Computational Analysis for Medical Research in Genomics and Metagenomics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat)

vorgelegt von

M. Sc. Sina Beier
aus Bruchsal

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündl. Qualifikation: 12.12.2018
Dekan: Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter: Prof. Dr. Daniel H. Huson
2. Berichterstatter: Prof. Dr. Detlef Weigel

Abstract

The investigation of microbes in research has been changing with the rise of environmental sequencing from a view centered on an isolate microbe living in a laboratory setting to a broader view of microbial communities - microbiomes - as they thrive in their natural environment. Sequencing of an isolated organism generates essential insights and enables us to produce reference genomes and genomes annotation, which in turn let us compare different organisms through their genomes and study the metabolic pathways they utilize. In comparison to that, environmental sequencing does answer entirely different questions, and it does provide us with a much better view on how microbes live in their environment, what changes they undergo based on their interactions with their host, environment or with each other. It also enables us to study the genomes of microbes which previously could not be cultured in isolation and provides many more new possibilities to many different life sciences which are interested in the microbial community on earth.
One of these sciences with a significant influence on our daily lives is medicine. In this work, I present multiple projects where genomics, 16S rDNA analysis and metagenomics have helped medical research to gain insights on diverse topics and under varying conditions. During these projects, I have determined a recurring need for primary analysis of environmental data, which currently often only can be done by bioinformatics specialists or at least would need time-consuming efforts to study the necessary tools for scientists from other fields. As those scientists also spend a lot of time planning and conducting the experiments which have led to the generation of the data and verifying the findings, they often do not have the time necessary to pick up the knowledge required even for many basic steps for environmental sequence analysis.
To make metagenomic analysis more approachable for a variety of users, I have developed three pipelines for the fundamental analysis of environmental data - collected in the CommunAl toolkit - which require minimal hands-on effort and infrastructure to run the necessary analysis. The toolkit includes the alignment-based 16S rDNA analysis tool STARA, which is applicable on any type of sequencing available and can analyze all samples from a dataset from preprocessing over alignment to the taxonomic assignment in one run. The second member of the toolkit is MAPle, a pipeline for paired-end short-read WGS metagenomic sequencing analysis. Like STARA this also analyzes a dataset from preprocessing through to taxonomic and functional assignment. The taxonomic abundances determined by those two tools can be further investigated using TaxCo. TaxCo computes correlations between taxonomic abundances and numeric metadata and presents the results in tabular and graphic form.
With these tools, I hope to enable everyone involved in environmental data analysis to generate insights from any of their datasets, which in turn will hopefully help to make environmental sequencing - especially metagenomics -

a more feasible choice for everyone who could profit from the possibilities and insights it can provide.

Zusammenfassung

Mit dem Aufkommen der Mikrobiomsequenzierung hat sich die Sicht der Wissenschaft auf Mikroben verändert. Lange war sie fokussiert auf isolierte, im Labor kultivierte Mikroben, nun ist eine breitere Ansicht auf mikrobielle Gemeinschaften - die Mikrobiome - in ihrer natürlichen Umgebung möglich. Das Sequenzieren eines isolierten Organismus produziert essenzielle Einsichten und ermöglicht es, Referenzgenome und deren Annotationen zu erhalten, durch welche man verschiedene Organismen über ihr Genom vergleichen kann und die Stoffwechselwege untersuchen welche sie nutzen. Im Vergleich dazu beantwortet Mikrobiomsequenzierung ganz andere Fragestellungen und bietet uns einen deutlich besseren Ausblick darauf, wie Mikroben in ihrer Umgebung leben und welchen Veränderungen sie durch Interaktion mit ihrem Wirt, ihrer Umgebung oder untereinander unterliegen. Sie ermöglicht uns auch, die Genome von Mikroben zu untersuchen, welche zuvor nicht isoliert kultiviert werden konnten und bietet viele weiter neue Möglichkeiten für verschiedene Bereiche der Naturwissenschaften die sich für die Gemeinschaft der Mikroorganismen auf der Erde interessieren.
Eine dieser Wissenschaften die großen Einfluss auf unser tägliches Leben hat ist die Medizin. In dieser Arbeit präsentiere ich verschiedene Projekte bei denen Genomik, 16S rDNA Sequenzierung und Metagenomik geholfen haben, unter verschiedensten Bedingungen und in unterschiedlichen Themenfeldern der Medizin neue Erkenntnisse in der medizinischen Forschung zu schaffen. Während dieser Projekte konnte ich eine regelmäßige Notwendigkeit für grundlegende Analyse von Mikrobiomdaten erkennen, welche im Moment noch meist nur von Spezialisten aus der Bioinformatik ausgeführt wird oder zumindest viel Zeit für die Einarbeitung in die nötigen Softwaretools für Forscher aus anderen Fachbereichen erfordert. Da diese Wissenschaftler bereits viel Zeit damit verbringen die Experimente zu planen und durchzuführen, durch die die zu analysierenden Daten generiert werden und die Endergebnisse der Analysen zu verifizieren haben sie oft nicht die nötige Zeit um sich zusätzliches Wissen anzueignen, das für viele Schritte in der Mirkobiom-Sequenzanalyse benötigt wird.
Um die Metagenomanalyse zugänglicher für unterschiedliche Nutzer zu machen, habe ich drei Pipelines für grundlegende Datenanalyse von Mikrobiomdaten entwickelt, die im CommunAl Toolkit vereint sind. Diese erfordern nur wenig Userinteraktion und keine komplexe Infrastruktur um die notwendigen Analysen auszuführen. Das Toolkit beinhaltet das Alignment-basierte 16S rDNA Analysetool STARA, welches für jede Art von Daten die durch aktuell verfügbare Sequenziertechnologien generiert wurden anwendbar ist. STARA analysiert alle Proben eines Datensatzes von der Datenvorbehandlung über Alignment zu taxonomischer Zuordnung in einem einzigen Durchlauf. Der zweite Teil des Toolkits ist MAPle, eine Pipeline für die Analyse

von Metagenomdaten. Genau wie STARA analysiert MAPle einen Datensatz von der Datenvorbehandlung bis zu taxonomischer Zuordnung und funktioneller Annotation. Die taxonomischen Häufigkeiten, die von diesen beiden Tools bestimmt wurden können schließlich mit TaxCo weiter untersucht werden. TaxCo berechnet Korrelationen zwischen taxonomischen Häufigkeiten und numerischen Metadaten und präsentiert die Ergebnisse grafisch und in Tabellenform.
Mit diesen Tools hoffe ich jedem, der an Mikrobiomanalyse interessiert ist die Möglichkeit zu geben, Erkenntnisse aus seinen Datensätzen zu schaffen, welche dann hoffentlich die Mikrobiomsequenzierung – besonders Metagenomik – zu einer realistischen Option für jeden machen der von den Möglichkeiten und Erkenntnissen die diese Technologie bietet profitieren könnte.

Acknowledgements

At the end of a journey, it is time to thank your fellow travelers, who made the trip into the experience it has been. I can only mention a few of the companions that I have had along the road that lead me here. For everyone, I do not mention: You know that even if you are not in my thesis, you most definitely are in my heart.

The first one to join me on the winding paths to a Ph.D. was Prof. Dr. Daniel H. Huson, who not only offered me the opportunity to travel with him, but always was willing to either tread a path, walk next to me chatting or have my back, whatever I needed at different steps of the way.

I also want to thank my collaborators from the Institute of Microbiology and Hygiene. Special thanks go to Prof. Dr. Julia-Stefanie Frick for believing in the magic of bioinformatics and always treating me like a valued member of her group. I also want to thank Anna Lange for our shared adventures from project planning to publication, with all the ups and downs this entailed.
I was lucky to have even another group which welcomed me in their middle, and I am grateful for all the long discussions and learning opportunities that Dr. Erwin Bohn, Dr. Monika Schütz and Janina Geissert have offered me. I did learn a lot about science, interdisciplinary teams and keeping a smile through whatever is going on from all of you on the winding road through our work together.
Further on I want to thank my collaborators from Immunology - Prof. Dr. Alex Weber - and Industry - Dr. Isabell Flade - who offered me the opportunity to participate in their work and learn about many different aspects of bioinformatics.

Of course I will not forget to include my fellow travelers from the group of Algorithms in Bioinformatics and many other people who do work on Sand for laughs and cakes in good times and the deep (coffee) conversations anytime, for group meetings and hikes and evening beers, for in-depth scientific discussions and joking around - often during the same conversation. You were always there for me when I needed it, and I do appreciate that very much. I am grateful for being so lucky that going to work always included going to meet friends.

Along the road, there were also many more people who did cheer me on. I am grateful to have close friends who have been there for me through good and bad times. I am thankful that you all allowed me to be myself, listened to me, distracted me when I needed it and never hesitated to give me a good nudge forward when it was the time for it.

The people who have followed my travels for the longest are of course my family, especially my parents. Thank you for being my safety net and my cheerleaders. I would not be where I am without your support and encouragement.

Last but not least, thank you, Alex, for believing in me when I do not, making me smile in any situation and loving me when I need it the most.

"Good company in a journey makes the way seem shorter."

Izaak Walton

"If you don't like bacteria,
you're on the wrong planet."

Stewart Brand

Contents

List of Figures

List of Tables

1

Introduction

Genomics and metagenomics for medical research and clinical use are fields on the fast track of science. While there is still much to learn about the human genome, while we still have no idea how to actually quantify assignments from metagenomic samples, while millions of microbial strains live in the grey zone of "unknown environmental sequence", medically relevant applications of sequencing have gone into overdrive fueled by public interest, funding possibilities and, of course, the needs and hopes of each patient. With fundamental research still on its way, we now try to generate actionable information in a field that is based on ongoing research. Actionable information, a term out of the business and data security context and knowledge research, means the information - although based on very fundamental knowledge in our case - has to be relevant, timely, accurate, complete in respect to the goals and ingestible. With the ascent of personalized medicine, this term is now seeping into medical terminology. It is an important goal to transform patient data into actionable information using modern technologies and automated analysis methods. Improving diagnostics and personalizing treatment decisions is where genomic and metagenomic analysis already fits closely into the daily routine of healthcare. To make sequence analysis a routine task for people who are not specialists, the analysis methods have to mature from a highly specialized research method to a widely applicable tool which assures reproducible and problem-tailored analysis, not only for research questions but further on also for every single patient's data.
All of these developments have been made possible by the broad availability of sequencing technology. Quality and output of sequencing are now usable for large research projects, while cost and usability have improved to bring it onto the benchtop and thus into the clinic. DNA sequencing started in 1970 [82], first using two-dimensional chromatography to determine separate sequences, but soon being developed into chain-termination sequencing by Sanger et al. in 1977 [68], which means it took only 17 years from understanding the structure of DNA to starting to determine existing DNA sequences. The first automated

sequencer was developed by Applied Biosystems who already in 1986 released their 370A DNA Sequencer, which is sometimes called the "first generation" of automated sequencing technology. The development slowed down a little until the Human Genome Project in 1990 added selection pressure to this evolution by increasing the need for high-throughput data generation. The newfound goal and competition ultimately led to the development of next-generation - originally called "second generation" - sequencing which became available commercially with the release of the 454 GS20 sequencers by 454 Life Sciences in 2005. Randomly sheared so-called "shotgun" DNA input generating large numbers of short reads has stayed at the top of the field for many years before long read sequencing methods finally proved stable and affordable enough to start taking over. Currently, short read sequencing is still the method of choice for clinical use, where low error rates usually are of utter importance. It will be interesting to see how Oxford Nanopore's long read technology will change this fact in the future by making sequencing available without the need of large, complicated and expensive technology and providing fast real-time data analysis for diagnostic purposes.

In the past ten years, sequencing technology has developed from a novel research method to a universal standard in many areas of life. One could go so far as to propose that nowadays every single person's life is to some extent influenced by the possibilities and knowledge that genome sequencing provides. It changed the food we eat by revolutionizing agriculture and breeding; it helped to find new possibilities to manufacture and degrade components used in industry, it teaches us new insights into evolution and helps to protect the environment and the diversity of species. Moreover, it significantly influences developments in health care, providing new diagnostics and treatments and introducing new fields like precision medicine or personalized medicine.

With reduced cost, increased throughput and an ever-increasing choice of technologies generating longer reads, needing less DNA input and reducing error as much as possible, and the technology being increasingly portable, sequencing is now a standard part of our toolkit in all aspects of life that benefit from knowledge about living organisms.

The development of methods to store, handle and analyze these massive amounts of data need to be improved by far to effectively and accurately leverage the information hidden in them. This work will focus on current tasks connected with analyzing sequencing data, especially tasks specific to analysis which are relevant in the context of medical research and the future of healthcare.

Human hereditary diseases have long been at the forefront of genomic research. Studying inheritance and predisposition for diseases has been done using analysis of family trees long before genome sequencing became a possibility. Sequencing the human genome has made it possible to take the step from studying hereditary disease to being able to find new mutations in a

genome which are related to diseases.
However, the use of genomics does not end with studying and diagnosing genetic disorders. In immunology, genomics is helping to understand the genetic background of our immune system. Infectious disease research does now rely significantly on the availability of reference genomes for pathogens, whether they are viruses, bacteria, amoebae or eucaryotic parasites. Reference genomes are also an essential factor in any study involving model organisms for human health, often utilizing mice, but also rats, monkeys and recently the greater wax moth. One major reference genome sequencing effort - the Human Genome Project - did not only facilitate and kickstart the development of next-generation sequencing (NGS) technology; it also started the use of whole-genome studies as a popular tool in any field of biology-related research. From microbes to large plant genomes, from small research labs to large leading industry players, having a reference genome for selected taxa of interest became a necessity and community projects were started to sequence populations of model organisms or many different species from interesting taxa. The need for new and better reference genomes resulted in many projects to generate genomic information for selected taxa. The 1000 (human) genomes project [7] directly inspired some of them, like the 1001 genomes of *Arabidopsis thaliana* varieties sequenced by the TAIR project [2] or a collection of 1002 yeast genomes planned to be released in the future. Newer projects aim for even more genomes to be collected, as high throughput and low cost have made that feasible. Now it is even possible to aim for example for 5000 insect genomes (i5k) [26]. Last but not least, there is an ever-growing flood of microbial genomes sequenced, unfortunately without a reasonable community effort to curate and connect this data.
Microbial genomes of all kinds are very important not only in medicine but in general for understanding the living organisms surrounding us and how they influence our life. Microbes are all around us, inside our bodies, inside our food. They colonize everything we touch and everything we ingest. Some are pathogens, and others help other organisms out in a seemingly selfless way. Microbes might be small, but the truth is that they are the main players in the endless balance of life and death on earth. They prime a child's immune system to set it up for a healthy life and they at the same time are the driving factor of decay in a dead tree. Humans had started to utilize microbes long before they knew of their existence, using yeasts to process food and brew drinks and bacteria to allow fermentation. They used bacteria and fungi to produce fertilizer, and some fungi even were used in early warfare and medicine, reportedly sedating an enemy or patient.
All this time, the existence of microbes was mostly unknown. The invention of the microscope in the 17th century was the necessary step which enabled researchers to observe microbes and led to their identification as living organisms. With the knowledge about microbes also came the first knowledge about

pathogens causing infectious disease. Quickly their status developed from a hidden helper into a threat to any living organism. The good reputation of yeast and other helpful microbes of daily life was overshadowed for a long time by the frightening presence of pathogens. Microbes where even called an "enemy of nature" by Thomas Sydenham, a 17th-century physician.
Metagenomics is bringing genome sequencing from isolated organisms to communities found in any possible environment. Sequencing these communities introduces specific challenges. The first one is of course that we cannot know which and how many different organisms and species are present, which of them have been successfully sequenced and which extracted sequence fragments do belong to the same genome. While we can extract a general taxonomic picture as well as functional assignments from a metagenomic analysis, it is hard to connect the two and generate actual "genomic" information. Metagenomic assembly and long-read sequencing improve this bottleneck, but so far not in a way which would make it possible to create a set of fully annotated genomes for the whole sequenced community of an environment. Hence, many questions often go unanswered.
Still, genomics of isolated organisms will not be able to provide those answers easily as well, as some organisms are hard to culture as an isolate and if they are cultured, information about the community they live in might be lost, which can be of major importance.
In comparison, human genomics, while it is also far from a solved case, is already developing into a valid diagnostic tool in modern medicine. Sequencing panels for known disease variants are available to determine the cause of an individuals disease and aid treatment decisions through this knowledge. The fact that they now have a known reason for their symptoms can be very beneficial for the patient, even if they do not gain insights for better treatment through it. With whole exome sequencing becoming increasingly feasible in a clinical context, the possibilities of finding previously unknown disease-causing variants have increased. The challenge has now switched from generating the data to extracting the needed information and protecting patient privacy. Genomics can be used in many areas of medical research and healthcare. From studying outbreaks like the Ebola outbreak of 2013 - 2016 [65], developing novel drugs and vaccines, diagnosing hereditary disease to many applications in cancer research and cancer medicine, genomics has already improved the lives of many patients. It is an increasingly common tool in medical research, becoming more and more usable as a diagnostic tool and an aide for important treatment decisions and will potentially generate novel treatments targeting the genome directly, for example utilizing the CRISPR mechanism to target specific genes.
Metagenomics has not yet become a full staple in medical research, but it will undoubtedly do so in the future. Finding new antibiotics, studying a pathogen in its natural environment or detecting unknown pathogens in a patient sample

are growing research fields. However, there are also increasing efforts in getting to know the commensal microbes which influence our health more subtly. These microbes have been shown to prime the human immune system, improve digestion, fight off pathogens and influence weight amongst many other things. The complex interplay of beneficial, commensal and pathogenic microbes has to be unraveled to improve healthcare and use metagenomics as a diagnostic tool as well as develop drugs which do not only interfere less with our healthy microbiome but might even utilize it positively or enhance it. The delicate balance of microbes in our gut, mouth, nose, genitals and on our skin are key factors in human health. The future of medicine will be influenced a lot by findings of metagenomic studies and the possibilities of the microbiome as a diagnostic marker.

Sequencing technology has been praised over and over as the technology that will help us cure cancer, develop new and more effective vaccines and introduce countless possibilities of personalized medicine in every hospital's diagnostic and interventional repertoire. Unfortunately, the advancements so far have rarely reached a broad usage outside of the research community. We need to prepare methods to be applicable not only by large specialized clinics but in any general hospital to have a public benefit from it. While sequencing technology is on its way to bench-top machines providing high-throughput data fast and with decreasingly complex preparation, bioinformatics analysis is still either very basic or has to be provided by specialists. As long as not every small town has their own genome center, this is not a feasible solution for broad healthcare and only applicable outside of an emergency situation, where it is critical that every result instantly reaches the treatment team of a patient. These results should also be provided in a manner they can be presented to the patient to build trust in the treatment and the diagnostic method used. It has been shown in placebo studies that this trust of a patient in their healthcare providers increases the effectivity of many therapies and the general well-being of the patient.

In the last years, there have also been exceptional efforts in public science outreach to bring knowledge about the benefits of sequencing and the importance of microbes for human health and disease to the public. Probiotics, books about the communities living in everybody's guts or mouths and citizen science projects which offer the ability to sequence parts of an individuals genome or different microbial communities in and on their body have become a part of life for many. The world is finally moving away from the idea that microbes are in general to be avoided as they are pathogens forward to the holistic view that there is also a large group of commensal and beneficial microbes which positively influence our daily life. This development is beneficial for medical professionals as it makes it easier to explain, for example, the responsible use of antibiotics and the benefits of not overdoing hygiene to the point where it strips the skin from its ability to keep up a healthy selective barrier. The

awareness of the public seeps back into medicine as it is paired with better acceptance of treatments and trust in the diagnostic method, better compliance with medication and of course a steady influx of finances based on the popularity of the subject.
This thesis will present projects of medical research in genomics in Part I and metagenomics in Part II, and the tools that we developed for those projects. All of the tools are meant to be applicable in a clinical setting, providing reproducible and problem-tailored analysis while being easy to use and presenting the results in a well ingestible format. After providing background information on sequencing and the main methods and tools used in the projects and pipelines in Chapter 2, this work will discuss assembly, annotation, and study of a commensal microbe for medical research in Chapter 3. I will present work on metagenomic analysis through the study of a mouse infection model of *Yersinia enterocolitica*, where both 16S rDNA sequencing and WGS metagenomics have been used to investigate the differences between the microbiota of mice succumbing to or surviving the infection in Chapter 5.
Metagenomics of the gut microbiota of patients undergoing stem cell therapy for acute myeloid leukemia (AML) in the ImMiGeNe project and the fully automated analysis pipeline developed for this project will be shown in Chapter 6. The metagenomic projects will finish off with an example of a community science project called the TüBiom Project, where the general healthy population could get their gut microbiome analyzed through 16S rDNA sequencing. I will explain the analysis pipeline for this project in Chapter 7.
The last Chapter 8 of Part II introduces the pipelines I developed for the analysis of environmental data throughout the projects. Those pipelines have been adapted for a more general use case and together form the CommunAl toolkit. These tools cover 16S rDNA analysis of single and paired reads in Section 8.1 using the STARA pipeline, taxonomic correlation analysis for results from STARA, MEGAN or other tools in Section 8.2 using TaxCO and finally different types of metagenomic sequence analysis in Section 8.3 which are provided by the MAPle tool.
Finally, I will present my general conclusions in Part III.

2

Background

2.1 Generations of Sequencing

The two first DNA sequencing methods developed in 1977 by Sanger [68] and Gilbert & Maxam [55] were both time-consuming and needed experts reading out the result from a gel. Sequencing was thus an expensive effort in every regard and only available for the few who could afford the cost. Only later when capillary sequencing became available around 1990, a few centers in the world could routinely sequence DNA and RNA using Applied Biosciences 370A and its successor model, the 373 Automated DNA Sequencer which got available in 1991. The availability of automated sequencing led to the first big efforts in eukaryotic genome sequencing to be possible, including the Human Genome Project sequencing DNA from 13 individuals to generate a human reference sequence [19] and the first individual human genome sequence of J. Craig Venter [51]. The development of PCR in 1983 and the availability of *taq* polymerase [25] enabled this step, as it made amplification of DNA on a large scale possible. Capillary sequencing is still used to generate comparatively long sequences with low error rates, which is useful to verify sequences or to sequence targeted genes over and over in different organisms for comparative efforts.

During the race for the human genome between J. Craig Venters' Celera and the Human Genome Project, Celera developed the method of Whole Genome Shotgun (WGS) sequencing as it was originally proposed for the Human Genome Project by Gene Myers instead of sequencing artificial clones of the partial sequence as done in BAC (bacterial artificial chromosome) sequencing. This involved breaking up the large chromosomal DNA into short, randomly cut pieces which would improve sequence quality and speed with the available technologies, but created the need for more powerful assembly algorithms to put the resulting small pieces back together.

Next Generation Sequencing (NGS) emerged with the availability of 454

pyrosequencing in 2005 and the introduction of Illumina using sequencing-by-synthesis in the same year. Since the commercial launch of the Illumina Genome Analyzer in 2006, the rise of NGS was better than it could have been expected using Moores Law. Indeed, the cost per base of sequencing started to drop significantly, making the goal of a 1000$ human genome sequence looking much more attainable. Next-generation sequencing brought us the second individual human genome of James Watson [81], which was made public only days away from J. Craig Venters genome in 2007.
After four more years, PacBio joined the field releasing the first PacBio RS in late 2010 and thus introducing high-throughput long read sequencing. Oxford Nanopore announced their portable MinION technology already two years later in 2012. Both long read technologies started with high error rates and their general applicability to many sequencing use cases was doubted.
PacBio reached a competitive error rate with the introduction of the v4 chemistry and the Circular Consensus sequencing (CCS) method. This method includes a basic self-correction directly into sequencing by using a consensus of up to 8 passes (iterations of sequencing) of the same sequence template, reducing the error rate down to as low as 2%. Oxford Nanopore similarly started improving with updated versions of their flowcells, but also with the improvements of their HMM-based base-calling algorithm using the input from the MinIon Access Program (MAP) project and the later influx of data after the release of the MinION and the developer program. Both companies encourage the developer community especially in academic bioinformatics using developer programs and offering workshops and meetings not only for users but also specifically for people developing tools related to their technology. This inclusion of the community immensely helped the speed of development in the field of long read sequencing.
A common classification of sequencing methods is to separate them into four generations of sequencing, with the first being Sanger sequencing and similar chain-termination methods leading up to the fourth being accounted to *in situ* sequencing [56]. Currently, the second (short read) and third generation (long read), often summarized as next-generation sequencing (NGS), are the most commonly used technologies.

2.2 Use Cases of Sequencing

Nowadays, sequencing exists in many different variations of the actual technology and protocols for DNA preparation and sequencing to use in different situations are available. This variety has made it a widely available tool with very many use cases. The topics presented here are selected based on the methodology used for the projects presented in Chapters I and II. They are

by no means a complete representation of the possibilities of sequencing.

2.2.1 Targeted Sequencing

Before high-throughput sequencing became cost-efficient and new analysis tools made it feasible to investigate the generated data in a reasonable timeframe, targeted sequencing methods were developed to reduce the needed sequencing depth while providing an opportunity for highly specific analysis of a sequence of interest. Targeted sequencing adds a selection step to the sample processing. If this step has the right specificity and sensitivity, analysis results can improve through the better signal to noise ratio achieved by targeting. The method usually involves using specific primers to sequence genes related to some trait or disease or a physical capture method which will select sequence with specific physiochemical features to be sequenced. Examples for targeted sequencing are Whole Exome Sequencing (WES) for eukaryotic genomes but also selective sequencing of rDNA genes for taxonomic analysis of microbial genomes in environmental samples.

