied to sequences of any source, as long as reference sequences are available. This makes GASiC particularly appealing for metagenomic analyses, where large fractions of the analyzed community may be uncategorized or a mixture of viral and bacterial sequences may be present.

We demonstrated that the common practice to only consider uniquely matching reads for abundance estimation can be heavily misleading. The high genomic similarity of the two bee viruses VDV-1_{VVD} and VDV-1_{DVD} yields relatively low numbers of unique reads for both of them, although VDV-1_{DVD} was the most abundant genome in the dataset.

One obvious drawback of GASiC is its need for reference sequences. Especially in complex metagenomic datasets, typically not all constituents are sequenced or even known. We identified four typical scenarios when GASiC can be applied: *i)* when the metagenomic community is well known from previous studies and comprehensive reference databases are available. This can be the case in metagenomic time series experiments, where the same community is sequenced repeatedly to observe temporal changes in the relative abundances of species. *ii)* GASiC can be used to identify genomes present in a metagenomic dataset, when the community structure is not precisely known, but exhaustive databases of reference sequences are available. We demonstrated that GASiC still provides reliable estimates when more genome sequences are added to the reference set; this is particularly interesting for diagnostic settings of well specified organisms and also for future applications since the number of available reference genomes increases rapidly. *iii)* GASiC can be applied when the scope of the study is to estimate abundances for a well-known closed subset of sequences, i.e. a set of sequences that has a sufficiently high genomic distance to all other genomes, such that the probability of falsely aligning reads to sequences of the closed subset is very low. We observed that unknown sequence reads with low similarity do not diminish GASiCs accuracy. These closed subsets can be obtained, for example, by clustering sequences by similarity or using tools such as MEGAN to carefully pick references by hand. And, *iv)*, GASiC is applicable in experiments with high sequencing depth or low community complexity, such that a preceding assembly step could directly deliver the references for quantification (Iverson et al., 2012). We demonstrated this (see Table A.6) by replacing the *E. coli* genome in the mixed *E. coli*/STEC dataset experiment by contigs readily assembled from *E. coli* reads and obtained GASiC estimates similar to using the *E. coli* reference.

We see difficulties for the application of GASiC when the reference set composition is insufficient; e.g., when the dataset contains reads of a novel species that is highly similar to an existing species or a known species obtained novel genomic fragments via gene transfer (STEC) or recombination (DWV/VDV-1). We also expect problems in precisely estimating abundances in small datasets containing high numbers of species, which is often the case for traditional Sanger sequencing experiments. However, the quality check step in GASiC displays warnings when the risk of misinterpretation of results arises and thus serves as an automated indicator of these situations.

Scenario *iv)* is particularly interesting as it is applicable when a metagenomic community is barely known, which is the case in many metagenomic studies. Yet, a complete assembly of all constituents of the sample is unrealistic, even in the case of a community with low complexity. However, GASiC is able to estimate abundances of single assembled contigs or groups of contigs when algorithmically treated as a discrete ”species”. For example, rough estimates of species (groups of contigs) abundances or abundances of single genes (encoded on the contigs) can be obtained in this way. This concept can also be applied to fragments of genomes, as for example to fragmented RNA viruses or functional units in the genome. As observed in the viral RNA quantification experiment, quantifying complete genomes may be prone to errors when recombination occurred. Quantification of fragments may lead to more meaningful results if the recombinant genomes are not known. Nevertheless, it is not directly possible to detect recombination events with GASiC, although highly differing abundance estimates of fragments may be a sign for recombination.

We conclude that GASiC is a highly accurate and robust tool for genome abundance estimation and detection on the species level in metagenomic datasets. The similarities of reference genomes, being the main source of ambiguities in most metagenomic methods, are used directly to correct observed abundances. No prior information is needed for the analysis apart from the reference genomes, making GASiC suitable for a broad range of applications. GASiC reduces quantitative error by as much as 60% over the best existing approaches for complex mixtures and quantitatively distinguishes even highly related organisms with more than 95% sequence similarity. We obtained accurate estimates on both viral and bacterial datasets from different sequencing platforms. Furthermore, we observed that GASiCs abundance estimates conform with virus levels obtained with qRT-PCR. This indicates that additional PCR based quantification may become unnecessary if NGS data is available.

2.5. Application to metaproteomics: Pipasic

In contrast to classical proteomic approaches, metaproteomics and environmental proteomics aim at deciphering the interplay of different organisms contained within an environmental sample (Muth et al., 2013). In short, one could say that metaproteomics relates to proteomics like metagenomics relates to genomics. While many goals and strategies correlate for metagenomic and metaproteomic approaches, several distinct differences are noteworthy. In metaproteomic approaches, expression levels are analyzed and thus quantitative measures differ even for proteins from a single organism. This can be highly insightful for functional analyses (Muth et al., 2013), but poses an additional challenge for data analysis. Further, the ambiguity of peptide identifications (Seifert et al., 2013; Hettich et al., 2013; Muth et al., 2013) is even more pronounced than the problem of ambiguous read mappings in metagenomics, since peptides are commonly shorter than sequencing reads and thereby less likely to be unique. Additionally, one spectrum can not only match to several peptides occurring in multiple proteins of the same organism, but may match to proteins in different organisms. This is particularly common for closely related organisms with sufficient sequence similarity and for well-conserved proteins. Consequently, this problem hinders the correct identification and quantification of the species present in a sample.

However, the basic problem of assigning spectra to reference proteomes and subsequently using this information for quantification correlates well to the problem we encountered in metagenomics. There, we developed the GASiC algorithm, which estimates species abundances even for highly related species with high fractions of shared reads. Our idea was to transfer the concept behind GASiC to metaproteomics.

Therefore, we developed Pipasic (peptide intensity-weighted proteome abundance similarity correction) as a tool for metaproteomic species detection and abundance estimation. Pipasic uses all peptide identifications available, not only unique peptides, and generates a strain-specific, quantitative output without resorting to a lower phylogenetic resolution. Further, Pipasic avoids potential bias by estimating the similarity only for expressed proteins, which may correlate with the state of conservation of proteins.

2.5.1. Pipasic method

The Pipasic method workflow is outlined in Figure 2.5. The workflow is similar to the GASiC workflow (see Figure 2.1), we will briefly discuss the differences here.

The analog to the read mapping step in GASiC is the **peptide identification** step. Here, the peptide spectra in the metaproteomic dataset are searched separately against each reference proteome using an appropriate database search algorithm and a standard decoy database strategy is employed to ensure specificity (Bradshaw et al., 2006). The choice of peptide search tool is not restricted; we tested searches with InsPecT (Tanner et al., 2005), Sequest/Tide (Diament and Noble, 2011) and BICEPS (Renard et al., 2012).

A **similarity estimation** strategy as in GASiC is not feasible for Pipasic. The available tools for spectra simulation and identification are very slow compared to their genomic counterparts. Therefore, we employed a weighted string comparison approach, which approximates the probability of assigning a spectrum originating from one organism to another one and is able to incorporate differing expression levels.

We regard the reference proteomes as sets of protein sequences, i.e. sets of strings.

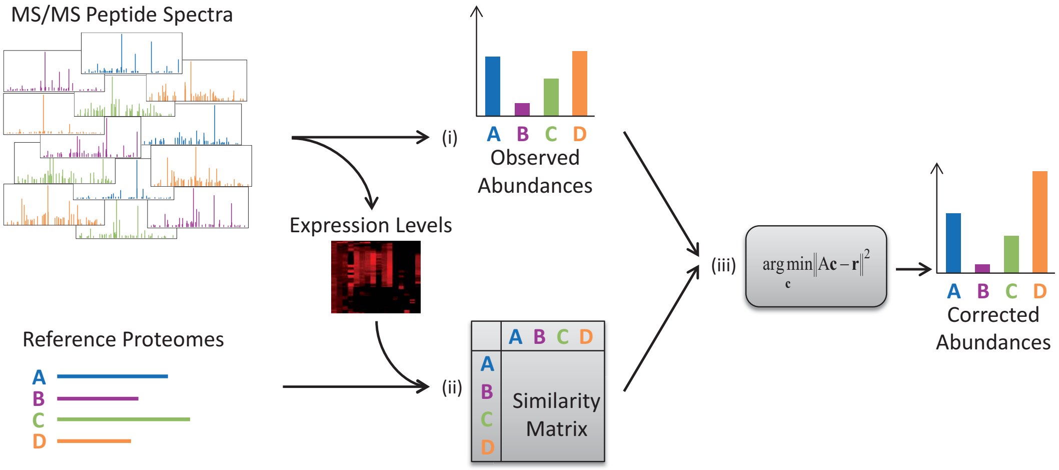


Figure 2.5.: Pipasic method overview. Pipasic involves three main steps: (i) Peptide identification; here metaproteomic peptide spectra are identified by a database search. The number of matches to the proteomes are the observed abundances. (ii) Similarity estimation; the similarities between the reference proteomes are calculated and stored in a similarity matrix. This incorporates the adjustment to only regard expressed proteins and to weight them according to their expression level. (iii) Similarity correction; the observed abundances are corrected using the similarity matrix yielding corrected abundances.

Since the proteins in the experiment are typically digested into tryptic peptides before the spectra are acquired, we perform an *in-silico* digestion of the reference proteomes, yielding a list of short peptide strings for each proteome. In order to account for the amino acids with indistinguishable masses, we replace all occurrences of I by L and Q by K.

A reference proteome often contains proteins that were not expressed or measured in the experiment. We reflect these particular effects in the similarity estimation by introducing weights for all peptides. The weight w_p for the tryptic peptide p in proteome $\mathcal{P}_i$ (total: N reference proteomes) is calculated as follows:

1. Assign a preliminary weight $\tilde{w}_p$ to each peptide p: add $\frac{1}{N_p}$ to $\tilde{w}_p$ for each spectrum that was identified with p, where N_p is the number of peptides the spectrum can be identified with.

2. For each protein $P \in \mathcal{P}_i$, set the peptide weights $\hat{w}_p$, $p \in P$, to the average preliminary peptide weight: $\hat{w}_p = \frac{\sum_{q \in P} \tilde{w}_q}{|P|}$.

3. Normalize the sum of all weights to one: $w_p = \frac{\hat{w}_p}{\sum_{q \in \mathcal{P}_i} \hat{w}_q}$.

The matrix entry a_{ij} of the weighted similarity matrix A is calculated by summing over the weights of the peptides in $\mathcal{P}_j$ that were found in $\mathcal{P}_i$:

$$a_{ij} = \sum_{p \in \mathcal{P}_j} w_p \text{ if } p \in \mathcal{P}_i.$$

Once the similarity matrix coefficients a_{ij} are calculated, the true proteome abundances c_i can be estimated from the relative observed abundance r_i of proteome $\mathcal{P}_i$ with the same **similarity correction** strategy as in GASiC, see Section 2.2.1 for the mathematical details.

2.5.2. Pipasic results

We conducted two experiments to evaluate the impact of the various algorithmic steps. In the first experiment we evaluated Pipasic using ground truth data and demonstrate that Pipasic provides more accurate results with regard to identification and quantification than the analysis with MEGAN and based on unique peptides. We mixed two pure proteomic MS datasets of highly similar proteomes in 11 predefined ratios (similar to Section 2.3.2) and challenged the tools to correctly estimate the fraction of each proteome in the dataset. We used two different but closely related cowpox virus strains with available reference proteomes: *Krefeld* (Kre) and *Brighton Red* (BR) (Doellinger et al., 2014).

We processed the 11 datasets with Pipasic using InsPecT for peptide identification. For the analysis with MEGAN, we searched the peptide sequences with BLASTP in the reference proteomes. Figure 2.6 shows the abundance estimates of Pipasic with and without expression correction (see Penzlin et al. (2014) for more details). Figure 2.7 (a) shows the output of MEGAN for the dataset containing 10% Krefeld and 90% Brighton Red spectra. Figure 2.7 (b) shows the estimated abundances of both strains for each dataset. The figures show that Pipasic provides much more accurate abundance estimates than the MEGAN-based approach. Even in the case of pure datasets there is still a significant number of spectra matching uniquely to the absent species (about 15%). Here, the Pipasic estimates (solid lines) are much closer to the ground truth.

In the second experiment, we demonstrated the applicability of Pipasic to a more complex Acid Mine Drainage (AMD) metaproteome (Denef et al., 2010). AMD biofilms are bacterial communities in a highly acidic environment and their composition is well understood, however, AMD communities are not as complex as other microbial communities.

We downloaded the freely available metaproteomic spectra and the corresponding protein database. We manually divided the database into six reference proteomes: *Leptospirillum* group II and III (Lepto2 and Lepto3), *Ferroplasma acidarmanus*

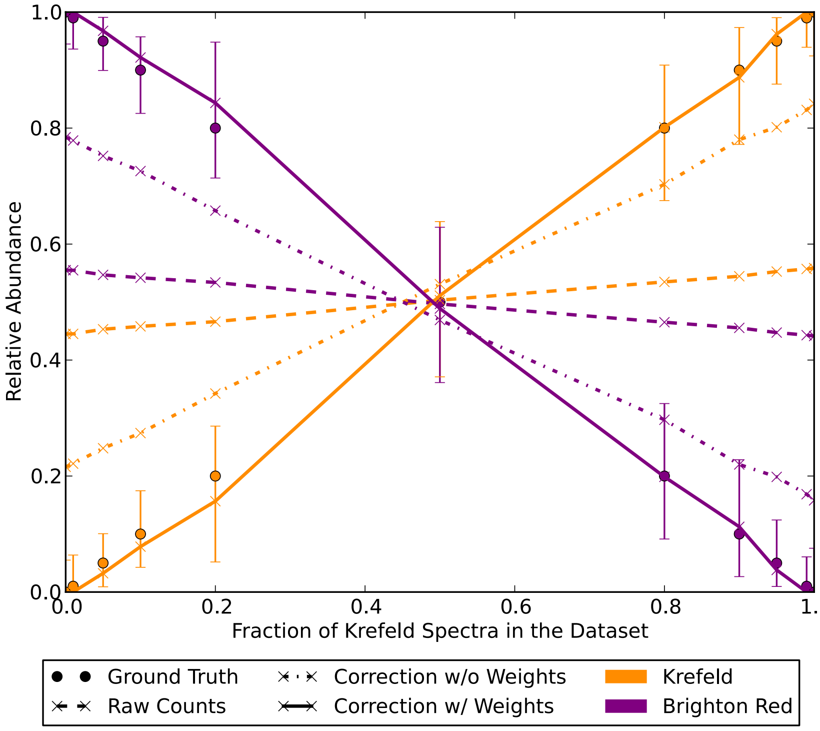


Figure 2.6.: Effect of Pipasic correction: the relative abundances of 11 mixed cowpox virus Krefeld/Brighton Red datasets were corrected with Pipasic without and with expression correction. The observed abundances (dashed lines) are insufficient estimates for the true abundances (solid dots): in the extreme cases of pure Krefeld or Brighton Red datasets the absent virus still receives 45% abundance. The unweighted correction (dash-dotted line) improves on this, but best results are obtained using the expression-weighted similarity matrices (solid line). The error bars indicate the 95% confidence interval after 100-fold bootstrapping.

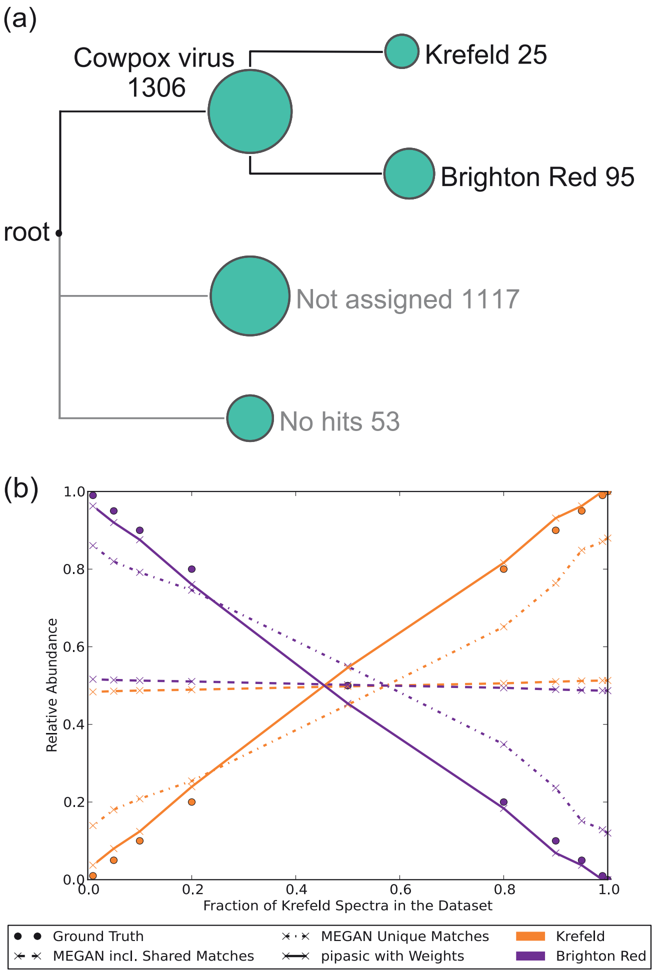


Figure 2.7.: Comparison of Pipasic and MEGAN on the cowpox virus datasets. (a) MEGAN output for the 10% Krefeld / 90% Brighton Red dataset. The size of the circles is log-proportional to the number of assigned spectra, visualizing that the majority of spectra was assigned to the higher level Orthopoxvirus node. The leaves, representing Krefeld and Brighton Red, obtained relatively few spectra: Only 8.4% of all matches were unique. (b) Comparison of MEGAN and Pipasic on all 11 mixed cowpox virus datasets. For MEGAN, the number of unique and shared matches (dashed lines) shows almost no difference between the two proteomes since the number of unique matches is very low. The number of unique matches (dash-dotted lines) provides abundances closer to the ground truth, but Pipasic (solid lines) yields the best estimates.

Type I and II (Fer1 and Fer2), G-plasma and others, such as contaminants and unassigned archaea and bacteria. Then we searched the spectra in the reference proteomes with Tide (Diament and Noble, 2011) and counted the number of matching spectra. We applied Pipasic on the results to obtain the corrected abundance estimates.

Table 2.5.: Acid Mine Drainage dataset abundance estimation. The peptide spectrum matches (PSM) were counted for each proteome and subsequently corrected with Pipasic using a weighted similarity matrix. The results show a strong relative correction for the highly similar Fer 1/2 and only a small relative correction for Lepto 2/3 and G-Plasma.

	Fer1	**Fer2**	**Lepto2**	**Lepto3**	**G-Plasma**	**Other**
Observed PSMs	195	189	4,470	2,014	692	87
Pipasic Estimate	111	88	4,281	1,655	671	32
Rel. Correction	43.1%	53.4%	4.2%	17.8%	3.0%	63.2%

The results of this experiment are shown in Table 2.5. Here, the effect of the correction is not as pronounced as in the previous experiment due to the relatively low similarity values (maximum 0.21 compared to 0.92, see Figure 2.8). Lepto3 receives the strongest absolute correction (-359 PSMs) due to the protein sequence similarities with Lepto2, which receives very low relative correction. Fer1 and Fer2 have the highest proteome similarities in this experiment (0.21/0.19) and their abundances were reduced in sum by 48.3%. G-plasma has the least similarity to the other proteomes (less than 0.04) and therefore receives only very little correction by 3%.

This demonstrates that Pipasic can handle real metaproteomic data and the calculated estimates are in agreement with the expectations. The two main groups Fer1/2 and Lepto2/3 receive abundance corrections within each group, but not between the groups. This is noteworthy since we did not require any prior information other than the reference proteomes and shows that the similarity estimates reflect the nature of the microbial community.

2.5.3. Discussion

The experiments indicate that Pipasic can be used for reliably identifying and quantifying the contributions of organisms and functional units even in cases when – as in the cowpox virus data experiment – 92% of all expressed tryptic peptides are identical. In particular, Pipasic allows having a phylogenetic resolution down to the strain level, which is inherently not feasible for LCA approaches for highly related species. This is also clearly visible in the comparison with MEGAN on the cowpox virus strain data (see Figure 2.7).

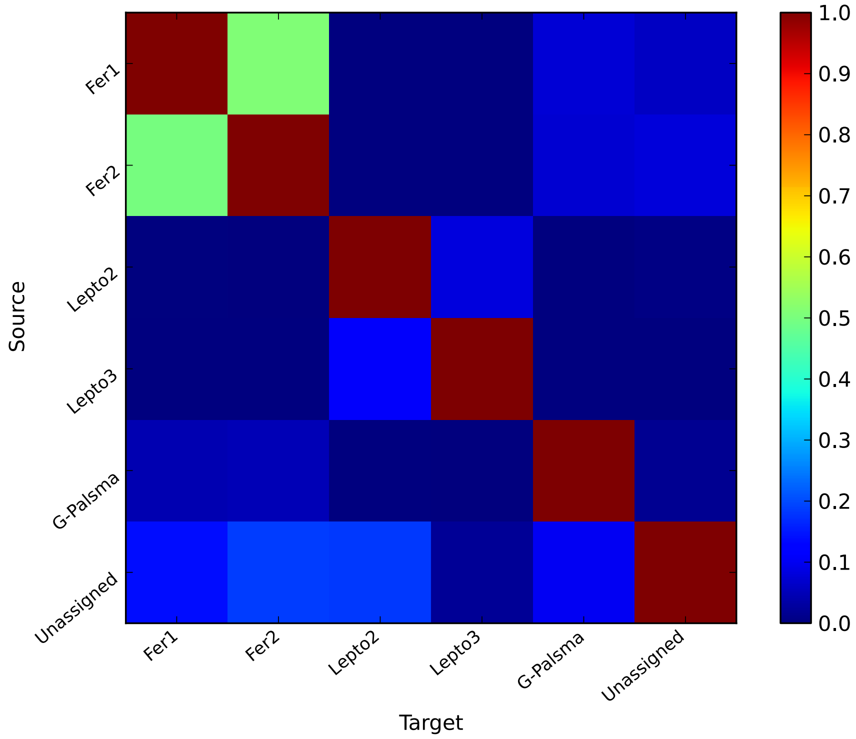


Figure 2.8.: Pipasic similarity matrix with data weighting for the AMD experiment. The matrix entries encode the probability that a peptide in a source proteome can be found in a target proteome, modulated by the metaproteomic data. Here, we see that the intra-group matrix coefficients for the Fer and Lepto group are greater than the inter-group coefficients. This means in practice that Pipasic corrects abundances within but not between the two groups. It is noteworthy that the matrix coefficients can be asymmetric, which has the effect that abundance can be shifted from one proteome to another rather than correcting both proteomes equally.

Given its reliability, Pipasic is preferable to approaches relying solely on the analysis of unique peptides. Figure 2.7 indicates the risk of analyzing unique peptides for highly related strains. Even though the overall number of identified peptides per species is above 1,000, the number of unique peptides remains low due to the sequence similarity. Thus, only few peptide identifications out of a thousand decide on the identification of a species when relying on unique peptides. Even in cases when the ground truth contains 0% spectra of the Krefeld strain, MEGAN finds 17 unique peptides; this effect was also observed when using the more conservative OMSSA (Geer et al., 2004) search engine instead of InsPecT. These may incorrectly be interpreted as proof of the presence of the Krefeld strain. However, given that the original peptide identification search was conducted at a 5% false discovery rate and given the large number of spectra searched, these identifications are indeed incorrect. Since Pipasic leverages the computed similarity and the shared peptides into the analysis, it is less at risk to overvalue these incorrect identifications and correctly reduces the presence of the Krefeld strain in this example down to a level where it cannot be distinguished from a 0% presence.

Pipasic computes its similarity correction adjusted to the expression level of proteins. This step highlights a major difference between metaproteomics and metagenomics: although the main idea of the metagenomic method could be applied in metaproteomics, the method itself must be tuned to the underlying difference in the biological data.

One general difficulty for Pipasic is the dependence on the completeness of the provided reference proteomes. Thus, any quantification or identification by Pipasic is also at risk of only reflecting the available reference proteomes. This problem is common to both the Pipasic and the GASiC approach, but with the advent of current MS technologies, the reference databases will grow continuously and the problem will be of less relevance in the future.

3. Fitting mixtures of discrete distributions to genome coverage depth profiles

When NGS reads are mapped to a reference genome, the genome coverage depth contains valuable information about the dataset, the reference genome itself, and the mapping process, and the coverage depth is easily accessible. Therefore, it is frequently consulted in bioinformatics analyses to improve decisions in algorithms or to provide meaningful information to the user. For instance, experimental design methods (Löwer et al., 2012) guide the experimentalist to achieve a specific average sequencing depth. After sequencing, the obtained reads can be mapped to a reference genome. Quality control tools (García-Alcalde et al., 2012; DeLuca et al., 2012) analyze the mapping data and report measures such as coverage depth information, mapping quality, or error rate to the user. For example, Qualimap (García-Alcalde et al., 2012) visualizes the depth profile and the coverage depth over the whole genome together with the GC-content, which allows detecting biases in the sequencing process. If no reference genome is available, the reads can be assembled to complete genomes or at least longer contiguous sequences (contigs). The latter is nowadays possible for metagenomic datasets. The assembler MetaVelvet (Namiki et al., 2012) uses the coverage depth information in the de Bruijn graph to connect contigs of similar depth, as they are more likely to belong to the same organism. In addition to these examples, local coverage depth information is also used for detecting copy number alterations in genomes, e.g. (Miller et al., 2011).

Despite these versatile applications of genome coverage depth, a vast amount of information commonly remains unused. Most current methods either use the average coverage depth over a certain sequence (Löwer et al., 2012; DeLuca et al., 2012) or describe the coverage depth profile using a single probability distribution such as the negative binomial (Miller et al., 2011) or gamma (Hooper et al., 2010) distribution. Yet, to the best of our knowledge, more complex models such as mixtures of distributions are not employed to fit genome coverage depth profiles. We suggest that more complex models can improve current methods and can open doors for new analysis strategies.

We see an application of complex coverage depth distribution models in metagenomics, where reference-based methods have become increasingly popular with the advent of high-throughput sequencing technologies (Mande et al., 2012). However, there are two major problems with reference genomes: First, the process of assembling and finishing reference genomes is time consuming and cumbersome and many reference genomes remain unfinished in the draft stage with varying qualities depending on the used sequencing technologies (Mavromatis et al., 2012). Draft genomes are typically a set of assembled contigs, where many contigs may be erroneous or, if assembled from metagenomic data, belong to different organisms. The second problem is of biological nature; evolution in the microbial world proceeds at high pace due to short replication times and new subtypes or even species emerge perpetually. This causes different microbial species to have high genomic similarities. Therefore, the coverage depth is generally far from homogeneous when mapping metagenomic reads to a reference genome; describing it with a single uni-modal distribution would not be appropriate. Here, more complex models can have the power to disentangle and quantify different contributors to the genome coverage depth.

Therefore, we developed a framework for fitting complex mixtures of probability distributions to genome coverage depth profiles. The proposed method has four steps (see Figure 3.1): After mapping a set of sequence reads to a reference genome, the mapped reads are analyzed and a genome coverage depth profile is constructed. Then, a mixture model of customized probability distributions is fitted to the profile using an iterative procedure derived from the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), seeking to identify and distinguish different contributors in the profile. Further analysis on the fit parameters can then be used to answer questions about the reference genome and the mapping process, such as the validity of a reference genome or the occurrence of multiple related organisms in one dataset. The presented method is not a new invention in itself, but rather a combination of established statistical methods, which we demonstrate to be useful for analyzing genome coverage depth profiles. The novelty of this contribution is the composition of the mixture models, the adaptation of the EM algorithm for discrete probability distributions, and the subsequent analysis steps.

GCP: Genome coverage depth profiles When reads are mapped to a reference genome, the per-base coverage depth for each position in the reference genome is given as the number of reads covering that position. We term the *histogram* over all per-base depths the *genome coverage depth profile* (GCP). A GCP encapsulates valuable information about the relation between the reference genome and the genome(s) contained in the dataset. In the following, we solely operate on GCPs, as they provide a condensed view on the mapping of reads to a reference genome.

A GCP can take shape in various ways: First, if the reference genome matches perfectly to the reads contained in the dataset, the genome is homogeneously covered and the GCP consists of a uni-modal distribution, as depicted in Figure 3.2 (a). In

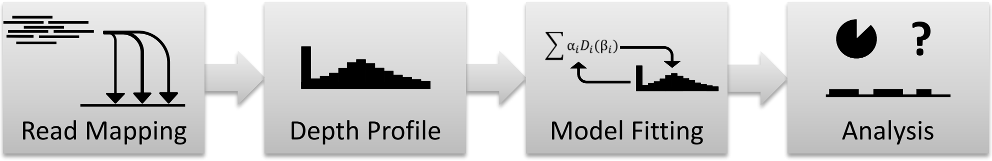


Figure 3.1.: Method overview. Starting with a set of reads mapped to a reference genome, we construct a genome coverage depth profile and fit a mixture of probability distributions to the profile. This procedure is the basis for subsequent analysis steps concerning the reference genome, the mapping process, and the read dataset.

reality, the reads and the reference genome will differ due to mutations and errors in the reads. Therefore, the differing parts of the reference genome will not be covered by reads and lead to an excess of zero-depth counts in the GCP, as shown in Figure 3.2 (b). Note that the distribution has a tail at low coverage depths, which we discuss in Section 3.3.2. As a third type (shown in Figure 3.2 (c)), a reference genome may have an overall low coverage depth as well as positions differing from the reads. Then, positions with zero depth may be caused either by a locally differing sequence or the position was not covered by chance due to the statistical fluctuations in the sequencing depth. In addition to the three simple types, a GCP can also be a more complex combination of coverage depth distributions, as shown in Figure 3.2 (d). In this example, the dataset contained two genomes, A and B, with differing sequencing depths. Both genomes share parts with the reference genome and also have similarities among each other.

3.1. Discrete probability distributions for GCPs

In this section, we give a short overview of probability distributions that we consider relevant for describing GCPs. The simplest assumption we can make is the *random sampling* property of shotgun sequencing devices, meaning that we assume a uniform distribution of the reads over the genome. When reads are mapped to a genome, the coverage depth of each position follows a **Poisson** distribution $P(x|\lambda)$. The Poisson distribution is well-studied and has one parameter λ, which simplifies fitting observed distributions. However, the Poisson distribution is often too narrow for fitting real genome coverage depth distributions, in particular for metagenomic data. This effect is called over-dispersion and occurs frequently in biological data. A common way (Bliss and Fisher, 1953) to account for over-dispersion is to assume that the Poisson parameter λ is distributed according to a second distribution. When λ is assumed to be gamma distributed, we obtain a **negative binomial** distribution $NB(x|a, b)$,

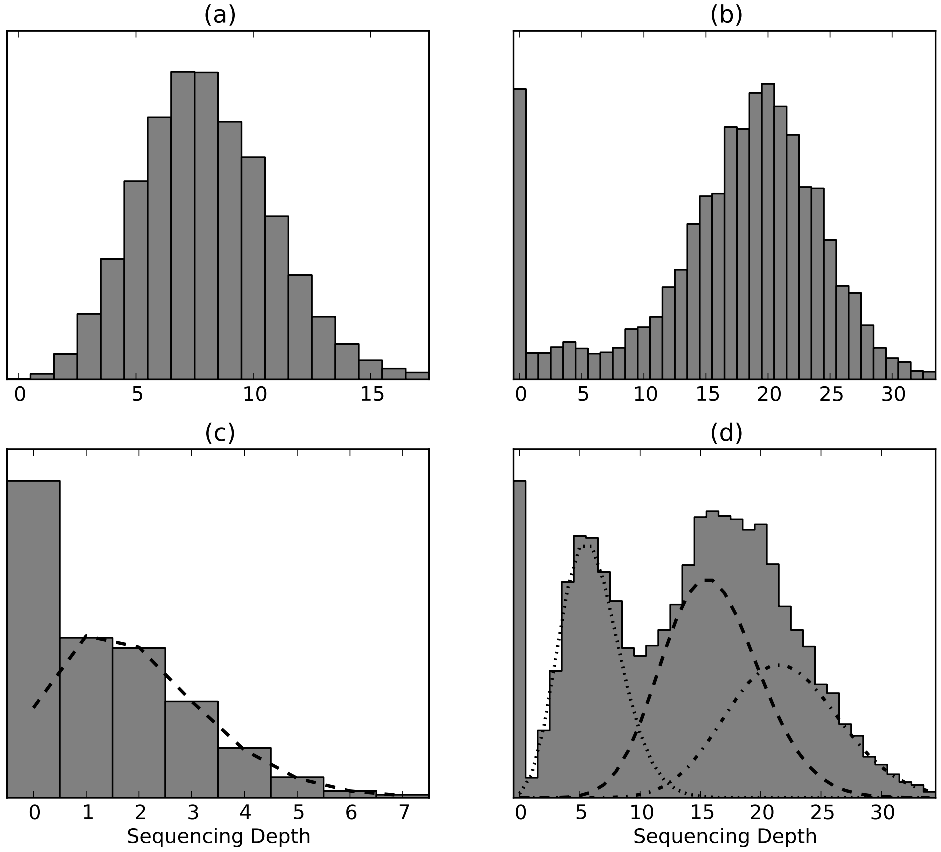


Figure 3.2.: Exemplary GCPs after mapping a set of reads to a reference genome. (a) Reads matching perfectly to reference genome. (b) Reference genome partially covered by reads; the covered areas have high sequencing depth. The areas where of the genome that had no similarity to any of the reads manifest in an excess of zeros in the GCP. (c) Reference genome partially covered by reads with low coverage depth. The dashed curve has a non-zero value for uncovered positions and thus adds to the number of positions that were not covered due to disagreement with the reads. (d) Reference genome obtaining reads from two organisms A and B with different abundances, yielding a mixture of four distributions: (i) a zero-distribution for the parts of the reference not covered by reads, (ii) the coverage depth caused by organism A (mean=6×, dotted curve), (iii) the coverage depth caused by organism B (mean=16×, dashed curve), and (iv) the coverage depth where A and B map to the same position (mean=22×, dash-dotted curve).

which has two shape parameters. Therefore, parameter estimation is harder than for the Poisson distribution. However, the negative binomial turned out to model GCPs well for both high and low coverage depths and has been used, for example, in differential expression analysis (Anders and Huber, 2010).

In Figure 3.2 (b), we observed a tail on the low-depth end of the main distribution. The magnitude of the tail depends on the fragmentation of the reference genome with mutations, as we discuss in detail in Section 3.3.2. The shape of the tail is determined by the original parent distribution, therefore the tail distribution has no own shape parameters and can rather be considered as an extensional distribution. We implemented the **Poisson tail** and the **negative binomial tail** in our framework.

Finally, we employ the **zero** distribution $z(x)$, which is useful to describe the excess of uncovered positions. The zero distribution has probability 1 at zero and 0 everywhere else. It is static as it has no shape parameters, but proves its usability in combination with other distributions. The zero-inflated Poisson and zero-inflated negative binomial are defined as mixed distributions of zero and Poisson or zero and negative binomial, respectively. These zero-inflated (ZI) distributions were used, for example, to model the number of defects in manufactured items (Lambert, 1992), but can also be applied for GCPs: the areas where the reference genome agrees with the reads in the dataset yields a coverage depth distribution according to Poisson or negative binomial, the areas with disagreement are modeled by the zero distribution.

We want to regard mixtures consisting of more than one probability distribution and write the joint distribution function as

$$f(x, \alpha|\beta) = \alpha_0 \cdot z(x) + \sum_{i=1}^{k} \alpha_i \cdot D_i(x|\beta_i), \tag{3.1}$$

where k is the number of non-zero distributions, α_i $(i = 0..k)$ are the non-negative mixing coefficients that sum up to 1 and give a weight to each distribution. D_i is either a Poisson, a negative binomial, or a tail distribution with the corresponding set of parameters β_i. $z(x)$ is the zero distribution. Fitting these mixture models to data cannot be done directly, but requires an iterative method, as described in the following section.

3.2. Fitting mixtures of discrete distributions

The following sections describe an iterative algorithm to fit mixtures of discrete probability distributions to a histogram of observed count data, such as the GCP or the RDP (see Section 3.4). However, we focus on GCPs in the following and discuss the modifications necessary to fit RDPs later.

The algorithm starts with a set of initial parameters for the distributions D_i that can either be defined by the user or estimated from the GCP. The algorithm repeatedly computes the so-called expectation step followed by an adjustment of the parameter set. With every iteration step, the algorithm improve the accuracy with which the model in Equation (3.1) describes the data, i.e. the likelihood of the data given the model with parameters after iteration $t+1$ should be greater than or equal to the likelihood after iteration t. This assumption is guaranteed if maximum likelihood estimation is used when adjusting the parameter set. In this case, the procedure is known as the EM algorithm (Dempster et al., 1977). The iteration is stopped when an accuracy threshold is reached, e.g., the change of the likelihood drops below a predefined value. The two steps of the iteration are described in the following.

3.2.1. Expectation step

Following the initialization, the expectation step estimates conditional probabilities identical to the EM algorithm: Using the current set of parameters $\beta^{(t)}$, for every coverage depth value x that occurs in the GCP and $0 \leq i \leq k$, we compute the probability that x belongs to distribution D_i, i.e.

$$\gamma_i(x) = \frac{\alpha_i^{(t)} D_i\left(x, \beta_i^{(t)}\right)}{\sum\limits_{j=1}^{k} \alpha_j^{(t)} D_j\left(x, \beta_j^{(t)}\right)} .$$

Here, the zero distribution is included as D_0 and has an empty set of parameters $\beta_0^{(t)}$. With these depth-wise probabilities and the number of occurrences n_x of each coverage depth value x, we re-estimate the mixing coefficients α. In a genome of length L, let C_l denote the coverage depth of position l, $0 \leq l \leq L-1$. Using the vector $C = (C_0, \ldots, C_{L-1})$, we set

$$\alpha^{(t+1)}(i) = \frac{1}{L} \sum_{l=0}^{L-1} \gamma_i(C_l) = \frac{1}{L} \sum_{x=0}^{\infty} \gamma_i(x) \cdot n_x .$$

The sum reduces to finitely many terms since n_x is zero for all x greater than the maximum depth observed. The second representation of the above sum drastically reduces the required computation time: By merging the n_x terms for each coverage depth value x to just one, we have to sum up much fewer terms compared to the summation over every single base.

3.2.2. Parameter estimation step

In this step, we optimize the parameter set $\beta^{(t)}$ by fitting the mixture model with respect to the previously calculated mixing coefficients $\alpha^{(t)}$. When fitting the distributions D_i, we have to decide whether to use moment-based or maximum likelihood estimates. Using the method of moments, for 1-(2-) parametric distributions we take the sample mean (the sample mean and the variance, respectively) and calculate the distribution parameters from these moments. For the Poisson distribution, the parameter λ is set to the sample mean, which is identical to the maximum likelihood estimate. The parameters a and b of the negative binomial distribution are estimated by making use of their relationship to the mean μ and variance var

$$\mu = \frac{a \cdot b}{1 - b}$$

and

$$var = \frac{a \cdot b}{(1 - b)^2} \ .$$

Maximum likelihood estimation directly selects a set of parameters $\beta^{(t)}$ that maximizes the likelihood observing the data given the set of parameters. We can utilize the maximum likelihood estimator for the Poisson (same as method of moments) and the negative binomial distribution. Yet, for the negative binomial distribution, there does not exist a closed form of the maximum likelihood estimator and requires application of, e.g., Newton's method. Due to the nature of our data there is no ultimate solution: Either method might be more suitable depending on the situation. The method of moments proves to yield similar results in most cases for the negative binomial distribution and it is numerically more robust and straightforward. The zero and the tail distributions do not require parameter estimation as the zero has no shape parameters and the tail distribution inherits the parameters from the parent distribution.

Possible applications of the introduced fitting framework are presented in the following sections.

3.3. Genome validity and genome fragmentation

3.3.1. The genome validity

In a standard scenario we have a reference genome available and a possibly unknown organism in a biological sample that was subject to genome sequencing. As depicted in Figure 3.3 (a), the unknown organism may have some parts of its genome sequence that agree with parts of the reference genome. Further, there are parts in the reference genome that do not agree with any part in the unknown genome and vice versa. Then, we define the *genome validity* score (val) as the fraction of the reference genome that has a counterpart in the unknown genome as shown in Figure 3.3 (b). In other words, the genome validity is the coverage of the reference genome if the genome sequence of the unknown genome would be aligned to the reference genome. However, the true validity of the reference genome is not directly observable, since the unknown genome is realized as a set of short reads, which may not cover all common parts of the reference genome (see Figure 3.3 (c)).

The naïve way to estimate the genome validity is to map the reads to the reference genome and measuring the fraction of the genome that was covered by reads. This estimate can be sufficiently good for high genome sequencing depths, such as sequencing depths above 10×. Here, the likelihood that a location shared between both genomes remains uncovered is negligibly small. Almost all sites on the known genome not covered by reads can be considered to be different from the unknown genome. In contrast, for very low abundances, the probability that a position is not covered by reads although it is shared by both genomes can not be neglected anymore. Let us assume, for example, a simple model where the coverage depth per position over the genome follows a Poisson distribution. While the probability of not covering a position at 10× sequencing depth is 0.0045%, it rises to 13.5% for 2× depth and 36.8% for 1× depth.

The iterative algorithm described previously can improve on the naïve approach and provide reliable estimates for much lower sequencing depths. Depending on the coverage depth distribution, we can fit a mixture of a Poisson or negative binomial distribution and a zero distribution to the GCP. The contribution of the zero distribution should then roughly correspond to the fraction of the reference genome that has no counterpart in the unknown genome(s). Therefore, we calculate the genome validity val as

$$val = 1 - \alpha_0,$$

where α_0 is the mixing coefficient of the zero distribution in the model. This calculation has a clear advantage over the naïve approach: at low sequencing depths, the probability that a position is not covered by chance (and not due to dissimilarity) is high and the naïve approach is at risk of overestimating the fraction of the genome with no counterpart. In contrast, the mixture model approach makes use of the positions with higher coverage depth to estimate the probability of obtaining not covered positions by chance and thus provides more realistic and more reliable estimates.

3.3.2. Genome fragmentation estimation

In addition to the pure similarity of the sequencing reads and the reference genome, we can also determine how fragmented the reference genome is with respect to the

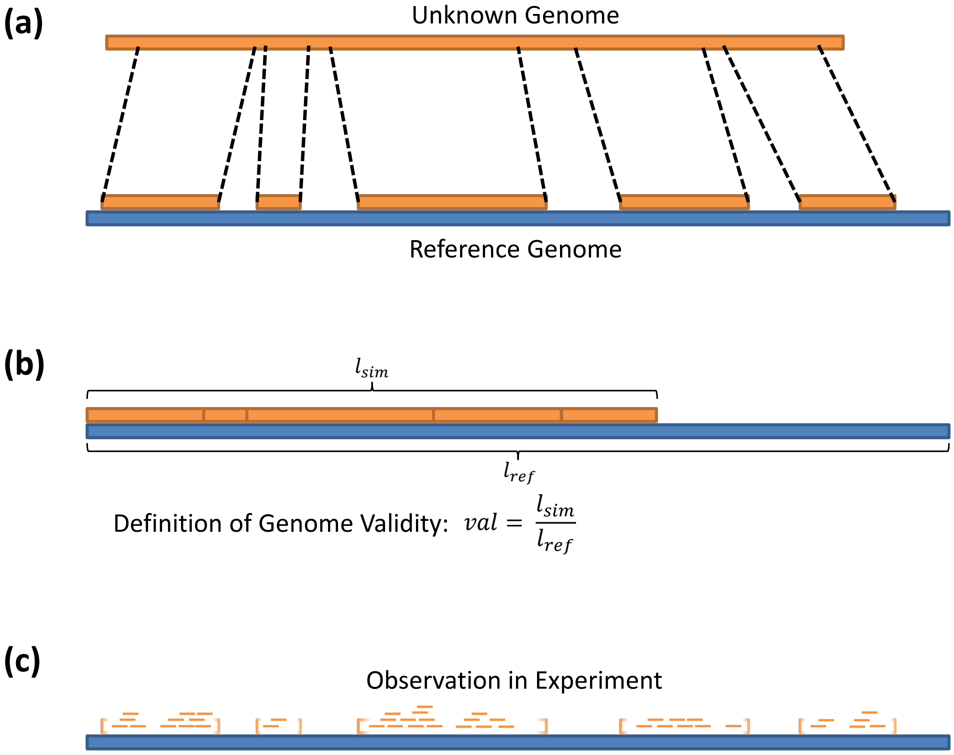

(a)
Unknown Genome
Reference Genome
(b)
l_{sim}
l_{ref}
Definition of Genome Validity: $val = \frac{l_{sim}}{l_{ref}}$
(c)
Observation in Experiment

Figure 3.3.: Definition of genome validity. (a) As a starting point, we consider an available reference genome (blue) and an unknown genome (orange), where some parts of the unknown genome agree with the reference genome. (b) The genome validity of the reference genome with respect to the unknown genome is defined as the fraction of the reference genome that agrees with the unknown genome. (c) Since the unknown genome is often realized as a set of sequencing reads obtained by sequencing a biological sample, the validity of the reference genome with respect to the unknown genome can be estimated from the sequencing reads mapped to the reference genome.

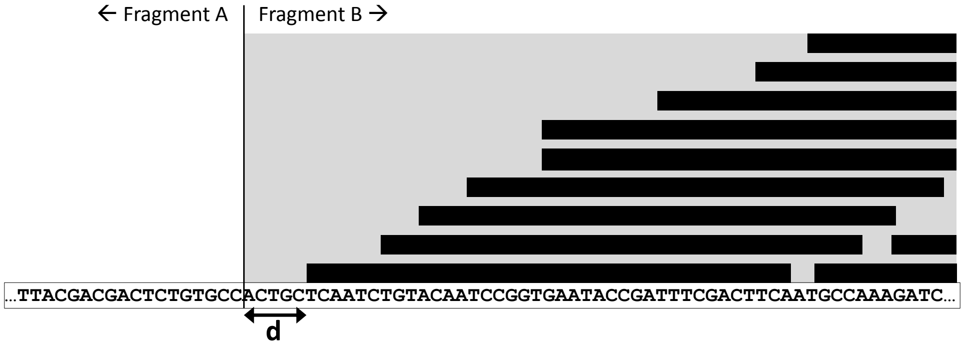

← Fragment A
Fragment B →
...TTACGACGACTCTGTGCCACTGCTCAATCTGTACAATCCGGTGAATACCGATTTCGACTTCAATGCCAAAGATC...
d

Figure 3.4.: Coverage depth effects at the border of zero depth (A) and non-zero depth (B) genome fragments. The coverage depth increases approximately linearly from the fragment border to the center of fragment B. The first read of the non-zero fragment starts behind the fragment border with a delay d. The delay is a function of the density of starting reads, i.e. the coverage depth and the read length. In highly fragmented genomes, the delayed start of the first read yields an excess of uncovered positions.

reads. With fragmentation, we mean the number of contiguous sequence fragments in the genome that conform with the reads. For example, single nucleotide polymorphisms, insertions, or deletions in the reads with respect to the reference genome can be the cause for genome fragmentation. Consider two genomes A and B, which share parts of their genomic sequences: the fraction s_{AB} of genome A can be found in genome B and a fraction s_{BA} of genome B can be found in genome A. The parts shared by both genomes typically do not form one contiguous sequence, but are fragmented by insertions or mutations. When reads of genome B are mapped to genome A, a fraction s_{AB} of genome A can be covered by reads. Under ideal conditions, we will observe a homogeneous coverage depth in the center of the shared sequence fragments but linearly decreasing coverage depth flanks on the fragment borders, since the reads do not map to areas behind the fragment border (compare Figure 3.4). At the fragment borders, the coverage depth decreases linearly from a maximum coverage depth to zero over a distance of one read length RL.