16S rDNA sequencing Sequencing only the 16S rDNA of bacteria by using specific primers which match to one of the conserved sequences in this gene and then sequencing into one of the variable sequences of it can be used to obtain a taxonomic profile of an environmental sample. With the correct choice of the variable sequence to be studied and long enough reads to include a useful amount of informative bases, many bacteria can be identified or classified up to a given taxonomic rank using this method. It is still not possible to identify all of them to genus level, much less species level though. Better databases to compare the sequences to and to produce markers for marker-based identification methods to improve our knowledge. On the other hand, they increase the likelihood to find more different species with similar 16S genes, making the identification less specific.
The use of 16S sequencing in modern research has become a topic of controversy. With more advanced options like whole genome shotgun (WGS) metagenomic sequencing available at a feasible effort and price point, it is now debated that 16S sequencing is becoming obsolete. There are of course apparent limitations of 16S sequencing, which will not provide information for the full taxonomic composition of a sample, as it is unable to identify eukaryotic and fungal species in the sample. Nor is it able to generate information on the functional capacity of the studied community. However, there are also clear benefits. 16S rDNA databases include more specific information about the taxonomy than protein or genome databases, which contain many hypothetical proteins and sequences of unknown origin. Even if the sequence cannot be classified, the fact that is is from a 16S gene means it is of bacterial origin.

The fact that this is a targeted sequencing method ensures that most of the sequencing output will be used in the final analysis results, whereas in WGS metagenomic sequencing, depending on the analysis a high percentage of reads will not be informative enough to be included in the results. While it does cost much more money per sample to do WGS metagenomic sequencing, in some cases up to 60 percent of the generated sequence might not be beneficial to the analysis.
Many of the limitations of 16S sequencing including the inability to have absolute instead of just relative quantitative results or the inability to determine cause-and-effect relationships [38] also do apply to WGS metagenomic sequencing. Especially when sequencing environments that have not been studied as extensively as the human or mouse gut, it can be more feasible and informative to utilize this technology, to get a first overview of the community and diversity of an environment.

2.2.2 Shotgun Sequencing of Short Reads

Unlike targeted sequencing, shotgun sequencing of short reads will generate random subsequences of the input DNA. DNA has to be broken into a set of small fragments by using methods like enzymatic shearing or sonification. These fragments then undergo size selection if necessary, to produce an input library of a fixed fragment length range. Thus, as the fragment length is known, paired-end or mate-pair sequencing can produce sequence reads from the ends of fragments which have a known distance from each other. The longer the fragment, the larger this distance will be, but also the standard deviation of size selection is generally higher for longer fragments. Hence, short insert sizes for paired-end reads provide a higher certainty of the distance than long distance mate-pair reads.
Technologies like Illumina or IonTorrent can then be used to sequence the resulting fragments into a set of fixed length output reads. Roche 454 sequencing did produce output reads of varying length but is now obsolete as a technology, and I will therefore not further discuss it in this work. Short read technologies can produce high-throughput data in a feasible timescale and for continually lowering prices, which makes short read sequencing feasible to generate high coverage sequencing of isolate genomes or to generate high sequencing depth for environmental samples. It also makes it possible to have multiple samples sequenced in the same run or in a short time, which gave rise to massive projects sequencing, for example, hundreds of isolates for one bacterium or time series data for multiple individuals' microbiomes.
Sequencing errors in short reads are typically randomly distributed substitution errors, and homopolymer stretches. For genome assembly, we can easily detect those errors and remove them utilizing the knowledge achieved from high coverage. In 16S sequencing and metagenomic samples, this can not be

done as easily, and the errors can lead to incorrect placement and classification of the reads. However, if we set alignment or clustering parameters appropriately, that should not have a significant effect on the resulting overall abundance measures.
Short read sequencing does present problems for assembly if genomes include repeats or copy number variants (CNVs). In the sequencing of *B. vulgatus* mpk in Section 3 this is represented by the many mobile elements found in the genome which are both paralogs of each other and can be located in different positions for different organisms in the selected and sequenced colony of bacteria. Thus, long read sequencing was chosen to be able to produce a high-quality result.

Assembly of Reference Genomes Medical research and diagnostics are very dependent on the availability of good reference genomes both from microbes (commensal bacteria or pathogens) and model organisms. Human and mouse references are abundant and consistently improved, and many common pathogens have also been well studied in the past. What is still lacking are references for the many commensal microbes associated with a human host and alternative model organisms.
Assembly methods have to be chosen according to the read length, sequencing error model, genome coverage achieved by the sequencing, availability of paired reads or mate pairs, genome complexity and other parameters.
Current projects to sequence new bacteria are often done using short reads to be cost effective and quick to process. Large eucaryotic genomes or genomes including many repetitive and mobile elements cannot be well assembled only using short reads, and thus sequencing of mate-pairs and long reads is often employed. These longer sequences can then be used in addition to short reads in a hybrid assembly or independently - as is done increasingly often and described in 2.2.3.

Metagenomic Sequencing Metagenomic sequencing using short read technology enables researchers to conduct large-scale studies including a plethora of individual samples and can enable hospitals to sequence many patients samples in one sequencing run. Using high-throughput sequencing technologies which now provide read lengths of up to 300 bp and paired reads, for example, Illumina MiSeq or HiSeq, these datasets can be sequenced in a reasonable timeframe and for a feasible cost. With the barrier of generating a data set much lower than just five years ago, the bottleneck of analysis is now becoming more important.
WGS metagenomics does study not only the taxonomic composition of the samples but also their functional potential, thus providing a lot of additional information compared to 16S sequencing. It will also cover all organisms found

in the environment and is not selective for microbes only. However, for successful metagenomic sequencing, experiment planning is incredibly important. Experiment design must be adapted so that significant results can be produced and sufficient material, as well as informative metadata, can be collected. Preliminary knowledge about the expected diversity of the sequenced microbiome as well as potential problems with DNA extraction or other crucial steps of sample preparation has to be used to determine the best approach. Metagenomic analysis is highly dependent on the knowledge stored in databases and sensitive to errors introduced by contamination and mistakes in the database used. This contamination is an increasing problem, as databases are growing exponentially and curation cannot be done in a timely way anymore.

2.2.3 Long-read Sequencing

Long read sequencing technologies like PacBio SMRT sequencing and Oxford Nanopores MinION do not provide reads of a fixed length. They can generate reads ranging from very short to rather long in the same run. DNA extraction has to be quite different compared to short read sequencing to be able to achieve the longest reads. Extracting high molecular weight (long) DNA fragments is not problematic for most microbes, but can be tricky in eukaryotes, especially plants. The cell lysis step needs to be strong enough to release the DNA, but not to fragment it into small pieces at the same time. If this step is done well, both SMRT and Nanopore sequencing can provide long reads, with SMRT sequencing resulting in half of the reads being longer than 30 kb (for around 400 000 reads per SMRT cell) and Nanopore sequencing now generating reported read length of up to 800 kb .

Assembly of Reference Genomes In genome assembly, long reads are helpful in spanning repeat regions and reducing assembly complexity through the availability of large genome stretches in one read. A drawback of the technologies is their comparably high error rate and different error pattern to short reads. While short read errors are usually small random substitutions and can be corrected using appropriate algorithms, typical long read sequencing errors are insertions and deletions of multiple bases. These indels generate frameshift errors in assemblies and do complicate alignment. It is tough to distinguish them from actual biological sequence variation, as there is usually not a high coverage of the same sequence available to correct for these errors. In PacBio SMRT sequencing, the circular consensus sequencing (CCS) method can be used to highly reduce the error rate by providing iterations of sequencing the same fragment and reporting a consensus of all iterations. CSS generates a higher quality sequence but reduces the maximal read length significantly.

Another option in case of genome sequencing is to sequence with comparatively high coverage and use this coverage to self-correct the longest reads which are then used for further analysis or to sequence additional short reads which can then be used for error correction of the long reads before or after assembly.
If long reads are determined to be too error-prone for assembly, they can also be used to scaffold and close gaps in a finished short read assembly, thus reducing the effect of the errors in the reads on the assembly.

Metagenomic Sequencing For metagenomics, the main promise of long reads is the possibility to detect genes in context, which will greatly improve for example pathogen detection, but also taxonomic classification. Having multiple genes on the same read provides the context for a more specific determination of taxonomy as well as information on features that require multiple genes or complete operons in the same organism. On the other hand, the indel errors complicate the alignment step especially for alignment to a protein database because they incorporate frameshifts and it is impossible to decide if any insertion or deletion has biological meaning or is a technical artifact. Frameshifts can lead both to missing protein annotations or errors in classification.

2.3 Software Tools

This section is a short introduction to the software tools which we used in the projects of Chapter I and II. They are not a complete representation of the range of tools available but represent what was chosen to be used in the specific projects. I will describe additional information on the tools and their usage in the context of the respective analysis or pipeline where necessary.
Tools are presented in groups of similar use, with preprocessing tools including quality control, filtering and trimming of raw reads as well as merging overlapping paired reads. Assembly includes tools for short read and long read assembly, sequence alignment introduces tools based on different alignment algorithms, metagenomics will describe the use of MEGAN for metagenomic analysis, and finally, I will introduce genome comparison and visualization tools.

2.3.1 Preprocessing

Quality control: FastQC [3] For quality control of short reads produced by Illumina sequencing, FastQC offers a full workflow of tests including sequence quality, read length distribution, GC content, known contaminants or overrepresented k-mers. It was developed for Illumina HiSeq quality control

and hence is not optimized for MiSeq technology. It assumes that reads are not from targeted sequencing or multiple organisms, so the test pass and fail conditions do not necessarily apply in all cases. Still, the various statistics provided can inform the user and guide parameter selection for preprocessing and analysis of the data.
We use FastQC in all analyses described in this work for quality control of raw data and often additionally after different steps of preprocessing to track the improvements of the remaining reads adequately. It is also included in the STARA and MAPle pipelines in Chapter 8.

Merging reads: FLASH [54] Merging or extending paired reads in FastQ format is done with FLASH. Compared to other tools for merging reads, FLASH can be parametrized for the minimum expected overlap and maximum allowed mismatch ratio. It will always return the reads merged by the maximal overlap, as it starts from a "full overlap", where the shorter read is fully overlapping the longer (or reads of the same size completely overlap). It will then reduce the overlap by 1 bp each step until it reaches an overlap that does satisfy the mismatch ratio. If this overlap also satisfies the minimal overlap parameter, the read is merged. FLASH can merge comparably large percentages of the input reads in a feasible time.
FLASH is used in Chapter 5 and 7 to merge paired-end 16S sequencing reads which have been sequenced with an overlap to generate long informative sequences and in Section 8.1 as part of the STARA pipeline.

Quality trimming: prinseq-lite [71] Prinseq has been designed to provide multiple quality control steps and preprocessing of metagenomic data. In this work, I use this tool for standard quality trimming of raw paired- and single-end reads in all projects as well as filtering sequences for a minimal length where needed. It runs both on FastQ and FastA input files. Paired-read trimming results in output filtered for full pairs which pass the trimming and filtering step, thus avoiding singles in further analysis steps or the need to filter for complete pairs before continuing analysis manually.
We use prinseq-lite in Chapter 5, 7 and 6 and as part of the STARA and MAPle pipelines in Chapter 8.

2.3.2 Assembly

Long read assembly: PBCR pipeline [44] If a high coverage of SMRT sequencing reads is available, the PBCR pipeline offers the option to self-correct the reads before assembling (selfPBCR), but if this is not possible, additional short reads can instead be used. The corrected reads can then be filtered - for high coverage, it is advised to reduce the coverage of corrected long reads

to about 25x coverage of only the longest corrected reads to reduce assembly complexity. The filtered reads are assembled using the Celera assembler (since 2017 replaced by the fork of the Celera assembler optimized for PacBio and Nanopore sequences - Canu), which was initially developed to work with high-quality Sanger sequencing reads. This pipeline includes one of the few PacBio error correction algorithms available in early 2016 which does work without additional short read sequencing needed.
The selfPBCR pipeline was used in Chapter 3 to correct and assemble SMRT sequencing reads starting with a 330-fold estimated raw read coverage of the genome available from sequencing the library on five SMRT cells.

Short read assembly: SPAdes [10] SPAdes is a *De Bruijn*-based assembler for bacterial (or in general short) genome data. It extends the basic *De Bruijn* assembly by using k-bimer adjustment, where k-bimers are paired k-mers with a known distance using read context and paired-read information. SPAdes also provides options to run the first contig assembly steps on multiple k-mer sizes, then automatically selects the best result and uses this selection for scaffolding. It is a fast and reliable assembler for short read bacterial genome sequencing. To follow the trend of using long read sequencing for genomes that otherwise cannot be adequately assembled, hybridSPAdes [4] has been developed to use for short read and PacBio or Oxford Nanopore hybrid assembly.
SPAdes was used to assemble contigs from re-sequencing of *B. vulgatus* mpk colonies to study the evolution of the strain in Chapter 3.

2.3.3 Sequence alignment

Alignment to a DNA reference: MALT [33] Alignment of short reads against a nucleotide database like the 16S Microbial database or NT from NCBI or against reference genomes can be done using MALT. MALT offers multiple alignment modes including global, semi-global and local alignment. It does provide options for taxonomic classification of the aligned reads if NCBI accession numbers are available in the database or a matching synonym file is provided. For the taxonomic placement, it uses the lowest common ancestor (LCA) algorithm as implemented in MEGAN.
MALT is used in Chapter 7 to align merged 16S sequencing reads to a 16S sequence database and classify them taxonomically. It is also used in Part II in general, to align metagenomic reads to a host reference genome and filter out aligned reads to reduce coverage of host-based sequences in the further analysis. It is part of the STARA and MAPle pipelines described in Chapter 8.

Alignment to a protein reference: DIAMOND [15] With DIAMOND, short sequence reads can be aligned to a protein database with similar specificity and sensitivity to BLASTX, but up to 20 000 times faster. This enables alignments for analysis of large metagenomic datasets to be done in a reasonable time frame. DIAMOND now also includes a frameshift-aware mode, which is useful for alignment of long reads. The alignments are returned in the binary Diamond Alignment Archive (DAA) format, which can be directly read into MEGAN (see 2.3.4) for further analysis and visualization.
I use DIAMOND in Part II for sensitive alignment of trimmed metagenomic reads to a protein database and in the MAPle pipeline in Section 8.3.

Long sequence alignment: bwa-sw [53] The bwa-sw algorithm is based on Burrows-Wheeler-Alignment (BWA) and was initially optimized for contigs and scaffolds from short read assembly, but can also be used to align long reads. It is especially suitable to align long reads or assembled contigs and scaffolds to a reference genome for variant detection.
In Chapter 3 it is used to align contigs from different *B. vulgatus* mpk colonies to the *B. vulgatus* mpk reference genome.

2.3.4 Metagenomics

Sequence classification: MEGAN [35] For taxonomic and functional classification of short and long reads from environmental samples, MEGAN 6 has been developed. It can import alignments provided by the user in many standard formats or full results of other classification tools in formats like Biom or basic tab-separated format. If alignments are provided, it will place the reads on the NCBI taxonomy using the lowest common ancestor (LCA) algorithm and classify them functionally if possible by using their protein database matches to map against functional ontologies. All results are mapped through NCBI accession numbers, so if alignments have been done against non-NCBI databases, a valid synonym mapping file has to be provided to assign them properly. Functional assignment can be done using GO [6] terms placed on the InterPro [76] ontology (InterPro2GO), KEGG [41, 42], EggNOGG/COG [36, 63] or the SEED [62] database. Additional mapping files can be included to, for example, map the reads onto protein families from PFAM [28, 27] or the antibiotic resistance information from the CARD [37] database.
I use MEGAN 6 for all metagenomic classifications in Part II and in the MAPle pipeline in Section 8.3 specifically.

Detecting ribosomal small subunit sequences: Metaxa [13] Metaxa 2 uses Hidden Markov Models (HMMs) to detect reads from ribosomal RNA genes. It can detect sequences from the small subunits (SSU), large subunits

(LSU) and other given barcoding genes. Identified sequences are extracted and identified as either Bacteria, Eukaryota, mitochondrial or chloroplast and can also be classified further by using BlastN, megablast or UCLUST [24]. I use Metaxa 2 to detect SSU reads in metagenomics datasets with the MAPle pipeline in Section 8.3, but do not use it to provide taxonomic classification.

2.3.5 Genome comparison

Whole genome comparison: Mauve [21] Mauve generates alignments of multiple related genomes or chromosomes. It primarily focuses on detecting structure variations in the genomes. Mauve uses Locally Collinear Blocks (LCBs) to align the genomes to each other. An LCB is a region where at least two of the compared sequences share a homologous sequence without any recombinations. Mauve thus generates not necessarily a sequence alignment - although the homology is of course based on sequence similarity - but a structural alignment of multiple genomes. It also provides visualization of the alignment and the possibility to search for features in one genome and determine if they can be found in other genomes and if they are found in a similar or different LCB or gene neighborhood. It is very useful in comparing multiple strains of bacteria from the same species or genus, but does get less helpful for evolutionary less related genomes.
Mauve also provides additional scripts, including the Mauve Contig Mover, which can be used to order contigs or scaffolds from a new assembly guided by one selected reference genome. Ordering by a reference can, of course, mask structural differences between the genomes if the assembly is very fragmented but works well for relatively connected assemblies, as assembled sequences will not be broken in the process, only aligned to the reference and iteratively reordered until the lowest number of LCBs needed for alignment is reached.
Mauve was used in Part I to order the contigs of *B. vulgatus* mpk.

2.3.6 Visualization

Genome Annotation Editor: Artemis [67] Artemis does read in sequence and annotation data in different file formats including Genbank or GFF. It presents the sequence and annotations in a searchable and editable format. It also allows converting between different formats and export features in many formats. I use Artemis to filter and select genes by keywords or features and export them in the necessary format for further analysis. The main limitation is its inability to read in files based on multiple sequences like multi-entry Genbank files. Hence it is only useful to be used on complete chromosomes, not with fragmented assemblies.
Artemis was used in Part I to selectively extract protein sequences, inspect the annotations and to manually add a predicted protein to the annotation.

Adaptable Visualizations: Circos [**47**] Circos was designed to visualize and compare circular genomes or circular representations of chromosomes of multiple organisms. The basis of each visualization done with Circos is an outer circular ideogram ring, which can be extended with different visualizations like heat maps or scatter plots. Positions on the outer rings can be connected pairwise by lines and ribbons. A set of configuration files has to be generated to determine the basic ideogram, circular visualizations, and connections. Those files can be automatically generated from specific input files, which makes Circos a powerful but visually pleasing visualization tool.
I used Circos in Part I not only to visualize the Genome of *B. vulgatus* mpk but also other types of genomic data by writing python scripts which generate the necessary files from the available input data.

Part I

Genomics

3

Assembly and annotation of *Bacteroides vulgatus* mpk

In this chapter, I will describe the in-depth study of the genome from a commensal bacterium. The study was done in collaboration with the Department of Medical Microbiology in the Institute of Hygiene of the University Hospital Tübingen.
Providing well-annotated reference genomes for pathogens has been one of the first occasions where medicine did benefit from sequencing technology. Now that genome sequencing, assembly, and annotation are feasible for larger and more complex genomes this can be extended to sequencing known beneficial and commensal microbes in the hope to provide a better picture of the whole microbial community interacting with us on a daily bases. Those bacteria can then be studied in detail to learn about their interactions with other microbes, their host, and their environment.

This Chapter is based on the project published in the following publication:

A. Lange, S. Beier, A. Steimle, I. B. Autenrieth, D. H. Huson, and J.-S. Frick. Extensive mobilome-driven genome diversification in mouse gut-associated Bacteroides vulgatus mpk. *Genome Biol. Evol.*, 8(4):1–34, 2016

It provides an example of studying a commensal bacterium which has shown beneficial influences on its host's immune system and to prevent the outbreak of inflammatory bowel disease (IBD) in susceptible mice infected with pathogenic *E. coli* mpk. *Bacteroides vulgatus* is a common and highly abundant member of the human and mouse gut; thus it is relevant to have the annotated genome of this strain as a reference for future annotation and metagenomic studies. The task was to get a good representation of a genome with high genome plasticity through use of long-read sequencing and to annotate the resulting draft genome from comparatively few closely related genomes available.

While many bacterial genomes can be assembled to at least high-quality draft genome status using short-read sequencing with reasonable coverage, some bacterial genomes fail to be assembled well enough this way. The results, in this case, are draft sequences broken in an unusefully large number of contigs and potentially including mis-assemblies. One reason why this is the case can be high genome plasticity. In procaryotes, genome plasticity is driven by a multitude of transposons, insertion elements and other mobile proteins which can move through the genome and in-between organisms, carrying other sequences with them and leading to sequence duplication, rearrangement and horizontal gene transfer (HGT) events. The family of Bacteroidetes has been shown to harbor many of these mobility factors [61]. This plasticity is one of the reasons they are so successful in colonizing a variety of different environments, being at the same time able to be highly adapted to the niche provided in each environment [83].
In the human microbiome, *Bacteroidetes* species are known as common gut commensals [64, 60]. They are generally found in high abundance in the intestinal tract of many healthy mammals [75, 46]. However, they also include obligate pathogens, which - especially if outside of their favored environment - can cause abscesses and other infectious diseases of their host. *Bacteroides vulgatus* is one of those chameleons. Already in the early days of human gut microbiome studies, it has been identified as a core species of many healthy human gut communities by Qin *et al.* (2010) [64] but is also an obligate pathogen which can cause abscesses and inflammation in other environments. The only available *B. vulgatus* reference at that time was *Bacteroides vulgatus* ATCC 8482 (also known as DSM 1447) which had been isolated from human feces. After the publication of the full genome of this strain in 2007 by Xu

et al. [83], very few new genome sequencing and assembly projects could be found in the available databases for this species for nearly ten years. This is remarkable for a bacterium which is found to live so close to humans and so abundant in human-associated microbiota. The available assemblies all proved to be unusually fractionated for a bacterial genome. These facts hinted that there must be some genome plasticity in the species which makes it hard to assemble the full genome from short reads, even after Illumina read lengths had significantly increased from the possible 75 bp in 2007 to 250, or 300 bp reads in 2016. Many other culturable bacterial species found in the human and mouse gut microbiome had a steady flow of new and updated reference genomes for different strains added to the databases in the meantime.
Bacteroides vulgatus mpk was isolated from the feces of healthy mice. It is prevalent in mouse gut microbiota. The strain has been shown to induce the mouse anti-inflammatory immune response [14, 57, 80]. By inducing this response, it can prevent *Escherichia coli*-induced colitis in a gnotobiotic interleukin-2-deficient mouse model [80]. This does make the mpk strain an interesting mouse model representative of the *B. vulgatus* species and a good candidate to choose as a reference genome, as the previously available reference strain was isolated from healthy humans. Comparing the human and mouse strains and having the mouse isolate available as a reference for research could lead to important insights.
This section will describe how we sequenced, assembled, annotated and analyzed the genome of *Bacteroides vulgatus* mpk. The analysis includes the examination of the annotated mobile elements, searching for paralogs in the genome to determine duplication events of mobile elements and study genome plasticity and comparison to other strains of *B. vulgatus* and to other *Bacteroidetes* in general.

3.1 Sequencing and Quality Control

It was decided to use PacBio SMRT sequencing to enable generation of a high-quality draft genome of *B. vulgatus* mpk. The long-read CCS technology promised to provide around 2500 bp long reads with good quality at the time of sequencing. We extracted DNA from the isolate and constructed a 10 kbp library. This library was sequenced on 5 SMRT cells using the PacBio RSII P4-C2 CCS protocol. This approach resulted in an estimated 330-fold genome coverage.
The raw sequences were assessed using FastQC under the premise that this tool was written for short read sequencing, especially Illumina error patterns and thus could only give a fundamental idea about the interpretation of the results. PacBio SMRT sequencing has much different error sources, and patterns and

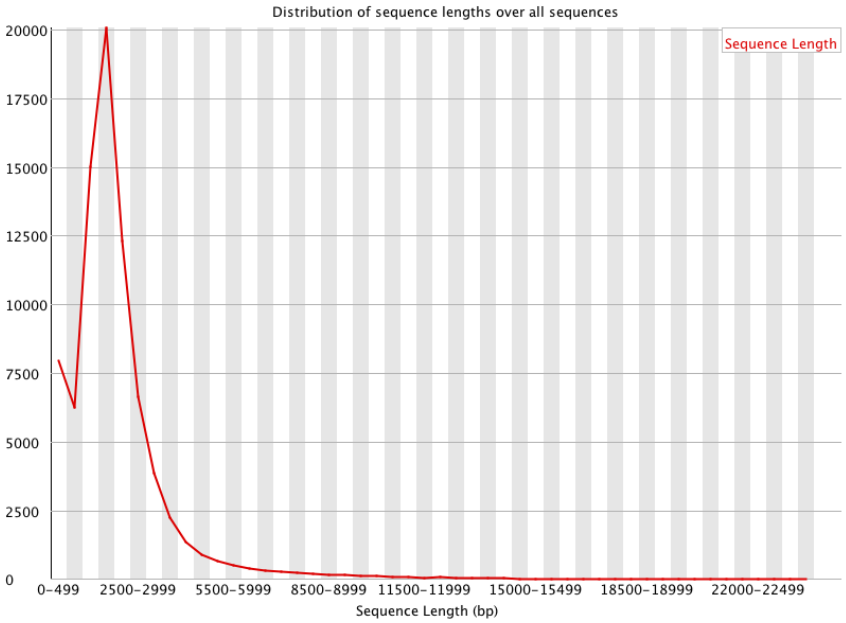

Figure 3.1: Length distribution of all PacBio SMRT sequencing reads for *B. vulgatus* mpk as provided by FastQC

quality values are not assigned in the same way and thus don't carry the same meaning. Most reads were between 1 000 and 2 500 bp long, the longest read had 23 971 bp, as shown in Figure 3.1.