To illustrate the effect of genome fragmentation to the GCP, we simulated two genomes A and A^* by randomly generating nucleotide sequences of length 1,000,000 bp and 100,000 bp, respectively. In order to create a fragmented genome B, we chopped genomes A and A^* into $frag$ fragments and assembled a new genome

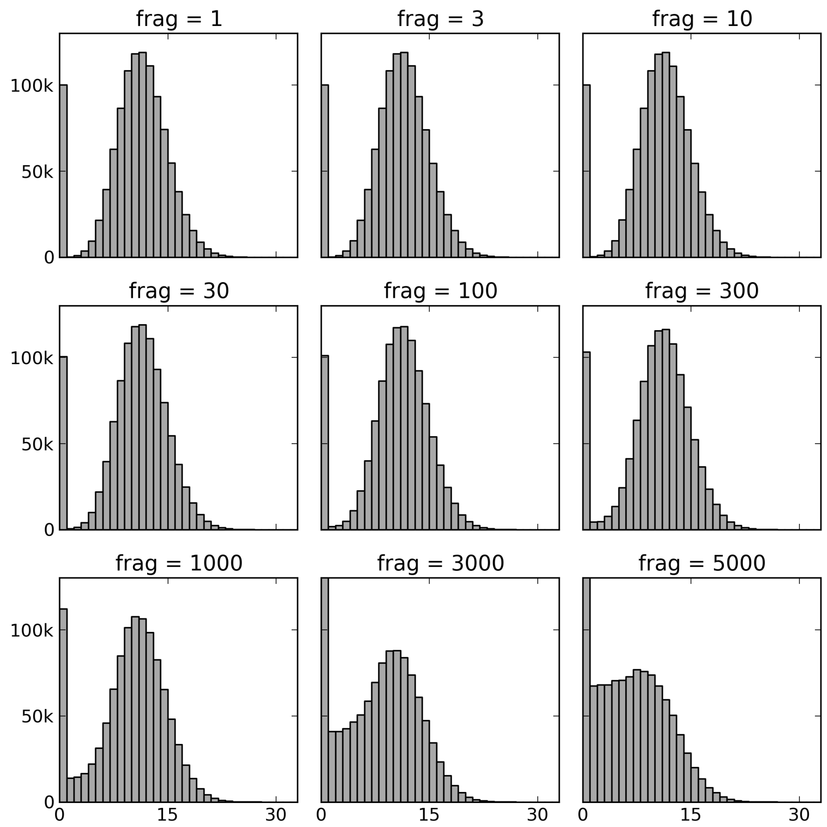


Figure 3.5.: GCPs obtained by mapping simulated reads to genomes with different degrees of fragmentation. Here, $frag = N$ denotes that the reads match to N unconnected sequence fragments in the target genome. The overall length of the sequence fragments is constant for all fragmentations. The coverage depth effects on the borders of the fragments increase as the number of fragments increases and therefore influence the distribution of values in the GCP.

by concatenating the fragments alternatingly. Note that the similarity of A to B is independent from the fragmentation. Then, we mapped 100,000 Illumina reads simulated from genome A to the fragmented genome B and created the GCP. Figure 3.5 shows the GCPs for a wide range of number of fragments. We observe a shift from higher coverage depths to lower depths in the profiles as the fragmentation increases, indicating that the border effects become increasingly important. Furthermore, the number of positions that are not covered increases with the fragmentation.

The linearly decreasing coverage depth flanks manifest in a low depth tail in the GCP, which over-disperses the parent distribution (the distribution in the center part of the shared fragments) to low coverage depths. To describe the tail distribution mathematically, let us assume that the coverage depth decreases linearly from a maximum depth c_{frag} in the center of the fragment to 1 at the border. In the GCP, we would observe a coverage depth distribution of the form

$$t(x|c_{frag}) = \begin{cases} \frac{1}{c_{frag}} & x \in [1, c_{frag}] \\ 0 & x \notin [1, c_{frag}] \end{cases} .$$

The parameter c_{frag} is the coverage depth in the center of the shared fragment and is therefore distributed according to the parent distribution D. The complete tail distribution therefore writes as

$$T(x|D) = \frac{1}{w} \sum_{c=x+1}^{\infty} D(c) \cdot t(x|c), \tag{3.2}$$

where w is a normalization constant, such that $\sum_{x=0}^{\infty} T(x|D) = 1$. We employed the Poisson and the negative binomial distribution as the parent distribution D, but any other discrete distribution could be used as well.

The tail distributions can be included in the mixture model in Equation (3.1) and their mixing coefficients can be estimated with the described iterative algorithm. We can use the relation of the mixing coefficients of the tail distribution to the corresponding parent distribution to estimate the number of fragments the shared part of the sequence is divided into. The average fragment consists of a center part with homogeneous coverage depth and the two flanks with linearly decreasing depth. The center part is represented by the parent distribution D, the flanking parts by the tail distribution $T(D)$. The length of the flanks is fixed by definition to one read length RL on each end of the fragment. The relation of the length of the center part of the fragment l_D and the flanks can be approximated by the relation of the corresponding mixing coefficients:

$$\frac{l_D}{2 \cdot RL} \approx \frac{\alpha_D}{\alpha_{T(D)}} .$$

The total length of the shared part is $(\alpha_D + \alpha_{T(D)}) \cdot L$ with the total genome length L and thus we find the relation between the number of fragments $frag$ and the mixing coefficients:

$$(\alpha_D + \alpha_{T(D)}) \cdot L = (2 \cdot RL + l_D) \cdot frag$$

and thus

$$frag = \frac{(\alpha_D + \alpha_{T(D)}) \cdot L}{2 \cdot RL + l_D} \approx \frac{(\alpha_D + \alpha_{T(D)}) \cdot L}{2 \cdot RL \cdot (1 + \frac{\alpha_D}{\alpha_{T(D)}})}$$

or more simple

$$frag \approx \frac{\alpha_{T(D)} \cdot L}{2 \cdot RL} \; . \tag{3.3}$$

A high fragmentation of the genome gives rise to additional correction terms for models with zero-inflation. Due to the excess of uncovered positions at fragment borders, which is most pronounced for genomes with partially non-zero, low ($1\times - 10\times$) coverage depth and uncovered positions elsewhere, we introduce a correction term z_{corr} for the mixing coefficient of the zero component of the model. The probability that a read starts immediately at the border of a shared fragment depends on the average center coverage depth c_{frag} of the fragment. On average, a read will start delayed with delay $\bar{d}$ after the fragment border (see Figure 3.4). The probability that a position is a starting position for one (or more) reads can be calculated under the assumption that the number of starting reads at each position is Poisson distributed:

$$p_{start} = \sum_{n=1}^{\infty} P(n|\lambda = c_{frag}/RL) \; ,$$

where P is the probability mass function of the Poisson distribution. The delay for the first read to start behind the border is negative binomially distributed and we can calculate the average delay as:

$$\bar{d} = \sum_{n=0}^{RL} n \cdot NB(n|1, 1 - p_{start}) \; ,$$

with the probability mass function NB of the negative binomial distribution, which takes the number of successes ($= 1$) the failure probability ($= 1 - p_{start}$) as parameters.

The average delay $\bar{d}$ can be used to correct the mixing coefficient of the zero distribution in the model:

$$z_{corr} = \frac{2 \cdot frag \cdot \bar{d}}{L} \; .$$

Taken together, we see that the presented iterative method can be used to fit mixtures of discrete probability distributions to histograms. We introduced a set of specialized functions to fit genome coverage depth histograms. The genome validity and the genome fragmentation, two quantitative traits of the mapping, can be calculated from the fitted distributions. However, these are only two examples and many other applications may be possible.

3.4. Extension to ultra-low coverage depths

As stated previously, calculating the GCP, i.e. transforming the observed position wise coverage depth values into the histogram of observed coverage depth values, summarizes the information hidden in the coverage depth of the genome. The previous section suggests that this condensed view may bring benefit for the analysis of sequencing data and that interesting features such as the genome validity can be derived from GCPs. However, this transformation comes with a loss of information. Before, we had one coverage depth value for each position in the genome, which is some million datapoints for a bacterial genome. In contrast, the number of datapoints in the GCP is the highest observed coverage depth on the genome. Even for ultra-deep sequencing, the number of GCP datapoints will not exceed several thousands. In the more common case of lower sequencing depths, the number of GCP datapoints becomes very critical, since all parameters in the mixed distributions in Equation (3.1) must be estimated from these datapoints. For example, the distribution consisting of zero, Poisson, and negative binomial with tail (`zpnt`) has 6 parameters and therefore requires that a sufficient number of positions with coverage depth 6× is observed in the dataset. In order to compensate for noisy data, higher sequencing depths are desirable. In the extreme case when the sequencing depth is so low that the reads do not overlap, the only possibility is to fit a zero-inflated Poisson (`zp`) model since the GCP contains only the datapoints 0 and 1, which allows determining solutions for two parameters. These parameters would be very error prone in practice due to noise.

In the following sections we develop a model similar to the GCP case. This model is designed for extremely low coverage depths and should therefore complement the existing model. The mathematical procedures, i.e. the modified EM algorithm, are identical; we only use different data to construct the histogram and fit different distributions to that histogram.

3.4.1. A new model for low coverage depths

Estimating the genome validity with the previously described approach can be problematic for very low coverage depths: here, the approach with calculating the GCP and fitting, e.g., a zero-inflated negative binomial distribution with tail (`znt`) model, is not appropriate since we do not expect to have sufficient datapoints for robust parameter estimation. To increase the number of datapoints, we can look at the starting positions of the reads mapped to the genome, as shown in Figure 3.6 (a).

Under the assumption that the sequencing depth is homogeneous over the entire genome, the probability that a mapped read starts at a certain position on the genome is equal for all positions. We denote this read start probability as p. Let us consider a read starting at position $X = 0$. The probability that the next read starts at position $X = k$ follows a geometric distribution with parameter p. This becomes clear when considering that if the *next* read starts at position $X = k$, no read starts at positions $X = 0..k-1$. Therefore, the distances between the starting positions of mapped reads are geometrically distributed.

In analogy to the GCP, we call the histogram of distances between read starting positions the Read Distance Profile (RDP). When reads of an organism are mapped to its reference genome, we can simply estimate the parameter p of the geometric distribution by making use of the formula for the expected value of the geometric distribution: $\bar{D} = E(X) = \frac{1-p}{p}$. However, if we expect more than one contribution to the RDP (e.g., reads from two different organisms map to the same genome), we have to set up a mixed model similar to Equation (3.1), consisting of two or more geometric distributions. In fact, our framework presented above can be adapted to fit RDPs simply by using models consisting solely of geometric distributions. Parameter estimation is identical to the GCP case.

Since the number of reads mapped to a genome is typically in the order of hundreds or thousands even in cases of ultra-low coverage depths and the distances are very large, we have increased the number of datapoints for parameter estimation significantly with this approach. We will show in the experiments that this approach can be applied at much lower coverage depths than the GCP approach.

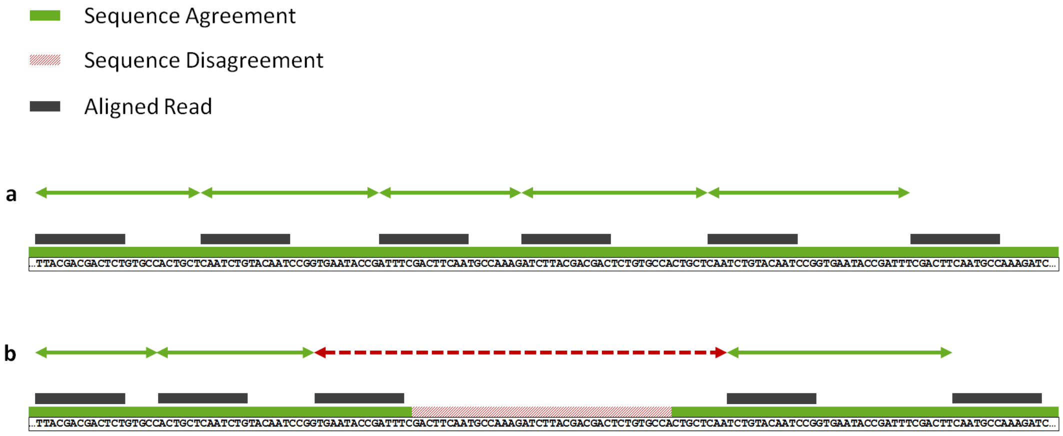

Sequence Agreement
Sequence Disagreement
Aligned Read
a
...TTACGACGACTCTGTGCCACTGCTCAATCTGTACAATCCGGTGAATACCGATTTCGACTTCAATGCCAAAGATCTTACGACGACTCTGTGCCACTGCTCAATCTGTACAATCCGGTGAATACCGATTTCGACTTCAATGCCAAAGATC...
b
...TTACGACGACTCTGTGCCACTGCTCAATCTGTACAATCCGGTGAATACCGATTTCGACTTCAATGCCAAAGATCTTACGACGACTCTGTGCCACTGCTCAATCTGTACAATCCGGTGAATACCGATTTCGACTTCAATGCCAAAGATC...

Figure 3.6.: Distances between read start positions. (a) An organism is sequenced with low and homogeneous sequencing depth and its reference genome is available. Then, the distances between neighboring read start positions in a shotgun sequencing experiment can be described by a geometric distribution. (b) When the reference genome differs from the sequenced organism, there will be islands of sequence agreement (green) divided by gaps of sequence disagreement (red). The distances between neighboring reads on the islands (green arrows) still follow a geometric distribution, disturbed by the distances spanning the gaps (red arrow).

3.4.2. Genome validity from RDPs

The genome validity introduced in Section 3.3.1 is a useful tool for estimating the similarity between a set of reads and a reference genome. However, since the estimation involves fitting complex models to GCPs, we expect decreasing accuracy for lower coverage depths. Here we show that it is also possible to calculate the genome validity from RDP and should therefore be accessible for much lower coverage depths.

Similarly to the role of the zero distribution in GCPs, we have to estimate the fraction of the genome that is not covered by reads. Therefore, we have to quantify the gaps between the parts of the genome that could potentially be covered by reads (see Figure 3.6 b). Here, we introduce a heuristic and assume that the gap lengths also follow a geometric distribution. This is motivated by the assumption that shorter gaps should be more frequent than longer gaps.

To calculate the genome validity, it is sufficient to fit two geometric distributions to the RDP: the first distribution fits the distances between reads lying on a contiguous sequence fragment of the reference (visualized by green arrows in Figure 3.6 b). The second distribution fits the distances between reads lying on neighboring sequence fragments divided by a gap of foreign sequence (red arrow in Figure 3.6 b). Let α_1 and p_1 denote the estimated parameters of the geometric distribution fitting the distances between reads on the same fragment and let α_2 and p_2 denote the estimated parameters of the geometric distribution fitting the distances between reads on neighboring fragments. Here, we also make the assumption that the distances between reads on the same fragment are typically lower than the distances between reads on neighboring fragments and therefore $p_1 > p_2$. One way to estimate the genome validity is to calculate the expected number of reads mapping to the genome under the assumption that the reference genome contained no foreign sequences and to relate this number to the observed number of reads R that were actually mapped to the genome. The expected number of reads is the genome length L divided by the expected distance between reads: $\bar{D} = \frac{1-p_1}{p_1}$. Therefore, we write the genome validity for low coverage depths as follows:

$$val_{lc} = \frac{R \cdot \bar{D}}{L} = \frac{R \cdot (1 - p_1)}{p_1 \cdot L}. \tag{3.4}$$

The coverage depth on the covered sequence fragments can be calculated using the (average) read length RL:

$$cov_{lc} = \frac{RL \cdot p_1}{1 - p_1}.$$

We test this approach on low sequencing depth data and present the experiments and results in the following sections. Further, the coverage depth and genome validity estimation from RDPs is an essential part of the MicrobeGPS tool presented in chapter 4.

3.5. Experiments and results

The framework described in the previous section provides a powerful tool for solving problems related to reference genomes and genome coverage depth distributions. Here we present four experiments that demonstrate the applicability of the framework in a metagenomic context. In the first experiment, we demonstrate that the proposed algorithm can fit complex mixtures of distributions to GCPs and evaluate the influence of the choice of model on the fit quality. In the second experiment, we demonstrate the robustness of the framework: quantitative traits (in this case the genome validity) can be estimated robustly from the GCP fits over a wide range of genome coverage depths. This is crucial in metagenomics, where the number of mapped reads per genome is typically very small, but can be high for single abundant species. In the third experiment, we apply the framework on real data and thereby illustrate a further application: we re-analyze data from a large scale human gut metagenomic study and compute the genome validity for a selected set of reference genomes. In the last experiment, we apply the alternative of GCPs for lower sequencing depths. We evaluate systematically in which scenarios the RDP approach is applicable to determine the local sequencing depth and genome validity.

3.5.1. Fitting complex mixture models

In this experiment, we evaluate the performance of the presented algorithm for fitting complex mixture models to multi-modal GCPs. Thus, we created a dataset with reads of two organisms sharing large genomic regions, *Escherichia coli* and *Shigella boydii*. We simulated 100,000 reads for *E. coli* and 600,000 reads for *S. boydii* with 75 bp length and Illumina sequencing characteristics using the Mason read simulator (Holtgrewe, 2010). These reads were then mapped to the *E. coli* reference genome with Bowtie (Langmead et al., 2009). We expected the genome to be homogeneously covered by the *E. coli* reads and locally by additional *S. boydii* reads. Yet, the number of *E. coli* reads could only account for 1.5× sequencing depth. This challenged the algorithm in two ways: First, the low *E. coli* sequencing depth caused a fraction of genome positions to be uncovered, yet, they should not be explained by a zero distribution since the organism in the simulated dataset agrees perfectly with the reference genome. Therefore, all positions that are not covered should be included in the distribution describing the *E. coli* coverage depth. Second, the *S. boydii* fragments with high sequencing depth produced a tail in the GCP, which overlapped with the *E. coli* distribution. For fitting, we used models consisting of three components: (i) a zero distribution (abbreviated by `z`), (ii) a

Poisson (`p`) or negative binomial (`n`) distribution for the *E. coli* reads and (iii) a Poisson, negative binomial, Poisson with tail (`pt`), or negative binomial with tail (`nt`) distribution for the *S. boydii* reads. The initialization was chosen such that component (ii) fitted the *E. coli* peak and (iii) fitted the *S. boydii* peak.

All models were fitted to the GCP using an accuracy threshold of 0.1% for the iteration and the zero-correction was calculated for the models with tail distribution. To compare the models by numbers, we calculated the Kolmogorov-Smirnov test statistic, the maximum absolute difference d_{max} between the observed and the estimated cumulative mass function.

Figure 3.7 depicts the fitted distributions of selected mixture models, detailed results about the mixing coefficients and fit errors are listed in Table 3.1. The results show a prominent difference between models with and without tail: Models with tail fit the observed GCP much better (average $d_{max} = 0.0022$) than the models without tail (average $d_{max} = 0.0073$). The simplest model, `zpp` (see Figure 3.7 a), yields the highest fit error of all models ($d_{max} = 0.0141$). For the models without tail, the fit error decreases as the model complexity (i.e. the number of parameters to fit) increases. The difference of the fit error between the models with tail is overall lower than between the models without tail: The lowest fit error among the models with tail is achieved by `zpnt` ($d_{max} = 0.0018$), the highest by `znnt` ($d_{max} = 0.0026$). In particular, the fit error does not decrease with increasing model

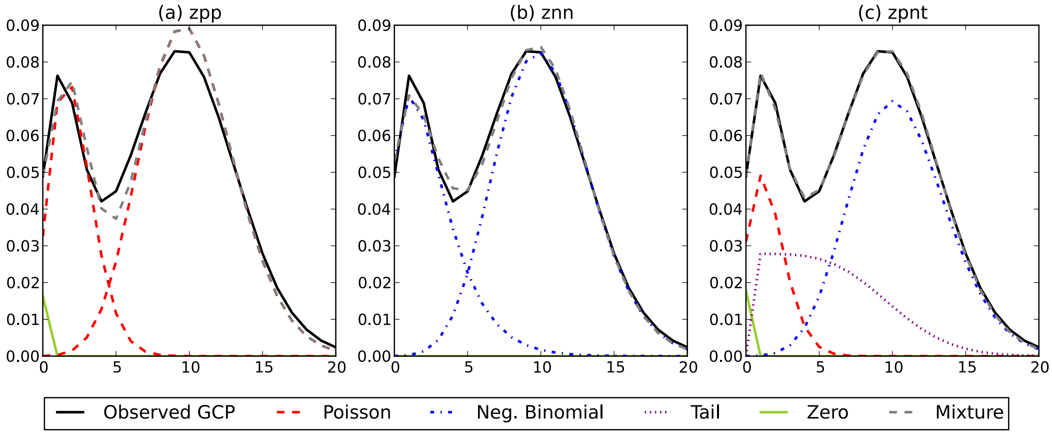


Figure 3.7.: Influence of the choice of mixture model for fitting GCPs. Three exemplary models are shown: (a) zero and two Poisson distributions (`zpp`), (b) zero and two negative binomial distributions (`znn`), (c) zero, Poisson and negative binomial distribution with tail (`zpnt`). Model (c) yields the lowest fit error ($d_{max} = 0.0018$), but is more complex than models (a) and (b). Model (a) has the lowest complexity, but yields the highest fit error ($d_{max} = 0.0141$).

Table 3.1.: Fitting complex mixture models to GCPs. *E. coli* and *S. boydii* reads were simulated and mapped to an *E. coli* reference genome. We fitted eight different models consisting of zero, Poisson, negative binomial, Poisson tail and negative binomial tail to the GCP to estimate the distribution parameters.

	Distribution 1	**Distribution 2**	**Distribution 3**	**Distribution 4**	**Fit Error**
Model 1	**Zero** $\alpha = 0.0163$	**Poisson** $\alpha = 0.2713$ $\lambda = 2.12$	**Poisson** $\alpha = 0.7123$ $\lambda = 10.09$	[not used]	0.0141
Model 2	**Zero** α= 0.0	**Negative Binomial** $\alpha = 0.3302$ Mean = 2.59	**Poisson** $\alpha = 0.6698$ $\lambda = 10.32$	[not used]	0.0056
Model 3	**Zero** $\alpha = 0.0119$	**Poisson** $\alpha = 0.2595$ $\lambda = 1.95$	**Negative Binomial** $\alpha = 0.7285$ Mean = 9.96	[not used]	0.0059
Model 4	**Zero** $\alpha = 0.0$	**Negative Binomial** $\alpha = 0.3049$ Mean = 2.27	**Negative Binomial** $\alpha = 0.6951$ Mean = 10.17	[not used]	0.0035
Model 5	**Zero** $\alpha = 0.0$	**Poisson** $\alpha = 0.1496$ $\lambda = 1.58$	**Poisson** $\alpha = 0.5665$ $\lambda = 10.57$	**Tail** $\alpha = 0.2656$ 8,278 Fragments	0.0020
Model 6	**Zero** $\alpha = 0.0$	**Negative Binomial** $\alpha = 0.1523$ Mean = 1.59	**Poisson** $\alpha = 0.5699$ $\lambda = 10.56$	**Tail** $\alpha = 0.2596$ 8,092 Fragments	0.0022
Model 7	**Zero** $\alpha = 0.0$	**Poisson** $\alpha = 0.1507$ $\lambda = 1.57$	**Negative Binomial** $\alpha = 0.5654$ Mean = 10.56	**Tail** $\alpha = 0.2659$ 8,289 Fragments	0.0018
Model 8	**Zero** $\alpha = 0.0$	**Negative Binomial** $\alpha = 0.1545$ Mean = 1.60	**Negative Binomial** $\alpha = 0.5699$ Mean = 10.54	**Tail** $\alpha = 0.2577$ 8,032 Fragments	0.0026

complexity. Furthermore, the model fits with tail are highly similar: besides the similar fit error, they also have almost identical mean values μ for the two non-zero distributions (distribution (ii): $1.57 < \mu < 1.60$; distribution (iii): $10.54 < \mu < 10.57$).

The relative sizes of the tail distributions are on par with the other distributions, indicating a high degree of fragmentation of the *E. coli* genome compared to *S. boydii*. The number of *S. boydii* fragments in *E. coli* can be estimated via Equation (3.3); depending on the model, there are between 8032 and 8289 *S. boydii* fragments in the *E. coli* genome. The contribution of the zero distribution is estimated to exactly zero in all models except `zpp` and `zpn`.

Further experiments using more complex models (e.g., `znnnt`) does not reduce the fit error. The spare distributions either take the same shape as one of the two original distributions or their mixing coefficients are reduced to zero, depending on the start parameters.

This experiment shows that our algorithm can fit complex mixture models to GCPs accurately. Best results are obtained when the complexity and the selected distributions in the model match the data, but more complex models do not decrease accuracy and should thus be chosen when in doubt. The low fit errors of the models with tail distribution support the usefulness of the tail distribution concept. Although our iterative algorithm is not guaranteed to converge to an optimal solution as EM does, we see that the fit results are highly similar, in particular for the models with tail.

3.5.2. Influence of average coverage depth

In this experiment, we demonstrate the robustness of our framework over a wide range of coverage depths. Information about the genomes – both the source of the reads and the reference – derived from GCPs should generally not be affected by the overall number of reads mapped to the reference genome.

We used the *Shigella flexneri* genome as reference and simulated datasets of short (75 bp) Illumina reads from the *E. coli* genome with Mason. The smallest dataset contained 1,000, the largest 10 million *E. coli* reads. We used Bowtie to map these datasets to the *S. flexneri* reference genome and fitted ZI Poisson and ZI negative binomial models, both with and without zero correction, to the GCPs. The genome validity was calculated for each model based on the fit parameters as described in Section 3.3.1. The true genome validity was estimated from the dataset with 10 million simulated *E. coli* reads to be 0.826. In this dataset, at least one read starts at each position in the *E. coli* genome due to the high coverage depth. Therefore, all fragments in the *S. flexneri* genome that are identical with *E. coli* and at least 75 bp long should be covered by reads.

The estimated validity of *S. flexneri* for the *E. coli* reads is summarized in Figure 3.8. Curve (a) shows the estimated validity for the ZI Poisson mixture models (without and with zero correction), (b) for the corresponding ZI negative binomial models. The estimated validity of the *S. flexneri* genome is close to the estimated true genome validity (gray line) for all mixture models when the number of reads is above 1 million. In the range from 100,000 to 1 million reads, the ZI Poisson with correction yields the best estimates; it keeps the estimates on an almost constant level. For the ZI negative binomial model, the correction has a much smaller influence and the estimates are slightly worse than the corrected Poisson model. In the low coverage depth regime (below 1.5× depth or 100,000 reads), the tail effect is not observed anymore and all models reduce to the ZI Poisson and ZI negative binomial, respectively. The Poisson model yields lower estimates as the number of reads decreases and therefore becomes increasingly unreliable. On the other hand, the negative binomial model yields relatively good estimates down to very low numbers of reads (approx. 10,000), which corresponds to approximately 0.2× coverage depth in the covered fragments.

The results of this experiment suggest two different strategies for the selection of the mixture models: For coverage depths below 1×, the plain ZI negative binomial distribution yields the best results and allows determining the validity with acceptable accuracy. For local coverage depths of 2× and above, the ZI Poisson model with zero-correction produces highly accurate estimates and outperforms all other models. There, the advantage of the Poisson model is two-fold: in addition to the better estimates, the Poisson model has one parameter less than the negative binomial model and parameter estimation is faster. Conclusively, we see that the estimated genome validity is largely independent of genome coverage depth and estimation is possible even below 1× depth.

3.5.3. Application to the human gut microbiome

The validity of reference genomes is crucial for a sound interpretation of the data, in particular in metagenomics. Thus, we estimated the validity of genomes on real metagenomic data. The work by Qin et al. (2010) serves as a test case; they sequenced the metagenomic communities in fecal samples of 124 European individuals on the Illumina platform and conducted exhaustive analysis to provide insight into the composition of genes and bacterial species in the human gut. As one result, they report a list of 75 prevalent bacterial species, the *common core*, which were present (genome coverage $> 1\%$) in a large number of individuals. We obtained the original reference genomes of the common core and selected 17 genomes that were originally found in all 124 individuals with at least 1% coverage. The metagenomic reads of individual `MH0012` were downloaded from the corresponding EBI database (accession numbers: `ERX004076`–`ERX004082`).

In this experiment, we estimated the genome validity of the selected reference

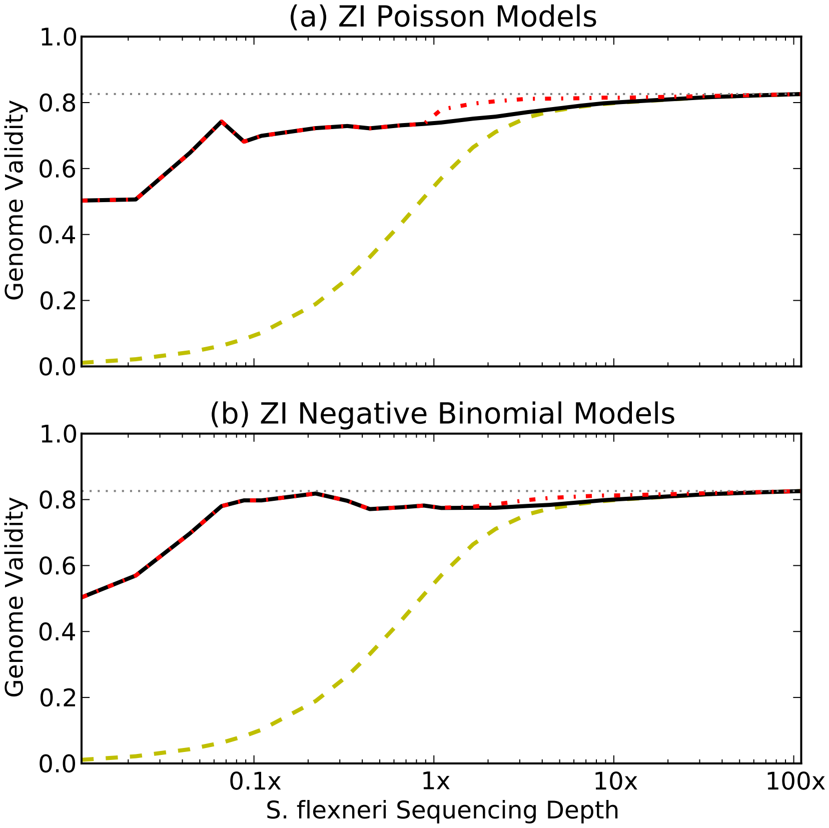


Figure 3.8.: Estimating the genome validity. The charts show the estimated fraction of the *S. flexneri* genome that is similar to the *E. coli* genome, i.e. the genome validity, depending on the number of *E. coli* reads mapped to the *S. flexneri* genome. We used zero-inflated Poisson (a) and negative binomial (b) mixture models (see text). Each model was fitted without (black solid line) and with (red dash-dotted line) zero-correction. The dashed line is the fraction of the genome that was covered by reads. The gray dotted line is the estimated true genome validity.

genomes with respect to the metagenome of individual MH0012. The 93 million paired-end Illumina reads were mapped to the selected reference genomes using Bowtie 2 (Langmead and Salzberg, 2012) and the GCP was calculated subsequently for each genome. In the next step, we fitted a mixture model of a zero distribution, two negative binomial distributions (with maximum likelihood estimation), and a negative binomial tail distribution to the GCPs. We preferred the negative binomial over the Poisson distribution here, since we expected over-dispersion due to a high biological variability in the metagenomic data. Two negative binomial distributions were chosen with genomic similarities in mind, where one distribution should fit the matches from the correct species and the other distribution should account for the noisy matches obtained by organisms with partial sequence similarity. The fit error was calculated as in the first experiment.

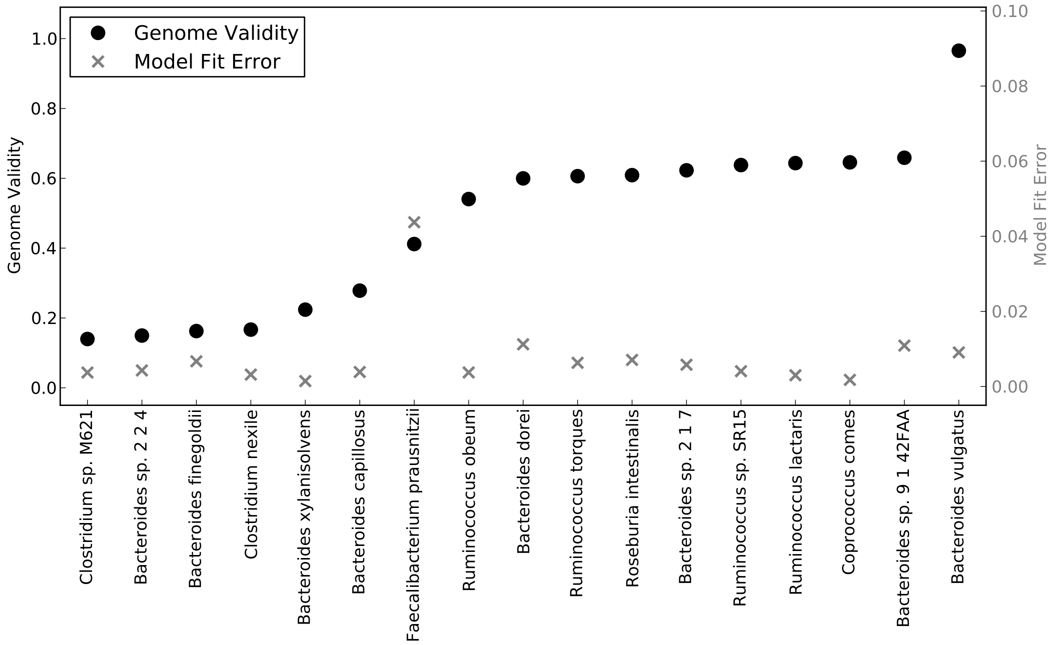


Figure 3.9.: Genome validity for human gut metagenome. The validity of 17 prevalent bacterial species with respect to one metagenomic human gut sample was estimated by fitting the GCPs as described in 3.3.1. The validity (dots) ranges from 0.140 for *Clostridium sp. M621* to 0.965 for *Bacteroides vulgatus*. A lower fit error (crosses, right axis) indicates a more trustworthy validity estimate.

The genome validity scores of the 17 selected reference genomes are shown in Figure 3.9 and ranges from 0.140 (*Clostridium sp. M621*) to 0.965 (*Bacteroides vulgatus*). All genomes have a moderate average coverage depth (min. 8×, max.

47×) and the coverage depth has only low correlation to the genome validity (Pearson correlation coefficient: $r = 0.34$). The fit error is below 0.02 for all genomes, which indicates that the assumed model is sufficient for the complexity of the data. The only exception is *Faecalibacterium prausnitzii*: the GCP was too complex for the assumed model and required an additional negative binomial component.

Considering that *E. coli* reads mapped to a *S. flexneri* genome yield a validity of more than 0.8, as demonstrated in the previous experiment, the numbers observed in this experiment are rather low. One reason may be that the selected reference genomes originally served as representatives for clusters of similar genomes. Furthermore, most reference genomes were sets of separate contigs, indicating that the reference genomes could be incomplete or have low quality. This is prototypic for metagenomics, as the majority of bacteria is still not or only poorly sequenced, such that a reference genome with low validity may be not a good, but the best possible choice. The only high quality reference genome (no contigs) is *B. vulgatus*, which achieves by far the highest validity ($val = 0.96$).

A similar picture can be observed on the full set of 75 genomes (see Figure 3.10); the genome validities are in the range from 0.013 (*Enterococcus faecalis*) up to 0.998 (*Clostridium leptum*). Interestingly, four out of the seven best scoring genomes are high quality genomes and the other three have less than 100 contigs. Manual validation confirmed the high validity and showed homogenous coverage depth over the genomes, only interrupted by very small gaps and single high coverage depth parts. On the other hand, the genome with the worst score – the gut bacterium *E. faecalis* – is also a high quality genome. Manual validation showed that *E. faecalis* was not covered homogeneously. This underlines the features of the genome validity: to achieve a high validity, reference genomes must be homogeneously covered by reads and must have high quality. A high genome quality but a low and inhomogeneous coverage depth, indicating that the species itself is not present, is correctly penalized by a low score.

3.5.4. Genome validity at low coverage depths

In this experiment, we evaluate how well the genome validity can be calculated for genomes with very low coverage depth using the distance-based approach presented in Section 3.4. The previous experiments have shown that the genome validity is a useful and robust measure for quantifying the divergence between a reference genome and the sequenced organism. However, the genome validity could only be calculated robustly for coverage depths as low as 0.2×. In principle, the distance-based approach should be able to estimate the validity for even smaller coverage depths, therefore, we assessed its robustness systematically.

We simulated two random genome sequences: a genome A with fixed length 4,600,000 bp (approximate length of the *E. coli* genome), which serves as ref-

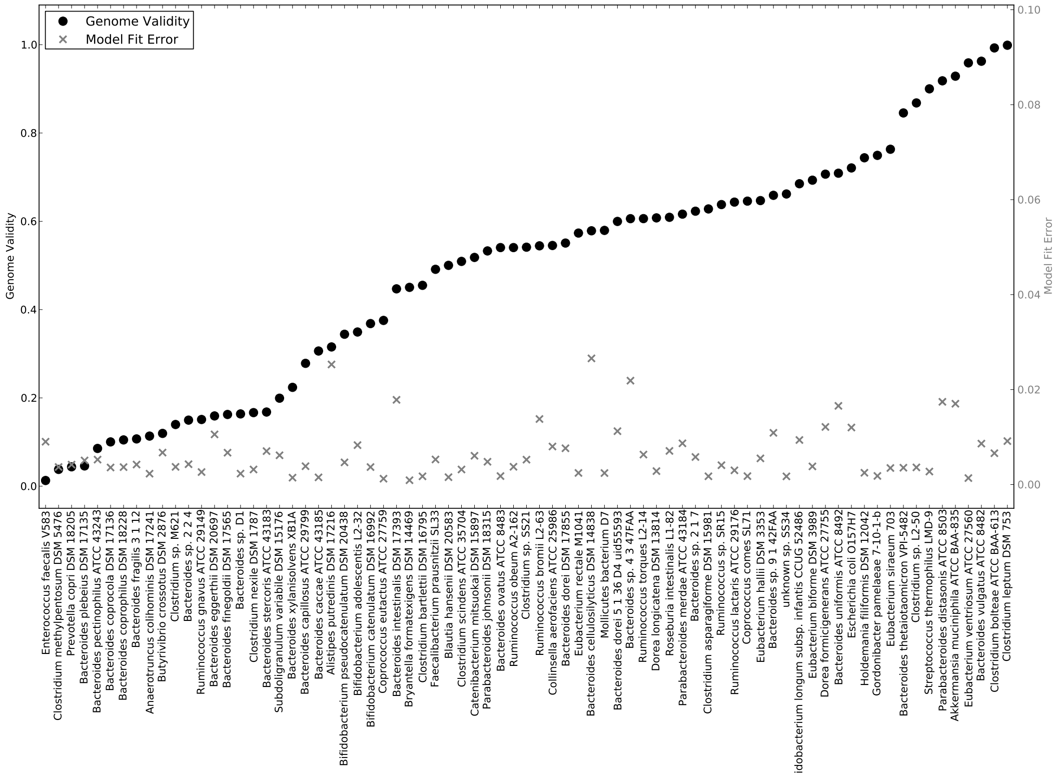


Figure 3.10.: Genome validities for the complete set of human gut reference genomes. The validity of all 75 bacterial species with respect to one metagenomic human gut sample was estimated by fitting the GCPs as described in 3.3.1. A lower fit error (crosses, right axis) indicates a more trustworthy genome validity estimate. Here, we observe a higher number of reference genomes with very low validity than in the set of 17 prevalent genomes.

erence genome, and a smaller genome B with varible length. The sequence of genome B was cut into 20 fragments that were integrated into genome A. The length of genome B was chosen such that A had a defined validity val for the reads sequenced from B. We simulated pairs of genomes A and B with validities $val = 0.8, 0.6, 0.4, 0.2, 0.1$. For each validity, we simulated sets of 72 bp reads from genome B, such that the sequencing depth on B was $cov = 5\times, 2\times, 1\times, 0.5\times, 0.2\times, 0.1\times, 0.05\times, 0.02\times, 0.01\times, 0.005\times, 0.002\times, 0.001\times$.

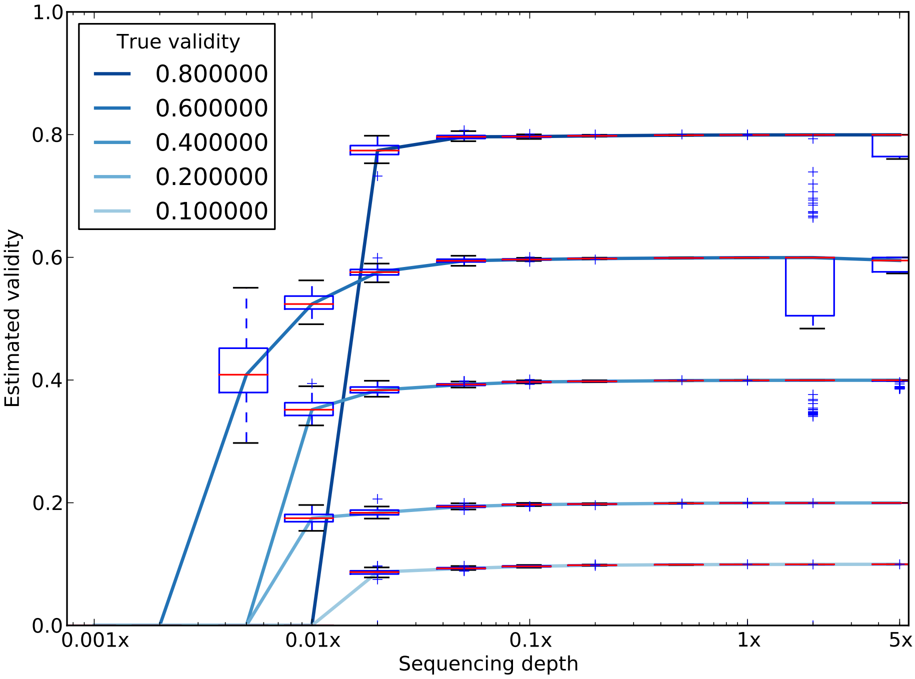


Figure 3.11.: Validity estimation at low coverage depths. We simulated pairs of reference genomes and datasets with fixed coverage depth and reference genomes with defined genome validity with respect to the dataset. Genome validities were estimated for each combination using the distance-based approach. The distance-based and true validities agree if the coverage depth exceeds a certain minimum depth. This minimum depth depends on the validity, where the intermediate valitities (0.2 and 0.4) can be robustly estimated at much lower depths than the extreme validities.

For each combination of ground truth and coverage depth, we calculated the genome validity and the estimated coverage depth using the distance-based approach and the equations (3.4) and (3.4.2). The experiment was repeated 100 times. The estimated validities are shown in Figure 3.11. Here, we see that the quality of the validity estimate depends both on the coverage depth and on the validity itself.

Intermediate validity scores (0.2 and 0.4) can be correctly estimated at much lower coverage depths than more extreme validities. For 0.2 validity, a minimum of 0.01× coverage depth is required to estimate the validity with below 10% error in 50% of all cases. In general, a minimum sequencing depth of 0.02× was sufficient to estimate the genome validity with less than 10% error.

This result is remarkable when we consider that the coverage-depth-based approach applied in Section 3.5.2 only provided stable estimates down to 0.2× coverage depth. This means that the new approach lowered this limit by up to a factor of 20 compared to calculating the validity from the coverage depth profile. For a genome with 0.2 validity and 72 bp reads, this means that only one read every 7,200 bp is required to estimate the validity correctly.

There are two effects that influence the required minimum coverage depth: For low validities, only small fractions of the genome are covered with reads at all and the covered regions are smaller than for high validities. At very low coverages, it becomes less likely that two or more reads are mapped to one contiguous fragment; at this point, we start to observe mostly distances between reads on different fragments, which makes it impossible to infer the local coverage depth on the covered fragments. On the other hand, for high validities, the gaps between the covered fragments become very small. If the coverage depth is too low, the distances between neighboring reads may become larger than the gaps between covered fragments, our assumption fails and makes it impossible to mathematically distinguish the distribution of gaps between reads on the same fragment from the gaps between reads on different fragments. We presume that our approach of calculating the genome validity is close to the technical limit that can be reached with any approach based on read mapping.

3.6. Discussion of results

We introduced GCPs as a means to extract quantitative information from mapping data. By fitting mixtures of probability distributions to the GCP, we obtain valuable information about the reference genomes and the mapping process, such as the fraction of the genome that could not be covered by reads or if there is more than one organism contributing to the observed coverage depth. This makes the proposed framework a powerful tool for the analysis of mapping data without restriction to a specific application.

The introduced genome validity score is a simple, yet powerful measure for how well a reference genome fits to the mapped reads. Especially in metagenomics, reference genomes are typically not required to fit perfectly to the data. Nevertheless, the degree of divergence should not become too large. As one example, we observed a validity score of 0.82 in the experiment in Section 3.5.2, where we mapped *E. coli* reads to a *S. boydii* genome. This illustrates a relatively high taxonomic divergence between data and reference despite a high validity score. We assessed validity scores in a real metagenomic experiment conducted by Qin et al. (2010) and observed surprisingly low scores for genomes that were originally considered to be present in the dataset; only 9 of 75 reference genomes achieved scores higher than 0.8. This is an imposing example for high discrepancy between metagenomic data and reference genomes, which we presume to be a common challenge of metagenomic experiments. One of the major reasons might be the quality of the reference genomes: as microbes from metagenomic experiments are typically not cultivable, their genomes must be assembled from environmental samples, which is significantly more complicated and error-prone than assembly from pure samples. In the experiment at hand, 37 of 75 reference genomes consisted of more than 100 (up to 1,700) separate contigs, only six genomes were one contiguous sequence. The framework proposed and applied in this work makes these flaws quantifiable.