3.2 Preprocessing and Assembly

Then I ran the reads through the PBCR self-correction pipeline [44]. The resulting corrected reads were downsampled to 25-fold coverage, based on advice given in the PBCR documentation. Only the longest reads needed to provide 25-fold coverage of the genome were selected to be used in the assembly step. This selection is made to improve assembly speed and quality. The subsampling step is especially crucial for this assembly, as it includes mobile elements. Keeping shorter reads covering only one of the mobile elements would make it impossible to place them correctly in the assembly and likely cause misassemblies of the genome.
The resulting filtered reads were assembled using Celera Assembler version 8.1 [58]. At the time of this project, Canu [45] was not yet available as a better option for long-read assembly. This assembly resulted in 33 contigs. The unfiltered set of all corrected reads was mapped against these 33 assembled contigs using bwa-sw [52] to check for potential misassemblies and low coverage contigs. 25 short contigs with a mapped read coverage of 10-fold

or lower were identified and removed from the assembly, as their coverage was much lower than the expected 25x fold. This filtering left eight long, well-covered contigs for the further steps. The alignments to these contigs identified 7 points of misassembly, where the contigs had to be split, resulting in the final curated assembly output of 15 contigs with a maximal length of 1 525 634 bases and a minimal length of 3 984 bases. Another alignment of the corrected reads to those curated 15 contigs showed that 96% of the original unfiltered corrected reads aligned to these 15 contigs, providing proof that these contigs accounted for a significant amount of the data and that no part of the *B. vulgatus* mpk genome was missed or removed through the filtering steps before and after assembly.

3.3 Draft genome finalizing

The 15 filtered contigs where compared to the only other available full reference genome of *B. vulgatus* ATCC 8482 [83] and sorted based on this reference using the Mauve ContigMover, a script provided by the Mauve alignment tool [21, 66]. The ordered contigs where checked for overlaps between consecutive contigs and either merged based on these overlaps or joined by inserting a spacing sequence of 100 Ns if we could find no overlap of two consecutive contigs. The *Bacteroides vulgatus* mpk draft genome is 5.1 Mbp long and has a GC content of 42.2%. These statistics compare well to the reference strain ATCC 8482 which is also 5.1 Mbp long and has a GC content of 42%.

3.4 Functional Annotation

Functional annotation of genomes is heavily based on the knowledge we already have, as we take information about predicted genes from comparison with sequences in protein databases. In this case, with only one other completed genome from the same species available, this information will be less detailed than for a bacterial strain with many other well-studied annotated genomes from the same species. However, the genus *Bacteroides* is important to human health and many of the human host associated *Bacteroides* strains have good reference genomes available. Our information will, therefore, be based mainly on those annotations.
For this genome annotation, we used three different annotation tools and curated all of the information gained through them into one final annotation of the draft genome. The three tools are RAST (based on the SEED database)

[8], BaSys [79] (based mainly on the UniProt database and additional bacterial model organism annotations) and xBase [17] (using a direct comparison with the *B. vulgatus* ATCC 8482 reference, which was originally annotated using multiple tools for gene finding and public databases for functional annotation). This way, the final annotation incorporates three different references or reference databases, but only two different methods for gene finding. RAST and BaSys both use GLIMMER [23] to predict protein-coding genes, while xBase uses a full pairwise genome alignment. Because of that, gene predictions are potentially better from RAST and Basys, as they are based on the actual input sequence with all mutations and potential frameshifts, while xBase could incorrectly predict protein boundaries. On the other hand, for proteins unique to the species, xBase functional annotation is potentially more specific than the comparisons to two large protein databases.
The merging and curating process of annotations was guided by weights for the different annotation methods and strong and weak removal criteria. RAST annotations were given the highest weight, both for selection of gene locus (start and stop position) and functional annotation. xBase was the second choice for functional annotation, while BaSys was the second choice for gene locus. The second choice was selected if the first choice was not available, because there was no annotation from this tool overlapping with the current locus or the locus was not functionally annotated. After merging the three annotations into one based on those weights, genes were filtered out and removed from the genome annotation if they matched either all three strong removal criteria or a total of four strong or weak criteria together.
The strong removal criteria are

- gene length of under 150 bp
- annotated as hypothetical protein by all methods which detected a gene at this locus
- no protein homology in comparison to selected related strains of other *Bacteroides*

The weak removal criteria are

- suspicious start codons
- overlapping another already accepted gene
- opposite reading direction (strand) to all neighboring genes

The final annotations also include Gene Ontology (GO) terms as provided by BaSys, if the accepted functional annotation matches the BaSys functional annotation. The resulting annotated draft genome can be found in the NCBI

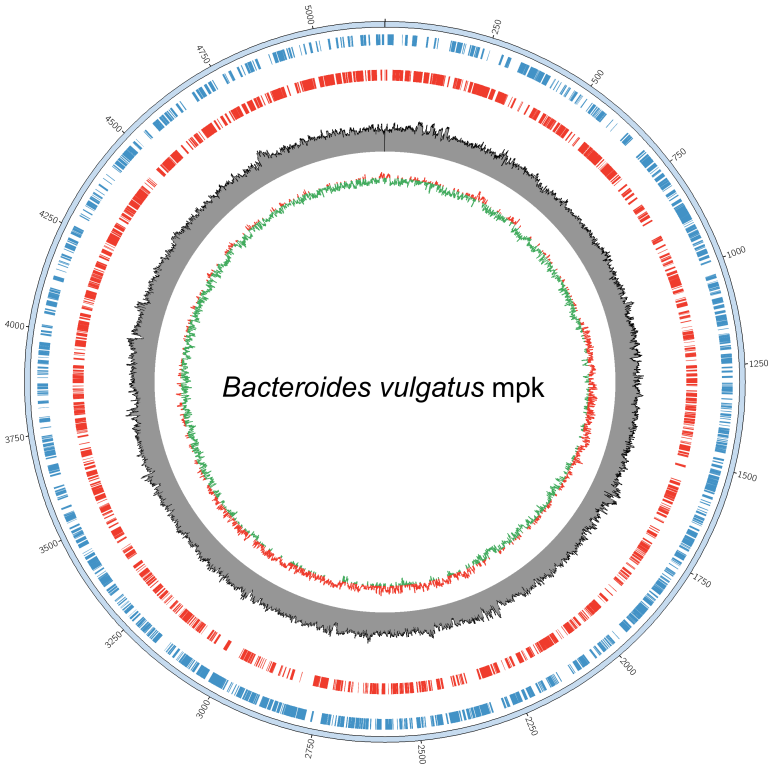

Figure 3.2: Circular representation of the *B. vulgatus* mpk genome, visualized using Circos [47]. Blue features represent forward strand annotations; red features reverse strand annotations. The grey curve represents GC content, and the green and red curve represents positive vs. negative GC skew.

GenBank database under the accession CP013020.1 and depicted in Figure 3.2, a second version of the genome annotated by the NCBI Prokaryotic Genome Annotation Pipeline is found under accession NZ_CP013020.1 and has been added to the RefSeq database, which means the genome assembly passed the RefSeq filter for assembly quality. The genome of *Bacteroides vulgatus* strain mpk codes for 4233 proteins, 79 tRNAs, and 21 rRNA genes in 7 operons. It harbors, as was expected, a high number of mobile elements, also in direct comparison to the reference strain ATCC 8482 (see Table 3.1). Other strains available only reached scaffold level (10 strains) or even just contig level (5 strains) assembly status on the NCBI genome overview.

	ATCC8482	ATCC8482 RefSeq	mpk	mpk RefSeq
CDS	4065	4226	4233	4304
tRNA	85	83	79	79
rRNA	21	22	21	22
transposase	89	77	133	113
integrase	17	27	33	29

Table 3.1: Comparison of annotations for *B. vulgatus* strains ATCC 8482 and mpk and their NCBI Prokaryotic Annotation Pipeline (RefSeq) reannotations. Coding sequences and genes annotated as transposase or integrase were counted using the Artemis Genome Browser [74] to search for relevant keywords.

3.5 Comparison to other strains of *B. vulgatus*

Currently, there is only one *Bacteroides vulgatus* genome available on GenBank that is counted as complete - which is the ATCC 8482 strain, and the mpk strain is the only other sequence which is seen as a complete chromosome. The estimated genome sizes for *B. vulgatus* strains range from 3.5 Mbp to 5.4 Mbp. There are only four strains (including the already mentioned ATCC 8482 and mpk) which could be assembled into less than 20 scaffolds. These four most connected assemblies have an average genome size of 5,09 Mbp, average GC content of 42,13% and code for 4324 genes on average, of which 3901 on average are protein-coding. The average for protein-coding genes is brought down by the NCBI annotation of the *B. vulgatus* mpk RefSeq version, which assigned an unusually high number of genes (793, thereof 714 based on frameshift errors) as pseudogenes, probably because of problems with frameshift errors common to PacBio sequencing changing the stop codons significantly. Using our original annotation instead of the RefSeq version, the average would have been 4082 protein-coding genes.
Both of the genomes with complete chromosomes have an original annotation and a RefSeq version of the annotation available. Comparing these annotations (see Table 3.1) shows small differences between the different annotation methods, but generally more transposases and integrases incorporated in the mpk strain genome.

3.6 Genome comparison to *Bacteroidetes*

Proteins from the *B. vulgatus* mpk annotation were extracted and compared to proteins from eight other *Bacteroidetes* genomes, which were *Bacteroides vulgatus* ATCC 8482, *Bacteroides dorei* isolates HS1_L_3_B_079 and

HS1_L_1_B_010, *Bacteroides xylanisolvens* XB1A, *Bacteroides fragilis* 9343 and YCH46, *Bacteroides thetaiotaomicron* VPI-5482 and *Parabacteroides distasonis* ATCC 8503. We chose the other *B. vulgatus* and the two *B. dorei* because they are the closest related strains and species available for *B. vulgatus* mpk. The others are reference genomes for common human or mouse commensal *Bacteroides* species with *Parabacteroides distasonis* as an outgroup inside the *Bacteroidetes*, all other strains being *Bacteroides*. Orthologs between the proteins from these genomes to the predicted proteins from *B. vulgatus* mpk were defined to be pairwise matches of at least 50% identity with the alignment covering at least 60% of the longer protein and a blast match e-value of under 0.0001. These hits were further filtered to have minimum 90% as well as 98% protein sequence identity to select closely related proteins between the strains and species. These numbers were selected to represent orthologs both with low variability (98% identity required), but also orthologs for proteins which are generally more variable (only 90% identity required).
The result of this comparison is that when filtering for minimum 50% identity, *B. vulgatus* mpk has 831 unique proteins with no orthologs found in this group of related genomes. The genomes from both *B. vulgatus* and both *B. dorei* strains together share unique 691 genes which cannot be found in the other *Bacteroidetes* genomes.

3.7 Studying *B. vulgatus* genome plasticity

In this project, we assumed that the various abilities of *B. vulgatus* mpk as a commensal, obligate pathogen and as protective against *E. coli* induced colitis are based on its high genome plasticity, which is promoted by the multitude of mobile elements incorporated in the genome. There are conserved regions, where the incorporation of a mobile element would be lethal to the organism, and variable regions which allow mobile elements to move through the genome and in many cases also change their copy number. It is interesting to see that there are so many mobile elements known in bacteria, as bacteria have a high density of genes and few and short intergenic regions where mobile elements could jump into without causing damage to genes or important regulative sequences.
As most of the genes that differ between the *B. vulgatus* strains - and also between *B. vulgatus* mpk and other *Bacteroides* species - belong to mobile elements, we focused on studying these elements and how they were both horizontally transferred from other bacteria and have moved through the *B. vulgatus* mpk genome, generating paralogs of the same mobile element in different positions.
B. vulgatus mpk includes a complete type I-C CRISPR/Cas system with CRISPR repeats and twelve spacer sequences. This system was first iden-

tified in *Bacillus halodurans* C-125 [59], and can also be found in *Bacteroides dorei* and some *B. vulgatus* strain assemblies, but not in the reference strain *B. vulgatus* ATCC 8482. The function of this specific type of CRISPR/Cas system is not well studied, and the twelve spacers did not show significant similarity to known CRISPR target databases, thus it is not known if this system is targeting bacteriophages, other bacteria or potentially even some mobile elements that had previously been transferred to the genome.
We first identified the CRISPR/Cas system through the annotated genes which showed a large operon of Cas genes. I confirmed this finding using the online tools CRISPRloci and CRISPRmap [50, 1]. They identified the type of system, helped find the missing annotation of the Cas2 gene and detected the spacer and repeat sequences. The annotation of Cas2 could be further strengthened by comparing the genome sequence to Cas2 proteins from the NCBI NR database, and led us to manually adding the Cas2 gene annotation to the draft genome.
Comparing the CRISPR spacers to the full NCBI NT database with an e-value cutoff of 0.05 does not return any significant non-self hits other than one match for the first spacer against the *B. dorei* CL03T12C01 strain assembly. Hence it is not possible to predict any targets for this CRISPR system. It can, however, be assumed those are functional spacers which target unknown mobile elements [72].
B. vulgatus and in general *Bacteroidetes* genome plasticity has been shown previously to be driven by conjugative transposons which enable horizontal gene transfer [20, 73]. These transposons can carry antibiotic resistance genes amongst other things. The complete conjugative transposon found in *B. vulgatus* mpk is closely related to transposons from *B. xylanisolvens* strain XB1A and *B. dorei* isolate HS1_L_1_B_010, which hints that the transposon is transferred regularly in-between different *Bacteroides* species. It is very similar to the conjugative transposon CTn341 [9], which was found first in a human isolate of *B. vulgatus* and carries a tetracycline resistance. *B. vulgatus* ATCC 8482, in comparison, does contain some conjugative transposon proteins but lacks important parts of the full transposon, making them potentially non-functional.
Analysis of orthologous proteins showed that the complete conjugative transposon of *B. vulgatus* mpk does carry an insertion of 11 additional proteins compared to the original CTn341 (see Table A.1), but also lacks a protein with homology to TetQ. The inserted proteins include transcriptional regulators, a glyoxalase-family protein and a β-lactamase gene. The complete insertion sequence is absent from any other *Bacteroides* species found in the NCBI database. Some of the proteins do have distant homologs in other *Bacteroides* species. As horizontal gene transfer (HGT) is often either neutral or even detrimental for the accepting organism, it is normally lost quickly after integration. Retaining the large conjugative transposon and a long insertion

sequence suggests that these - especially the additional β -lactamase gene - provide a fitness benefit to *B. vulgatus* mpk. Antibiotic resistance experiments did prove that it has an additional cephalosporine resistance compared to other *B. vulgatus* strains [49]. We assume that the β-lactamase gene in this insertion could be responsible for this additional resistance.
Mobile elements not only move between species in HGT events, but they can also be copied and moved inside one genome, creating copy number variations of proteins. We analyzed the predicted proteins in *B. vulgatus* mpk for paralogy, similar to a comparative analysis of homology to proteins from other organisms. Through this similarity analysis, we could identify two insertion elements which occurred multiple times in the genome with no or few changes to the included gene sequences.
IS21-like elements usually consist of two open reading frames (ORFs). We could identify two types of IS21-like elements in the genome. One we found in three almost identical copies and another lightly mutated one. The less similar copy is located downstream of two conjugative transposon proteins. One of the identical copies is directly next to the conjugative transposon described earlier. As IS21-like elements are not known to include additional genes when jumping to another position, we assume that this particular copy was generated through a different duplication mechanism.
The second IS21-like element is different from the first one, but all 13 copies in the *B. vulgatus* mpk genome have very high sequence similarity. This specific pair of IS21-like ATP-binding protein and transposon gene is only found once in *B. vulgatus* ATCC 8482.

Through our paralogy and synteny study of the genome, we also found another potential IS element, consisting of a transposase always paired with similar genes annotated as a hypothetical protein. This pairing is found 18 times in the *B. vulgatus* mpk genome and also six times in the *B. vulgatus* ATCC 8482 genome. Using multiple sequence alignment and clustering with ClustalΩ we determined that the 24 hypothetical proteins from these pairs in both genomes group into four clusters of high sequence similarity. The four clusters show partial similarity in-between each other (see Figure 3.3), so could share similar function to some extent.

3.8 *B. vulgatus* mpk genome evolution

The prevalence of mobile elements in high copy numbers in the *B. vulgatus* mpk genome led to the idea of attempting to study their movement and change in copy number through a short-term mutation experiment. In this experiment, the original culture of the *B. vulgatus* mpk strain, which we sequenced and studied in detail before, was passed through different environments for six weeks. The bacteria were either colonized on an agar plate or passaged through

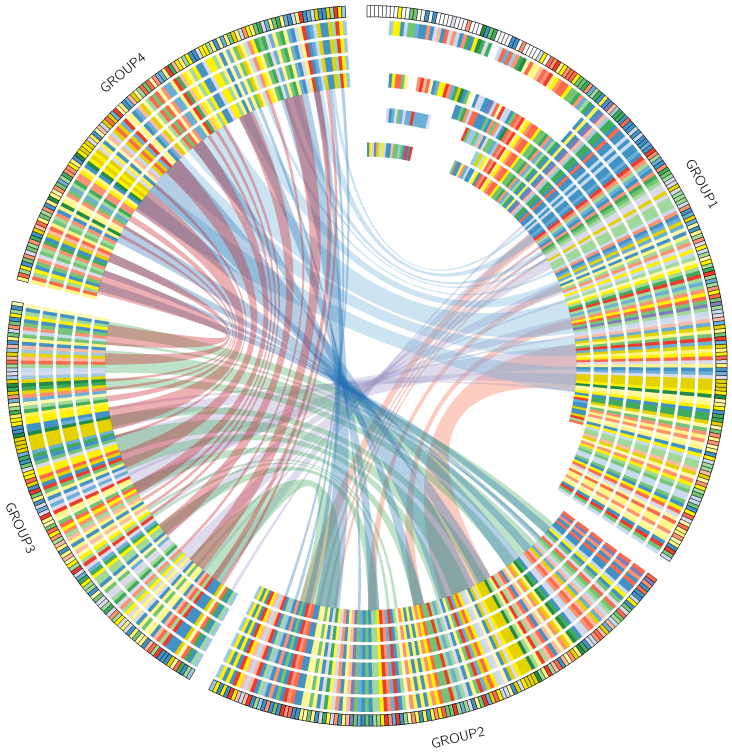

Figure 3.3: Mapping of the four clusters of hypothetical proteins associated with transposases in *B. vulgatus* mpk and ATCC 8482. The multiple sequence alignments of the sequences in the cluster are depicted as the outer rings; colors stand for amino acids. Consensus sequences for each cluster form the outmost ring. Matches from a multiple sequence alignment of the consensus sequences are depicted as links between the groups, where each pair of groups is shown in a different color.

the gut of a germ-free mouse by oral gavage. Both settings also included colonizing only with *B. vulgatus* mpk, co-colonization with *Escherichia coli* mpk and colonization together with a stable defined minimal mouse gut microbial flora as described by Uchimura *et al.* (2016) [78]. Isolating *B. vulgatus* mpk out of any of the samples is possible through selection by antibiotic resistance. After six weeks, the colonies were isolated and sequenced on an Illumina MiSeq, assembled using SPAdes [10], and I mapped the resulting scaffolds onto the *B. vulgatus* mpk genome using bwa-sw [53]. The analysis showed that the time frame of the project was not enough to witness a significant change in the genome, in particular, the copy number of the mobile elements. The methodology of using short reads resulted in very fragmented assemblies, and changes in the location of the mobile elements might have been only present in a part of the bacterial population. Thus, the changes we were looking for were masked through the coverage of bacteria without changes and the complexity of assembling short reads over those repeated sequences. This problem is potentially only to be solved using a longer experimental time frame and using single-cell sequencing to determine the exact changes in single organisms. When using single-cell sequencing, unassembled reads could have been simply mapped to the reference genome to track variation in the population.

3.9 Evolutionary dynamics of *B. vulgatus*

Garud *et al.* (2017) [29] studied the evolutionary dynamics of multiple prevalent bacterial strains from the human gut microbiome by variant calling them in a panel of time series stool samples of healthy humans sequenced with high coverage. As an example of a highly prevalent and abundant species with a fairly divergent genome, they used *B. vulgatus*. They found two distinct clades of potential *B. vulgatus* subspecies which can be found in many different gut metagenomic samples. Some of the samples also had high within-host polymorphism of *B. vulgatus* which could contain lineages from both clades, showing that these clades were not derived through full isolation by any external factor like host location or diet.
They also focused on short-term inter-host differences of the *B. vulgatus* population and showed while it is usually distinctly lower than in-between host differences in successive time points, in some cases, there were significant changes in the nucleotide differences in successive time points from the same host. These time points were still all at least multiple months apart - the exact number for the study of short-term changes is not clear from the publication - and thus investigated over a longer timescale than the experiment discussed in Section 3.8. They assume that the bigger changes are caused by replacement events where another lineage from the full population of *B. vulgatus* replaces the most prevalent lineage in the host.

A different assumption using the knowledge of *B. vulgatus* genome plasticity could be that, while the constant smaller changes are usually not prevalent in the whole population, environmental events could cause selective pressure which enforces adaptation. While other less adaptive species would become less abundant or even be replaced by different better-adapted species of their genus, *B. vulgatus* can retain its prevalent role in the microbiome while adapting to the change as a species.

4

Conclusions from Part I

Bacterial genome assembly and annotation are widely seen as a straightforward problem in comparison to similar projects for larger, eukaryotic genomes. The attempts where no satisfactory results can be generated are either scrapped or in the best case deposited in databases without further analysis. Nevertheless, it could be beneficial for the future of bacterial genomics, which influences medicine, metagenomics and often provides the simple examples for testing new algorithms to study the cases where standard sequencing, assembly, and annotation pipelines still fail to provide satisfactory results for a bacterial species in more detail.

With the rise of metagenomics and the low cost of sequencing bacterial isolates, increasing amounts of bacterial sequence are generated but never thoroughly analyzed. Especially if standard available methods and tools fail, the knowledge we could gain through an in-depth study of the genome is often not generated. With the rise of SMRT sequencing and high quality reads generated from the CCS protocol, new possibilities have opened up to successfully study the neglected bacteria where short read sequencing and assembly tend to fail. This development also showed that methods for assembly of long reads needed to be updated and developed further, to enable the development of new standard tools like PBCR together with Canu for PacBio error correction and assembly. With the improvements of Canu specifically for long read sequencing compared to the original Celera Assembler, this could potentially have led to a better assembly of *B. vulgatus* mpk from the beginning and less need for manual curation, making the project more feasible to repeat on other bacterial strains with high genome plasticity.

Working through adding those additional strains and species to the databases and improving on the available references will then, in turn, improve our abilities to annotate new genomes by increasing the potential number of closely related genomes and annotations already available for comparison. The growing number of well-curated references will also provide a better basis for comparative genomics and the study of genome evolution in bacteria. With sequencing

single bacterial organisms instead of full colonies using single-cell sequencing technology, better studies of the genome plasticity and evolution of bacteria will be possible. Being able to study those genomes will be crucial to improving our knowledge on how pathogenicity works and how commensal bacterial shape their environment while also being shaped by it.
While the number of new bacterial strains that are sequenced is still on the rise, currently the increase of actual knowledge generated by this sequences is much lower than it could be if proper methods, tools, and curation would be used. Instead, often another sequence is generated to merely be added to the flood which is producing increasing contamination and errors in the databases we rely upon so heavily.

Part II

Metagenomics

In this part, I will present three different projects studying environmental microbiomes to generate medical information as well as three pipelines developed during those projects. The projects study the gut microbiome of either mice or humans under various premises and using different technologies and methods. To produce analyses for all of the use cases, I have developed pipelines which can be used to reproducibly analyze samples coming in over a period of time and producing comparable results. I will describe each of the projects background and specific needs, the bioinformatics pipeline used for the analysis and present exemplary results and further analysis that can be done with the output of these analysis pipelines.

This part is based on analyses done using MEGAN 6, as published in the following publication:

D. H. Huson, S. Beier, I. Flade, A. Górska, M. El-Hadidi, S. Mitra, H.-J. Ruscheweyh, and R. Tappu. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS computational biology*, 12(6):e1004957, 2016

MEGAN 6 is the newest version of the MEGAN tool for metagenomic classification, analysis, and visualization. Sequences can be classified taxonomically based on the NCBI taxonomy and functionally, based on several functional ontologies available. The user can upload metadata for all samples which enable further analysis. Large datasets with high numbers of samples can be compared and visualized. MEGAN 6 supports both short read and long read modes for classification and can also read in output from other classification tools for further analysis and comparison.

The three different projects presented differ in their setup and goals. Firstly, the study of *Yersinia entrocolitica* infection in a mouse model in Section 5 was an investigative study, using both 16S rDNA sequencing and Whole Genome Shotgun (WGS) metagenomics to study changes in the gut microbiota of a mouse model after infection with different substrains of *Yersinia entrocolitica*. The goal of this study was to investigate how the microbiota change taxonomically and functionally after infection and how the original microbiota of the mouse and its changes do affect the outcome of infection.
The second project described in Section 6 presents a full metagenomic pipeline set up for a medical study investigating the gut microbiome of stem cell transplant patients before and after their transplant and compare them with their

donor's microbiome. This setup had a specific infrastructure available in the University Clinics of Tübingen as a basis and is developed to be used in a study as an investigative tool, but also potentially after the successful end of that study in daily clinical healthcare, using the results from the study to improve donor selection and transplant success rate for hematopoietic stem cell transplants.
The last project presented in this chapter is entirely different in the underlying idea and experimental setup. Section 7 is about an analysis pipeline for a community science project called the "TüBiom project." I developed a stable and fast analysis pipeline for this project, where interested participants can sample their gut microbiota and send them in for a free 16S rDNA analysis. The results were presented to the participants over a web service, where they can compare their gut community to the average microbiota of other participants with selected features. The project aimed to collect data for the study of the general microbiota of the population to promote a better understanding of its variability in health and disease.
The pipelines I have developed during these projects have been continuously adapted and made more generally applicable. They are based on freely available tools and run on any Linux system. The pipelines are collected in the CommunAl toolkit for microbial community analysis which I describe in Chapter 8. This part ends with a conclusion from the projects and analyses which I have presented in the final Chapter 9.