The first experiment showed that the iterative algorithm is able to fit complex mixtures of highly specialized probability distributions to GCPs. The impact of the tail distributions became apparent, as they significantly reduced the fit error and prevented overfitting with too many distributions. The second experiment showed that quantities calculated on fitted GCPs are robust towards influences of the average genome sequencing depth. There, we observed stable estimates of the validity score over a wide range of sequencing depths, starting at average depths about 0.2×. Yet, the iterative algorithm encounters limitations in extreme cases, for example when the average coverage depth is very low but locally extremely high, as it occurs when a genome is not present in the data, but shares a common gene with other present genomes. Then, the algorithm may fail to fit the low-depth distribution as intended by the user, but tries to fit the extremely high noise contributions. In other cases, the standard start parameters are inappropriate, such that the algorithm ends up in a local probability maximum instead of fitting the distribution as intended. These problems demonstrate that visual inspection of the fit is necessary, which is supported by the framework. Common strategies used for the EM algorithm are also possible, such as the initialization with different or manually determined starting parameters.

Here we focused on applications in metagenomics, however, the information obtained by fitting the GCP is by no means limited to metagenomics but can be used for other purposes, such as experimental design and coverage depth estimation (Hooper et al., 2010), the detection of copy number variations (Miller et al., 2011) or metagenome assembly (Namiki et al., 2012). As an example, metagenomic sequencing experiments can be designed in a way, such that the validity score can be calculated robustly for reference genomes with a certain minimum abundance in the sample. The minimum amount of sequencing required can be found by finding the minimum required sequencing depth for a robust validity score calculation in a simulation-based experiment, as presented in Section 3.5.2. Tools for estimating species abundances in metagenomic data, such as GRAMMy (Xia et al., 2011), GASiC (Lindner and Renard, 2013), or READSCAN (Naeem et al., 2012), can make use of the validity score to more precisely estimate the abundance of the organism truly contained in the dataset, if the used reference genomes have a low validity. One possible application in metagenomics is presented in the following chapter, where the information from the GCPs can be used to estimate the evolutionary distance of unknown organisms in the data to known organisms by mapping the reads to the known genomes and calculating validity scores. In connection with taxonomic information, the validity score can be used to narrow the truly contained organism down to a certain area of a taxonomic tree by excluding reference genomes yielding a lower validity score.

4. Characterization of known and unknown microbes in metagenomes

Recent advances in experimental and computational technologies have increased the number and diversity of sequenced microbial organisms. Today, single cell sequencing (Mason et al., 2012; Dodsworth et al., 2013) and metagenome assembly (Luo et al., 2012; Namiki et al., 2012) allow extracting the genomic sequences even of uncultivable bacteria. With the increasing number of microbial reference sequences, reference-based metagenomic analysis methods have become significantly more powerful and popular (Segata et al., 2012; Francis et al., 2013; Bonfert et al., 2013).

Although the taxonomic resolution of reference-based methods in metagenomic experiments is higher with whole genome sequencing than with other strategies such as 16S rRNA (von Mering et al., 2007) or composition-based taxonomic profiling (Simon and Daniel, 2011), these methods encounter a different problem: the reference genome databases are still far from complete and – due to continuous evolution – will never be. This means in practice that the often proposed species or strain level accuracy (Francis et al., 2013) is only achieved if sufficient sequenced strains of the organism of interest are available. Otherwise, these methods are at risk of suggesting accuracy to the user that is not justified by the underlying reference data when they report the presence of a species in the database that merely happens to be the closest sequenced relative to the organism in the sample. For example, the NCBI bacterial genomes contain the *Akkermansia muciniphila* ATCC BAA-835 genome, which is the only representative of the class *Verrucomicrobiae* in the database. If a related *Verrucomicrobium* is sequenced, current tools are likely to report *A. muciniphila* ATCC BAA-835 without warning the user that the identified strain may have considerable difference to the true organism.

MetaPhlAn (Segata et al., 2012) is a fast and popular taxonomic profiling approach that maps metagenomic reads to a set of selected marker sequences. These marker sequences are carefully selected, such that they are unique for each organism in the database. A read can only match to one marker and can therefore be assigned to a distinct organism. Therefore, the abundances of organisms can be easily estimated by extrapolating from the number of reads hitting the respective marker sequences. Together with the small size of the marker sequence database, this makes MetaPhlAn very fast while the accuracy is comparable to other reference-based methods. However, since whole genomes are reduced to short marker sequences, there is no possibility to detect or even quantify differences between the sequenced organism and the reference. Organisms that contain marker sequences of different reference strains (e.g., due to horizontal gene transfer) show up in the results multiple times and it is not possible to detect such cases. Another popular approach is Pathoscope (Francis et al., 2013), a recent and powerful method that analyzes read alignments to whole genomes with particular focus on reads mapping to multiple genomes. The program calculates a probability for each read alignment that is used in a Bayesian mixture model. The reads are then reassigned to their most likely origin by optimizing the model parameters by expectation-maximization. This allows Pathoscope to differentiate between highly similar strains even in cases with very low sequencing depth. Although Pathoscope is able to identify the closest related reference when the true genome is not present in the database, it is not immediately clear if the reported identification is a perfect match.

Here, we present MicrobeGPS, a tool that accurately identifies microbial organisms in metagenomic sequencing data and quantifies their distances to known reference genomes, thereby uncovering potential error sources. In contrast to current methods, which typically seek reference genomes present in the data, MicrobeGPS approaches the problem from a biological perspective and finds microbial organisms that are then described with suitable reference genomes. Here, a microbial organism is characterized by its sequencing depth, in a similar fashion as the composition-based AbundanceBin (Wu and Ye, 2011) method. MicrobeGPS searches for unique source reads (USR) that are likely to originate from a region in a microbial organism that cannot be found in the other organisms in the sample, i.e. map to genomes with the same sequencing depth. Based on the USR information, MicrobeGPS creates clusters of reference genomes supporting the same candidate organism. The supporting genomes determine the taxonomic affiliation of the candidate. The available quality measures, such as the genome validity (Lindner et al., 2013), the distribution of read mapping error, and the homogeneity of the read distribution, quantify the genomic divergence between candidate and supporting genomes. This observation driven approach in combination with the quality measures makes MicrobeGPS unique in the sense that the tool reports highly accurate results when suitable reference genomes are available. Otherwise, it describes the contained organism using the closest related known reference genomes providing a sound quantification of sequence disagreement. The graphical user interface simplifies data analysis and interpretation, providing browsable results, color representation of quantitative traits, as well as interactive graphs and taxonomic trees.

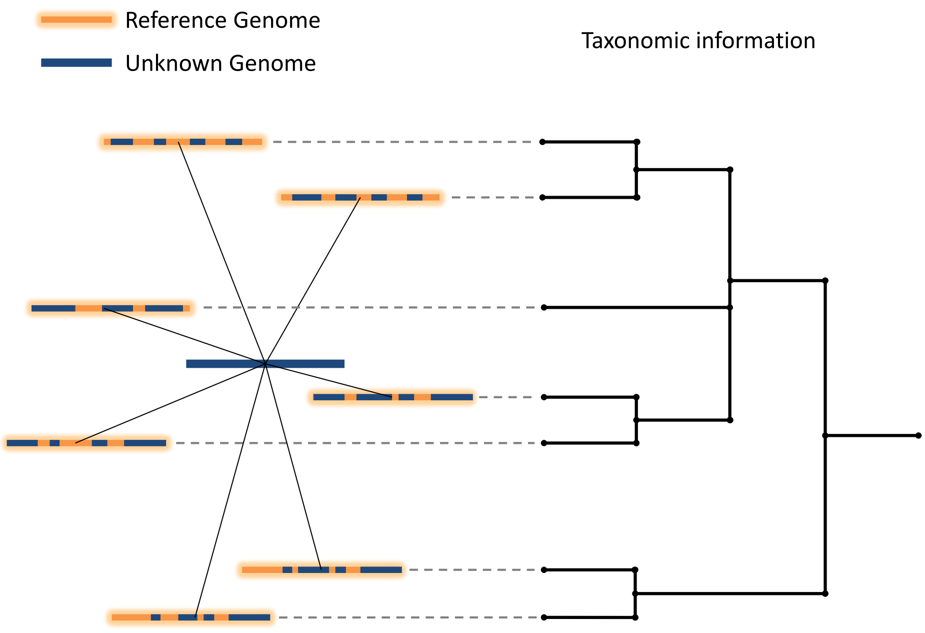


Figure 4.1.: The GPS principle. The taxonomic *location* of an unknown organism (blue) is estimated by calculating distances to already known and taxonomically classified reference genomes (orange). As the unknown organism is typically realized as a set of reads in a sequencing experiment that are mapped to the reference genomes, we use the genome validity (Section 3.3.1) as distance between the unknown organism and the reference genomes.

4.1. Characterizing unknown organisms with known genomes

The term GPS in MicrobeGPS refers to the Global Positioning System, where known positions of satellites in space are used to calculate a position on earth's surface. We transfer this principle to metagenomics, where we use known and characterized reference genome sequences as satellites to taxonomically locate unknown organisms (see Figure 4.1).

A metagenomic dataset contains sequencing reads derived from a mixture of microbial organisms living in an environmental community. The number of reads from each organism is assumed to be proportional to its genome length and its abundance in the community. Furthermore, we assume that the reads are sampled from random positions in the genome such that we would expect a homogeneous sequencing depth over the genome sequence.

4.2. The MicrobeGPS algorithm

Our MicrobeGPS approach to the taxonomic profiling problem consists of two parts. First, we developed an algorithmic recipe for the analysis of metagenomic datasets, which follows the GPS principle described in Section 4.1. Second, we provide an implementation of the MicrobeGPS algorithm including a graphical user interface (GUI). The following sections describe all parts of MicrobeGPS.

4.2.1. Read mapping and local sequencing depth estimation

MicrobeGPS is related to the similarity-based or reference-based taxonomic binning methods (Simon and Daniel, 2011), which means that it uses a database of existing and characerized reference genomes. As for all reference-based methods, the performance of MicrobeGPS depends on the quality and suitability of the database. Therefore, it is essential to select a database of reference genomes that comprises all taxa that are expected in the sample. In contrast to other tools, it is not necessary for an organism in the sample to have an exact match in the database, as MicrobeGPS is able to quantify the divergence between organism and reference genome. However, it is helpful if reference genomes of at least some related organisms are present. Since it is often not known which organisms can be found in a metagenomic sample, we recommend using broad and extensive databases, such as provided by the NCBI (Pruitt et al., 2007, 2014) or HMP (Nelson et al., 2010). However, current databases tend to vary strongly in their taxonomic resolution, i.e. some parts of the taxonomy contain large numbers of reference genomes while there is only one genome per family or genus present in other parts. Although the high numbers of reference genomes in some taxonomic units (e.g., 62 *E. coli* genomes in NCBI Bacteria) are beneficial for the taxonomic resolution, more balanced databases may be preferable for explorative tasks.

Similarly to other methods in this genre, MicrobeGPS requires the metagenomic reads to be mapped to the set of reference genomes. Considering the large reference genome databases that contain thousands of genomes and the vast amounts of data produced by current sequencers, it makes sense to use a fast read mapper that is capable of reporting all or the N best read mappings. Since MicrobeGPS makes heavy use of reads mapping to multiple reference genomes, it is not possible to use read mappers that only report the best match for each read. To increase sensitivity, we recommend reporting all read matches up to a feasible mapping error rate. We had good experience with Bowtie 2 (Langmead and Salzberg, 2012), NGM (Sedlazeck et al., 2013) and Masai (Siragusa et al., 2013), but other tools supporting the SAM format as output are also feasible.

MicrobeGPS takes the SAM files produced by the read mapper as input. Since these files can be very large, the user can apply a read filter and a genome filter to discard undesired reads or genomes. MicrobeGPS allows excluding reads from the analysis that originate from highly conserved regions, such as 16S rRNA reads. Since these reads can be mapped to almost all bacterial genomes, they are not considered as informative in the MicrobeGPS setup. In fact, the high numbers of reads mapped to such conserved regions in the genomes can suppress the faint signals of very low abundant organisms, which are often only represented by some tens to thousands reads. Since the size of the conserved regions is small compared to the typical genome sizes in bacteria, the positive effect of this filter can outbalance the loss of information that inherently comes with discarding reads. To further reduce the computational effort, MicrobeGPS can discard reference genomes without sufficient support by matching reads or a highly uneven read distribution on the genome. A trivial filter excludes all reference genomes from further analysis that have no reads assigned. More restrictive filter criteria can be designed based on heuristics. For example, excluding reference genomes that obtained no unique read matches can help removing large fractions of reference genomes that were only hit due to very small local similarities in ubiquitous genes. However, this filter is critical if there are many highly related genomes in the dataset: then, it is likely that all reads of a truely present organism match to multiple genomes. Another frequently occuring case are genomes that only share small parts with the organisms in the sample; unfortunately, it is not possible to exclude them reliably by their abundance when there are sufficient reads mapping to these small matching sequence parts. However, if the reference genome was truely present in the sample, we would expect the reads to be distributed more or less evenly over the genome. In the described case, the majority of reads are mapped to a small region on the genome. This difference can be detected by a homogeneity measure borrowed from hypothesis testing: The Kolmogorov-Smirnov test (Massey Jr, 1951) is used to test whether two distributions are the same. Here, we test if the distribution of read start positions is compatible with a uniform distribution and calculate the Kolmogorov-Smirnov test statistic. In our opinion, this test statistic (or the corresponding p-value) is very suitable for identifying reference genomes that can be discarded from further analysis. However, the filter settings should be checked and adapted for each analysis separately and, when in doubt, should be set conservatively.

In the next step, we estimate the local sequencing depth of each genome by fitting a zero-inflated Poisson distribution with fragmentation tail (`zpt`, see chapter 3) to the GCP, the histogram of the observed sequencing depths on the genome. For genomes with very low numbers of reads, we fit three geometric distributions to the RDP, the histogram of the distances between neighboring reads (see Section 3.4).

With this strategy, MicrobeGPS is able to recover the original sequencing depth of an organism down to 0.05× even if the closest related available reference genome has only low similarity. As described in Lindner et al. (2013), this step also calculates the genome validity for each reference genome, which serves as distance measure in the later analysis steps. The genome validity is the fraction of the reference genome that would be covered by reads under the assumption of infinitely high sequencing depth. Thus, the genome validity measures the similarity of the reference genome to the closest related organism in the sample.

4.2.2. Unique source read extraction and clustering

Once the local sequencing depths are estimated, MicrobeGPS searches for unique source reads (USR). USR are reads mapping exclusively to reference genomes with similar local sequencing depth, i.e. the maximum difference between the sequencing depths on the target genomes of a read does not exceed a user defined limit. USR are likely to originate from a genomic region of an organism in the sample that is unique with respect to all other organisms in the sample: if the read would originate from a non-unique genomic region and could be assigned to multiple organisms, it is likely that these organisms have different sequencing depths. However, it is still possible that two or more organisms have the same sequencing depth such that false USR are identified. In these cases, there may be reads that map to reference genomes representing different organisms in the sample, for example when a common conserved gene is sequenced in two organisms with the same sequencing depth. However, the number of these reads should be low compared to true USR. Therefore, we relax the USR criterium in the following clustering step.

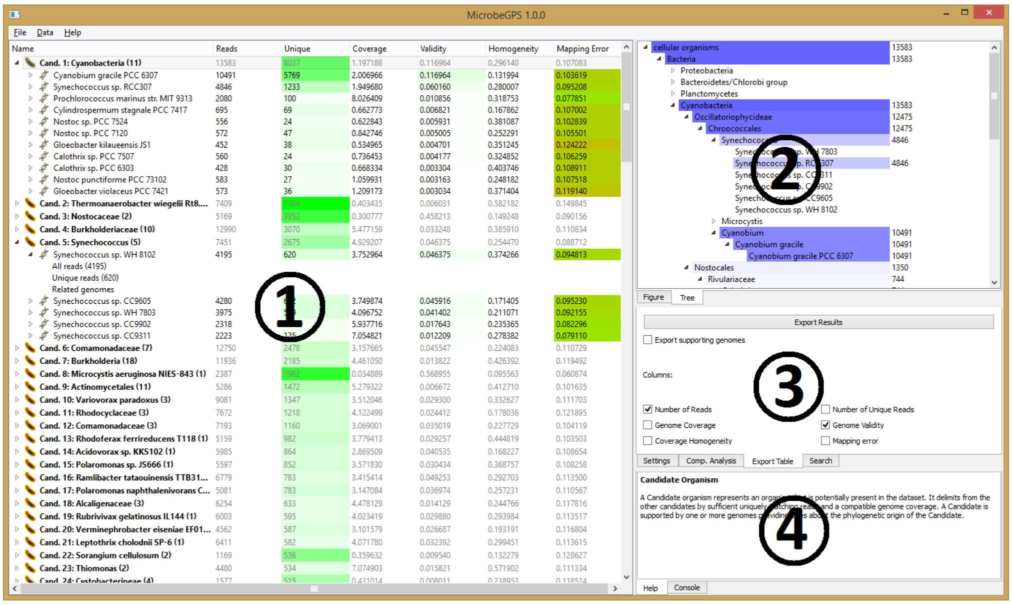
MicrobeGPS clusters reference genomes sharing high numbers of USR using a greedy strategy: a reference genome is compared to all existing clusters and the fraction of shared USR out of all shared reads are computed. If the reference genome shares sufficient USR with existing clusters, MicrobeGPS assigns the genome to the cluster with the highest overlap. Adjusting the minimum required overlap allows accounting for organisms with similar sequencing depth, as the number of USR shared with a reference genome of a different organism is typically low compared to all USR. If no suitable cluster can be identified based on the USR, MicrobeGPS searches for clusters with a very high fraction of shared reads. Finally, a new cluster is created if the genome cannot be assigned to any existing cluster based on the shared USR or all shared reads. This greedy clustering strategy has two advantages over existing clustering schemes: First, the number of clusters is not required beforehand as for k-means (MacQueen, 1967). Second, it is not necessary to calculate a full distance matrix between the reference genomes as for hierarchical clustering (Johnson, 1967), which is computationally expensive. Clustering can be sped up if scientific names or genome identifiers are available for the reference sequences. Then, taxonomically related genomes are clustered previous to other genomes, reducing the number of necessary comparisons. Compared to other traditional clustering approaches, such as hierarchical clustering, our approach was designed with focus on low run time rather than mathematical exactness.

Each cluster is built up of one or more reference genomes and represents one organism in the sample, identified by its sequencing depth. Discrimination between organisms with similar sequencing depth is achieved by requiring a minimum overlap of reference genomes for clustering. This overlap criterion ensures that only highly similar organisms with similar sequencing depths are at risk to be falsely regarded as one organism. The distances between the identified candidate organisms and their associated reference genomes are given by the genome validity scores and allow estimating the taxonomic identity of the organism.

The clustering step has quadratic complexity in the number of reference genomes for the worst case, when each reference genome is used to create a new cluster. However, with reasonable choice of parameters (e.g., default), the complexity lies between linear and quadratic in practice, depending on the structure of the reference genome collection and the composition of the microbial community. With the number of reads M and the number of reference genomes N, the upper bound of the computational complexity of MicrobeGPS is $O(MN^2)$. However, due to the greedy algorithm, the complexity approaches $O(MN)$ in practice.

4.2.3. Result visualization

We implemented the described algorithm as a platform independent Python program (Van Rossum and Drake Jr, 1995). To maximize usability, we created a graphical user interface (GUI) that allows to run the whole program with few mouse clicks and visualizes the results. The GUI features a step-by-step guide through the complete analysis pipeline, where the user can adjust all relevant parameters. Default values are set for each parameter to calculate conservative community composition estimates, i.e. the algorithm discards as few reference genomes as possible. In order to reduce the run time and to increase the specificity of the results, the user can tune the parameters manually. Single analysis steps can be repeated using different parameters without rerunning all preceding steps. The main window (Figure 4.2) visualizes the results and allows the user to inspect the detected candidate organisms in detail. It is divided into four parts (numbering according to Figure 4.2): The **data panel** (1) lists all organisms identified by MicrobeGPS. The list provides information about all quantitative measures of each candidate and is ordered descendingly by the number of unique reads. Color coding highlights particularly trustworthy candidates. Each candidate can be expanded to show the list of supporting reference genomes and the corresponding measures. Furthermore, MicrobeGPS can calculate lists of the mapped reads, unique reads and other genomes with shared reads for



Name	Reads	Unique	Coverage	Validity	Homogeneity	Mapping Error
Cand. 1: Cyanobacteria (11)	13583	8037	1.197188	0.116964	0.296140	0.107083
Cyanobium gracile PCC 6307	10491	5769	2.006966	0.116964	0.131994	0.103619
Synechococcus sp. RCC307	4846	1233	1.949680	0.060160	0.280007	0.095208
Prochlorococcus marinus str. MIT 9313	2080	100	8.026409	0.010856	0.318753	0.077851
Cylindrospermum stagnale PCC 7417	695	69	0.662773	0.006821	0.167862	0.107002
Nostoc sp. PCC 7524	556	24	0.622843	0.005931	0.381087	0.102839
Nostoc sp. PCC 7120	572	47	0.842746	0.005005	0.252291	0.105501
Gloeobacter kilaueensis JS1	452	38	0.534965	0.004701	0.351245	0.124222
Calothrix sp. PCC 7507	560	24	0.736453	0.004177	0.324852	0.106259
Calothrix sp. PCC 6303	428	30	0.668334	0.003304	0.403746	0.108911
Nostoc punctiforme PCC 73102	583	27	1.059931	0.003163	0.248182	0.107518
Gloeobacter violaceus PCC 7421	573	36	1.209173	0.003034	0.371404	0.119140
Cand. 2: Thermoanaerobacter wiegelii Rt8....	7409	7304	0.403435	0.006031	0.582182	0.149845
Cand. 3: Nostocaceae (2)	5169	3952	0.300777	0.458213	0.149248	0.090156
Cand. 4: Burkholderiaceae (10)	12990	3070	5.477159	0.033248	0.385910	0.110834
Cand. 5: Synechococcus (5)	7451	2675	4.929207	0.046375	0.254470	0.088712
Synechococcus sp. WH 8102	4195	620	3.752964	0.046375	0.374266	0.094813
All reads (4195)						
Unique reads (620)						
Related genomes						
Synechococcus sp. CC9605	4280	[illegible]	3.749874	0.045916	0.171405	0.095230
Synechococcus sp. WH 7803	3975	[illegible]	4.096752	0.041402	0.211071	0.092155
Synechococcus sp. CC9902	2318	[illegible]	5.937716	0.017643	0.235365	0.082296
Synechococcus sp. CC9311	2223	125	7.054821	0.012209	0.278382	0.079110
Cand. 6: Comamonadaceae (7)	12750	2478	3.157665	0.045547	0.224083	0.110729
Cand. 7: Burkholderia (18)	11936	2185	4.461050	0.013822	0.426392	0.119492
Cand. 8: Microcystis aeruginosa NIES-843 (1)	2387	1962	0.034889	0.568955	0.095563	0.060874
Cand. 9: Actinomycetales (11)	5286	1472	5.279322	0.006672	0.412710	0.101635
Cand. 10: Variovorax paradoxus (3)	9081	1347	3.512046	0.029300	0.332627	0.111703
Cand. 11: Rhodocyclaceae (3)	7672	1218	4.122499	0.024412	0.178036	0.121895
Cand. 12: Comamonadaceae (3)	7193	1160	3.069001	0.035019	0.227729	0.104119
Cand. 13: Rhodoferax ferrireducens T118 (1)	5159	982	3.779413	0.029257	0.444819	0.103503
Cand. 14: Acidovorax sp. KKS102 (1)	5985	864	2.869509	0.040535	0.168227	0.108654
Cand. 15: Polaromonas sp. JS666 (1)	5597	852	3.571830	0.030434	0.368757	0.108258
Cand. 16: Ramlibacter tataouinensis TTB31...	6779	783	3.415414	0.049253	0.292703	0.113500
Cand. 17: Polaromonas naphthalenivorans C...	5081	783	3.147084	0.036974	0.257231	0.110567
Cand. 18: Alcaligenaceae (3)	6254	633	4.478129	0.014129	0.244766	0.117816
Cand. 19: Rubrivivax gelatinosus IL144 (1)	6003	595	4.023419	0.029880	0.293984	0.113517
Cand. 20: Verminephrobacter eiseniae EF01...	4562	587	3.101579	0.026687	0.193191	0.116804
Cand. 21: Leptothrix cholodnii SP-6 (1)	6411	582	4.071780	0.032392	0.299451	0.113615
Cand. 22: Sorangium cellulosum (2)	1169	536	0.359632	0.009540	0.132279	0.128627
Cand. 23: Thiomonas (2)	4480	534	7.074903	0.015821	0.571902	0.111334
Cand. 24: Cystobacterineae (4)	1577	515	0.431014	0.008011	0.238953	0.118514

cellular organisms	13583
Bacteria	13583
Proteobacteria	
Bacteroidetes/Chlorobi group	
Planctomycetes	
Cyanobacteria	13583
Oscillatoriophycideae	12475
Chroococcales	12475
Synechococcus	4846
Synechococcus sp. WH 7803	
Synechococcus sp. RCC307	4846
Synechococcus sp. CC9311	
Synechococcus sp. CC9902	
Synechococcus sp. CC9605	
Synechococcus sp. WH 8102	
Microcystis	
Cyanobium	10491
Cyanobium gracile	10491
Cyanobium gracile PCC 6307	10491
Nostocales	1350
Rivulariaceae	744

□ Export supporting genomes

Columns:

☑ Number of Reads
□ Number of Unique Reads
□ Genome Coverage
☑ Genome Validity
□ Coverage Homogeneity
□ Mapping error

Candidate Organism

A Candidate organism represents an organism that is potentially present in the dataset. It delimits from the other candidates by sufficient uniquely matching reads and a compatible genome coverage. A Candidate is supported by one or more genomes providing clues about the phylogenetic origin of the Candidate.

Figure 4.2.: Screenshot of MicrobeGPS main window. After running the MicrobeGPS algorithm, the tool visualizes the results in four panels: (1) Data panel. (2) Visualization/taxonomy tree panel. (3) Settings/modules panel. (4) Console panel.

each reference genome on the fly, allowing the user to browse the metagenome interactively. The **visualization/taxonomy tree panel** (2) shows information about the currently selected candidate in the data panel. It can either show interactive charts (depending on the current selection) or the location of the currently selected element in the taxonomic tree. The **settings/modules panel** (3) contains one settings tab and optionally further module tabs. While the settings tab allows to control basic settings of the program, the module tabs represent external tools that can be added by the user. The **console panel** (4) displays information about the currently selected item in the GUI and therefore increases the usability of the tool.

4.2.4. Availability

The platform independent Python source code is freely available from `https://sourceforge.net/projects/microbegps/`. Furthermore, we created standalone versions for Linux and Windows platforms.

To obtain full functionality, MicrobeGPS requires additional information about the reference genome database and the taxonomic affiliation of the genomes. We already include this information for the NCBI bacterial genomes database. Instructions for using MicrobeGPS for different reference databases can be found in the documentation.

4.3. Experiments and results

In this section, we present experimental results that demonstrate on the one hand that MicrobeGPS provides more accurate community composition estimates than previous approaches and on the other hand that MicrobeGPS provides a new quality in analyzing microbial communities. The former is demonstrated on artificial metagenomic data allowing comparison of different tools to a gold standard. For the latter, we reanalyze two different real microbial communities that are challenging for other tools and therefore highlight the benefits of MicrobeGPS.

4.3.1. Comparison on artificial mock community

We compared MicrobeGPS to other methods on the Mock Community (MC) metagenomic dataset provided by the Human Microbiome Project (HMP) consortium (Methé et al., 2012) that is an in vitro synthetic mixture of 21 known bacteria created to establish sequencing standards within the HMP. Originally, two mixtures were created, one with even abundance profile, i.e. all organisms are about equally abundant, and one with staggered abundance profile, where the abundances spread over several orders of magnitude. We used the staggered data set because it resembles a natural abundance distribution and is more challenging for the tools. The Illumina sequencing data of the staggered abundance distribution mixture are available from NCBI SRA (accession SRX055381) and contain 7.9 million 75 bp reads. We analyzed the community composition of the MC dataset with the state-of-the-art methods Pathoscope (Francis et al., 2013) and MetaPhlAn (Segata et al., 2012), and compared results with MicrobeGPS. Both MicrobeGPS and Pathoscope build upon the alignment of the metagenomic reads to a database of microbial reference genomes. As reference genome database, we used the HMP (Nelson et al., 2010) and the NCBI (Pruitt et al., 2007, 2014) bacterial reference genomes for both tools. MetaPhlAn comes with its own curated set of marker genes for identification. The reads were mapped with Bowtie 2 (version 2.1.0 with parameters `--fast -p 12 --no-unal -k 60`) to both reference databases for further processing with Pathoscope 1.0 and MicrobeGPS 1.0.0 (Linux binaries). We used default parameters for Pathoscope and configured MetaPhlAn 1.7.7 to report species level abundance estimates. MicrobeGPS was configured to discard reads mapping to more than 50 references and consider only references with 10 or more unique reads. Pathoscope reported abundance estimates for 828 (NCBI) and 690 (HMP) reference sequences, MetaPhlAn reported abundance estimates for 32 species in the precomposed database. MicrobeGPS reported 24 (HMP) and 23 (NCBI) candidate organisms. For comparison, we selected the N most abundant identifications of each tool and compared the results to the ground truth. For each $N \leq 60$, we calculated the sensitivity, false positive rate and precision for the selected set. The receiver operator characteristic (ROC) for each setup was obtained by plotting the sensitivity against the false positive rate. Additionally, we calculated the F-measure (Van Rijsbergen, 1979) – the harmonic mean of sensitivity and precision – for all N and report its maximum for each setup.

One benefit of MicrobeGPS over Pathoscope becomes directly apparent when looking at the program output: MicrobeGPS reports a brief list of candidate organisms that are assumed to be present in the sample. Pathoscope reports substantially longer lists of identified (and sometimes highly similar) reference genomes, for which it is not clear if they are all in the sample or only some of them; therefore, we restricted analysis to the 60 most abundant organisms for the comparison. For both NCBI and HMP references, MicrobeGPS reports better results than Pathoscope, both in terms of F-measure and AUC (see Figure 1). Best performance is achieved with NCBI genomes (F=0.93, AUC=0.90), where MicrobeGPS correctly identified 19 among the first 20 reported organisms. MetaPhlAn (F=0.84, AUC=0.87) is comparable to MicrobeGPS with HMP reference genomes (F=0.84, AUC=0.84) and better than Pathoscope (F=0.71, AUC=0.80) while being significantly faster (2:19 min) than MicrobeGPS (31:37 min) and Pathoscope (15:08 min) on the same hardware.

The main error sources for MicrobeGPS are missing genomes in the reference databases and false positive identifications. For example, HMP lacks *Streptococcus pneumoniae* and *Deinococcus radiodurans* reference genomes. For *D. radiodurans*,

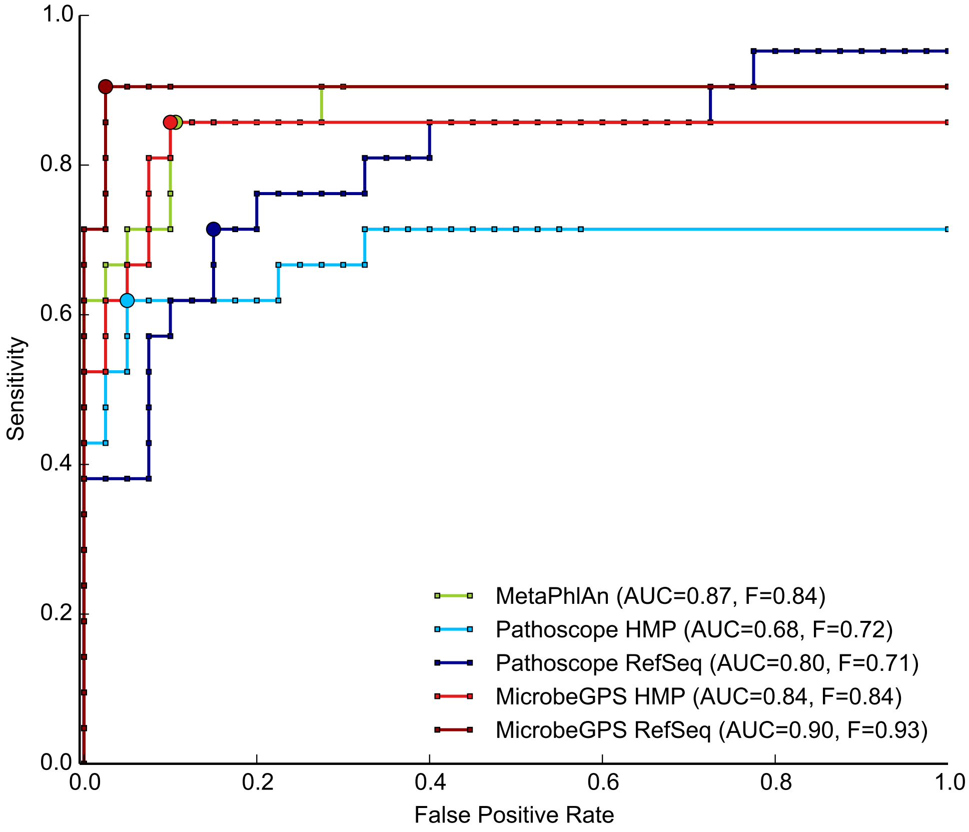


Figure 4.3.: Evaluation of taxonomic profiling tools on *in vitro* metagenomes. We compared MicrobeGPS to Pathoscope (Francis et al., 2013) and MetaPhlAn (Segata et al., 2012) on the *in vitro* HMP mock community dataset (Methé et al., 2012) with known composition. The number of true and false positive identifications among the top N ($N \leq 60$) reported organisms were counted and ROC curves were calculated. The large circle indicates the point of maximum F-measure for each method. MicrobeGPS in combination with the NCBI reference genomes provides the best results, both in terms of F-measure and AUC.

both MicrobeGPS and Pathoscope report the closest available relative, *Deinococcus deserti*. However, from the Pathoscope output it does not become clear that *D. deserti* is not a perfect match. In contrast, MicrobeGPS reports a low genome validity score (0.02) and high mapping error rate (0.12) here, indicating that the candidate is related but not identical to *D. deserti*. The number of false positive identifications for MicrobeGPS is lower than for Pathoscope or MetaPhlAn. Additionally, the false identifications receive low quality scores and can therefore be spotted easily in the graphical user interface. However, we observed one exception, *Stenotrophomonas maltophilia*, which was detected by all three tools and received a considerably high validity score (0.76) in MicrobeGPS, but was not listed in the MC dataset ground truth. Therefore, we speculate that a bacterium closely related to *S. maltophilia* was actually present in the sample, probably as contamination.

While several *Streptococci* were present in the dataset and reported by the tools, the very low abundant *S. pneumoniae* was neither detected nor explained by a closely related genome. Instead, MicrobeGPS assigned the *S. pneumoniae* genome to the *Streptococcus mutans* candidate such that it could only be identified as separate organism via manual inspection of the results. Although MetaPhlAn should in principle be able to recover such situations, it fails to make the score and only reports one out of three *Streptococci*.

Taken together, the experiment showed that MicrobeGPS is able to estimate microbial community compositions more accurately than previous approaches. In cases where the identified organism did not agree with the reference genomes, MicrobeGPS quantified the divergence between organism in the sample and reference genome with the genome validity score. MicrobeGPS therefore allows differentiating between cases where the suitable reference genome was available and cases where a distantly related genome had to be selected as representative. This differentiation is not possible using MetaPhlAn or Pathoscope.

4.3.2. Application to human gut microbiome

We further analyzed three human gut metagenomes with MicrobeGPS to evaluate its potential on real data. The datasets (IDs 1122, 2535, 2638) were acquired from diarrhea patients during the Shiga-toxigenic *Escherichia coli* (STEC) outbreak in Germany, 2011 (Loman et al., 2013) and were downloaded from NCBI SRA (accessions ERX237457, ERX234998, ERX237461). The datasets contained between 332,257 and 879,176 paired-end reads with length 2×151 bp. Clinical tests identified a *Clostridium difficile* infection in dataset 1122 and high abundances of STEC in datasets 2535 and 2638. We reanalyzed the three datasets and used the NCBI bacterial genomes as reference database. For testing purposes, we also removed the STEC reference sequence from the set of reference genomes. The reads were mapped with Bowtie 2 using the same settings as in the first experiment. MicrobeGPS was configured to report genomes with at least one uniquely matching read.

MicrobeGPS presented 41 candidate organisms for dataset 1122, indicating a higher complexity than the MC dataset. In accordance with the diagnosed *C. difficile* infection, we found a candidate supported by five *C. difficile* genomes with the closest related reference genome being *C. difficile* Bl9. However, the reported genome validity is low (0.23) and suggests that the infecting strain differs from all strains available in the database. Higher genome validity is reported for candidates supported by *Alistipes finegoldii* (0.81) or *Bacteroides vulgatus* (0.79). The most abundant (*A. finegoldii*, 232,090 reads) and the least abundant (*Eggerthella lenta*, 80 reads) highly relevant candidates differ by a factor of 2,901 in their abundance. Nevertheless, the typical gut bacterium *E. lenta* is a highly valid candidate since most reads mapped uniquely and the reads are homogeneously spread over the genome (Kolmogorov-Smirnov test p-value below 0.05).

The most abundant candidate in datasets 2535 and 2638 was closely related to *E. coli,* however, none of the supporting references could be identified as perfect match since STEC was not part of the reference database. When added to the database, STEC was the most abundant supporting genome and was identified as almost perfect match (genome validity: 0.99), as shown in Figure 4.4. This is contrasting to dataset 1122, where the *E. coli* candidate has lower abundance and the supporting reference genomes had lower validity; the STEC genomes were not reported as supporting genomes for the *E. coli* candidate. Since *E. coli* is a common human gut bacterium, we expect that the non-pathogenic *E. coli* was also present in the datasets 2535 and 2638, but was assigned to the same candidate as the highly abundant STEC due to the high sequence similarity. This challenging scenario could benefit from follow-up analyses with tools specialized to differentiating highly similar organisms in metagenomic samples, such as GASiC (Lindner and Renard, 2013). Although no *C. difficile* infection was diagnosed for both samples, MicrobeGPS identified candidates mainly supported by *C. difficile* in both datasets, similarly to dataset 1122. This may not necessarily be wrong since *C. difficile* is a common gut bacterium, but shows that using this set of genomes does not allow the distinction between pathogenic and non-pathogenic *C. difficile* strains.

Altogether, this experiment demonstrates the ability of MicrobeGPS to handle real data and to identify the correct strain if sufficient reference data is available. Otherwise, MicrobeGPS locates the candidate as good as possible using the available genomes.

4.3.3. Reanalysis of Lake Lanier metagenome

Our method is not restricted to the analysis of human-associated microbiomes and is particularly suitable for the exploration of communities with few known reference genomes. Therefore, we reanalyzed the Lake Lanier freshwater metagenome

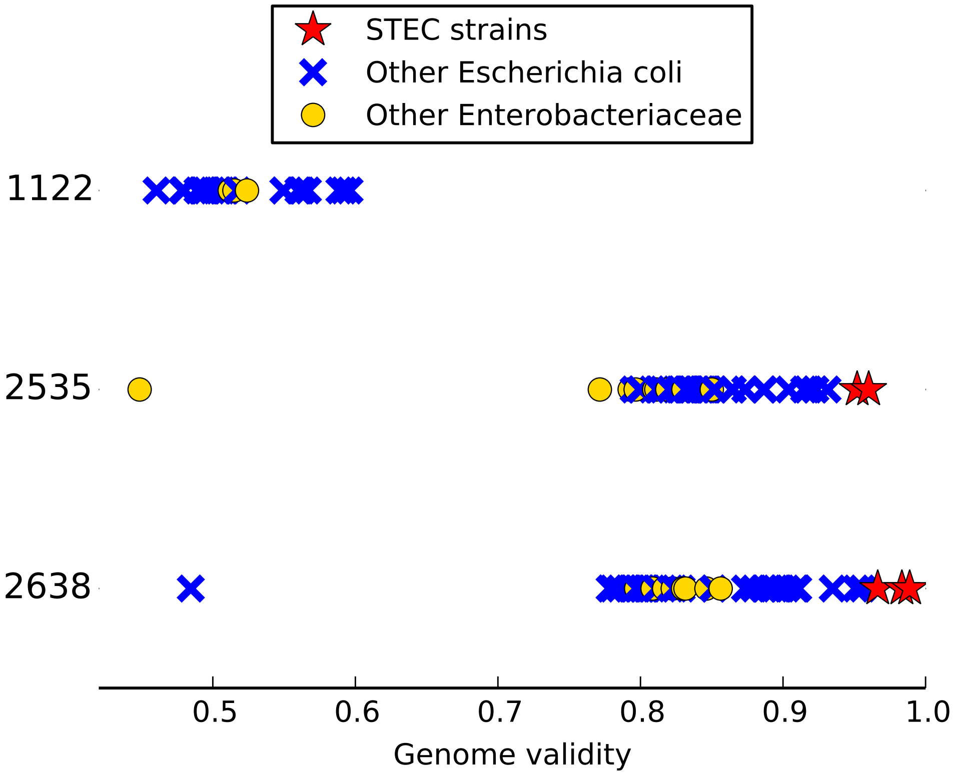


Figure 4.4.: Identification of STEC strains in human gut metagenomes. Three human gut metagenome datasets (Loman et al., 2013), one of them without (1122) and two with STEC infection (2535, 2638), were analyzed with MicrobeGPS. MicrobeGPS identified one *E. coli*–related candidate in each dataset. In the datasets with STEC infection, MicrobeGPS finds the highest validity scores for different STEC strains, showing almost perfect agreement between the sequenced organism and reference genome. In dataset 1122, MicrobeGPS only finds other *E. coli* strains with much lower validity scores, rightly indicating that this sample was not infected with STEC.

datasets (AUG1, AUG2, SEPT, NOV) (Oh et al., 2011; Poretsky et al., 2014), a series of four datasets from the same location at different time points. In the original study, the authors assessed the community composition by means of 16S rRNA gene amplicon sequencing and by assembling the metagenomic sequencing reads into contigs and subsequently identifying genes in the sequences. These genes were then searched in databases of all sequenced bacterial and archaeal genomes. Here, MicrobeGPS offers a third way, since the metagenomic sequencing reads are used directly to infer the composition of the community. Therefore, we downloaded the published datasets from NCBI SRA (accessions SRX039150, SRX039152, SRX039381, SRX039382). The dataset sizes were between 13.6 million (SEPT) and 17.1 million (NOV) 2×101 bp paired-end reads. The NCBI bacterial genomes served as reference database for MicrobeGPS and reads were mapped with Bowtie 2 using the same settings as in the previous experiments. MicrobeGPS was used with default settings.

These datasets posed a particularly challenging problem to MicrobeGPS since only about 1% of the reads in each dataset could be mapped to the NCBI bacterial genomes database, indicating that the freshwater metagenome is still far less studied than other environments such as the human microbiome.

MicrobeGPS identified between 165 and 238 candidate organisms per dataset from 15 bacterial phyla (see Figure 4.5), indicating a much higher community complexity than the MC or STEC datasets. However, the MicrobeGPS quality measures clearly pointed out that the majority of detected candidates diverge strongly from the genomic material in the sample: The highest observed validity score was 0.45 (*Anabaena* Sp. 90 in AUG1), with the majority of scores being below 0.05. This indicates that the available reference genome sequences are not suitable for species accurate identification and the low scores warn the user that each individual candidate should be treated with caution.

Nevertheless, when looking at a more coarse level, we obtained more meaningful results and our observations largely agreed with the results presented in the original studies. Both approaches reported *Proteobacteria* as the most abundant phylum and all originally reported phyla were also identified by MicrobeGPS. Our overall estimated abundances show patterns similar to the assembled contigs approach presented in Poretsky et al. (2014) (Figure S4). Also the temporal variations could be reproduced with MicrobeGPS: while the relative abundances of *Proteobacteria*, *Actinobacteria* and *Verrucomicrobia* were relatively stable over all datasets, we also observed a significant drop of abundance for *Cyanobacteria* in the NOV dataset and an increase of *Bacteroidetes*. *Planctomycetes*, which were hardly detected in the 16S analysis, show highest abundances in the SEPT dataset in both the assembly-based and MicrobeGPS analysis.

However, we also observed that the structure of the reference genome database influenced the relative abundance quantification of the different phyla. The 16S

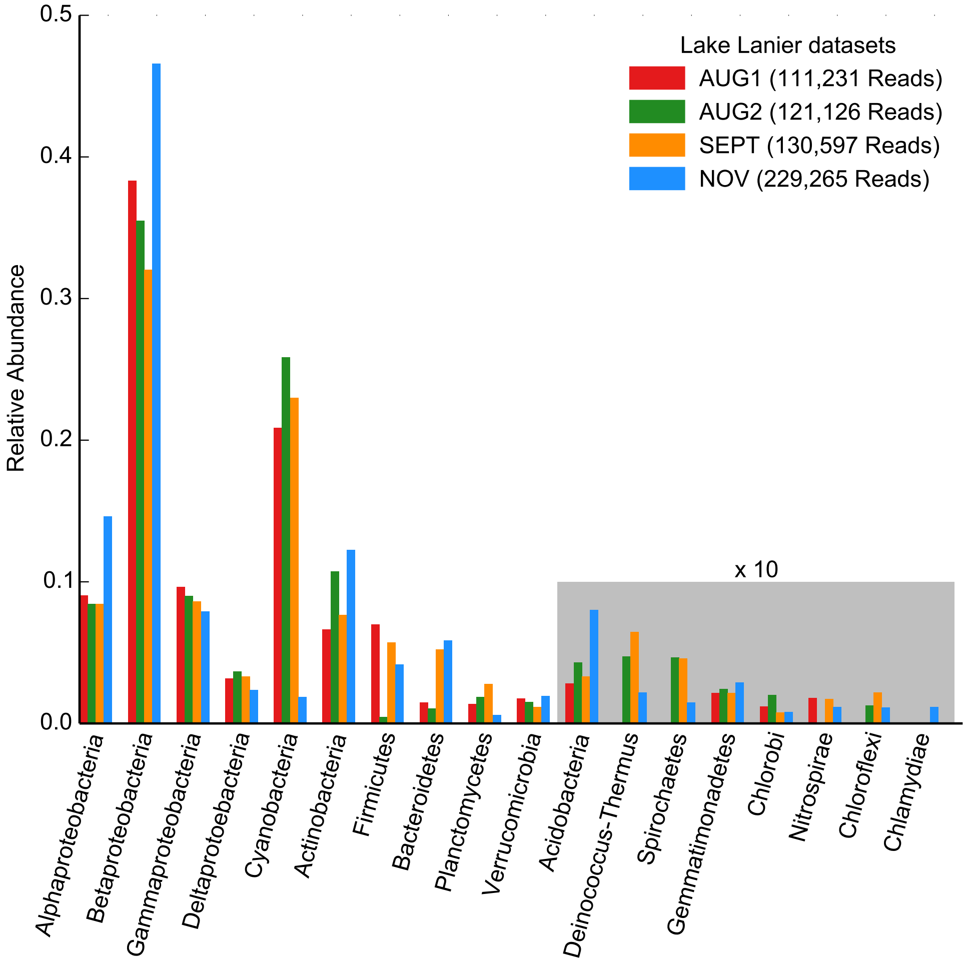


Figure 4.5.: High level taxonomic analysis of freshwater metagenome. We estimated the composition of the four Lake Lanier metagenomic time series datasets (Oh et al., 2011) on the phylum level (*Proteobacteria* were expanded to the class level), since species-accurate identification was not possible due to low coverage depth and insufficient similarity to reference genomes. MicrobeGPS detected all phyla reported in the original study based on 16S amplicons and assembled contigs with comparable abundances and showed temporal abundance shifts similar to the other approaches.

approach detected higher fractions of *Cyanobacteria* and *Verrucomicrobia* than MicrobeGPS, but lower fractions of *Proteobacteria.* We attribute this skew to the uneven representation of the phyla in the reference database: the database contained 1104 *Proteobacteria* genomes, but only 71 *Cyanobacteria* and 4 *Verrucomicrobia* genomes. This limitation is inherent to reference-based approaches and was also observed in the assembly-based analysis (Poretsky et al., 2014).