5

Changes of the mouse gut microbiome during infection with *Yersinia entrocolitica*

As a part of the Priority Program SPP1656 "Intestinal Microbiota - a Microbial Ecosystem at the Edge between Immune Homeostasis and Inflammation" funded by the German Research Foundation (DFG), this project, named "Interaction between Yersinia enterocolitica, the intestinal microbiota, and the host: From molecular analysis to therapeutic intervention", aimed at studying mechanisms of colonisation resistance against infection with *Yersinia enterocolitica* in mice.
To be able to study the changes of microbiota during *Yersinia* infection, we collected time-line samples from feces of infected and control mice as well as singular time point samples from the illeum of these mice. We collected general metadata about the mice and infections, and the samples were analyzed both using 16S rDNA sequencing and WGS metagenomics.
This project is a prime example of the importance of including bioinformaticians in experiment planning and the various problems that can still arise and complicate the analysis of sequencing data. Unfortunately, it was infeasible to do the experiments in a setting optimal for statistical analysis for multiple reasons. First, it is impossible to get actual timeline datasets from one mouse from the illeum, as mice have to be sacrificed to collect the microbiota from the illeum. As the fecal microbiome and the illeum microbiome can differ significantly, it is hard to make a reliable connection between results from fecal samples over time and illeum at the last time point, even if they would come from the same individual. Secondly, as it is not feasible to have mice from different treatment and sampling groups in the same cages throughout a more extensive experiment, it is also not possible to distinguish significant effects of treatment from a confounding cage effect. This problem is exacerbated by the fact that mice are coprophagic. We tried to counteract this through extensive

co-housing of all mice before starting the experiments, but we cannot ignore this effect in the interpretation of the results.
Before we started to conduct the larger metagenomic studies, we conducted a preliminary investigative 16S rDNA sequencing experiment. In this experiment, 60 samples were generated using 454 sequencing of fecal samples from mice infected with different strains of *Yersinia enterocolitica*. One strain is the wild-type (WT), the other three were mutants of this strain with potentially lower virulence. The mutants used in any of the following experiments are knockouts of known virulence proteins. The mutant ΔYadA, called A0 for short, is deficient of the Yersinia adhesion protein YadA, which enables it to attach to a host cell. The Δ pYV515, called pYV for short, mutant is deficient in a virulence plasmid, thus not able to secrete the Yersinia Outer Proteins (YOPs). Lastly the Δirp1 mutant, called irp1 for short, lacks one gene of the yersiniabactin biosynthetic gene cluster.
The preliminary data was analyzed using QIIME and did confirm the assumption that the oral gavage of *Y. enterocolitica* leads to changes in the gut microbiome of the infected mice. Also, the changes were different for the different mutants, so we assumed they were not only based on on the inflammation of the gut, which is much less severe in the mice infected with any of the mutants. As the experiment did include only limited replicates, it had low statistical significance. It also only covered a very limited timescale (one, two and three days after infection), did not monitor the microbiota before infection and included no strong negative control. Hence I will not describe the analysis and results of this experiment in detail but focus on the later experiments in which we incorporated the knowledge gained.
The first major experiment of this project was a competition experiment, where we colonized mice with both wild-type and one of the mutants of *Y. enterocolitica* to see the fitness of both substrains in direct comparison. We used this investigative experiment to test multiple improvements to the experimental and sequencing strategy and analysis before moving further to the main experiment.

5.1 Investigative Competition Experiment

The setup of the competition experiment included 56 mice which were either infected with a mixture of *Y. enterocolitica* wild-type (WT) and ΔYadA or wild-type and Δ pYV515, kept wholly untreated or received oral gavage of Phosphate-buffered saline (PBS). PBS (isotonic saline solution) injection is a negative control for infection, but generates the same stress level for the animal to gavage of *Yersinia* and is thus an accurate negative control for infection. It rules out changes in the microbiota caused by handling and stress or the amount of fluid given through the gavage are the primary factor

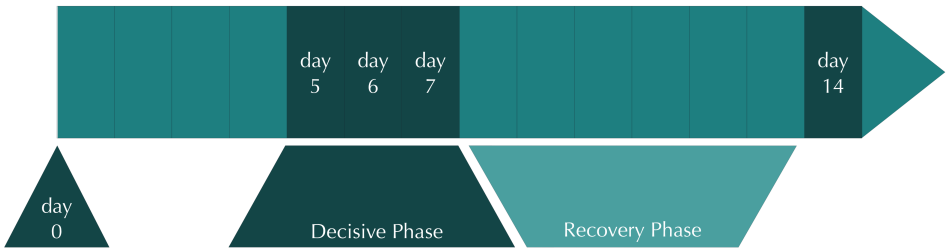

Figure 5.1: Timeline of the competition experiment. Two weeks of co-housing the mice were done before the experiment, utilizing the mouse coprophagy to even out the microbiome at the beginning of the experiment. Day 0 samples were collected to confirm the success of the co-housing strategy. Additional samples were taken on day five to seven and two weeks after infection.

in the difference between test and control samples. The infected mice were specific pathogen free (SPF), so known not to have *Y. enterocolitica* or other pathogens that would influence the course of infection before the experiment started.

We collected samples before infection(day 0) and five, seven and fourteen days after infection (day 5, day 7, day 14) (see Figure 5.1). As we already knew that there was high variability in the microbiome due to the onset and peak of inflammation in the first three days after infection, those days were omitted. The main changes between recovering mice or mice succumbing to the infection should present between day five and seven, were succumbing mice usually reach the point of fast weight loss while recovering mice can already hold their weight or start gaining. Day 14 would then show the microbiota of recovering mice when they should be fully recovered from the acute infection, even though some recovering mice can still excrete low abundances of *Y. enterocolitica* in their feces at that time. As multiple mice were suffering from fast weight loss between day five and seven, some additional mice had to be sacrificed on day 6, adding another timepoint (day 6) to the later analysis.
Numbers of mice in the different groups are shown in Table 5.1.

The samples sequenced were selected from these available samples. 16S rDNA extracted from feces or luminal content of the ileum or both locations for some mice and was sequenced on an Illumina MiSeq resulting in 42 samples of either feces or luminal content of the ileum from 30 different mice as shown in Table 5.2.

It is clear from comparing the tables that we were still expecting most of the differences in the microbiota to present in day 14 mice which had either started to hold or gain weight again or were still suffering from weight loss.

Infection	day 0	day 5	day 6	day 7	day 14
uninfected	3	0	0	0	6
WT+ A0	-	1	0	1	15
WT+ pYV	-	0	2	2	10
PBS	4	2	1	2	7

Table 5.1: Numbers of mice for each group in the competition experiment. The experiment included 56 mice, 16 infected with PBS, 9 uninfected, 17 infected with the A0 mutant and 14 infected with the pYV mutant. Some additional mice had to be excluded early in the experiment in the case when oral gavage was not successful.

Infection	day 0	day 5	day 6	day 7	day 14
uninfected	3/0	0/0	0/0	0/0	0/0
WT+ A0	-	1/0	0/0	1/0	4/4
WT+ pYV	-	0/0	2/0	2/0	4/0
PBS	4/4	2/0	1/0	2/0	4/4

Table 5.2: Number of sequenced samples from the competition experiment. Counts are given as Luminal/Fecal for each timepoint. The samples were taken from 30 different mice, so some samples can be directly compared between the two locations.

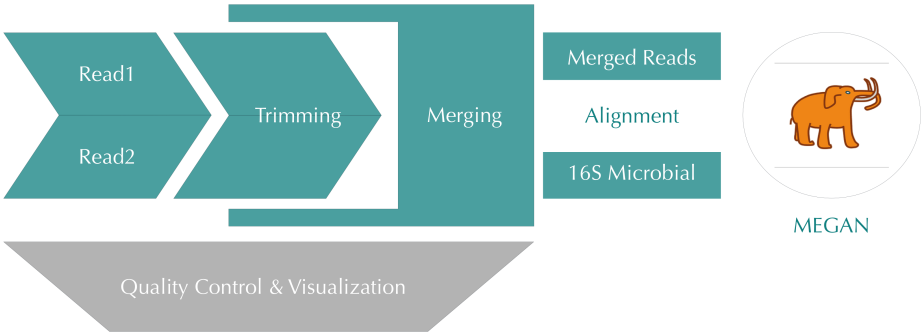

Figure 5.2: Basic 16S rDNA analysis pipeline for paired-end Illumina MiSeq data as it was used in this project. Quality Control was done repetitively for raw, trimmed and merged reads. Trimming and filtering were done using both reads of a pair, so only complete read pairs were used in the merging and further analysis. Alignment of the sequences was done against the NCBI 16S Microbial database.

This experiment made clear, that the actual mice surviving until day 14 all are recovering and that the essential changes in microbiota responsible for the outcome of infection would probably present some time between day 3 and day 7. During this experiment, mice were sampled either selectively before infection or kept until day 14. Only mice which had to be killed because of the severe progression of disease were sampled on other days. The mice from the control group for day 5, 6 and 7 were selected to be sampled as a direct comparison to the severely affected ones. This practice, of course, led to a lack of statistical significance through replicates for analysis of any of the mid-infection samples. As the primary goal of the experiment was to track the abundance of the different substrains of *Y. enterocolitica* in the feces by counting colony forming units (CFU), we had to proceed in this way to be able to run the experiment ethically, including a minimal amount of animals.

For this experiment, a different analysis pipeline (see Figure 5.2) was used than for the preliminary data. Sequencing was done on an Illumina MiSeq and provided 2x 250 bp reads with an overlap of roughly 210 bp, resulting in merged reads of 290 bp. In hindsight, a much smaller overlap would have been sufficient to merge most reads and the resulting longer sequences could potentially have been classified more specifically. The stepwise analysis of these samples was the basis of the 16S analysis pipeline I later developed into the TüBiom pipeline (see 7) and STARA (see 8.1).

Preprocessing I ran Quality Control for raw, trimmed and merged reads using FastQC [3] and the FastX toolkit [32] for additional plots. Trimming was done with prinseq-lite [70], ensuring an average quality of 30 over a window size of 15 bases. Compared to the read length, this was a strict quality trimming which I decided to apply because the MiSeq chemistry at that time did have comparatively poor base quality, which was significantly dropping at the end of the reads. As we had a large overlap, trimming the reads this strictly was not problematic in regard to being able to merge the reads. The trimming step solved the problem of decreasing quality very well, and the reads could be merged using FastQ-join, which is a part of the ea-utils toolset [5].

Alignment The merged reads were aligned to the 16S Microbial Database (downloaded September 2015) using the MALT aligner [33] in semi-global mode. MALT also provided a first taxonomic classification and thus returned an RMA file. The classification parameters are not relevant for the final classification here, as at that time the available MALT version did not support all available parameters for the LCA, and thus I chose to do a more detailed reclassification with MEGAN and MALT only provided alignment information in RMA format.

Classification Taxonomic classification of the aligned reads was computed using the LCA implementation of MEGAN 6. The minimal bit score was set to 50, the maximal e-value was set leniently to 1.0, top percent of the score to keep a hit was 10%, and the minimum percentage of reads in the sample to support a taxon was 0.005. The most critical parameter for this analysis was selecting the 16S Percent Identity filter which would only assign a sequence to a taxon if it had a specific minimal identity to the sequences from this taxon. I chose the percentages following the common sequence identities expected for 16S rDNA on each taxonomic level; from 99% for the same strain over 97% for the species up to 80% percent to be at all assigned as bacteria. With these parameters and preprocessing steps, around 60% of each samples' raw read pairs could be assigned to a microbial taxon.

5.1.1 Results

The numbers of assigned sequences from each of the 42 sequenced samples do vary considerably. If we define 40 000 sequences as a minimum for usable analysis of a sample and 60 000 sequences as the optimum, there are only nine optimally covered samples and six more with sufficient sequencing depth, and of the other 27 samples, nine even have under 20 000 sequences assigned.
We only took fecal samples from PBS treated mice and some mice treated with WT+ A0 at day 14, so they could not be used to compare the different

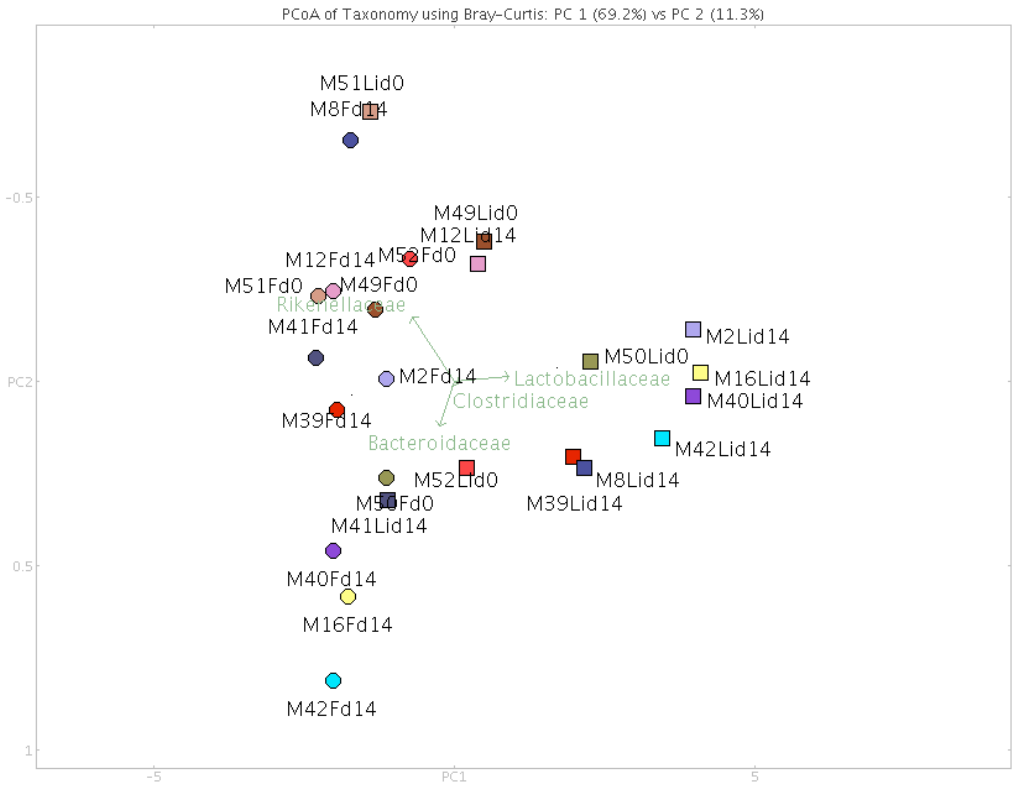

Figure 5.3: PCoA analysis of the fecal and ileal microbiota from 12 mice. The analysis was conducted on a projection of the taxonomic classification onto the family level. Fecal samples are shown as circles, ileal samples as squares. Each color represents one mouse. The Biplot (green arrows) depicts the taxa which separate the samples the most on the given coordinate.

mutants or compare healthy, recovering and succumbed mice. We only used those samples to compare the luminal content of the illeum to fecal samples from the same mouse at the same time point.

Figure 5.3 shows that the ileal samples, both from PBS treated and infected, but successfully recovered mice have an increased abundance of *Lactobacillaceae* and *Clostridiaceae* compared to fecal samples. Fecal samples on the other hand have higher relative abundances of *Bacteroidaceae* and *Rikenellaceae*. The most prevalent taxa from these families in the samples are *Bacteroides* and *Alistipes* for the fecal samples and *Lactobacillus* and the *Clostridiales* genus *Candidatus Arthromitus* in the ileal samples respectively.

Comparing all 29 ileal samples available provides a better resolution to compare the mice suffering from a stronger infection and succumbing to it to the healthy, untreated mice and the ones who can recover from infection with the same type and amount of pathogen. Here the most significant differ-

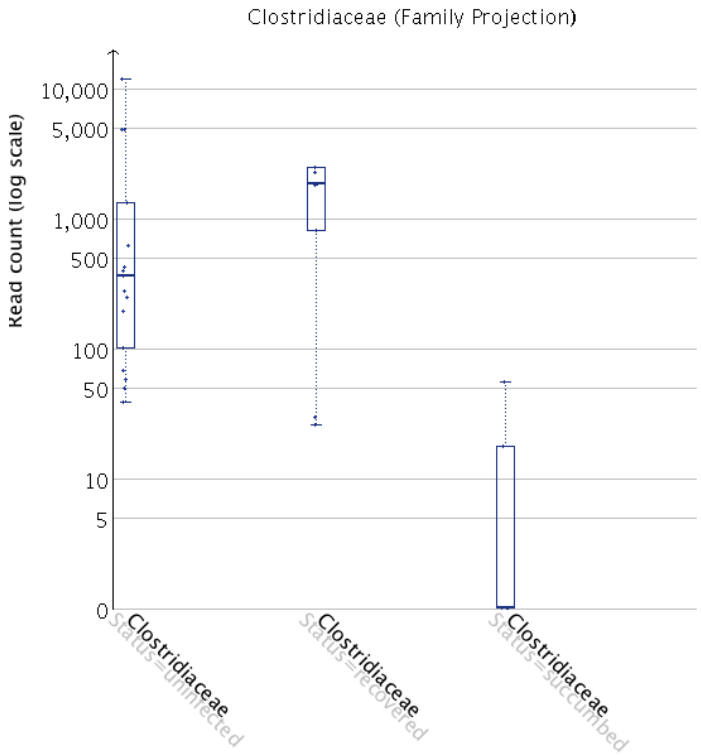

Figure 5.4: Distribution of *Clostridiaceae* in samples grouped by status (uninfected/PBS, succumbed to infection or recovering from infection). Counts are taken from a projection of the taxonomic classification to family level and drawn in logarithmic scale.

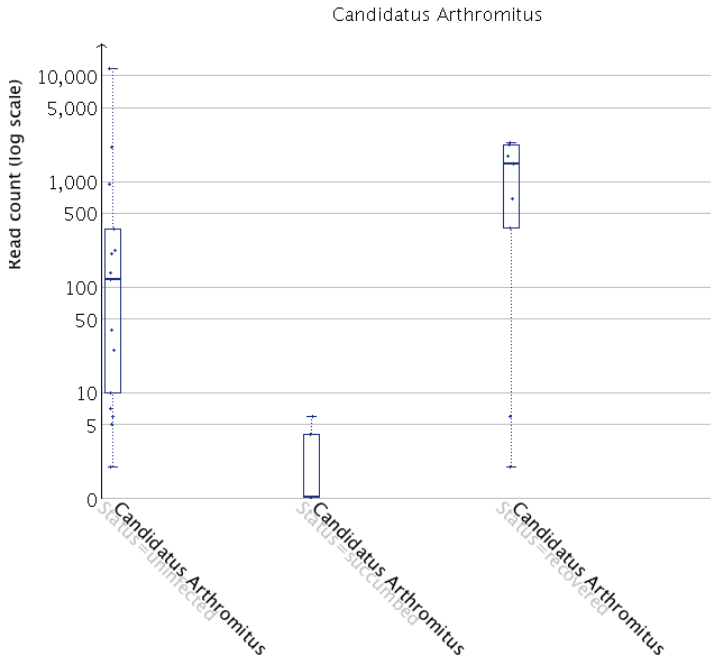

Figure 5.5: Distribution of *Candidatus Arthromitus* in samples grouped by status (uninfected/PBS, succumbed to infection or recovering from infection). Counts are taken from a normalized comparison of all ileal samples and drawn in logarithmic scale.

ence between the groups lies in the family of *Clostridiaceae* (see Figure 5.4). *Clostridiaceae* are significantly reduced in succumbing mice in comparison to both healthy and recovering mice. As previously mentioned, the *Clostridiaceae* are mainly represented by the genus *Candidatus Arthromitus* in our samples.

Candidatus Arthromitus is a member of the group of segmented filamentous bacteria (SFB). SFB from human and mouse gut were assigned to be *Candidatus Arthromitus* based on their morphology similar to the original *Candidatus Arthromitus* found in arthropod guts. More recent studies of their 16S rDNA and genome have led to the conclusion that the mammal gut SFBs are a different genus, belonging to the *Clostridiaceae* and called *Candidatus Savagella* [77]. Unfortunately, the NCBI taxonomy which MEGAN does use to classify and assign names to the taxa still used the name of *Candidatus Arthromitus* for this genus at the time of analysis. We have therefore decided to generalize and call them SFB.

Additionally to the SFB as a potentially positive influence on the chance of recovery, we also determined one possible negative influence. *Akkermansia muciniphila* is a species which has very variable occurrence and abundance in healthy mice in general. Without the inflammation and infection as a background, *Akkermansia* does not seem to be a problem for the mouse's

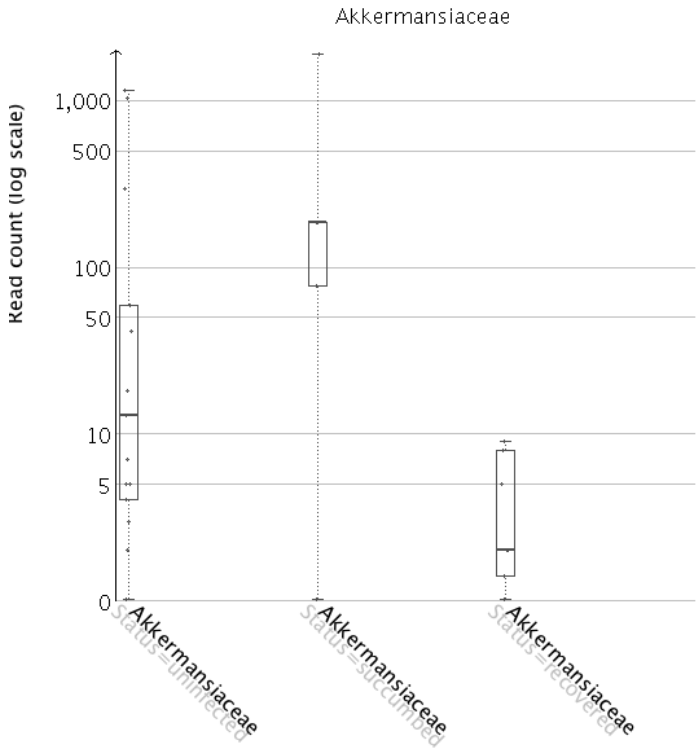

Figure 5.6: Distribution of *Akkermansiaceae* in samples grouped by status (uninfected/PBS, succumbed to infection or recovering from infection). Counts are taken from a projection of the taxonomic classification to family level and drawn in logarithmic scale.

health. However, for the additional infection with *Y. enterocolitica*, *Akkermansia muciniphila*, most prevalent member of the family of *Akkermansiaceae* in the mouse gut, could be a risk factor for a more severe course of infection, as seen in Figure 5.6.

Both SFB and *Akkermansia* have one thing in common: They can be found on or even inside the mucus on the intestinal wall. *Akkermansia muciniphila* degrades this mucus. In humans, it has been found to be beneficial for weight-loss as well as potentially helping to decrease inflammatory immune responses [30]. It is assumed to increase the thickness of the gut wall, as the host has to produce more mucin which is used by the bacterium as a source of nutrition. In our experiments, the presence of *Akkermansia* had an opposite effect, making it more likely for the mice to succumb to the infection. On the other hand, it is unclear if they do succumb to a stronger inflammation of the gut or potentially a generalized infection with *Y. enterocolitica*, which can spread to the spleen and other organs. With the degradation of mucus, it could be easier for *Y. enterocolitica* to travel through the mucus and leave the gut through

the intestinal wall, which then leads to a spread of the infection.
SFB are known to stimulate their hosts' innate immune response and through this can have a protective role for their host [77]. They also form filaments close to the epithelium of the mouse and thus could help to stabilize the mucus and make it harder for pathogens to reach the intestinal wall by creating a mechanical barrier of filaments. Thus it is likely that they keep *Y. enterocolitica* from directly reaching the epithelium, which results in less contact with the immune system and lower immune response as well as lower likelihood to spread to other locations in the host and cause a general infection.
Detecting these two taxa as potentially influencing the course of *Y. enterocolitica* infection in mice is the main result of this experiment. However, we also used it to guide further experiment planning to get more statistically useful results and decisions on sequencing depth for the WGS metagenomic sequencing.