For comparison, we profiled the four Lake Lanier datasets with MetaPhlAn and compared the results to the analysis in the original study and to MicrobeGPS. We used the reference sequence database provided by MetaPhlAn in combination with the Bowtie 2 mapper, as suggested in the MetaPhlAn manual. MetaPhlAn was configured to report abundances for all taxonomic levels (`--tax_lev=a`). All datasets were analyzed separately.

Overall, MetaPhlAn reported fewer taxa than MicrobeGPS and the original study. Between six and nine different species per dataset from three to five different phyla were reported (see Table 4.1).

Table 4.1.: Reanalysis of the Lake Lanier metagenomic datasets with MetaPhlAn. The numbers are the estimated percentage abundances on the phylum level. A dash indicates that the phylum was not detected by MetaPhlAn in the dataset.

Dataset	AUG1	AUG2	SEPT	NOV
Cyanobacteria	73.3	59.9	59.3	3.6
Proteobacteria	14.2	19.9	30.7	74.8
Bacteroidetes	9.5	10.1	10	14.9
Chlamydiae	3	-	-	1.7
Actinobacteria	-	7.1	-	5
Chloroflexi	-	3	-	-

In contrast to MicrobeGPS, MetaPhlAn detected only 3–5 phyla and 6–9 different species with varying abundances in the datasets without providing any information about the accuracy of the results. These results suggest a community complexity far lower than what one would expect for a freshwater metagenome and than what was observed in the original study and the MicrobeGPS analysis. Furthermore, most MetaPhlAn abundance estimates do not coincide with the other approaches. For example, *Cyanobacteria* abundance was estimated between 59% and 74% in the first three datasets, whereas *Actinobacteria* were only detected in AUG2 and NOV and *Verrucomicrobia* were not found at all.

This experiment showed that MicrobeGPS estimates microbial community compositions similarly to manual assembly and 16S based approaches even on challenging datasets, where other methods have severe problems.

4.4. Discussion of results

Reference-based taxonomic profiling is currently the most specific way of assessing the composition of microbial communities as it allows – in principle – strain-accurate identification of organisms. However, current tools tend to overestimate their accuracy as they report a specific strain or species as present when the dataset in fact contains a previously unknown, related organism. We demonstrated that in practice the user has no possibility to differentiate between cases with strain accurate identification and identification of a related organism.

Therefore, we introduced MicrobeGPS, a novel approach to the taxonomic profiling problem, which is beneficial in two ways. First, MicrobeGPS provides more accurate community composition estimates than other reference-based methods. Second and more importantly, MicrobeGPS calculates quality measures for each detected candidate organism, allowing the user to judge the quality and reliability of the identification. Organisms that are not represented in the reference database can be evaluated critically. The supporting reference genomes provide valuable information about their taxonomic affiliation. Here, MicrobeGPS is far ahead of related tools that only report estimated abundances for hundreds of genomes without any other quality information, and thus contributes to the trustworthiness of already powerful reference-based metagenomic analyses.

The genome validity turned out as a valuable measure of sequence disagreement, quantifying similarity between the organism in the sample and the available reference genome. Compared to the average nucleotide identity (ANI, see Konstantinidis and Tiedje, 2005), a common measure for genomic similarity that only operates on the regions shared between genomes, the genome validity is designed to compare an existing genome to a set of sequencing reads and is particularly applicable when the genome shares only small regions with the sequenced organism. When the only similarity between the sequenced organism and the reference genomes is a single gene transferred with only very few mutations from one of the species available as reference to the organism in the sample, the ANI would measure a high identity as it only considers regions shared between the genomes. In contrast, the genome validty would be close to zero since only a very small part of the whole reference genome is actually present in the dataset.

Our experiments demonstrated that MicrobeGPS provides more concise and accurate results than previous approaches on the *in vitro* microbial community data with known composition. On real datasets, MicrobeGPS is able to provide strain accurate identifications for well-studied species (such as *E. coli*) with sufficient reference genomes and at the same time coarse identification of organisms where no perfectly matching reference genome is available. When no accurate identification on low taxonomic levels is possible, as in the Lake Lanier experiment, the user can analyze the dataset on a higher level (e.g., *phylum*). Here, MicrobeGPS produces community composition estimates that are comparable to established approaches involving assembly of the sequence reads into contigs or the analysis of 16S rRNA amplicon data.

The benefits of MicrobeGPS come at the cost of an increased run time and memory footprint. On the same hardware Pathoscope is about 50% faster than MicrobeGPS. Further, the interactive graphical user interface requires that all mapping information is kept in memory. Memory consumption strongly depends on the number of shared read matches; in our experiments we observed peak memory consumptions between 1 GB (Lake Lanier datasets) and 38 GB (MC dataset). Thus, smaller datasets can be processed on a laptop computer while larger datasets may require a larger (e.g., workstation) computer.

Despite the significant improvement on the status quo, MicrobeGPS still suffers from a problem common to all reference-based taxonomic profiling approaches – the influence of the reference database. Especially in the two real data experiments, we saw that MicrobeGPS is able to identify on species or strain level when the database contains sufficient reference genomes, but only finds distantly related organisms if no suitable reference genome is available. Furthermore, unbalanced databases skew the abundance estimates, where taxonomic groups with many reference genomes appear more abundant than groups with few reference genomes. However, MicrobeGPS reduces the influence of the reference genome database in comparison to other recent methods and makes the problem of missing reference genomes tractable.

While reference based-taxonomic profiling could previously only be applied to microbial communities with known structure, where the reference genomes of the most dominant organisms are known, our experiments demonstrate that MicrobeGPS can also be used for taxonomic profiling of less explored communities such as the Lake Lanier metagenome. This result in combination with the growing databases of reference genomes could pave the way for the application of reference-based taxonomic profiling beyond applications such as clinical diagnostics (Francis et al., 2013). Compared to traditional methods such as 16S rRNA profiling or lowest common ancestor approaches, we showed that our method has the advantage that strain specific identification is possible in principle. However, when reference genomes are missing, MicrobeGPS can fall back to more coarse identifications. As for all existing reference-based methods, our approach may still not be applicable to microbial communities with extremely complex composition and only few known genomes, such as soil metagenomes.

5. Summary and conclusion

NGS-based analysis of metagenomic samples has been established in the last years as the method of choice for the analysis of environmental microbial communities. In particular whole genome sequencing has proven to be useful for the unbiased analysis of metagenomic samples. The high throughput of modern NGS devices made it impossible to analyze the experimental data manually, therefore the demand for algorithmic and computational approaches has grown massively in the last decade. The main challenges in metagenomics include the taxonomic identification and quantification of organisms in a sample (taxonomic profiling), the assembly of novel genomes from metagenomic sequencing reads, and the functional and biological interpretation of the data. Particularly taxonomic profiling – one of the most fundamental tasks in metagenomics – can not be considered as a solved problem: reference-based taxonomic profiling, which is the most precise profiling approach, still struggles with highly similar organisms and incomplete or skewed genome databases.

In this work, we addressed current challenges in reference-based taxonomic profiling of metagenomic samples. In chapter 2, we presented the GASiC method that allows the identification and relative quantification of highly similar organisms in metagenomic datasets when the reference genomes are available. We also showed that the idea behind GASiC is applicable in a metaproteomic setting. A novel framework for the analysis of coverage depth profiles was presented in chapter 3; as one immediate application, we developed the genome validity score, which measures the similarity of a reference genome to the true organism in the sample. Based on the validity score, we developed MicrobeGPS (chapter 4), a novel taxonomic profiling approach that can handle incomplete and skewed reference genome databases.

Especially for clinical applications (e.g., metagenomics-based diagnostics) accurate detection and identification results are indispensable. However, since previous metagenomic methods were not able to differentiate between highly similar strains in metagenomic samples, not to mention the simultaneous identification and relative quantification of multiple highly similar genomes (e.g., pathogenic and non-pathogenic strains) in one sample, metagenomics has rarely been used for clinical applications. Our method GASiC approaches the problem, it allows to accurately quantify even highly similar organisms in metagenomic samples. To achieve this, GASiC first estimates the similarities between the available reference genomes by simulating reads for each reference genome and mapping these reads to all other genomes. Based on these similarities, GASiC corrects the number of reads obtained by mapping the metagenomic reads to the reference genomes. Thus, corrected abundance estimates for each genome are provided. Our experiments demonstrated that correct abundance estimation is even possible for reference genomes with 96% sequence identity and that the abundance estimates are more accurate than by any other method. The experiment, in which we used *in silico* mixtures of the non-pathogenic *E. coli* strain DH10B and the pathogenic STEC strain, highlighted the potential use of the method for strain accurate identification and quantification of pathogens in clinical applications. However, also other applications requiring strain accurate quantification can benefit from GASiC, as demonstrated in the experiments: the method can both handle very simple communities consisting of few, highly similar organisms, as well as complex communities consisting of more than one hundred organisms. Furthermore, it is largely independent of the type of input data and mapper characteristics.

The idea behind GASiC, using simulations to estimate the similarities between reference genomes and correcting the observed abundances correspondingly, worked out well in metagenomics. However, this is by no means limited to metagenomics and we successfully transferred the concept to metaproteomics, where we developed the Pipasic method. The mass spectra can be seen as analogue to the sequencing reads and the reference proteomes represent the reference genomes; together with mass spectra simulators and spectrum search tools, the complete GASiC pipeline could be rebuilt for metaproteomics. For practical purposes, we replaced the simulation and spectrum searching step for the similarity estimation by a faster weighted string searching approach. Our experiments showed that Pipasic is as accurate as its metagenomic counterpart and works reliably on real metaproteomic datasets.

Furthermore, we developed a framework for fitting discrete probability distributions to genome coverage depth histograms (GCP), which was presented in chapter 3. When mapping NGS reads of an organism to its reference genome, we would expect to observe a single Poisson distribution in the GCP given ideal conditions (even sequencing depth, no GC-bias, no sequencing errors, no repeats, etc.). However, we obtained more complex distributions when using real data, which can be modeled by a mixture of multiple distributions. With our framework, we were able to fit the observed GCPs using both established and new distributions tailored to the needs in GCPs. The framework offers a wide range of possible applications. For example, the parameters of the fitted distributions provide information about the sequencing depth, the degree of fragmentation of the reference sequence, or the number of organisms in the sample that contributed to the coverage depth profile of the genome. As one immediate application, we introduced the genome validity score, which measures the distance of a reference genome to the closest related organism in a metagenomic dataset. We demonstrated that this score can be calculated robustly by our framework, even for extremely low sequencing depths of approximately 0.01×. The validity score proved to be useful for quality control in metagenomics: by reanalyzing previously published metagenomic data, we found that a large fraction of the representative genomes used in the original study were not suitable to represent the organisms in the dataset. This shows that the genome validity is a useful measure for the fidelity of a reference genome for a given dataset. We recommend calculating the genome validity as a default quality control step when metagenomic reads are mapped to reference genomes. This helps identifying possible errors or prevent misinterpretation due to unfeasible reference genomes reliably at an early stage of the analysis.

Based on the genome validity score and the framework developed in chapter 3, we created a novel taxonomic profiling approach, called MicrobeGPS. The idea behind MicrobeGPS differs from all previously published taxonomic profiling tools, because the algorithm focuses on the organism in the sample rather than on the available reference genomes. We demonstrated on multiple datasets that MicrobeGPS clearly outperforms the detection rates of other methods both in terms of sensitivity and specificity. The experiments proved that MicrobeGPS can overcome the typical difficulties in metagenomic analyses: Both very high abundant and very low abundant organisms were correctly detected in the datasets, with abundances differing by a factor of 2901 in one observed case. Furthermore, when no suitable reference genome was available, MicrobeGPS recruited the closest related available reference genomes to describe the organism in the sample. The distance of these supporting reference genomes to the organism was measured by the genome validity, such that it was immediately visible to the user that the true reference genome was missing. Skews in the databases of available reference genomes were compensated by creating clusters of different sizes representing the organisms in the sample: we observed clusters with sizes between one single genome and 80 genomes, depending on the number of available reference genomes in the taxonomic vicinity of the organism. A unique feature of MicrobeGPS is the genome validity that measures the distance of every reference genome to the candidate organism it is supporting. Together with the graphical user interface and the additional quality measures, the user has the possibility to spot unreliable candidates that may require further analysis. This has not been possible so far, since previous tools either reported plain lists of estimated numbers of assigned reads or abundances, or did not provide as meaningful results for each genome as MicrobeGPS does.

Future research While the approaches presented in this thesis describe possible solutions for major problems in metagenomic data analysis and taxonomic profiling, our experiments and results also revealed that on the one hand the presented solutions are not applicable in all situations and may require further development, on the other hand new problems arose and showed that there is more research required on the path to robust and reliable taxonomic profiling.

While GASiC showed high accuracy in abundance prediction at reasonable runtime for few, highly similar reference genomes, we also observed that the approach cannot be scaled reasonably to large reference genome databases since the run time for simulating the similarity matrix grows quadratically; in the experiment with 113 reference genomes, the calculation took about 2 days. Reference genome databases consisting of thousands of genomes require different approaches. One possible solution is based on the observation that most entries of the similarity matrix are exactly or close to zero, representing highly dissimilar pairs of genomes. Filtering these pairs by setting them directly to zero could greatly improve the runtime. As a possible approach, one could make use of taxonomic information associated with the reference genomes. For each reference genome, the similarity to other genomes could be calculated in an ordered way, starting with the closest related organisms and then proceeding with the more distantly related ones. Similarity estimation could be stopped when the observed similarity drops below a user defined threshold close to zero. Setting the remaining similarities to zero could potentially reduce the number of similarity estimations dramatically. When no taxonomic information is available, similarity estimation could be sped up by estimating the similarity using a fast (but in most cases inaccurate) comparison method, e.g., based on k-mers, and only calculating the simulation-based similarity if the genomes are sufficiently similar according to the k-mer similarity. While GASiCs main purpose is identifying and quantifying genomes present in a dataset, it does not provide classification on the single read level. This means, after estimating the abundances, the information which read belongs to which reference genome is not given. Other tools such as Pathoscope, exclusively rely on read reassignment. Since this information can be used for applications such as variant discovery, it could be beneficial to use the GASiC abundance estimates to reassign the reads to the most likely reference genome. However, this would require developing a new read assignment model that incorporates the GASiC abundances. Finally, we observed that the major drawback of GASiC is its dependence on matching reference genomes: the abundance estimates are skewed when only a closely related genome is available instead of the exactly matching genome. To widen GASiCs scope, one could apply a strategy similarly to the weighted string matching used in Pipasic. The areas that are actually covered by reads could be assigned a higher weight in the similarity calculation. On the one hand, this strategy has the potential to improve the abundance estimates, as we have seen for Pipasic (Penzlin et al., 2014), and possibly relax the need for an exactly matching reference genome, on the other hand quality control could be improved by assigning confidence estimates to single genomes based on the genome validity.

As we have seen in chapter 3, the GCPs observed in experiments with real data can be very complex and involve multiple different distributions. While it turned out to be sufficient to use a fixed set of distributions for calculating the genome validity, the number and type of suitable probability distributions may vary from case to case in other applications. Therefore, a general method that finds the minimal set of probability distributions, being sufficient to describe the GCP, would be desirable. While it is possible to automatically determine the optimal number of Gaussian distributions in a mixture model for the continuous case (Zhang et al., 2003), this has not yet been applied to discrete distributions of different types. Further research in that direction would not only contribute to the general understanding and applications of the EM algorithm, but could also simplify the development of novel applications for our framework, e.g., an automated detection of the number and abundances of organisms with reads mapping to the same genome. We also see options for further improvements by combining two alternatives we developed in our framework: currently, we fit the mixture model either to the histogram of coverage depth values or to the histogram of distances between read start positions. An algorithm that simultaneously fits distributions to both complementary representations of the read mapping and links the parameters of the involved distributions might lead to improved and more stable results.

The development of the taxonomic profiling tool MicrobeGPS was intended to approach the problem of missing reference genomes. Although we demonstrated that our community composition estimates are more precise and meaningful than the previous approaches, we observed that the results are still not satisfying for complex and underexplored communities, such as the Lake Lanier freshwater metagenome. For the analyzed datasets, MicrobeGPS found between 165 and 238 candidate organisms; however, most of them had low validity and were only supported by one reference genome. Since MicrobeGPS marked these candidates as unreliable, the results on the species level were not very informative and satisfying. As long as there are not sufficient reference genomes available for the given dataset, MicrobeGPS could be complemented by other profiling strategies that do not rely on reference genomes. Since only about 1% of all reads in the datasets could be mapped to reference genomes, the unmapped reads could be used to assemble contigs. These contigs could be either treated as novel unknown species and roughly placed in the taxonomy by calculating similarities to existing genomes or could be used to predict genes that can be searched in protein databases. Another approach to circumvent the loss of information caused by unmapped reads could be searching read alignments with very sensitive and error tolerant methods, such as BLAST (Altschul et al., 1997). In addition to the increased runtime, one would have to take special care of reads mapped with high error rates and one would probably incorporate the mapping quality into the calculation of the genome validity.

As an overall picture, we see that missing reference genomes still represent a major challenge for taxonomic profiling. Although the number and diversity of reference genomes will increase within the next years due to technological advances, we think that making use of different taxonomic profiling techniques, such as reference-based, taxonomy independent, and protein-based approaches could drastically improve taxonomic profiling on complex metagenomes. Therefore, integrating the different data sources will be a major challenge in the near future until new biological or technical approaches, such as single cell or nanopore sequencing, are capable to obtain longer and more accurate genomic sequences from metagenomes.

A. Appendix

Table A.1.: List of data accessions used in the GASiC experiments

E. coli experiment

Name	Resource	Access Date	Accession
E. coli DH10B	NCBI RefSeq	22.11.2011	NC_010473.1
S. flexneri	NCBI RefSeq	07.12.2011	NC_004337.2
E. fergusonii	NCBI RefSeq	07.12.2011	NC_011740.1
K. pneumoniae	NCBI RefSeq	21.11.2011	NC_011283.1
P. ananatis	NCBI RefSeq	07.12.2011	NC_013956.2
E. coli TY-2482	Internet	18.11.2011	see Lindner et al. (2013)
IonTorrent *E. coli* DH10B	Internet	30.09.2011	see Lindner et al. (2013)
IonTorrent *E. coli* TY-2482	NCBI SRA	04.10.2011	SRX072974

Virus experiment

Name	Resource	Access Date	Accession
Illumina dataset	NCBI SRA	20.12.2011	SRA020830
Deformed wing virus	NCBI Reference Sequence	20.12.2011	NC_004830.2
Varroa destructor virus-1	NCBI Reference Sequence	20.12.2011	NC_006494.1
VDV-1VVD	NCBI GenBank	21.12.2011	HM067438.1
VDV-1DVD	NCBI GenBank	21.12.2011	HM067437.1

Table A.2.: GASiC abundance estimation for 6 reference genomes. The 11 datasets contained *E. coli* and STEC reads with varying concentrations. The true concentration of the organism is given in the column Frac, the proportion of aligned reads is given in column Aln. Column GASiC is the abundance estimated by GASiC, together with the estimated P-value for the presence of the organism in the dataset.

	E. coli				**STEC**				***S. flexneri***			
Dataset	**Frac**	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**
1	0.00	0.509	0.000	1.00	1.00	0.698	1.000	0.00	0.00	0.475	0.000	1.00
2	0.01	0.512	0.000	1.00	0.99	0.697	1.000	0.00	0.00	0.476	0.000	1.00
3	0.05	0.526	0.043	0.00	0.95	0.699	0.957	0.00	0.00	0.484	0.000	1.00
4	0.10	0.547	0.107	0.00	0.90	0.703	0.892	0.00	0.00	0.494	0.000	1.00
5	0.20	0.592	0.225	0.00	0.80	0.719	0.774	0.00	0.00	0.521	0.000	1.00
6	0.50	0.713	0.541	0.00	0.50	0.746	0.458	0.00	0.00	0.586	0.000	1.00
7	0.80	0.826	0.819	0.00	0.20	0.761	0.180	0.00	0.00	0.643	0.000	1.00
8	0.90	0.859	0.913	0.00	0.10	0.759	0.086	0.00	0.00	0.657	0.000	1.00
9	0.95	0.874	0.958	0.00	0.05	0.758	0.041	0.00	0.00	0.664	0.000	1.00
10	0.99	0.888	0.992	0.00	0.01	0.758	0.007	0.90	0.00	0.670	0.000	1.00
11	1.00	0.891	0.999	0.00	0.00	0.758	0.001	1.00	0.00	0.672	0.000	1.00

E. fergusonii				***K. pneumoniae***				***P. ananatis***			
Frac	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**
0.00	0.188	0.000	1.00	0.00	0.014	0.000	1.00	0.00	0.005	0.000	1.00
0.00	0.188	0.000	1.00	0.00	0.014	0.000	1.00	0.00	0.005	0.000	1.00
0.00	0.191	0.000	1.00	0.00	0.014	0.000	1.00	0.00	0.005	0.000	1.00
0.00	0.197	0.000	1.00	0.00	0.015	0.000	1.00	0.00	0.005	0.000	1.00
0.00	0.208	0.001	1.00	0.00	0.015	0.000	1.00	0.00	0.005	0.000	1.00
0.00	0.235	0.001	1.00	0.00	0.018	0.000	1.00	0.00	0.006	0.000	1.00
0.00	0.259	0.001	1.00	0.00	0.020	0.000	1.00	0.00	0.007	0.000	1.00
0.00	0.265	0.000	1.00	0.00	0.020	0.000	1.00	0.00	0.007	0.000	1.00
0.00	0.268	0.000	1.00	0.00	0.021	0.000	1.00	0.00	0.007	0.000	1.00
0.00	0.270	0.000	1.00	0.00	0.021	0.000	1.00	0.00	0.007	0.000	1.00
0.00	0.271	0.000	1.00	0.00	0.021	0.000	1.00	0.00	0.007	0.000	1.00

Table A.3.: Abundance estimation in the presence of a distantly related abundant organism. The 11 datasets contained *E. coli* and STEC reads with varying concentrations as well as one very distantly related organism accounting for 50% of the total number of reads. The true concentration of the organism is given in the column Frac, the proportion of aligned reads is given in column Aln. Column GASiC is the abundance estimated by GASiC, together with the estimated P-value for the presence of the organism in the dataset.