5.2 Metagenomics of mice infected with *Y. enterocolitica*

For this study, we improved the experimental design based on the lessons learned from the evaluation of the preliminary data, competition experiment and other small experiments which did not include sequencing. Samples were taken 1, 3, 5, 7, 10 and 14 days after infection with *Y. enterocolitica* from the ileum of a group of sacrificed mice. We had to increase the size of the groups because we had determined in other experiments that about five to seven days after infection, about 20-25 % of the mice had to be sacrificed prematurely because of fast weight loss. Hence we started with enough mice in each group to still have a statistically significant group size on day 10 and 14. Optimally, each member of a group would be placed in a different cage to decrease the cage effect on the results, but this was deemed impractical in animal handling, and therefore most mice in one group would be in the same cage for the course of the experiment. That makes it mathematically impossible to distinguish a cage effect from an actual effect of the different time points and treatment groups in statistical analysis. Some mice were relocated between day five to seven to keep the numbers of mice for day 10 and 14 as equal as possible. Theoretically, this should not have been necessary with our plans for the loss of mice, but we tried to keep the number of mice in the experiment to a minimum for ethical reasons. Therefore we had chosen our group size to be five, so we could lose up to two mice without completely losing the ability to provide some statistics for our results.
A better course of action would have been to keep at least one or two control mice in each cage, optimally to have mice from different groups sharing a cage. Also, mice should have stayed in the groups they were assigned to at the

beginning of the experiment, while the groups should have been even bigger to have enough statistical power after potentially losing 20% of the groups for day 7, 10 and 14. However, this would have made the experiment unfeasible in the given setting for the number of mice involved and the difficulties in handling and tracking the metadata.
Unfortunately, as an additional problem, the DNA extraction from the ileum of the mice did not yield enough DNA for all of the samples as necessary for sequencing on an Illumina MiSeq. We did have all samples sequenced, even the ones with a DNA concentration deemed too low to produce usable results, but as expected many did not generate enough sequencing depth to be included in the analysis. Surprisingly in the later analysis, the good samples also turned out to include 95-99% mouse DNA which means that they generated very low coverage of the microbial community. These problems with the WGS metagenomics led to the decision to do additional 16S rDNA sequencing of the samples which I will describe in Section 5.3. This section will focus on describing the analysis, results, and problems of the WGS metagenomics from the same samples, comparisons will be shown with the description of the 16S sequencing results.
The samples were sequenced on an Illumina HiSeq, generating paired reads of 2x125 bp. Read pairs per sample ranged from 17 million to 55.5 million. After read trimming, the number of read pairs was reduced to 9.5 to 19 million reads. As we determined a high percentage of mouse sequences in a first investigative analysis, those reads had to be filtered out to be able to study the microbial sequences in more detail. The result of detecting the mouse reads was that the samples included 9 to 18.5 million mouse reads pairs, which makes up for 91-98% of each sample. Because the mouse reads were filtered by comparison to the C57BL/6 ("black 6") mouse reference genome, some mouse sequences would still be found in the remaining data.
The pipeline used for the analysis of these samples was the basis for the Host-Associated Data - module of the later Immigene pipeline described in Section 6.

Preprocessing Preprocessing for general WGS reads is not drastically different than for the 16S rDNA samples, other than adapted parameters and no merging of the reads, as I instead used MEGAN 6 paired-end mode for classification later on. I did the quality control for raw and trimmed reads with FastQC [3] and trimming with prinseq-lite [70]. As the raw reads already had a comparatively high per base quality over the full read length, filtering for a minimal length of 75 bp was the most restrictive preprocessing step, which also removed a lot of "reads" without any sequence which were included in the raw data as a result of adapter trimming done by the sequencing provider. As many of the samples had spikes of short read lengths in their read length distribution, we decided to radically filter them out to only retain reads with enough information content in both of the reads for use in paired-read MEGAN

classification.

Host filtering To reduce the number of host reads in the samples and get a more specific view of the microbial community, the trimmed and length filtered reads where aligned to the C57BL/6 mouse reference genome (version X) using MALT [33] in semi-global alignment mode and requiring at least 75% sequence identity and a maximum e-value of 0.01 to accept an alignment. The e-value might look large on first glance, but we are aligning against a comparatively small database, which does influence the computation of the e-value. Together, these parameters do filter out any reads which are a significant match to the mouse genome.The unaligned reads were then selected to be kept for further analysis.

Alignment The remaining read pairs were aligned using DIAMOND [15] in BlastX mode against the NCBI NR database (downloaded April 2016). DIAMOND was set to return up to 25 alignments per query. Further filtering was done during the classification step through the LCA parameters.

Classification Aligned read pairs were classified using MEGAN, providing taxonomic classification as well as the functional classification for Interpro2GO and COG/EGGNOG. They were classified using the naive LCA, with a minimal percent identity threshold of 30%, maximum e-value of 0.01 for accepted hits, 0.0001% of sequences in the sample assigned to a taxon to accept this taxon (MinSupport) and only hits with 5% of the top bitscores for all hits on a query sequence accepted as significant hits. Many of the samples still had a significant abundance of sequences matching Eukaryota, specifically mouse (*Mus musculus*) and usually a very low number of sequences assigned to Archaea. I decided to extract the sequences assigned to the domain Bacteria into separate RMA6 documents and re-classify the sequences contained in those files using the weighted LCA algorithm.
The weighted LCA will run the LCA two times, which can be time-consuming for large datasets. However, with the few bacterial sequences retained in our samples after the preprocessing and filtering steps, it was very feasible to run the weighted LCA. At first, MEGAN will try to find sequences which have unique hits to one taxon and thus can be directly assigned to that specific taxon. After the first LCA pass, all taxa will have the number of those uniquely classifiable sequences assigned as a weight. As those taxa are for sure found in the sample, during the second pass of LCA, the weight for a taxon will be factored in every match a sequence has to that taxon; thus more additional sequences will potentially be assigned to taxa with high weights from the first pass. This process can be beneficial especially for a dataset where we might not have enough sequencing depth to cover the full diversity, but would want

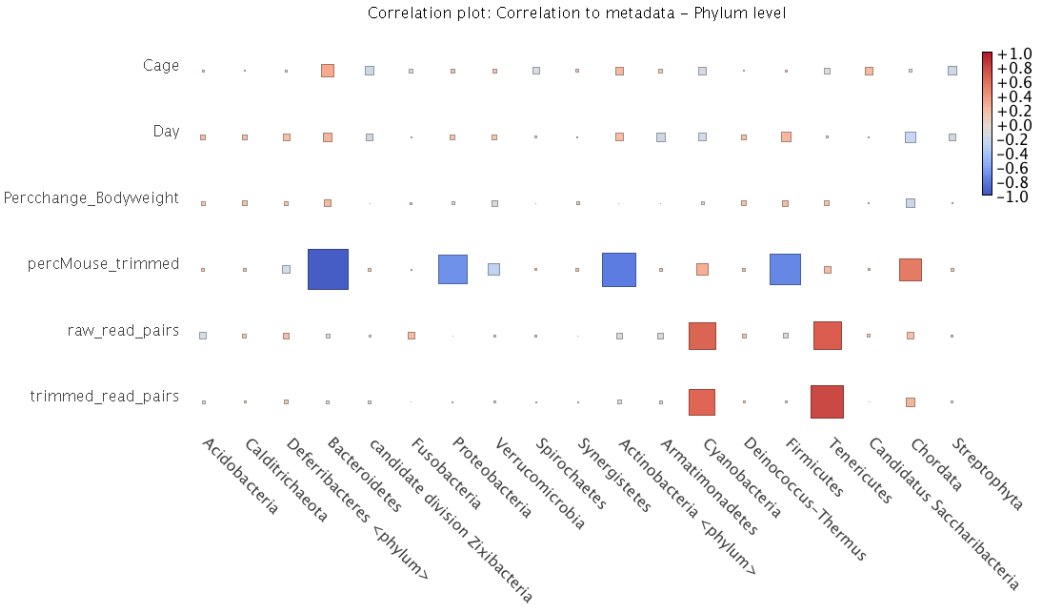

Figure 5.7: Correlation of Phylum level assignments to some of the available biological and technical metadata. These phyla were classified only from sequences assigned to the domain Bacteria from the original host-filtered reads, but re-classified, thus still contain some Chordata (mouse).

to find at least the taxa which are abundant in the sample with high sensitivity.

5.2.1 Results

The results presented are from the files including only the sequences assigned to Bacteria in the original metagenomic analysis results, to reduce the noise generated by the high numbers of mouse sequence in all samples.
Correlation of the taxonomic assignments on phylum level (see Figure 5.7) shows that the strongest negative correlations are to the percentage of mouse reads that were found in the sequences after the trimming step of the analysis. This means that the content of the host reads in the raw data still influences the results of bacterial assignment after two steps of filtering out the host sequences. I only found a strong positive correlation for *Tenericutes* and *Cyanobacteria*, which are correlated positively to the overall number of raw reads in a sample as well as the number of reads left after trimming, but before filtering out the host sequences.
There is no significant correlation of taxa on phylum level to the mouse cage, and time passed after inoculation or changes in body weight of the mice.

As the sequencing and the general number of sequences left after filtering out the host reads was not sufficient for further analysis, it was decided to re-sequence the samples using 16S rDNA sequencing. More results from this experiment will be presented in Section 5.3 in direct comparison to the results from the 16S sequencing analysis of the samples.

5.3 16S of the ileal samples

This analysis was done on the samples from the metagenomics experiment, but only 48 samples amplified enough bacterial DNA to be sequenced. Two of these samples also ended up having 0 reads assigned after preprocessing. 12 more samples had less than 20 000 reads assigned and were excluded from most analyses. Only 11 samples had more than 40 000 assigned reads, which would have been an acceptable minimal sample size for comparison. The low read counts available for analysis of these samples lead to the idea of adapting the STARA pipeline (see Section 8.1), so it would stop analysis of a sample in case there are too few sequences left to get usable results.
Our first idea, that mice with high flaring infection and resulting inflammation would include more mouse DNA in the sample and thus less bacterial DNA to be amplified was proven wrong by the fact that all but 3 of the samples under 20 000 assigned reads were either healthy control mice or mice with no weight loss before we took the sample. Only two of those samples were from mice showing significant weight loss. Analysis of the samples was done with an early version of the STARA pipeline, without the selective breakpoints for low count samples. Hence, all samples were analyzed using the same parameters, and the 14 samples with less than 20 000 reads had to be excluded from most further analysis. Comparative analysis using MEGAN does suffer in the presence of samples with extremely low assigned sequence counts, as the other samples will be heavily subsampled for normalized comparison.
Rarefaction analysis (see Figure 5.8) of all 48 sequenced samples shows that the taxonomic content of many samples could not be adequately covered. Samples under 10 000 assigned reads most likely are not depicting the diversity found in the probe.

Unfortunately, the results from metagenomic sequencing and 16S sequencing of the samples do not compare well at all. Figure A.1 and Figure A.2 show the relative abundance of the top 10 taxa on the family level from the metagenomic sequencing and 16S rDNA sequencing respectively. Of course, these assignments are based on different types of sequence (protein coding sequences and 16S genes) and have been assigned using different databases (NR and 16S Microbial). However, on the family level, it could be expected to at least find mostly the same taxa in the top 10, but those two analyses only share four of those. The assignments in MEGAN are based on the same

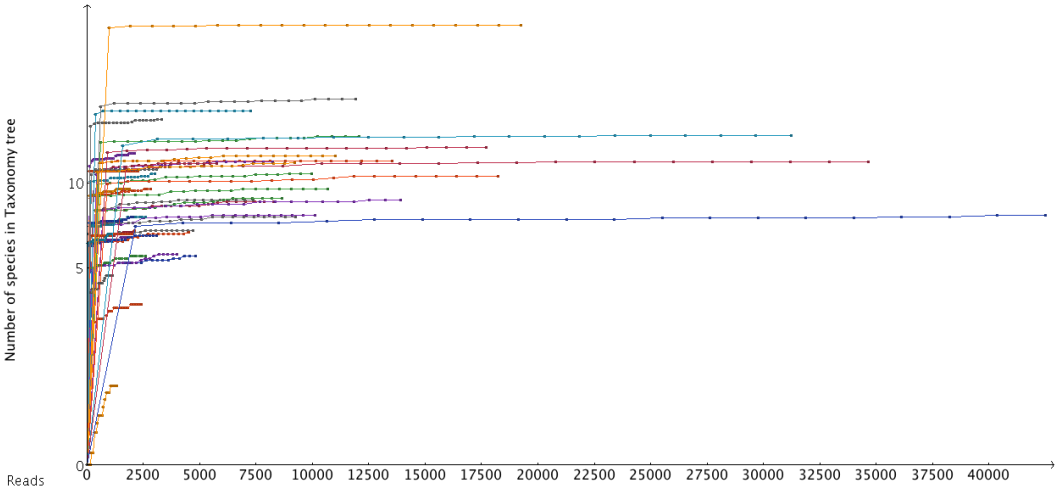

Figure 5.8: Rarefaction analysis of the 16S sequencing of the ileal samples from mice infected with *Y. enterocolitica*. Reads where sampled from Species assignments, hence the number of leaves in the given taxonomy tree represents the number of detected species. Number of leaves is shown on a logarithmic scale.

(NCBI) taxonomy, so differences in the taxonomy do not account for the difference in these analyses.

In general, there are relatively low counts of *Yersinia enterocolitica* or even the family of *Enterobacteriaceae*. These low abundances made it impossible to compare the pathogenic load to the outcome of the infection or other measured parameters. The obvious differences between these analyses led to the realization that it will not be informative to try any further analysis of the few samples which could be sequenced with sufficient depth. Also, these samples did not represent many of the different days after infection with enough replicates to do any multivariant or correlation analysis. Metadata for colony forming units (CFUs) of *Yersinia enterocolitica* did not correspond with the measured abundance of the taxon in the sequencing data.
Even after intensive testing and planning of the experiment over a timeframe of nearly two years, the resulting dataset could not provide the insights on colonization resistance against *Y. enterocolitica* we had hoped to be able to achieve. Complications - like the difficulties encountered during DNA extraction from the ileum, the fact that there cannot be a timeline for a single mouse because of the need to extract the samples after killing a mouse and the coprophagic behavior of mice leading to cage effects and the problematic infection which resulted in having to put animals out of their suffering before

the time point they were selected to represent - brought the metagenomics approach to a halt.
To show a direct comparison of samples with sufficient sequencing depth, five samples could be found where the metagenomic sequencing resulted in more than 200 000 sequences assigned to Bacteria, and the 16S rDNA sequencing had more than 40 000 assigned sequences, respectively. These samples are compared on the Class level in Figure 5.9. It shows that WGS metagenomic sequencing has more assignments for *Bacteroidia* and *Clostridia* in general, while 16S sequencing has more assignments for *Bacilli*, *Deltaproteobacteria* and *Erysipelotrichia*. The samples cannot represent much about the difference of samples during the infection with *Y. enterocolitica*, as mouse 62 and 66 are from the control groups and mouse 23, 43 and 44 from the same group of infected mice that have survived until day 14 and did recover from the infection (weight gain or only very little weight loss compared to before the inoculation). Comparison on Family level (see Figure 5.10) further depicts the differences between the samples from the same probes.

The overrepresentation of *Bacilli* is mainly made up of the genus of *Lactobacillus*, which has also been shown to be overrepresented in 16S data compared to WGS metagenomic sequencing by Jovel *et al.* [40]. However, they also found a consistent overrepresentation of *Clostridium* in their 16S samples, which were analyzed using different OTU-clustering methods. This result does not match our 16S samples which have an underrepresentation of the class *Clostridia* compared to the respective metagenomic samples. The assignment of *Clostridiales* and *Enterobacteriales* can be difficult using the V3 and V4 regions of the 16S rRNA, but as Jovel *et al.* used the V4 region and we used a primer for V3-V4, the representation of sequences from these taxa should be comparable between the different datasets.

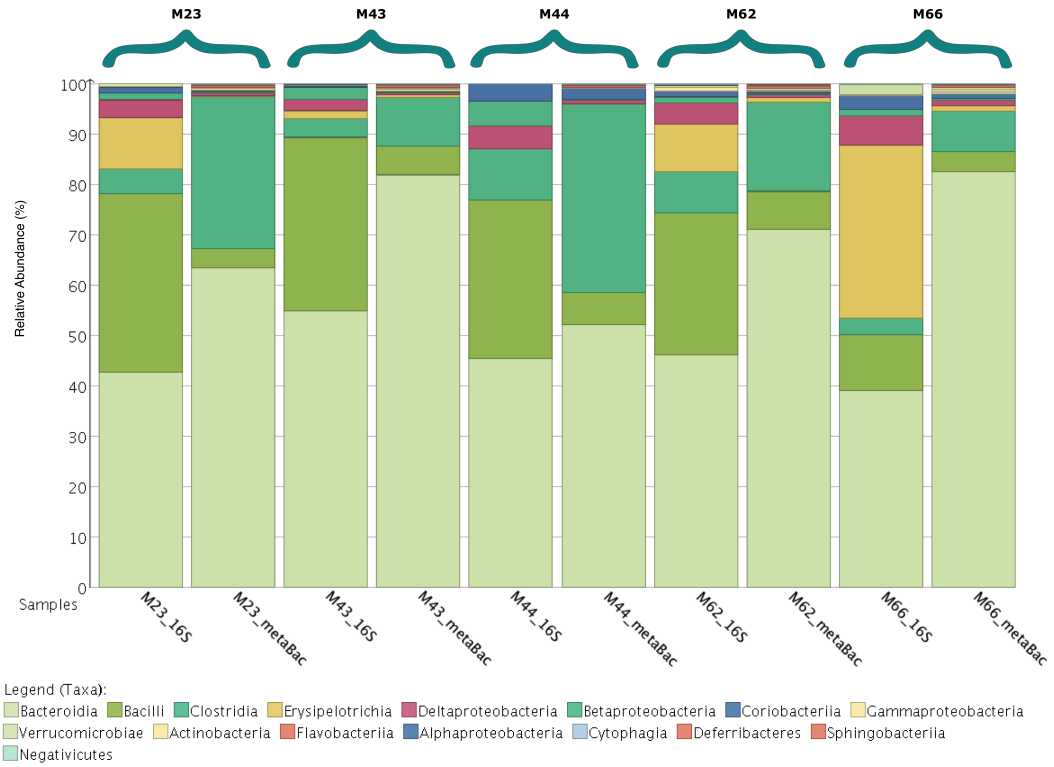

Figure 5.9: Relative abundances of taxa on class level for Mice 23, 43, 44, 62 and 66, both from WGS metagenomic and 16S sequencing based on a normalized comparison of all 10 samples.

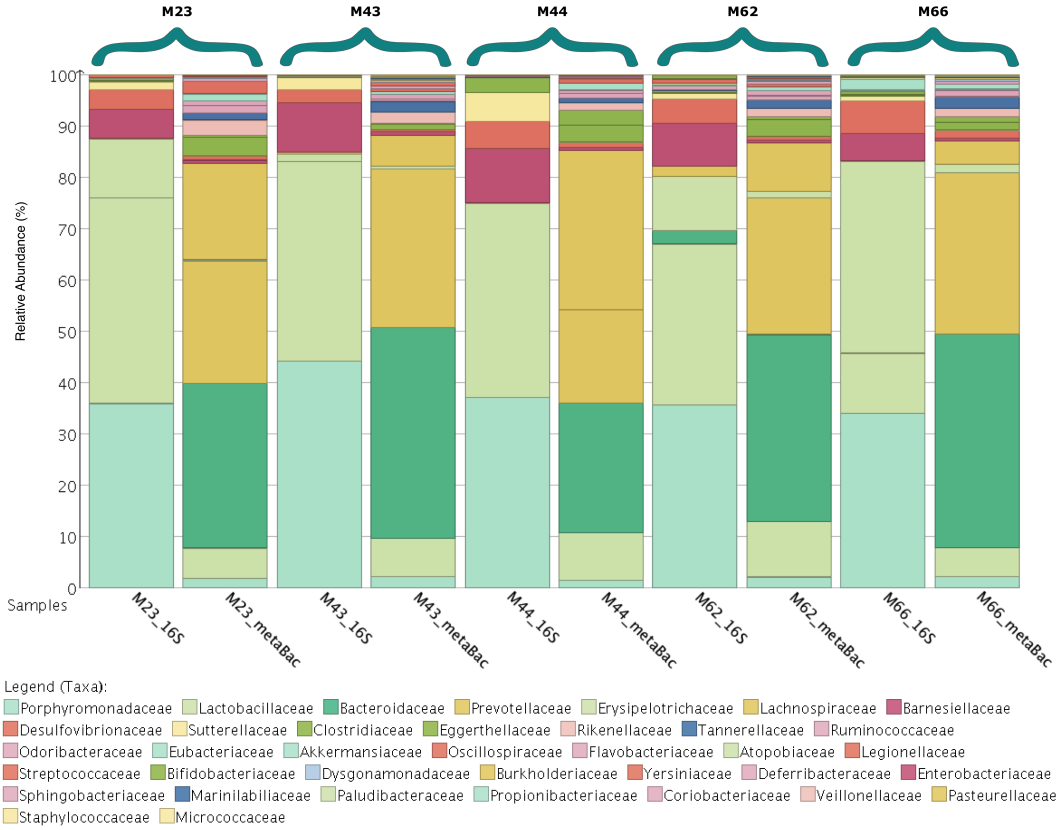

Figure 5.10: Relative abundances of taxa on family level for Mice 23, 43, 44, 62 and 66, both from WGS metagenomic and 16S sequencing based on a normalized comparison of all 10 samples.

6

Metagenomic analysis for the ImMiGeNe project

The ImMiGeNe Project in Tübingen is a collaborative project between the University and University hospital of Tübingen (UKT) together with the Tübingen Center for Personalized Medicine (ZPMT) and the Center for Quantitative Biology (QBiC). The goal of this interdisciplinary effort was to implement a reliable, well-designed pipeline to streamline high-throughput sampling, sequencing, analysis and integrative interpretation of clinical data collected from patients, integrating gut metagenome data, host immunogenic characteristics and clinical gut inflammatory biomarkers. This effort should help to decipher the complex interplay between the patient's immune system, gut microbiota and the influence of both on their disease and treatment.
The pipeline will first be used to study a cohort of patients undergoing hematopoietic stem cell transplantation (HSCT) as a treatment of acute myeloid leukemia (AML). The aim is to study for example the similarity of donor and patient gut microbiota before and after the transplant and how the microbiome influences the success of treatment, especially the occurrence rate of graft-versus-host disease.

6.1 The ImMiGeNe Metagenomics Pipeline

The analysis pipeline for the ImMiGeNe project was planned to be highly automated. This automation ensures both consistency and reproducibility as well as making it user-friendly for all types of future users. The pipeline is available as a docker container, together with other containers needed to generate the databases used for alignment. It will read in files from a selected input directory, automatically pair the files and run preprocessing and analysis on each sample based on the parameters given through a configuration file.

The result of this analysis is an RMA file for each sample which can then be either inspected using MEGAN or fed into further analysis pipelines depending on the available metadata and research interest.
The ImMiGeNe Analysis Pipeline is the basis of the general metagenomic analysis pipeline module without host filtering that is described in Section 8.3. It is also a further development of the metagenomic analysis pipeline described in Section 5.2, without the filtering for host reads and the re-classification using the weighted LCA, but with adapted parameters and extended logging of the progress.

6.2 Metagenomic sequencing of healthy human gut samples

To prepare and test the workflow of sample preparation, metadata collection, sequencing, and analysis, 51 samples of healthy volunteers as well as 16 blank controls from lysis buffer were sequenced using 2x 150 bp Illumina HiSeq sequencing. For these samples, metadata on the volunteer's health and general data was collected through a questionnaire, and immunological markers for specific genotypes of the TLR5, TLR2, NLRP6, and MB21D1 genes were determined.

Correlations of these genotypes to bacterial taxa and functional roles of proteins found in the samples show some known connections of host immunology to their gut microbiome. Figure 6.1 for example shows correlations of the immunological markers that we determined from the volunteer's blood samples to functions associated with Human Disease.
Reads assigned to human (*Homo sapiens*) are not found frequently in the data, even without filtering for host reads. The domain of Eucaryota accounts for less than 0.2% of assigned sequences in all of the samples.
So far, analysis of the ImMiGeNe samples cannot lead to insights specific for this project, as they are test runs on healthy human gut samples which are expected to be very different to samples from patients with AML. Further analysis in this project will be based on the metagenomic analysis pipeline presented here, but also include additional analysis steps which still have to be determined based on the diversity and sequencing depth which can be reached for the patient samples. To utilize the available samples, I used one healthy human gut sample from this project as an example for the general WGS metagenomic analysis pipeline MAPle in Section 8.3.

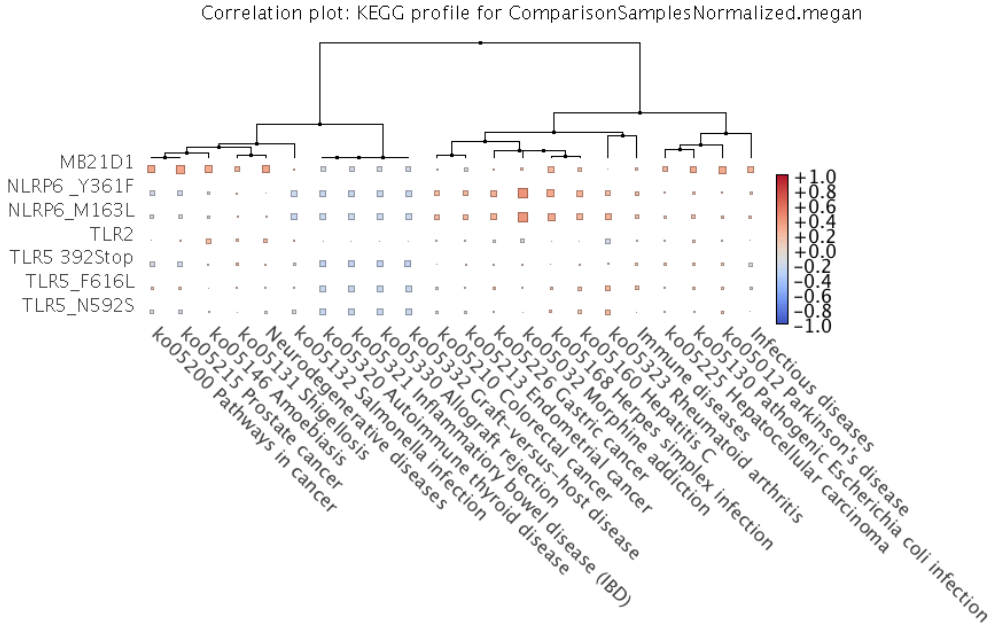

Figure 6.1: Correlations of immunological markers with selected Human Disease related functions as annotated by KEGG

7

Automated 16S rDNA analysis pipeline for the TüBiom project

This section is based on contributions to the TüBiom project, as published in the following Preprint:

S. Beier, A. Górska, P. Grupp, T. A. Harbig, I. Flade, and D. H. Huson. Bioinformatics support for the Tuebiom community gut microbiome project. *PeerJ Preprints*, pages 1–9, 2016

The TüBiom Project is a shared project between the CeMeT GmbH, the Department of Hygiene at the University Hospital of Tübingen and the Department of Algorithms in Bioinformatics at the Eberhard-Karls University Tübingen. The role of the Bioinformatics group in the project was to develop the database system and user interface as well as to develop the standardized analysis pipeline for the project. I provided an automated analysis pipeline with a gold standard configuration file which was adapted to fit the needs of the project we described in this publication.

7.1 The TüBiom Project

The TüBiom Project was designed as a community science project to give the public an opportunity to get insights into their gut microbiome for free while at the same time building a valuable reference database from a mixed population to study the variations in the microbial communities in the gut of healthy persons and also people with various disorders or diseases.
Participants could order a kit which provides them with all tools and explanations to sample their feces and is then sent back for sequencing. Sequencing was

done through a specific tested and standardized DNA extraction and library preparation protocol on a MiSeq, providing 2x 250 bp paired-end reads. The wet lab protocol used was optimized to both keep the cost feasible and generate good quality, reproducible and comparable results for every sequencing run. This way it was to make sure that the samples collected for the project will always be comparable to each other and can be fed into the database without additional normalization steps that would cause unnecessary bias.
We analyzed the samples through a fully automated pipeline which had been optimized for the provided input and required output. From the full taxonomic assignment, several different profiles for five selected taxonomic ranks (phylum, class, order, family, and genus) were computed. These taxonomic profiles were then fed into a database which constitutes the core of the project. From the information in the database, comparisons of single samples with averages of different groups of samples (for example all vegetarians, all females or everyone who took antibiotics in the last four weeks) were calculated and provided the basis for an interactive graphic report. These reports were provided to the participant on a website, who could compare their samples with anonymized averages of other groups of samples or, if they had provided multiple samples, with each other. The database was also designed to be the basis for further research on this population, providing insights into the microbiota of health and disease and the variability of them in different subgroups of the population.
The analysis was based on the sequencing of 16S rDNA, which gives information about the bacterial species found in a sample and their relative abundance. The project was not designed to study the functional content of the participants' samples, as this would not have been feasible to provide as a free of charge study on this scale and it would also complicate providing and roughly explaining the collected information to participants. The project provides participants in general with a simplified view of their microbiome, not a complete in-depth analysis. It also did not provide any diagnostic or medically predictive information to participants.

7.2 The TüBiom Analysis Pipeline

The TüBiom project needed an analysis pipeline which would be able to produce comparable results for all samples from many different sequencing runs done over a comparatively long stretch of time. This means, while the protocols and technology are fixed, there would be differences in many factors which have been shown to influence sequencing results (batches of chemistry, updates of the sequencer or even usage of different MiSeq machines, changes in the personnel processing the sample and more).
The analysis pipeline included automated preprocessing, based on the expected quality range of one of the standardized MiSeq runs, alignment of the prepro-

cessed reads against the 16 Microbial database of NCBI and processing the alignments into a classification of each read to a taxon using the lowest common ancestor (LCA) algorithm provided by MEGAN. Alignment and classification were guided by parameters which were selected based on the expected input and the needs of 16s rDNA analysis.
Read pairs were read in from the provided input directory and matched according to their file names. Sample identifiers in the filename are kept as identifiers throughout the pipeline, so any intermediate files and results always uniquely match with the raw input.

Preprocessing Raw reads underwent quality control using FastQC to keep track of the improvement of the data through preprocessing and to be able to determine the overall loss of data from raw input to assigned reads. The raw reads were then quality trimmed using fixed parameters according to the TüBiom standard with prinseq-lite [70]. Only pairs where both reads pass the filter after trimming were used for further analysis. Trimmed reads go through another run of quality control before being merged. We did TüBiom sequencing with a good overlap, so even trimmed reads usually could be merged without problems. If reads could not be merged or the merged sequence was under a length threshold of 75, we discarded them for lack of information content. Too short reads often don't get assigned with significant specificity or if they are, attract erroneous assignments. The filtered merged reads went through another quality control step before alignment.

Alignment Merged reads were aligned against the NCBI 16S Microbial Database. We made this decision for licensing and consistency reasons. The database used for the TüBiom Project was the version from September 2016. Other databases available had not been up to date at that time or were not available to license for the project.
Alignment was done using MALT [33] in semi-global mode. Semi-global alignment is appropriate for merged 16S rDNA sequencing reads: As we align to different 16S references which can be full length or start and end at any position of the 16S gene, but we align reads that only cover the V3-V4 variable regions, we need to allow gaps at the start or end of the alignment, to be able to place the read on the reference. Still, we do not want large gaps in the middle of the alignment, which would lead to many false positive alignments and alignment scores which do not adequately represent the similarity of the query sequence to the reference. As we used this similarity to place the merged reads taxonomically, alignment has to be as strict as possible. We also expected the full query sequence (merged read) to align to the reference. Otherwise, it would probably not a relevant match for our analysis. Hence, local alignments are also ruled out.

Classification MALT also provides a taxonomic classification for each input sequence that it computes using the LCA algorithm as it is implemented in MEGAN. Assignments were saved in the RMA6 file format for MEGAN so that they can be fed into the profiling scripts and entered into the TüBiom database. This classification is adapted to the needs of the TüBiom project. The pipeline could be run on other 16S samples sequenced with a comparable protocol but is not as widely usable as it would need to be as a general pipeline.

8

CommunAl - A toolkit for analysis of environmental data

The projects previously described did lead to three pipelines for data analysis which were continually developing during each project. The final pipelines I will present in this chapter are generalized versions I developed based on the different flows of analysis which were necessary for each of the projects and some additional features. These tools are now fully automated, but easily configurable pipelines for the primary analysis of environmental data. Of course, for most use cases, further analysis steps will be necessary, but those do differ from case to case and will usually have to be adapted and developed specifically based on the available input, metadata and questions that have to be answered using the data.
These three tools are designed to ease the repetitive chore of general data analysis which always has to be done for this type of data. If sequencing is consistently done with the same setup and depth for different samples and experiments, the same configuration files will often be able to be used repeatedly, or might only need minor adaptations. This provides hands-off reproducible data analysis and allows the people involved in the experiment to focus on the more specific parts of the analysis.
All of the presented pipelines share their name with a city, town or community and do provide microbial community analysis. They are called the CommunAl toolkit (Community AnaLysis).

8.1 STARA - A generalized 16S analysis pipeline

The original TüBiom pipeline provided an analysis that was tailored to the needs of the project and was run with fixed parameters every time and on spe-

cific types of input data. The future of the project is uncertain, and further developments of the original pipeline might not be made. To make this type of analysis usable for the general case of short read 16S rDNA sequencing, automatically providing a basic analysis and report on important statistics, I developed the STARA pipeline.
STARA stands for 16S based Taxonomic Analysis of Ribosomal Gene Abundance.
STARA can analyze data from single or paired-end reads in an automated way, following basic parameter settings that are provided by the user in a configuration file. It will run the analysis on all FastQ files found in the input directory. The user can set breakpoints for all samples, where sample analysis will be stopped in case certain thresholds number of sequences are not met, or the loss of input from the last processing step is unusually high. Hitting one of these breakpoints will return information for the user in the log file and continue with the next sample instead of spending time on analysis of a sample which will most likely not generate sufficient results. Those samples could potentially be salvaged later by using different parameter settings, or be excluded completely from the analysis.
The pipeline was designed for the use of standard Illumina short read sequencing data, but can also be run on long read samples without any changes to the pipeline, just by adapting the parameters adequately. However, it has so far not been tested on long reads, for lack of a suitable dataset. A flowchart describing the pipeline is shown in Figure 8.1.

STARA follows the workflow of the previously described TüBiom Analysis Pipeline from Chapter 7 but offers more options for input and analysis. It can accept single and paired-end reads. Already assembled or long reads can be used in the single mode. STARA also analyses the Quality Control (QC) output after every step of QC and reports the results, as well as stops analysis of the current sample if it does not match given thresholds.
A different approach for the pipeline would be to exchange the quality trimming and merging step. Paired reads from 16S rDNA usually start with a conserved region at the beginning of the forward read and sometimes also include part of a conserved region at the end of the fragment - which is the beginning of the reverse read - depending on fragment length and length of the variable region which has been sequenced. In this case, both ends of the merged read hold comparatively little information content for the analysis. It is important to merge as many of the reads as possible and do so correctly to achieve a higher information content for each sequence. If trimming is done before merging, low quality and thus potentially erroneous bases from the end of the reads - which is also the overlap region for merging - will be removed. If the remaining reads still have enough overlap to be merged this makes for a high-quality sequence in further analysis.
Merging before trimming, in this case, has the advantage of producing more

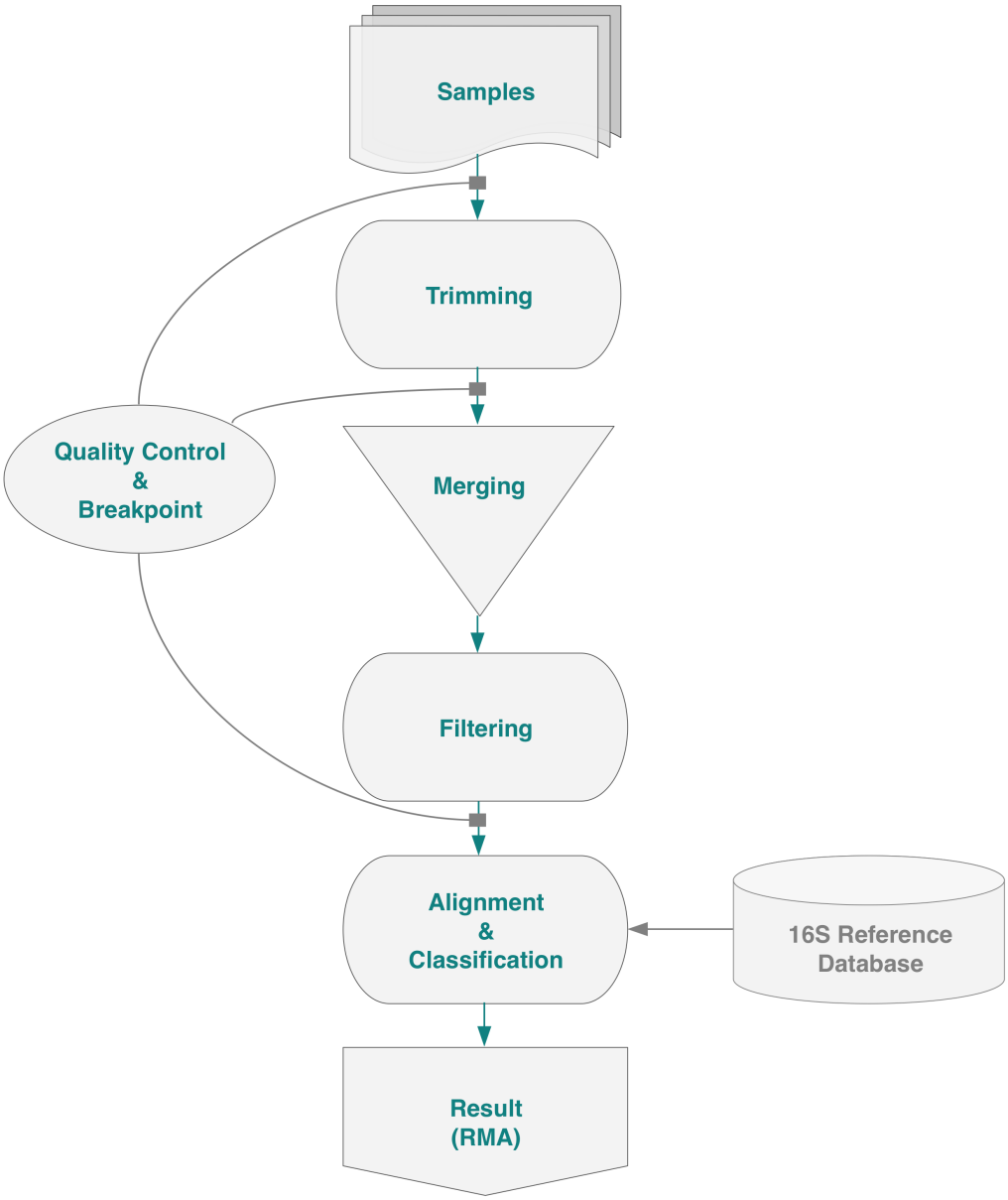

Figure 8.1: Flowchart of the STARA analysis pipeline. Merging can be skipped for single end sequencing or pre-merged input data, in that case, trimmed reads will directly go into the filtering step which removes sequences under a threshold for minimal length. Parameters for each step and the path to the database files are provided in a configuration file and can be adapted for different types and quality of the raw data.

merged sequences, but they can have low quality in the overlap region. If these merged sequences are strictly quality trimmed, they would be reduced to only the first part of the sequence before the quality dropped in the middle and are hence more likely to be filtered out based on the threshold for minimum sequence length. This would also happen if trimming was done first, as not enough of the overlap region would remain for merging. On the other hand, if the untrimmed reads led to a wrong sequence in the merge region, this might not be detected and removed after merging. This can lead to erroneous assignments (false positives) instead of missing assignments (false negatives) in the further analysis. It is desirable to keep the false positive rate as low as possible, hence STARA does quality trimming before merging.
To show an example of the analytic power of STARA on a short read dataset derived from real data and using current sequencing technology, I chose to analyze raw data from a recent publication by De Bruyn *et al.* [22]. The raw read data for this publication I retrieved from the European Nucleotide Archive (ENA) under the Project Accession PRJEB21337. I collected the metadata from the Supplemental File of the publication.
I ran the STARA pipeline with lenient parameters for 2x250 bp MiSeq sequencing data, including quality trimming in a window of 15 bases for the average quality of 30 and filtering for a minimal merged sequence length of only 75. From the results, I attempted to reproduce two graphs given in the publication. The first one are the Phylum level assignments for the Phyla found by De Bruyn *et al.*. STARA detected low counts of further phyla, but those account for less than 1% of the overall community in any of the samples, so where discounted for this plot. The result is shown in Figure 8.2 which does match Figure 4b from the original publication.

Similarly, Figure 8.3 is a PCoA Plot based on all taxonomic assignments made using STARA, which matches Figure 2b from the original publication nearly precisely.

The analysis from the original paper was done quite differently to what STARA provides. Preprocessing steps are comparable, but the analysis was done using OTU-clustering by UCLUST, and the taxonomy was assigned to these OTUs using the RDP database. Results from OTU-based and alignment-based 16S analysis are often said to provide different results. However, also results from single read analysis and merged sequences of the same dataset could provide different results.
To test the alignment-based approach and show the possibilities of running STARA in single-sequence mode, I repeated the analysis twice, with adapted parameters. The first version is again using the paired read information, but this time the length threshold is set up to 170 bp, as this is reasonably smaller than the maximal fragment length of 220 bp for this dataset but bigger than the average trimmed read length, which leads to mostly merged sequences being used for taxonomic analysis. I also adapted the minimum

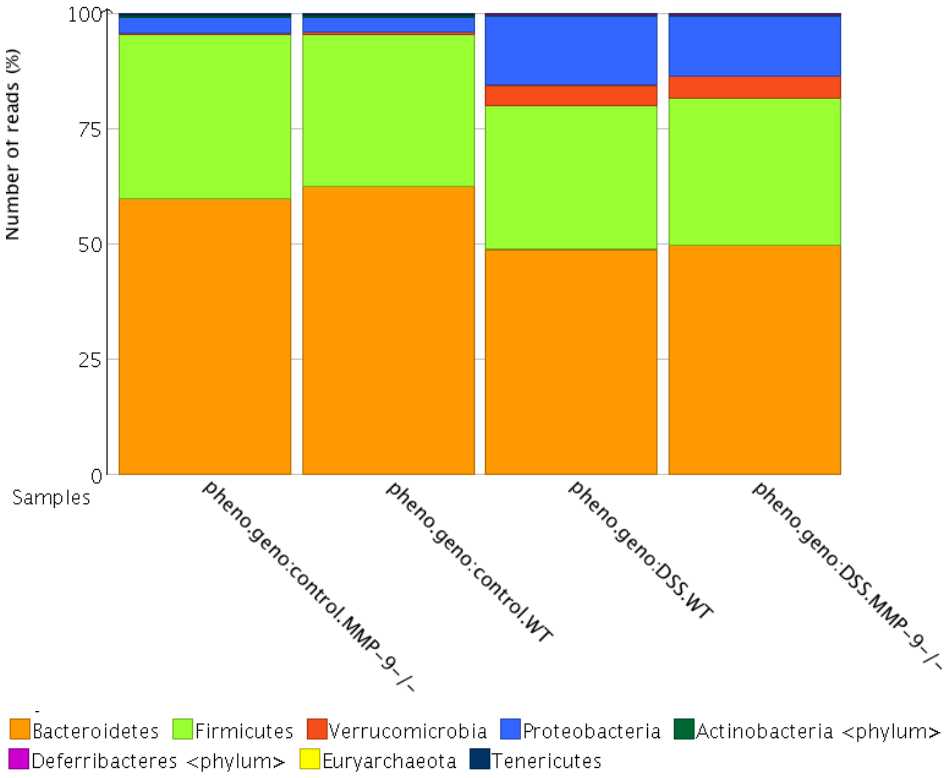

Figure 8.2: Relative abundances of the fecal microbiota from mice treated with DSS and control mice.

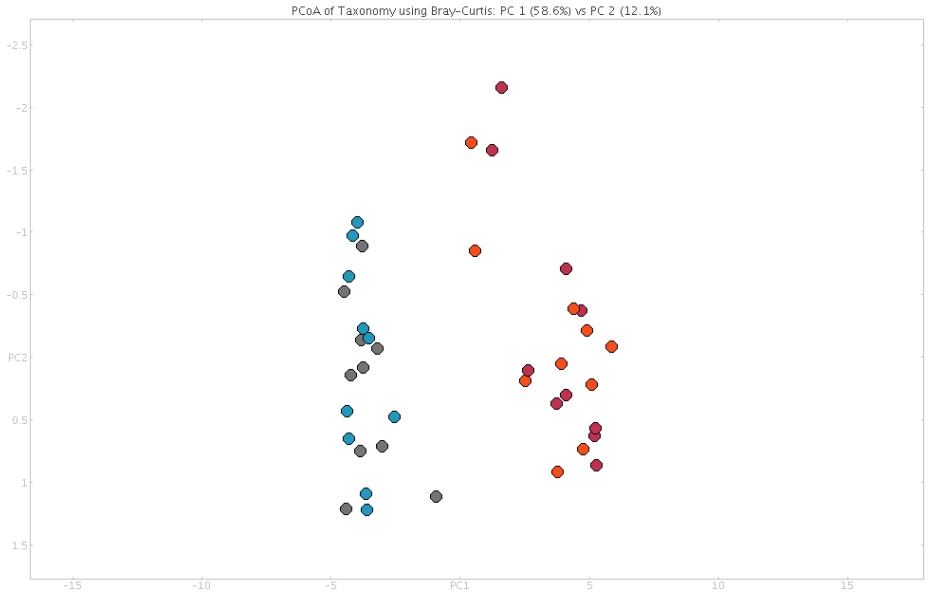

Figure 8.3: PCoA Plot based on Bray-Curtis dissimilarity for the data analyzed by STARA from [22]. Dark red circles represent wild type mice treated with DSS, orange represents MMP-9 -/- mice treated with DSS, wildtype controls are represented by grey circles and MMP-9 -/- controls by blue circles, respectively.

support percentage for MALT to 0.1, which means that 0.1 percent of all merged sequences have to be assigned to a taxon for it to occur in the final result. Reads assigned to a taxon with support under this threshold will be pushed up in the taxonomic hierarchy until they are placed on a node with sufficient support. The second version of the STARA analysis was treating the paired reads as single reads, generating two analysis for each sample, one from the forward and one from the reverse read. Using the same pipeline in both modes makes it possible to compare the results of only analyzing the forward or reverse read and the merged sequence. For the single-end run, I adapted the minimum length to 100 bp, all other parameters for trimming, filtering and assignment were kept the same to keep the results from the two runs as comparable as possible.

8.2 TaxCo - Correlation analysis for taxonomic data

A common question in environmental datasets is if and how abundances of a taxon change over time. Often it is studied how they change in relation to selected metadata. However, the microbes we measure are also part of a connected community. Changes of one taxon might directly or indirectly influence other taxa by changing the environment for example through the availability of nutrients, the host immune reaction or by directly suppressing the growth of other microbes. This interaction is tough to measure, as it theoretically needs a complete picture of the environment. One way to determine potential positive or negative interactions of microbes is to determine correlations. A common choice are co-occurrence measures, which decide correlation based on the occurrence of taxa in the same samples or the lack thereof. TaxCo (Taxonomy Correlations) instead computes three different correlation measures based on the relative or absolute abundance of all taxa for multiple samples.
It was initially developed to determine correlations between the changes in abundances over time by sorting the samples temporally. However, it can also be used for correlations of independent samples as long as they are sorted in a fixed order. The correlation is then not measured over the changes in abundance over time, but over the differences between samples.
TaxCo provides three correlation measures which are appropriate for the given data type: Spearman's rank correlation (Spearman's ρ), the Pearson correlation coefficient (Pearsons' r) and the Kendall rank correlation coefficient (Kendall's τ).
While Spearman's ρ is possibly the most commonly used of these correlation coefficients, Pearsons' r and Kendalls' τ are more appropriate for metagenomic data because we should not expect the relationships between correlating taxa

to be strictly monotonic, which is what Spearman's ρ does. Pearson's r is more general in that it does not expect monotonic relationships and Kendall's τ is especially robust to outliers, which are quite common in this type of data. For example, an outlier from a cage of mice would typically be the mouse lowest in the hierarchy and under the most stress, which can strongly affect their microbiota.
The TaxCo Pipeline is pictured in Figure 8.4. It can read in tabular output formats as they are provided by QIIME, mothur or MEGAN 6, and it does produce tables in tab-delimited text format, graphs in PDF format and graphs in GraphML format which can be used by many graph visualization tools, for example, Cytoscape.

Converting Input TaxCo can utilize input from different common sources. At the time of development, QIIME [16] and mothur [69] were the most commonly used 16S rDNA analysis pipelines. As TaxCo was at that time developed for the further analysis of the preliminary data from the *Yersinia enterocolitica*-project (see Chapter 5), it was designed to use the output from these tools as well as output which can be extracted from MEGAN. It could be adapted in the future to include other file formats for input, for example from QIIME 2.
TaxCo converts the different input formats into its own format *'taxcoIn'*, a basic tab-separated representation of the abundances on each taxonomic level.

Computing Correlations Ranked correlations between all pairs of taxa found in the dataset are computed using the scistats module from scipy [39]. The correlation coefficients provided also return the two-sided p-value for each test, which is used to filter the results by the p-value cutoff of TaxCo. Only correlations matching this cutoff are saved in the intermediate *'taxcoCor'* files generated by this step.

Correlating Metadata Metadata are read from the required tab-delimited metadata table which provides a numeric value for each attribute on each sample. All metadata attributes are correlated with all assigned taxa in the same way the taxa were correlated in the previous step and saved in the *'taxcoMeta'* file.

Generating Networks In the network generation step, the correlations are converted to edges between correlated taxa or metadata and all edges are assigned their color - blue for positive correlation and red for negative

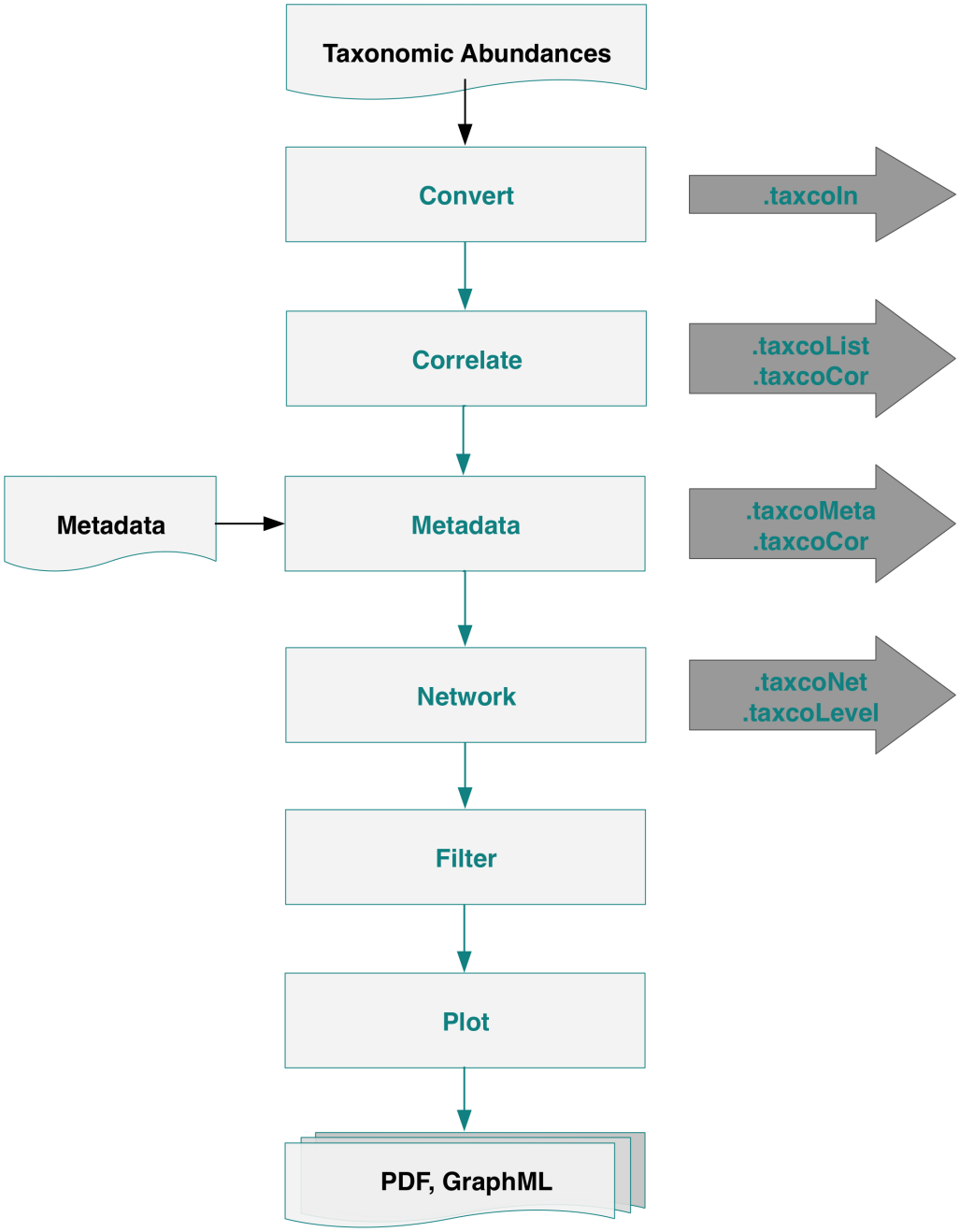

Figure 8.4: Steps of the TaxCo Pipeline to determine relevant correlations from taxonomic abundance data. Correlations are determined between taxa and between a taxon and given numeric metadata.

correlation - and edge weight, which is a scaled version of the absolute correlation value. These networks are saved in the *'taxcoNet'* file. Additionally, a *'taxcoLevel'* file is generated for each taxonomic level, which is only necessary for plotting.