	E. coli				**STEC**				***S. flexneri***			
Dataset	**Frac**	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**
1	0.00	0.2499	0.0000	1.00	0.50	0.3428	1.0000	0.00	0.00	0.2331	0.0000	1.00
2	0.01	0.2515	0.0000	1.00	0.50	0.3428	1.0000	0.00	0.00	0.2339	0.0000	1.00
3	0.03	0.2595	0.0366	0.00	0.48	0.3449	0.9634	0.00	0.00	0.2379	0.0000	1.00
4	0.05	0.2690	0.1045	0.00	0.45	0.3457	0.8955	0.00	0.00	0.2427	0.0000	1.00
5	0.10	0.2883	0.2219	0.00	0.40	0.3498	0.7781	0.00	0.00	0.2526	0.0000	1.00
6	0.25	0.3459	0.5569	0.00	0.25	0.3587	0.4431	0.00	0.00	0.2824	0.0000	1.00
7	0.40	0.4044	0.8394	0.00	0.10	0.3693	0.1606	0.00	0.00	0.3139	0.0000	1.00
8	0.45	0.4240	0.9209	0.00	0.05	0.3732	0.0791	0.00	0.00	0.3234	0.0000	1.00
9	0.48	0.4342	0.9585	0.00	0.03	0.3759	0.0415	0.00	0.00	0.3297	0.0000	1.00
10	0.50	0.4415	0.9920	0.00	0.01	0.3765	0.0080	1.00	0.00	0.3331	0.0000	1.00
11	0.50	0.4440	0.9984	0.00	0.00	0.3780	0.0016	1.00	0.00	0.3345	0.0000	1.00

Table A.4.: GASiC abundance estimation with missing STEC reference genome, using only *E. coli* and *S. flexneri* as reference genome. The 11 datasets contained *E. coli* and STEC reads with varying concentrations. The true concentration of the organism is given in the column Frac, the proportion of aligned reads is given in column Aln. Column GASiC is the abundance estimated by GASiC, together with the estimated P-value for the presence of the organism in the dataset. The abundance estimates by GRAMMy can be found in the last column.

	E. coli					***S. flexneri***				
Dataset	**Frac**	**Aln**	**GASiC**	**P**	**GRAMMy**	**Frac**	**Aln**	**GASiC**	**P**	**GRAMMy**
1	0.00	0.5092	0.6034	0.00	0.5429	0.00	0.4750	0.3966	0.00	0.4571
2	0.01	0.5121	0.6092	0.00	0.5424	0.00	0.4762	0.3908	0.00	0.4576
3	0.05	0.5264	0.6324	0.00	0.5509	0.00	0.4835	0.3676	0.00	0.4491
4	0.10	0.5470	0.6628	0.00	0.5603	0.00	0.4945	0.3372	0.00	0.4397
5	0.20	0.5918	0.7150	0.00	0.5805	0.00	0.5206	0.2850	0.00	0.4195
6	0.50	0.7132	0.8421	0.00	0.6450	0.00	0.5864	0.1579	0.00	0.3550
7	0.80	0.8261	0.9442	0.00	0.7248	0.00	0.6431	0.0558	0.00	0.2752
8	0.90	0.8586	0.9753	0.00	0.7597	0.00	0.6571	0.0247	0.01	0.2403
9	0.95	0.8743	0.9900	0.00	0.7808	0.00	0.6637	0.0100	0.59	0.2192
10	0.99	0.8876	0.9985	0.00	0.7998	0.00	0.6698	0.0015	0.99	0.2002
11	1.00	0.8913	0.9992	0.00	0.8063	0.00	0.6716	0.0008	1.00	0.1937

Table A.5.: Abundance estimation with missing STEC reference genome, using only *E. coli* and *P. ananatis* as reference genomes. The 11 datasets contained *E. coli* and STEC reads with varying concentrations. The true concentration of the organism is given in the column Frac, the proportion of aligned reads is given in column Aln. Column GASiC is the abundance estimated by GASiC, together with the estimated P-value for the presence of the organism in the dataset. The abundance estimates by GRAMMy can be found in the last column.

	E. coli					***P. ananatis***				
Dataset	**Frac**	**Aln**	**GASiC**	**P**	**GRAMMy**	**Frac**	**Aln**	**GASiC**	**P**	**GRAMMy**
1	0.00	0.5092	0.9988	0.00	0.9914	0.00	0.0048	0.0012	1.00	0.0086
2	0.01	0.5121	0.9988	0.00	0.9911	0.00	0.0048	0.0012	1.00	0.0089
3	0.05	0.5264	0.9989	0.00	0.9921	0.00	0.0049	0.0011	1.00	0.0079
4	0.10	0.5470	0.9989	0.00	0.9927	0.00	0.0051	0.0011	1.00	0.0073
5	0.20	0.5918	0.9991	0.00	0.9932	0.00	0.0054	0.0009	1.00	0.0068
6	0.50	0.7132	0.9995	0.00	0.9953	0.00	0.0062	0.0005	1.00	0.0047
7	0.80	0.8261	0.9997	0.00	0.9974	0.00	0.0069	0.0003	1.00	0.0026
8	0.90	0.8586	0.9998	0.00	0.9980	0.00	0.0071	0.0002	1.00	0.0020
9	0.95	0.8743	0.9999	0.00	0.9981	0.00	0.0072	0.0001	1.00	0.0019
10	0.99	0.8876	0.9999	0.00	0.9983	0.00	0.0072	0.0001	1.00	0.0017
11	1.00	0.8913	0.9999	0.00	0.9985	0.00	0.0072	0.0001	1.00	0.0015

Table A.6.: Abundance estimation using contigs from assembled STEC reads instead of the complete STEC reference genome. The 11 datasets contained *E. coli* and STEC reads with varying concentrations. The true concentration of the organism is given in the column Frac, the proportion of aligned reads is given in column Aln. Column GASiC is the abundance estimated by GASiC, together with the estimated P-value for the presence of the organism in the dataset. The abundance estimates by GRAMMy can be found in the last column.

	E. coli				**STEC contigs**				***S. flexneri***			
Dataset	**Frac**	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**	**Frac**	**Aln**	**GASiC**	**P**
1	0.00	0.5088	0.0000	1.00	1.00	0.6975	0.9998	0.00	0.00	0.4750	0.0000	1.00
2	0.01	0.5116	0.0001	1.00	0.99	0.6973	0.9997	0.00	0.00	0.4762	0.0000	1.00
3	0.05	0.5260	0.0391	0.00	0.95	0.6990	0.9605	0.00	0.00	0.4835	0.0001	1.00
4	0.10	0.5466	0.1035	0.00	0.90	0.7035	0.8958	0.00	0.00	0.4945	0.0001	1.00
5	0.20	0.5913	0.2210	0.00	0.80	0.7185	0.7778	0.00	0.00	0.5206	0.0003	1.00
6	0.50	0.7124	0.5336	0.00	0.50	0.7461	0.4611	0.00	0.00	0.5864	0.0034	0.97
7	0.80	0.8252	0.8092	0.00	0.20	0.7614	0.1815	0.00	0.00	0.6431	0.0066	0.83
8	0.90	0.8576	0.9009	0.00	0.10	0.7590	0.0874	0.00	0.00	0.6571	0.0088	0.68
9	0.95	0.8732	0.9447	0.00	0.05	0.7580	0.0436	0.00	0.00	0.6637	0.0087	0.68
10	0.99	0.8866	0.9791	0.00	0.01	0.7577	0.0084	0.70	0.00	0.6698	0.0094	0.62
11	1.00	0.8902	0.9857	0.00	0.00	0.7579	0.0022	0.97	0.00	0.6716	0.0089	0.67

Bibliography

E. Allen and J. Banfield. Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, 3(6):489–498, 2005.

S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

F. Angly, B. Rodriguez-Brito, D. Bangor, P. McNairnie, M. Breitbart, P. Salamon, B. Felts, J. Nulton, J. Mahaffy, and F. Rohwer. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC bioinformatics*, 6(1):41, 2005.

F. Angly, D. Willner, A. Prieto-Davó, R. Edwards, R. Schmieder, R. Vega-Thurber, D. Antonopoulos, K. Barott, M. Cottrell, C. Desnues, et al. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLOS Computational Biology*, 5(12):e1000593, 2009.

F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, 40 (12):e94–e94, 2012.

D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 36(suppl 1):D25–D30, 2008.

R. D. Berg. The indigenous gastrointestinal microflora. *Trends in Microbiology*, 4 (11):430–435, 1996.

M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1):26–31, 2009.

C. Bliss and R. Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176–200, 1953.

S. Boisvert, F. Raymond, É. Godzaridis, F. Laviolette, and J. Corbeil. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, 13(12): R122, 2012.

T. Bonfert, G. Csaba, R. Zimmer, and C. C. Friedel. Mining RNA–seq data for infections and contaminations. *PLOS ONE*, 8(9):e73071, 2013.

R. A. Bradshaw, A. L. Burlingame, S. Carr, and R. Aebersold. Reporting protein identification data: the next generation of guidelines. *Molecular and Cellular Proteomics*, 5(5):787–788, 2006.

M. Breitbart, A. Hoare, A. Nitti, J. Siefert, M. Haynes, E. Dinsdale, R. Edwards, V. Souza, F. Rohwer, and D. Hollander. Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Cienegas, Mexico. *Environmental Microbiology*, 11(1):16–34, 2009.

B. Chevreux, T. Wetter, and S. Suhai. Genome sequence assembly using trace signals and additional sequence information. In *German Conference on Bioinformatics*, pages 45–56, 1999.

J. E. Clarridge. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 17(4):840–862, 2004.

E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009.

E. F. DeLong, C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N.-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, et al. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311(5760): 496–503, 2006.

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

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

V. J. Denef, L. H. Kalnejais, R. S. Mueller, P. Wilmes, B. J. Baker, B. C. Thomas, N. C. VerBerkmoes, R. L. Hettich, and J. F. Banfield. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6):2383–2390, 2010.

B. J. Diament and W. S. Noble. Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research*, 10(9):3871–3879, 2011.

J. A. Dodsworth, P. C. Blainey, S. K. Murugapiran, W. D. Swingley, C. A. Ross, S. G. Tringe, P. S. Chain, M. B. Scholz, C.-C. Lo, J. Raymond, et al. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nature Communications*, 4:1854, 2013.

J. Doellinger et al. Cowpox virus mature virion proteome: Composition, ubiquitination and attachment. *Manuscript in submission*, 2014.

J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105–e105, 2008.

A. Döring, D. Weese, T. Rausch, and K. Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9(1):11, 2008.

J. Dröge and A. C. McHardy. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in Bioinformatics*, 13(6): 646–655, 2012.

B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

J. T. Ellis, R. C. Sims, and C. D. Miller. Monitoring microbial diversity of bioreactors using metagenomic approaches. In *Reprogramming Microbial Metabolic Pathways*, pages 73–94. Springer, 2012.

R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, et al. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl 1):D247–D251, 2006.

S. Fishman and A. Branch. The quasispecies nature and biological implications of the Hepatitis C virus. *Infection, Genetics and Evolution*, 9(6):1158–1167, 2009.

K. J. Forsberg, S. Patel, M. K. Gibson, C. L. Lauber, R. Knight, N. Fierer, and G. Dantas. Bacterial phylogeny structures soil resistomes across habitats. *Nature*, 509(7502):612–616, 2014.

O. E. Francis, M. Bendall, S. Manimaran, C. Hong, N. L. Clement, E. Castro-Nallar, Q. Snell, G. B. Schaalje, M. J. Clement, K. A. Crandall, et al. Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Research*, 23(10):1721–1729, 2013.

J. Gans, M. Wolinsky, and J. Dunbar. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, 309(5739):1387–1390, 2005.

F. García-Alcalde, K. Okonechnikov, J. Carbonell, L. Ruiz, S. Götz, S. Tarazona, T. Meyer, and A. Conesa. Qualimap: evaluating next generation sequencing alignment data. *Bioinformatics*, 2012.

L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *Journal of Proteome Research*, 3(5):958–964, 2004.

J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbrück, J. Reeder, B. Temperton, S. Huse, A. C. McHardy, R. Knight, I. Joint, et al. Defining seasonal marine microbial community dynamics. *The ISME Journal*, 6(2):298–308, 2012.

C. A. Heid, J. Stevens, K. J. Livak, and P. M. Williams. Real time quantitative PCR. *Genome Research*, 6(10):986–994, 1996.

R. L. Hettich, C. Pan, K. Chourey, and R. J. Giannone. Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Analytical Chemistry*, 85(9):4203–4214, 2013.

D. A. Hill, C. Hoffmann, M. C. Abt, Y. Du, D. Kobuley, T. J. Kirn, F. D. Bushman, and D. Artis. Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis. *Mucosal Immunology*, 3(2):148–158, 2010.

K. H. Hoffmann, B. Rodriguez-Brito, M. Breitbart, D. Bangor, F. Angly, B. Felts, J. Nulton, F. Rohwer, and P. Salamon. Power law rank–abundance models for marine phage communities. *FEMS microbiology letters*, 273(2):224–228, 2007.

M. Holtgrewe. Mason–a read simulator for second generation sequencing data. *Technical Report TR-B-10-06, FU Berlin*, 2010.

S. Hooper, D. Dalevi, A. Pati, K. Mavromatis, N. Ivanova, and N. Kyrpides. Estimating DNA coverage and abundance in metagenomes using a gamma approximation. *Bioinformatics*, 26(3):295–301, 2010.

A. C. Howe, J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown. Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences of the United States of America*, 111(13):4904–4909, 2014.

D. Huson, A. Auch, J. Qi, and S. Schuster. MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386, 2007.

V. Iverson, R. Morris, C. Frazar, C. Berthiaume, R. Morales, and E. Armbrust. Untangling genomes from metagenomes: Revealing an uncultured class of marine euryarchaeota. *Science*, 335(6068):587–590, 2012.

S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

K. T. Konstantinidis and J. M. Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, 102(7):2567–2572, 2005.

D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, 1988.

B. Langmead and S. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

B. Langmead, C. Trapnell, M. Pop, S. Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10 (3):R25, 2009.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078, 2009.

M. S. Lindner and B. Y. Renard. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Research*, 41(1):e10, 2013.

M. S. Lindner and B. Y. Renard. Metagenomic profiling of known and unknown microbes with microbegps. *PLOS ONE*, 10(2):e0117711, 2015.

M. S. Lindner, M. Kollock, F. Zickmann, and B. Y. Renard. Analyzing genome coverage profiles with applications to quality control in metagenomics. *Bioinformatics*, 29(10):1260–1267, 2013.

B. Liu, T. Gibbons, M. Ghodsi, and M. Pop. MetaPhyler: Taxonomic profiling for metagenomic sequences. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 95–100. IEEE, 2010.

B. Liu, T. Gibbons, M. Ghodsi, T. Treangen, and M. Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, 12(Suppl 2):S4, 2011.

N. J. Loman, C. Constantinidou, M. Christner, H. Rohde, J. Z.-M. Chan, J. Quick, J. C. Weir, C. Quince, G. P. Smith, J. R. Betley, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic Escherichia coli O104:H4. *Journal of the American Medical Association*, 309(14):1502–1510, 2013.

M. Löwer, B. Renard, J. de Graaf, M. Wagner, C. Paret, C. Kneip, Ö. Türeci, M. Diken, C. Britten, S. Kreiter, et al. Confidence-based somatic mutation evaluation and prioritization. *PLOS Computational Biology*, 8(9):e1002714, 2012.

C. Luo, D. Tsementzi, N. C. Kyrpides, and K. T. Konstantinidis. Individual genome assembly from complex community short-read metagenomic datasets. *The ISME Journal*, 6(4):898–901, 2012.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

S. S. Mande, M. H. Mohammed, and T. S. Ghosh. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, 13(6):669–681, 2012.

O. U. Mason, T. C. Hazen, S. Borglin, P. S. Chain, E. A. Dubinsky, J. L. Fortney, J. Han, H.-Y. N. Holman, J. Hultman, R. Lamendella, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *The ISME journal*, 6(9):1715–1727, 2012.

F. J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, 2007.

K. Mavromatis, M. Land, T. Brettin, D. Quest, A. Copeland, A. Clum, L. Goodwin, T. Woyke, A. Lapidus, H. Klenk, et al. The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. *PLOS ONE*, 7(12):e48837, 2012.

M. J. McInerney, J. R. Sieber, and R. P. Gunsalus. Syntrophy in anaerobic global carbon cycles. *Current Opinion in Biotechnology*, 20(6):623–632, 2009.

B. A. Methé, K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker, J. H. Badger, et al. A framework for human microbiome research. *Nature*, 486(7402):215–221, 2012.

C. Miller, O. Hampton, C. Coarfa, and A. Milosavljevic. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLOS ONE*, 6(1):e16327, 2011.

J. L. Mokili, F. Rohwer, and B. E. Dutilh. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, 2(1):63–77, 2012.

J. Moore, A. Jironkin, D. Chandler, N. Burroughs, D. Evans, and E. Ryabov. Recombinants between Deformed wing virus and Varroa destructor virus-1 may prevail in Varroa destructor-infested honeybee colonies. *Journal of General Virology*, 92 (1):156, 2011.

T. Muth, D. Benndorf, U. Reichl, E. Rapp, and L. Martens. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular BioSystems*, 9(4):578–585, 2013.

R. Naeem, M. Rashid, and A. Pain. READSCAN: A fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics*, 2012. doi: 10.1093/bioinformatics/bts684.

T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40(20):e155–e155, 2012.

K. E. Nelson, G. M. Weinstock, S. K. Highlander, K. C. Worley, H. H. Creasy, J. R. Wortman, D. B. Rusch, M. Mitreva, E. Sodergren, A. T. Chinwalla, et al. A catalog of reference genomes from the human microbiome. *Science*, 328(5981): 994–999, 2010.

H. Noguchi, J. Park, and T. Takagi. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids research*, 34(19):5623–5630, 2006.

S. Oh, A. Caro-Quintero, D. Tsementzi, N. DeLeon-Rodriguez, C. Luo, R. Poretsky, and K. T. Konstantinidis. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Applied and Environmental Microbiology*, 77(17): 6000–6011, 2011.

A. M. O'Hara and F. Shanahan. The gut flora as a forgotten organ. *EMBO Reports*, 7(7):688–693, 2006.

A. Penzlin, M. S. Lindner, J. Doellinger, P. W. Dabrowski, A. Nitsche, and B. Y. Renard. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics*, 30(12):i149–i156, 2014. doi: 10.1093/bioinformatics/btu267.

J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, et al. The NIH human microbiome project. *Genome Research*, 19(12):2317–2323, 2009.

R. Poretsky, L. M. Rodriguez-R, C. Luo, D. Tsementzi, and K. T. Konstantinidis. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLOS ONE*, 9(4):e93827, 2014.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(suppl 1):D61–D65, 2007.

K. D. Pruitt, G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, 42(D1): D756–D763, 2014.

J. Qin, R. Li, J. Raes, M. Arumugam, K. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

G. Reinert, D. Chew, F. Sun, and M. Waterman. Alignment-free sequence comparison (i): statistics and power. *Journal of Computational Biology*, 16(12):1615–1634, 2009.

B. Renard, M. Kirchner, H. Steen, J. Steen, and F. Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9(1):355, 2008.

B. Y. Renard, B. Xu, M. Kirchner, F. Zickmann, D. Winter, S. Korten, N. W. Brattig, A. Tzur, F. A. Hamprecht, and H. Steen. Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS). *Molecular and Cellular Proteomics*, 11(7):M111.014167, 2012.

C. S. Riesenfeld, R. M. Goodman, and J. Handelsman. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental microbiology*, 6(9): 981–989, 2004.

L. M. Rodriguez-R and K. T. Konstantinidis. Estimating coverage in metagenomic data sets and why it matters. *The ISME journal*, 2014.

S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno. SHRiMP: accurate mapping of short color-space reads. *PLOS Computational Biology*, 5(5):e1000386, 2009.

D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, et al. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLOS Biology*, 5(3):e77, 2007.

D. C. Savage. Microbial ecology of the gastrointestinal tract. *Annual Reviews in Microbiology*, 31(1):107–133, 1977.

P. D. Schloss and J. Handelsman. Toward a census of bacteria in soil. *PLOS Computational Biology*, 2(7):e92, 2006.

F. J. Sedlazeck, P. Rescheneder, and A. von Haeseler. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21): 2790–2791, 2013.

N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, 2012.

J. Seifert, F. A. Herbst, P. Halkjaer Nielsen, F. J. Planes, N. Jehmlich, M. Ferrer, and M. von Bergen. Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. *Proteomics*, 13(18-19):2786–2804, 2013.

I. Sekirov, N. M. Tam, M. Jogova, M. L. Robertson, Y. Li, C. Lupp, and B. B. Finlay. Antibiotic-induced perturbations of the intestinal microbiota alter host susceptibility to enteric infection. *Infection and Immunity*, 76(10):4726–4736, 2008.

C. Simon and R. Daniel. Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology*, 77(4):1153–1161, 2011.

E. Siragusa, D. Weese, and K. Reinert. Fast and accurate read mapping with approximate seeds and multiple backtracking. *Nucleic Acids Research*, 41(7):e78–e78, 2013.

E. Stackebrandt and B. Goebel. 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*, 44(4):846–849, 1994.

L. D. Stein. The case for cloud computing in genome informatics. *Genome Biology*, 11(5):207, 2010.

S. Tanner, H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, 77(14):4626–4639, 2005.

R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4 (1):41, 2003.

H. Teeling and F. O. Glöckner. Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Briefings in Bioinformatics*, 13(6):728–742, 2012.

The Human Microbiome Jumpstart Reference Strains Consortium. A catalog of reference genomes from the human microbiome. *Science*, 328(5981):994–999, 2010.

V. Torsvik and L. Øvreås. Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology*, 5(3):240–245, 2002.

V. Torsvik, J. Goksøyr, and F. L. Daae. High diversity in DNA of soil bacteria. *Applied and Environmental Microbiology*, 56(3):782–787, 1990.

G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004.

C. J. Van Rijsbergen. *Information Retrieval.* Butterworths, London, 2nd edition, 1979.

G. Van Rossum and F. L. Drake Jr. *Python reference manual.* Centrum voor Wiskunde en Informatica Amsterdam, 1995.

J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.

S. Vinga. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 15(3):376–389, 2014. doi: 10.1093/bib/bbt068.

S. Vinga and J. Almeida. Alignment-free sequence comparison – a review. *Bioinformatics*, 19(4):513–523, 2003.

C. von Mering, P. Hugenholtz, J. Raes, S. Tringe, T. Doerks, L. Jensen, N. Ward, and P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126–1130, 2007.

W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6578–6583, 1998.

C. S. Wong, S. Jelacic, R. L. Habeeb, S. L. Watkins, and P. I. Tarr. The risk of the hemolytic–uremic syndrome after antibiotic treatment of Escherichia coli O157:H7 infections. *New England Journal of Medicine*, 342(26):1930–1936, 2000.

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

J. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLOS Computational Biology*, 6(2):e1000667, 2010.

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

L. Xia, J. Cram, T. Chen, J. Fuhrman, and F. Sun. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLOS ONE*, 6(12): e27992, 2011.

Z. Xu, M. A. Hansen, L. H. Hansen, S. Jacquiod, and S. J. Sørensen. Bioinformatic approaches reveal metagenomic characterization of soil microbial community. *PLOS ONE*, 9(4):e93445, 2014.

Z. Zhang, C. Chen, J. Sun, and K. Luk Chan. EM algorithms for Gaussian mixtures with split-and-merge operation. *Pattern Recognition*, 36(9):1973–1983, 2003.

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind.

Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt. Die Bestimmungen der Promotionsordnung sind mir bekannt.

Martin Lindner, Berlin, 18.08.2014

Lebenslauf

Mein Lebenslauf wird aus Gründen des Datenschutzes in der elektronischen Fassung meiner Arbeit nicht veröffentlicht.