Filtering with cutoffs The network files still hold all computed correlations, but the filtering step now generates separate files for each absolute correlation value cutoff which has been set by the user. Often, useful cutoffs to get both an overview of all correlations and specifically select strong correlations are 0.3, 0.6 and 0.8. Low cutoffs safe more correlations in the filtered files, which could be useful for further analysis, but can be hard to plot. Strong correlation cutoffs return only the strongest correlations and often provide useful overview plots of the correlation network in the dataset even with the automated layout used by TaxCo.

Generating graphical output All generated (filtered and unfiltered) *'taxcoNet'* files are finally converted into a graph representation using the python package NetworkX [31], saved in GraphML format for use in a variety of available network visualization tools and automatically plotted using matplotlib [34].
An exemplary output of an automatically generated plot is shown in Figure 8.5.

Figure 8.5 shows the average strong negative correlation between Bacteroidetes and Firmicutes, which is common to be found in human gut metagenomic data, based on the dysbiosis of those two taxa. People usually have high counts in Bacteroidetes, but low numbers of Firmicutes or the other way round - which is thought to be associated with the general gut health and the weight of the individual. There is also a strong positive correlation between Chlamydiae and Nematoda, which is unfortunately caused by database contamination. These taxa often occur in samples from both human and mice, especially in the samples which have higher abundances of host reads. Potentially, sequences from these taxa have either contaminated sequences used for reference genomes of human and mouse, or they do share real sequence similarity. Either way, these are common mis-assignments caused by contamination and their correlation is caused by the confounding factor of the abundance of host sequence in the sample.
As it is often critically mentioned, correlation does not imply causation, especially in this setting using variable biological data with many unknown influences that could act as confounding factors. TaxCo visualizes the correlations as a network because the idea of using correlation analysis on this data

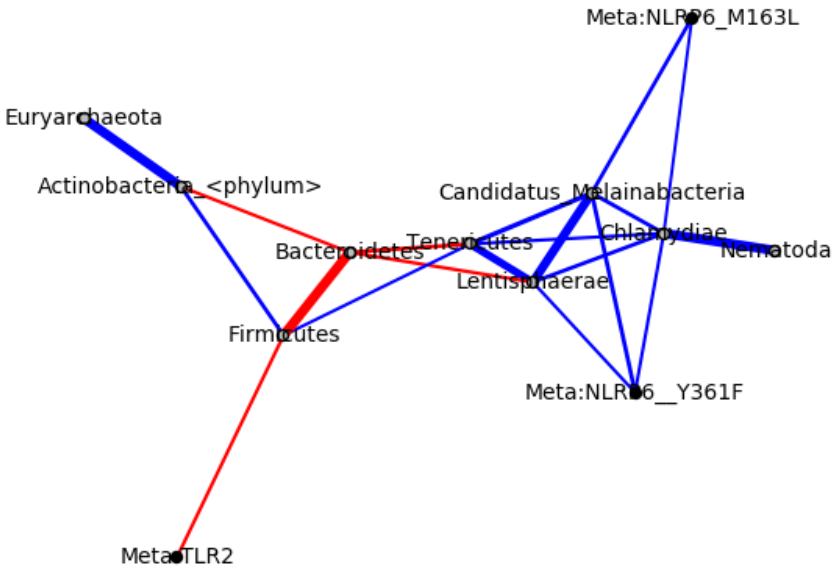

Figure 8.5: TaxCo Correlation Network plot on Phylum level for pearson correlation on the Immigene dataset. Red edges represent negative correlation, blue edges positive correlation. Edge width is relative to correlation strength. Correlations have been generated with a p-value cutoff of 0.05 and plotted automatically by TaxCo. Metadata is marked by "Meta"

is based on the notion that we are studying a community of taxa. Correlation between abundances of taxa can have an abundance of reasons. They can be caused by technical errors or contamination like in the case of Chlamydiae and Nematoda; they can be caused by confounding factors which were not measured, but also by any interaction between the taxa. Correlation networks provide a view of the interrelations of the community. The added correlation with metadata can potentially identify them as confounding factors, but also show over which taxa they doe influence the community or which taxa might influence changes in their environment. TaxCo is meant to identify sub-communities, basically functional clusters, of taxa in a dataset which then can be selected for further study. The visualization as an undirected network is meant to avoid producing assumptions on causation from the results.

8.3 MAPle - Metagenomic Analysis PipeLinE

MAPle stands for Metagenomic Analysis PipeLine. MAPle provides three different modules for the analysis of paired-end short read metagenomic data. All modules can be selected in the same run, but they will currently run consecutively for each sample. If no modules are selected, only the preprocessing steps are run. A further extension to make it applicable for single-end, assembled and long-read input and to parallelize the modules are planned. Another feature could be to provide adaptor removal in preprocessing, but as this step is often done by sequencing providers and directly on the machine, datasets where adapter sequences are common enough to cause problems are often generally unfavorable for analysis.

Preprocessing The preprocessing step is the basis for all three analysis modules of MAPle. Quality control for raw and trimmed reads is done using FastQC [3] and trimming with prinseq-lite [70]. Parameters for trimming can be set through the configuration file and include average quality, window size, and minimum read length to keep after trimming. Quality control includes a breakpoint at the end of preprocessing where analysis through the modules is only done if given thresholds for the minimal number of reads in the sample after preprocessing and maximal loss of reads compared to raw reads in percent during preprocessing are met.
After preprocessing, any of the modules can be run for analysis. Modules will run sequentially if multiple modules are selected, but they are essentially independent of each other, and this step could be parallelized in the future. The overall structure of the pipeline is shown in Figure 8.6, the structure of the modules is shown in Figure 8.7 and will now be described in further detail.

Basic Metagenomics The basic metagenomic analysis is done by aligning the trimmed reads from preprocessing directly against a protein database using DIAMOND [15]. Because of the speed of DIAMOND, it is possible to align metagenomic data against the full NCBI NR database in feasible time. The matches from alignment are saved in the DAA output format of DIAMOND. The DAA files are analyzed with the daa2rma tools from the MEGAN 6 Community Edition [35] which provides taxonomic classification with the LCA algorithm and functional classification based on mapping the assignments onto the functional ontologies supported by MEGAN though the appropriate mapping files. The output of this module is the resulting RMA file for each sample.

Host-associated Data Datasets with a known significant contamination by host reads or potentially even other known contaminants can be analyzed with the Host-associated Data module. Here a database of host (or contaminant) sequences has to be provided for MALT. This database can easily be generated from a FastA file of the sequences. Trimmed reads are first aligned against this filter database with comparatively strict parameters to avoid mis-mapping reads from the microbial community, and only the reads unmapped in this alignment are further analyzed. Host reads are kept separately and can be used for other further analysis. The filtered reads are again aligned against a protein database with DIAMOND and taxonomically and functionally assigned by MEGAN, as in the Basic Metagenomics module. The output is an RMA file of the assignments from non-host reads. Unfortunately, depending on the size of the filter database, the filtering step can be very time-consuming. This module is supposed to be a way of salvaging heavily contaminated datasets if necessary and should not be used as a standard. With current DNA extraction methods and sequencing, contamination for a standard metagenomic sample should be kept to a minimum and in the worst case be filtered before sequencing if possible, not to lose too much sequencing depth on those additional sequences. Faster alignment and classification of the reduced number of reads in a dataset that initially had large numbers of contaminant sequence can offset the time spent a little, but in general it is advised to use the Basic Metagenomics module for timely analysis and keep the filtering option as a last resort for samples which could not be sufficiently analyzed using that module.

Taxonomic Analysis The taxonomic analysis is of course also provided by both other modules, based on the analysis of the proteins matched to the

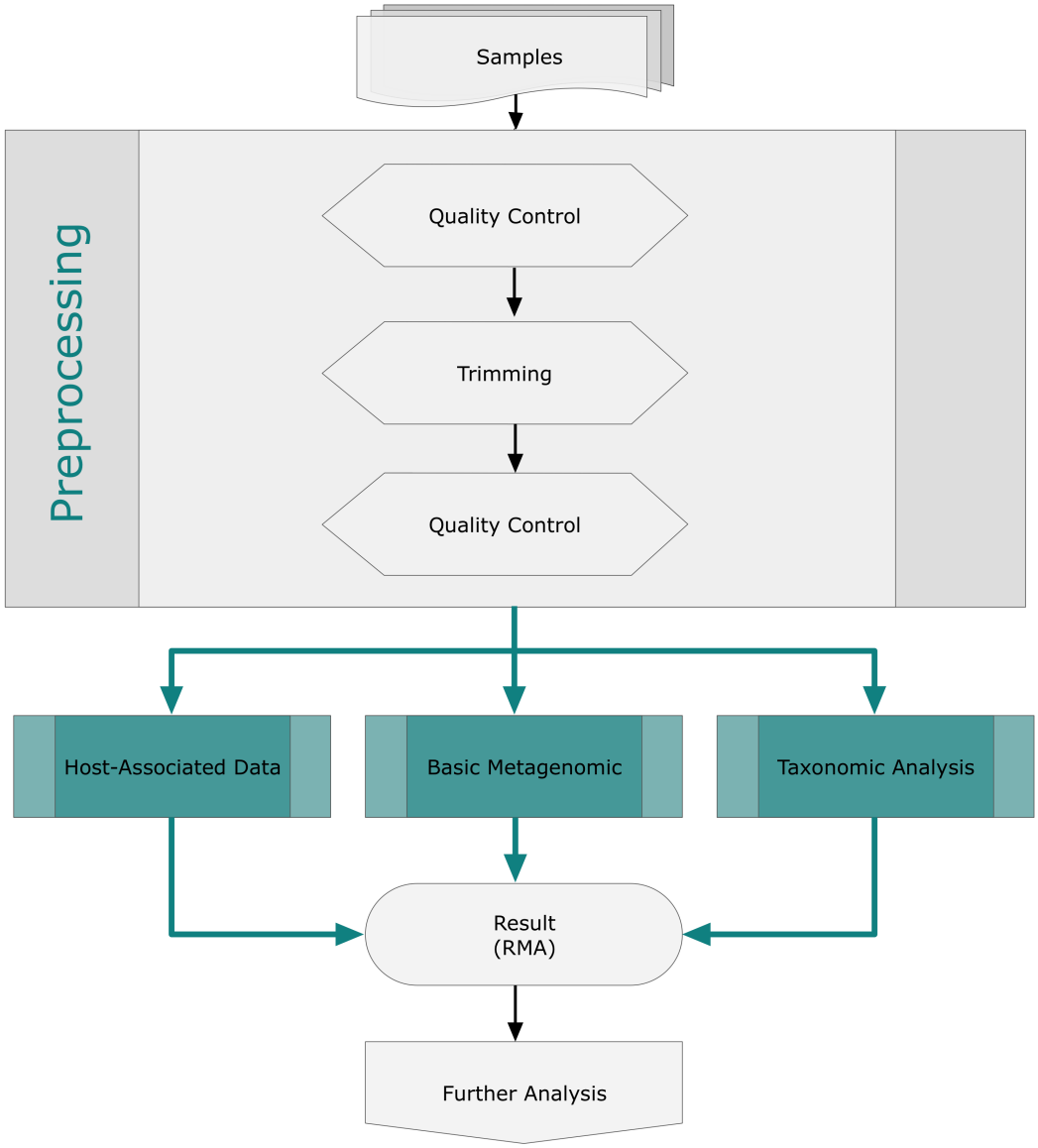

Figure 8.6: MAPle, including three modules which can be selected for analysis in parallel.

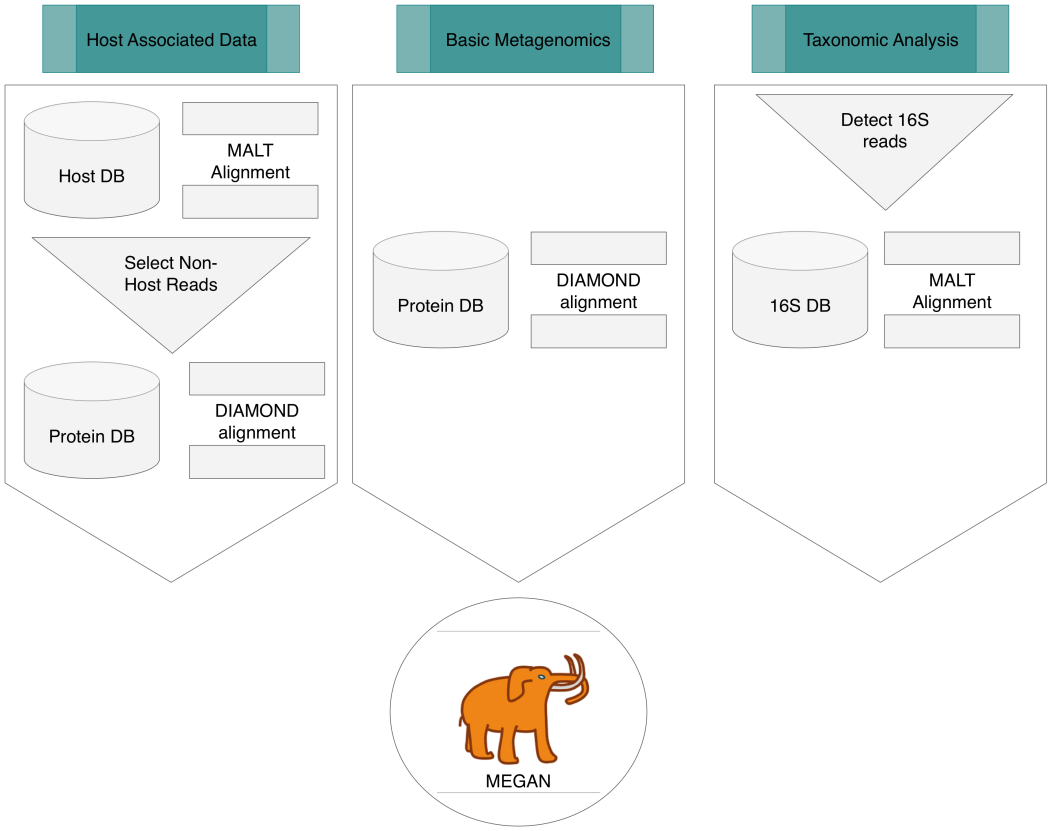

Figure 8.7: MAPle modules available for analysis

database. However, the Taxonomic Analysis module is very different from the two other modules. Here, reads from the genes coding for the small subunit (SSU) of the ribosomal RNA are filtered out from the preprocessed reads of each sample using the tool Metaxa 2 [13]. Metaxa uses HMMs to detect those reads and can provide all potential SSU sequences from the sample in a separate FastA file. These reads come not only from bacteria but potentially also from eukaryotic, mitochondrial and chloroplast rRNAs. MAPle compares those reads against a suitable database for taxonomic assignment. using MALT [33]. The database could, for example, be a standard 16S database, which would make this step comparable to the 16S analysis of STARA. Aligned reads are taxonomically assigned with the LCA algorithm, and STARA finally provides RMA output.

As all three modules provide the same output format, results from all modules on the same sample can easily be compared using MEGAN. An example of this is shown in Figure 8.8, which shows the relative abundance of the Top 10 Genera assigned by MAPle on the same healthy human gut sample using

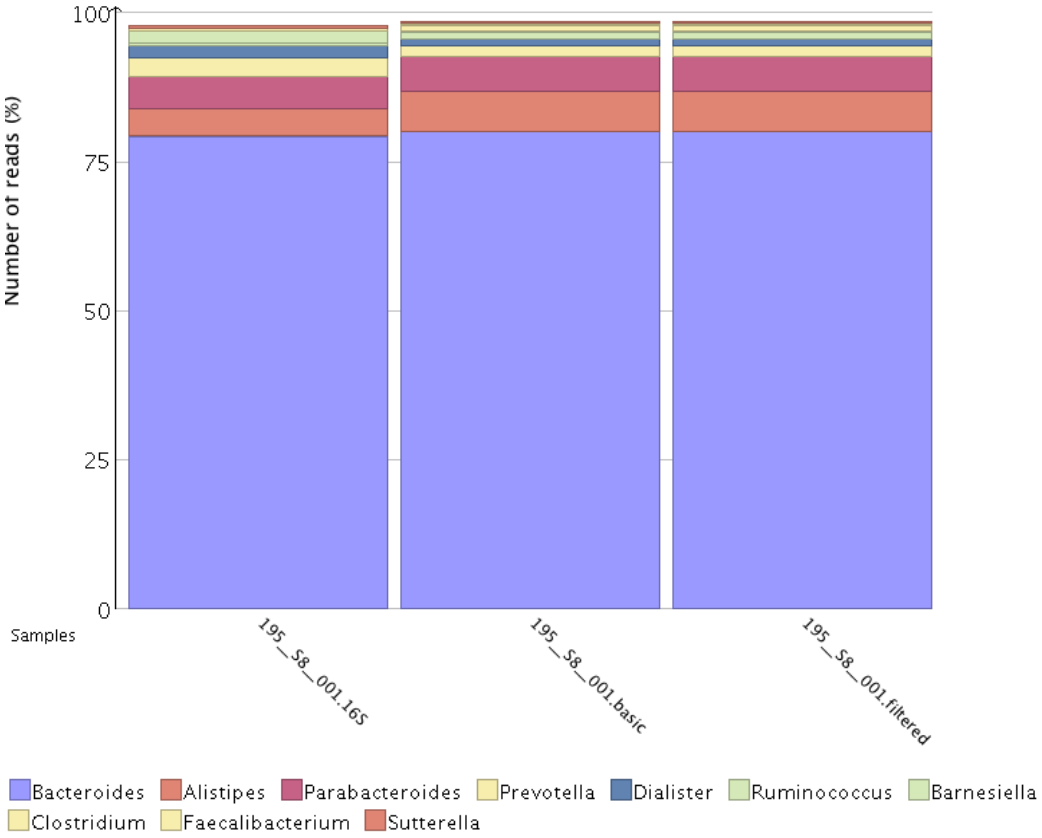

Figure 8.8: Relative abundances on genus level from MAPle analysis for one ImMiGeNe sample (healthy human gut) using all three modules.

the three different modules. In general, all three modules did assign most reads to Bacteria. The two metagenomic modules (Basic and Host-filtered) are very similar as this sample had very little host-reads which could be filtered.

The metagenomic and 16S based samples show a difference both in specificity and diversity. While the metagenomic samples do have a lot of additional species-specific assignments with very low read counts, the diversity calculated for the samples is generally higher for the metagenomic sample (see Table 8.1). Species-level assignments are generally hard to do based on partial 16S sequences, so the most relevant comparison for these samples is on the Genus level like it is done in Figure 8.8.
Overall, results from the different modules of MAPle are comparable under good conditions - meaning little host or contamination sequences to filter and enough 16S rDNA sequence coverage in the sample to enable a reasonable 16S sequence analysis. For samples which lack in coverage of the microbial

Diversity Index	16S module	Filtered module	Basic module
Shannon-Weaver	3.825	3.770	3.770
Simpson's reciprocal	11.165	10.977	10.978

Table 8.1: Diversity indices for the assignments on one ImMiGeNe sample (healthy human gut) using all three MAPle modules

community, the filtered module will, of course, provide different results to the basic module, as more reads of the input can be assigned to microbes and hence more microbial taxa will pass the minimal support parameter. For samples where previous 16S rDNA sequencing and analysis of similar datasets has given very different results compared to the taxonomic analysis from the WGS metagenomic sequencing, the additional 16S analysis module can provide insight on the cause of this problem - an actual difference in the sample or change caused by using different wet-lab protocols for preparation of the sequencing or potentially a technical difference caused by using different databases and the different information content of protein annotations compared to 16S assignment.

9

Conclusions from Part II

While metagenomic analysis is an important tool for the future of our journey to understanding how genomes influence and shape organisms and what part the environment plays in this as well as how organisms interact and co-develop in a shared environment, it is still often reduced to OTU-based 16S rDNA analysis. 16S sequencing is comparatively cheap and readily available, and the analysis can be done by available pipelines, even using web services like QIIME 2 which have proven to produce reproducible results. While this is surely a start in community analysis, it is heavily based on clustering algorithms and the selection of representative sequences from the OTUs. Sometimes assignments which could be made by comparing a single read to a database will be lost in this progress. As I have faced these problems through the course of different projects and especially during the planning phase of the TüBiom project, the decision to provide an automated alignment-based 16S analysis pipeline was clear. STARA aims to be at least as user-friendly as QIIME and mothur are and can be developed further with future user feedback.
However, 16S based analysis is not the future standard for analysis of microbial communities. It should and will be replaced by WGS metagenomic sequencing. WGS sequencing is still only feasible for studies which have access to a comparatively large infrastructure of sequencing providers, data management and storage and bioinformatics analysts. There are web services available for metagenomic data analysis (for example MG-RAST [43] or IMG-M [18]), but they require time-consuming uploads of the raw input data, which is usually not feasible for large datasets. Providing pipelines like MAPle does not negate the need for infrastructure and specialists able to proceed with the further analysis, but it reduces time spent on repetitive tasks and does provide a basic analysis which can potentially already provide the necessary insights. The pipeline can be used intuitively with little adaption to given configuration files necessary. Given the appropriate infrastructure, it will, of course, run faster, but it does run on a relatively minimal Linux

system setting, which would be feasible at least for the analysis of a few samples. Similarly, it does scale well to large datasets given the appropriate infrastructure. While analysis can take some time, the process is completely hands-off after starting and no unnecessary analysis time will be spent on samples with low sequencing depth.
For further analysis of both 16S or WGS metagenomic data, TaxCo is available to give a first overview on the potential interactions in the studied community, showing off the complexity of the microbiome as an interconnected system in contact with its environment.
These types of analysis could hopefully bring environmental data analysis "to the bench", where it can, for example, be useful in the investigative analysis of a patient sample directly in the clinic, without the need for bioinformatics specialists to be involved from the start. It hopefully also can be a basis of many more elaborate analysis pipelines which can utilize the generated data and further research, not only in medicine but also in many other (life) sciences.

Part III

Conclusions and Outlook

Sequencing has changed the world of microbe study just as the microscope did in the 17th century. As the bottleneck of genomics and metagenomics shifting away more and more from the generation of data to the analysis of an ever-increasing flow of data, automated pipelines for fundamental data analysis will keep gaining importance. Genomics is already providing many options for automated data analysis which give sufficient results, while WGS metagenomics is still at the very beginning of a development to be an easily attainable standard method for any study of microbes in their environment.
Following the development short read genomics has already made, 16S rDNA sequencing is now quickly developing to become a mainstream, investigative technique being used to gain insights from preliminary data instead of being the main tool of the trade in environmental sequencing, especially in a research setting. The ability to analyze the flood of data at least on a basic level with low effort enables large-scale experiments to be feasible to handle on a comparatively small infrastructure and helps bring all these techniques into the daily life of scientists and life science in general.
With the availability of metagenomics as a standard tool, modern health care will change drastically. We will be able to diagnose new diseases, and diagnosis of known diseases will be more efficient. Personalized treatment decisions will be made based on a patients microbiota, and we will both be able to learn how to preserve the healthy microbiome of a patient and even utilize it to fight disease. General well-being of the public will be improved by insights on how to fight obesity, indigestion or malnutrition in an optimal way for every patient. All of these developments though need the ability to automate as much of the data analysis steps as possible to be able to provide it without the barrier of sending a patient to a specialized center which has a large infrastructure and many specialists available. Metagenomics has to be brought to every hospital to make it available just like skin swabs or blood testing.
The pipelines I have designed throughout gaining experience with the needs of data analysis for medical research do not require large financial investment. The tools used are open source and free for academic use or can at least be easily licensed even for commercial use. They do run on any Linux based system, and the pipelines are designed to be intuitive and need very little user input which might need specialist knowledge. This way they can help bring metagenomics closer to the people who ask the questions, into daily healthcare situations, and in the future maybe even to citizen science.
Making the technology approachable for a broader spectrum of users has, of course, its downsides. With the flood of data being generated from genome sequencing, many bacterial genomes are often insufficiently studied, as they can be automatically assembled and annotated and then forgotten. The project in Part I of this thesis is an example of how this can fail, and more focused data generation and analysis can improve the knowledge gained compared

to standard automated - default - methodology. However, environmental sequencing is not yet close to this stage where datasets are cheap enough to be generated and forgotten. It is entering the scientific toolkit of many fields like medicine or agricultural science based on the promises fundamental research in the field has produced. These ground-breaking analysis though require a strong computer infrastructure and the involvement of a large interdisciplinary team which covers the experimental wet-lab stage, DNA extraction and sequencing, sequence analysis and finally a concerted effort of experiment planning before all of this can begin and interpretation of the data after it is finished.
Working with specialists and maintaining a complex infrastructure might be possible for large institutions, but many projects will not be started because the basic research, preliminary experiments, and analyses cannot be handled by the scientists before they can prepare and plan a larger effort. Here automated data analysis can already come in handy when sequencing could be outsourced to sequence providers.
Additionally, in a situation where already specific kinds of metagenomic datasets are produced on a regular basis - like they potentially would be in a hospital - the fundamental analysis of these datasets is repetitive and generally based on the same types of input data and output requirements. If the analysis of this data can be automated, more time for further analysis will be left because of the reduced hands-on time spend on the fundamental analysis.
The projects presented in this work provide an overview of different use cases of genomic and environmental sequencing in medical research, connecting the research to public outreach and education with the TüBiom project and potential application in diagnostic decisions and health care in the future of the ImMiGeNe project. Of course, they can also be applied to datasets outside of the medical field.
The tools I developed throughout working on those projects were able to provide all fundamental data analysis necessary, but they also have their limitations. With the future development of sequencing, they would, of course, need to be adapted and further developed. For example, when long-read sequencing in metagenomics enters general use, MAPle could be adapted to include a single read mode for short contigs or sequences from technologies which provide single-end reads with lengths under 1000 bp and a full long-read mode with frameshift-aware alignment and better classification adapted to the typical sequences provided by PacBio or Oxford Nanopore-like sequencing technologies. Full-length 16S rDNA sequences with low insertion- and deletion-error probability as they are generated by PacBio CCS sequencing can already be analyzed using STARA with appropriate parameters.
Bioinformatics has always been a field moved continuously by technological developments and participating in cutting-edge life science research. However,

it has also been seen as a "service science" supposed to provide results through the press of a button. I hope with the work I have presented here I can help bridge the gap between this two very different expectations, further cutting-edge science, and personalized healthcare by reducing the need to spend time and effort on the most fundamental tasks and thus enable bioinformaticians to get back onto the interesting new developments. instead of drowning in the never-ending flood of sequence data.

Appendices

Appendix A

Supplemental Material

locus tag	original annotation	RefSeq annotation
BvMPK_0107	alpha/beta fold family hydrolase	Annotated as pseudogene
BvMPK_0108	putative general stress protein	Not found
BvMPK_0109	Regulatory sensor-transducer, BlaR1/MecR1	TonB family protein
BvMPK_0110	Transcriptional Repressor CopY Family	BlaI/MecI/CopY family transcriptional regulator
BvMPK_0111	Glyoxalase family protein	VOC family protein
BvMPK_0112	hypothetical protein	(4Fe-4S)-binding protein
BvMPK_0113	Transcriptional regulator, AraC family	AraC family transcriptional regulator
BvMPK_0114	putative general stress protein	pyridoxamine 5'-phosphate oxidase
BvMPK_0115	Transcriptional regulator	general stress protein
BvMPK_0116	Beta-lactamase	class A beta-lactamase
BvMPK_0117	Transcriptional repressor, BlaI/MecI family	BlaI/MecI/CopY family transcriptional regulator

Table A.1: Functional annotations in the original genome and RefSeq annotation for the *B. vulgatus* mpk insertion in the CTn341 conjugative transposon

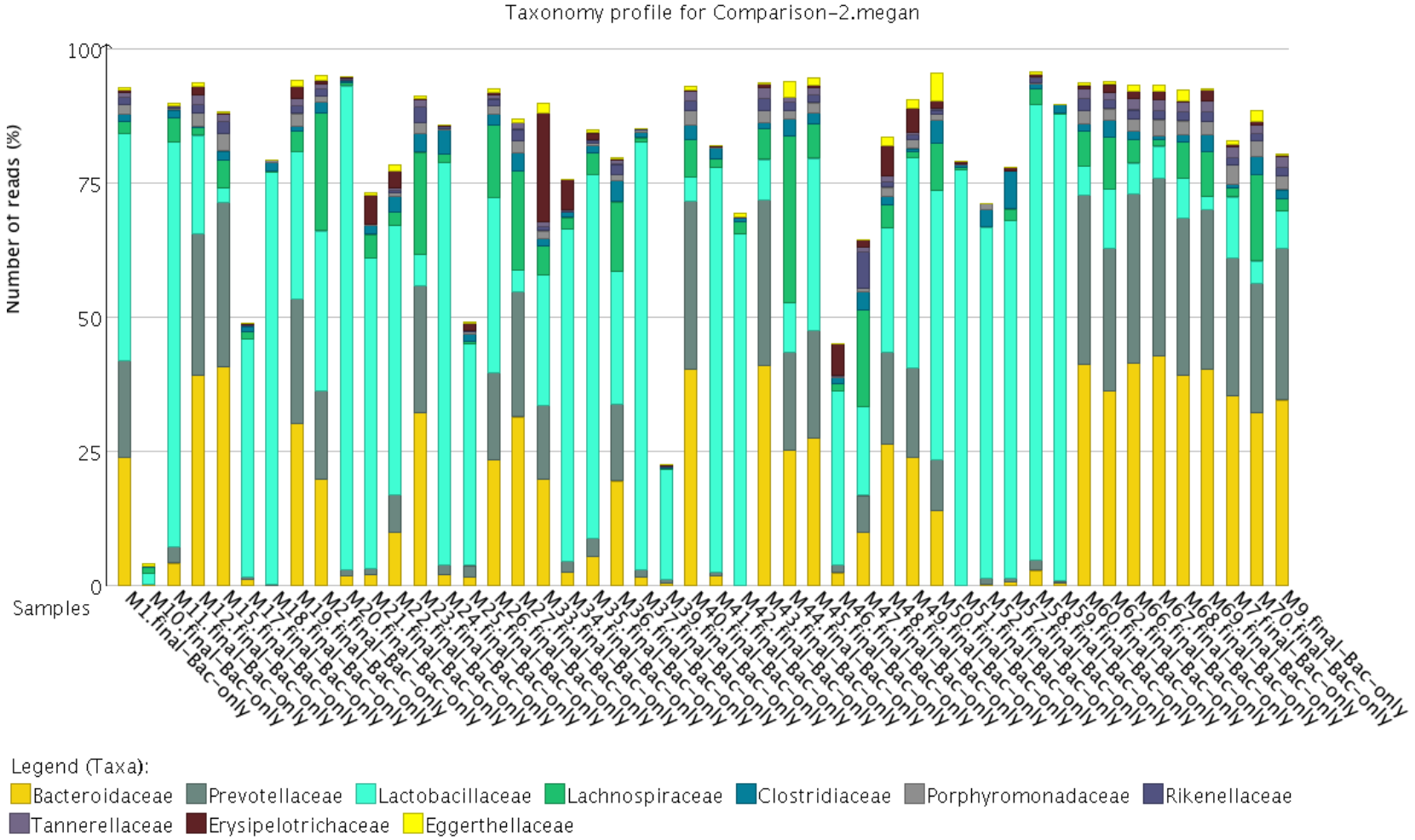

Figure A.1: Relative abundances of the top 10 family-level taxa as they were assigned from metagenomic sequencing. Abundances are generated only from the metagenomic reads assigned to bacteria, which means relative abundance means they are only relative to the total of bacteria, not the total of all metagenomic sequence assignments.

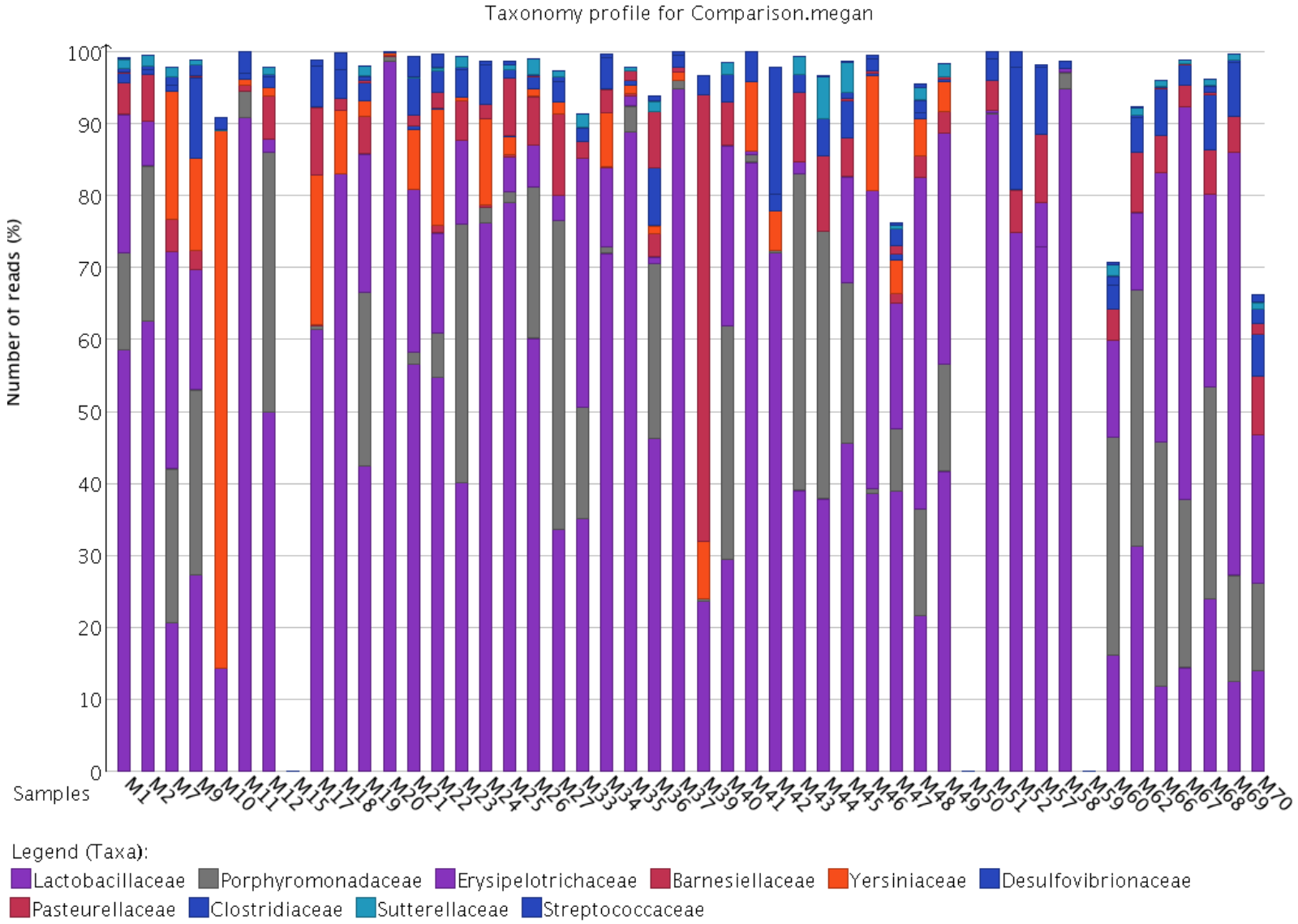

Figure A.2: Relative abundances of the top 10 family-level taxa as they were assigned from 16S sequencing

Appendix B

Contributions

Assembly and Annotation of *Bacteroides vulgatus* mpk

Sina Beier (SB) and Anna Lange (AL) contributed to this project. AL isolated the DNA and interpreted the analysis results. SB preprocessed and assembled the data, provided the annotations and genome comparison.

Studying the change of mouse gut microbiome during infection with *Yersinia enterocolitica*

Sina Beier (SB), Janina Geißert (JG), Monika Schütz (MS), Erwin Bohn (EB), Daniel H. Huson (DHH) and Ingo Autenrieth (IA) contributed to this project. MS, EB, DHH and IA conceived the project. SB, JG, MS, EB and IA planned the experiments, JG conducted the experiments and extracted the DNA. SB conducted the analysis and wrote the 16S analysis pipeline.

Metagenomic analysis for the ImMiGeNe project

Sina Beier (SB), Daniel H Huson (DHH), Alexander Weber (AW), Silke Peter (SP) and a lot of other people contributed to this project. DHH, AW and many others conceived the project. SB and AW planned the analyis pipeline, SB wrote the analysis pipeline and conducted the analysis. JG and SB interpreted the results.

Automated 16S rDNA analysis pipeline for the TüBiom project

Sina Beier (SB), Isabell Flade (IF), Daniel H Huson (DH), Ingo Autenrieth (IA), Matthias Willmann (MW), Anna Gorska (AG), Patrick Grupp (PG) and Theresa Anisia Harbig (TAH) contributed to this project. IF, DH, IA

and MW conceived and planned the project. IF conducted the sequencing and preliminary studies. SB developed the data analysis pipeline and analysed the data, AG and PG planned and developed the database system and the website. AG planned the visualisations, AG and TAH developed the visualisations.

Appendix C

The CommunAl Toolkit

The CommunAl toolkit source code is available on GitHub:

STARA

Available at:
`https://github.com/BioSina/STARA.git`

MAPle

Available at:
`https://github.com/BioSina/MAPle.git`

TaxCo

Available at:
`https://github.com/BioSina/TaxCo.git`

Appendix D

My Publications

S. Beier, A. Górska, P. Grupp, T. A. Harbig, I. Flade, and D. H. Huson. Bioinformatics support for the Tuebiom community gut microbiome project. *PeerJ Preprints*, pages 1–9, 2016

D. H. Huson, S. Beier, I. Flade, A. Górska, M. El-Hadidi, S. Mitra, H.-J. Ruscheweyh, and R. Tappu. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS computational biology*, 12(6):e1004957, 2016

A. Lange, S. Beier, A. Steimle, I. B. Autenrieth, D. H. Huson, and J.-S. Frick. Extensive mobilome-driven genome diversification in mouse gut-associated Bacteroides vulgatus mpk. *Genome Biol. Evol.*, 8(4):1–34, 2016

S. Beier, R. Tappu, and D. H. Huson. Functional Analysis in Metagenomics Using MEGAN 6. In *Functional Metagenomics: Tools and Applications*, pages 65–74. Springer International Publishing, Cham, 2017

A. Lange, S. Beier, D. H. Huson, R. Parusel, F. Iglauer, and J.-S. Frick. Genome Sequence of Galleria mellonella (Greater Wax Moth). *Genome announcements*, 6(2):e01220–17, 2018

Bibliography

[1] O. S. Alkhnbashi, F. Costa, S. A. Shah, et al. CRISPRstrand: Predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, 30(17):489–496, 2014.

[2] C. Alonso-Blanco, J. Andrade, C. Becker, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, 166(2):481–491, 2016.

[3] S. Andrews. FastQC - A quality control tool for high throughput sequence data.

[4] D. Antipov, A. Korobeynikov, J. S. McLean, and P. A. Pevzner. HybridSPAdes: An algorithm for hybrid assembly of short and long reads. *Bioinformatics*, 32(7):1009–1015, 2016.

[5] E. Aronesty. ea-utils, 2011.

[6] M. Ashburner, C. A. Ball, and J. A. Blake. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[7] A. Auton, G. R. Abecasis, D. M. Altshuler, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[8] R. Aziz, D. Bartels, A. Best, and M. DeJongh. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, 9(1):75, jan 2008.

[9] M. Bacic, A. C. Parker, J. Stagg, et al. Genetic and structural analysis of the Bacteroides conjugative transposon CTn341. *Journal of Bacteriology*, 187(8):2858–2869, 2005.

[10] A. Bankevich, S. Nurk, D. Antipov, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–77, may 2012.

[11] S. Beier, A. Górska, P. Grupp, T. A. Harbig, I. Flade, and D. H. Huson. Bioinformatics support for the Tuebiom community gut microbiome project. *PeerJ Preprints*, pages 1–9, 2016.

[12] S. Beier, R. Tappu, and D. H. Huson. Functional Analysis in Metagenomics Using MEGAN 6. In *Functional Metagenomics: Tools and Applications*, pages 65–74. Springer International Publishing, Cham, 2017.

[13] J. Bengtsson, K. M. Eriksson, M. Hartmann, et al. Metaxa: A software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 100(3):471–475, 2011.

[14] E. Bohn, O. Bechtold, N. Zahir, et al. Host gene expression in the colon of gnotobiotic interleukin-2-deficient mice colonized with commensal colitogenic or noncolitogenic bacterial strains: common patterns and bacteria strain specific signatures. *Inflammatory Bowel Diseases*, 12(9):853–862, 2006.

[15] B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(November):59–60, 2014.

[16] J. G. Caporaso, J. Kuczynski, J. Stombaugh, et al. QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature Publishing Group*, 7(5):335–336, 2010.

[17] R. R. Chaudhuri, N. J. Loman, L. A. S. Snyder, et al. xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic acids research*, 36(Database issue):D543–6, jan 2008.

[18] I. M. A. Chen, V. M. Markowitz, K. Chu, et al. IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research*, 45(D1):D507–D516, 2017.

[19] F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterson. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[20] M. J. Coyne, C. M. Fletcher, M. Chatzidiaki-Livanis, et al. Phylum-wide general protein O-glycosylation system of the Bacteroidetes. *Mol Microbiol.*, 88(4):772–783, 2013.

[21] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1403, 2004.

[22] M. de Bruyn, J. Sabino, D. Vandeputte, et al. Comparisons of gut microbiota profiles in wild-type and gelatinase B/matrix metalloproteinase-9-deficient mice in acute DSS-induced colitis. *npj Biofilms and Microbiomes*, 4(1):18, 2018.

[23] A. Delcher. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641, 1999.

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

[25] H. A. Erlich. Polymerase chain reaction. *Journal of Clinical Immunology*, 9(6):437–447, 1989.

[26] J. D. Evans, S. J. Brown, K. J. Hackett, et al. The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, 104(5):595–600, 2013.

[27] R. D. Finn, J. Mistry, J. Tate, et al. The Pfam protein families database. *Nucleic acids research*, 38(November 2009):211–222, 2010.

[28] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, et al. The Pfam protein families database. *Nucleic Acids Research*, 36(November 2007):281–288, 2008.

[29] N. R. Garud, B. H. Good, O. Hallatschek, and K. S. Pollard. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *bioRxiv*, page 210955, 2017.

[30] X. Guo, S. Li, J. Zhang, et al. Genome sequencing of 39 Akkermansia muciniphila isolates reveals its population structure , genomic and functional diverisity , and global distribution in mammalian gut microbiotas. *BMC genomics*, pages 1–12, 2017.

[31] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proc. SciPy 2008*, pages 11–15, 2008.

[32] G. Hannon. FastX Toolkit, 2009.

[33] A. Herbig, F. Maixner, K. I. Bos, A. Zink, et al. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv*, page 050559, 2016.

[34] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[35] D. H. Huson, S. Beier, I. Flade, A. Górska, M. El-Hadidi, S. Mitra, H.-J. Ruscheweyh, and R. Tappu. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS computational biology*, 12(6):e1004957, 2016.

[36] L. J. Jensen, P. Julien, M. Kuhn, et al. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36(Database):D250–D254, 2007.

[37] B. Jia, A. R. Raphenya, B. Alcock, et al. CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1):D566–D573, 2017.

[38] J.-H. Jo, E. A. Kennedy, and H. H. Kong. Bacterial 16S ribosomal RNA gene sequencing in cutaneous research. *J Invest Dermatol*, 136(3):e23–e27, 2016.

[39] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed ¡today¿].

[40] J. Jovel, J. Patterson, W. Wang, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, 7(APR):1–17, 2016.

[41] M. Kanehisa, S. Goto, Y. Sato, et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue):D109–14, jan 2012.

[42] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, page gkv1070, 2015.

[43] K. P. Keegan, E. M. Glass, and F. Meyer. *MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function*, pages 207–233. Springer New York, New York, NY, 2016.

[44] S. Koren, M. C. Schatz, B. P. Walenz, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693–700, jul 2012.

[45] S. Koren, B. P. Walenz, K. Berlin, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27:722–736, 2017.

[46] L. Krych, C. H. F. Hansen, A. K. Hansen, et al. Quantitatively different, yet qualitatively alike: a meta-analysis of the mouse core gut microbiome

with a view towards the human gut microbiome. *PloS one*, 8(5):e62578, jan 2013.

[47] M. Krzywinski, J. Schein, I. Birol, et al. Circos- An information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–45, 2009.

[48] A. Lange, S. Beier, D. H. Huson, R. Parusel, F. Iglauer, and J.-S. Frick. Genome Sequence of Galleria mellonella (Greater Wax Moth). *Genome announcements*, 6(2):e01220–17, 2018.

[49] A. Lange, S. Beier, A. Steimle, I. B. Autenrieth, D. H. Huson, and J.-S. Frick. Extensive mobilome-driven genome diversification in mouse gut-associated Bacteroides vulgatus mpk. *Genome Biol. Evol.*, 8(4):1–34, 2016.

[50] S. J. Lange, O. S. Alkhnbashi, D. Rose, et al. CRISPRmap: An automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Research*, 41(17):8034–8044, 2013.

[51] S. Levy, G. Sutton, P. C. Ng, et al. The diploid genome sequence of an individual human. *PLoS Biology*, 5(10):2113–2144, 2007.

[52] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, 2009.

[53] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

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

[55] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–4, 1977.

[56] M. Mignardi and M. Nilsson. Fourth-generation sequencing in the cell and the clinic. *Genome medicine*, 6(4):31, 2014.

[57] M. Müller, K. Fink, J. Geisel, et al. Intestinal colonization of IL-2 deficient mice with non-colitogenic B. vulgatus prevents DC maturation and T-cell polarization. *PLoS ONE*, 3(6), 2008.

[58] E. W. Myers, G. G. Sutton, A. L. Delcher, et al. A Whole-Genome Assembly of Drosophila. *Science*, 287(March):2196–2205, 2000.

[59] K. H. Nam, C. Haitjema, X. Liu, et al. Cas5d protein processes Pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg crispr-cas system. *Structure*, 20(9):1574–1584, 2012.

[60] Y.-D. Nam, M.-J. Jung, S. W. Roh, et al. Comparative analysis of Korean human gut microbiota by barcoded pyrosequencing. *PloS one*, 6(7):e22109, jan 2011.

[61] M. Nguyen and G. Vedantam. Mobile genetic elements in the genus Bacteroides , and their mechanism(s) of dissemination. *Mobile Genetic Elements*, 1(3):187–196, 2011.

[62] R. Overbeek, T. Begley, R. M. Butler, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, 33(17):5691–702, jan 2005.

[63] S. Powell, K. Forslund, D. Szklarczyk, et al. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, 42(D1):D231–D239, 2014.

[64] J. Qin, R. Li, J. Raes, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, mar 2010.

[65] J. Quick, N. J. Loman, S. Duraffour, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[66] A. I. Rissman, B. Mau, B. S. Biehl, et al. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, 25(16):2071–2073, aug 2009.

[67] K. Rutherford, J. Parkhill, J. Crook, et al. Artemis: Sequence visualization and annotation. *Bioinformatics*, 16(10):944–945, 2000.

[68] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, 1977.

[69] P. D. Schloss, S. L. Westcott, T. Ryabin, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–41, dec 2009.

[70] R. Schmieder and R. Edwards. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, 6(3):e17288, jan 2011.

[71] R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

[72] S. A. Shmakov, V. Sitnik, K. Makarova, et al. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mbio*, 8(November):1–18, 2017.

[73] N. B. Shoemaker, H. Vlamakis, K. Hayes, and A. A. Salyers. Evidence for extensive resistance gene transfer among bacteroides spp . and among bacteroides and other genera in the human colon. *Appl Environ Microbiol*, 67(2):561–8, 2001.

[74] P. Stothard and D. S. Wishart. Automated bacterial genome analysis and annotation. *Current Opinion in Microbiology*, 9(5):505–10, oct 2006.

[75] J. S. Suchodolski, J. Camacho, and J. M. Steiner. Analysis of bacterial diversity in the canine duodenum, jejunum, ileum, and colon by comparative 16S rRNA gene analysis. *FEMS microbiology ecology*, 66(3):567–78, dec 2008.

[76] The InterPro Consortium, T. K. Attwood, A. Bairoch, et al. InterPro - an integrated documentation resource for Protein Families, Domains and Functional Sites. *Society*, 16(12):1145–1150, 2000.

[77] C. L. Thompson, A. Mikaelyan, and A. Brune. Immune-modulating gut symbionts are not " Candidatus Arthromitus ". *Nature*, 6(1):6–7, 2013.

[78] Y. Uchimura, M. Wyss, and S. a. Brugiroux. Complete Genome Sequences of 12 Species of Stable Defined Moderately Diverse Mouse Microbiota 2. *ASM*, 4(5):4–5, 2016.

[79] G. H. Van Domselaar, P. Stothard, S. Shrivastava, et al. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Research*, 33(Web Server issue):W455–W459, jul 2005.

[80] M. Waidmann, O. Bechtold, J.-S. Frick, et al. Bacteroides vulgatus protects against Escherichia coli-induced colitis in gnotobiotic interleukin-2-deficient mice. *Gastroenterology*, 125(1):162–177, jul 2003.

[81] D. A. Wheeler, M. Srinivasan, M. Egholm, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, 2008.

[82] R. Wu. Nucleotide sequence analysis of DNA. *Nature New Biology*, 236(68):198–200, 1972.

[83] J. Xu, M. A. Mahowald, R. E. Ley, et al. Evolution of symbiotic bacteria in the distal human intestine. *PLoS biology*, 5(7):e156, jul 2007.

"Science is never finished."

Albert Einstein

Statement of Authorship

I hereby declare that I am the sole author of this dissertation and that I have not used any sources other than those listed in the bibliography and identified as references. I further declare that I have not submitted this thesis at any other institution in order to obtain a degree.

Ort, Datum Unterschrift