Standardisation and optimisation techniques in gut microbiome community analysis

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von

Frank Michael Nikolaus Stämmler

aus

Königstein im Taunus

im Jahr 2019

Der Promotionsgesuch wurde eingereicht am:
13.06.2019

Die Arbeit wurde angeleitet von:
Prof. Dr. Rainer Spang

Unterschrift:

Frank Michael Nikolaus Stämmler

I dedicate this thesis to my beloved family and friends who supported me throughout these years far beyond imagination. A special feeling of gratitude to my loving parents, Erwin and Elke, my sister Frauke and my girlfriend Katharina, for their never ending words of encouragement and support. Additionally, I thank Sanne and Felix for substantiating my decision to pursue a doctorate, by believing where the only boundary is in our heads. Also many thanks to my long year companion and dear friend Martin for pushing me when I was getting lazy and distracting me when I needed it the most.

Declaration in lieu of oath

I herewith declare in lieu of oath that I have composed this thesis without any inadmissible help of a third party and without the use of aids other than those listed. The data and concepts that have been taken directly or indirectly from other sources have been acknowledged and referenced.

The persons listed beneath have helped me to select and choose the following material gratuitously/for a consideration in the manner described in each case:

1. Prof. Dr. Dr. André Gessner and staff (Collection and extraction of microbiome samples and generation of raw sequencing data for the experiments of chapters two and three)

2. Prof. Dr. Ernst Holler, Dr. Daniela Weber and staff (Collection of human stool specimens from ASCT patients for the experiment of chapter two)

Other persons have not helped to produce this work as regards to its content or making. In particular, I have not used the services of any professional agencies in return for payment or those of other persons. Nobody has received payment in kind - neither directly nor indirectly - from me for any work that is connected with the content of this doctoral thesis.

This thesis has not been submitted, wholly or substantially, neither in this country nor abroad for another degree or diploma at any university or institute.

I declare in lieu of oath that I have said nothing but the truth to the best of my knowledge and that I have not withheld any information.

Before the above declaration in lieu of oath had been taken down, I was advised about the significance of a declaration in lieu of oath as well as the legal consequences of an incorrect or incomplete declaration.

Frank M.N. Stämmler
June 2019

Declaration of own contribution to presented academic manuscript

Chapter two of the presented thesis has been adapted by the doctoral candidate from an already published manuscript (Stämmler et al. [1]):
Stämmler F, Gläsner J, Hiergeist A, Holler E, Weber D, Oefner PJ, Gessner A, Spang R. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome*. 2016;4:28. doi:10.1186/s40168-016-0175-0.

I hereby declare that my contribution to this publication included the design of the validation experiment, bioinformatics analysis and preparation of figures. Additionally, I conceptualized the presentation of the results and drafted all sections of the manuscript. The sections regarding laboratory methods were outlined by me following the experimental protocols of the laboratory staff (named below). Furthermore, I filled the position of corresponding author of the publication and coordinated the authors.

All laboratory work for this publication was performed by Joachim Gläsner and Andreas Hiergeist, with support by Nadja Reul, Claudia Deinzer, Christine Irtenkauf and Holger Melzl at the Institute of Clinical Microbiology and Hygiene at the University Clinic Regensburg. This included animal housing, harvesting, sample preparation, sequencing and quantification of 16S-rDNA copies. Human stool samples for this manuscript were collected and provided by Prof. Dr. Holler and Dr. Weber and their staff at the Department of Haematology and Oncology at the University Clinic Regensburg.

Frank M.N. Stämmler
June 2019

Acknowledgements

Firstly, I wish to express my sincere gratitude to my supervisor Prof. Dr. Rainer Spang for his ambitious support and guidance throughout my doctorate, as well as his patience and motivation. Furthermore, I am thankful for him providing me a productive and fun place to work at the Department of Statistical Bioinformatics. Besides my supervisor, I would like to give a special thanks to my further advisor Prof. Dr. Dr. André Gessner for generously providing a vast amount of data and fruitful discussion. Also for nurturing my fascination for the human microbiome and its relationships to the host, as well as introducing me to the annual conference meetings "Microbiota and Host" at Seeon Monastery.

Also many thanks go to my thesis committee, mentors Rainer Merkl and Sven Rahmann, as well as to everyone at the Regensburg International Graduate School of Life Sciences (RIGeL).

A special thanks is directed to Joachim Gläsner and Andreas Hiergeist for all the helpful hours of discussions, explanations, encouragement, motivation and their efforts in the laboratory to create all the data which I was able to use during my thesis. In this context, I also would like to thank Nadja, Claudia, Christine and Holger for their excellent work throughout all experimental procedures of sample preparation and sequencing.

Furthermore, I would love to thank all collaboration partners during my time as PhD student and everything I have learned from them. Namely, Peter Oefner, Ernst Holler and Daniela Weber.

Above and beyond, I feel very grateful that I had the honour to work with such terrific colleagues and discussion partners at the Department of Statistical Bioinformatics including Claudio, Anton, Christian, Paula, Farhad, Thorsten, Julia, Nicholas, Michael G., Franziska G., Michael A., as well as Franziska T. and Martin P..

Summary

With the emergence of high throughput next-generation sequencing the importance of the human gut microbiota as regulators, modulators and maintainers of human health and disease became more and more imminent. Advances in sequencing in the last two decades enabled the analysis of the composition and dynamics of the gut microbiome in unprecedented resolution and complexity. Investigations of this complex community by marker gene studies allowed assertions on presence, absence and ecological dynamics of gut bacteria. Several studies discovered strong relationships between the gut microbiota and human health. Some of these bacteria are shown to be essential for daily life processes like digestion, nutrition uptake, pathogen resistance and immune maturation. Likewise, disturbances of this close relationship, called dysbiosis, have been found to be associated with diseases like diabetes, obesity, colon cancer and inflammatory bowel disease. All this renders the gut microbiome as a highly relevant target of research in medical diagnostics and microbiome community analysis a valid hypothesis building tool.

Nevertheless, the vast amount of different methodologies and lack of broadly accepted standards to create and handle gut microbiome abundance data complicates reproducible or replicable findings across studies. Especially in settings, where samples diverge significantly in their total biomass or microbial load, the analysis of the microbiome is hampered. Several efforts to allow accurate inter sample comparisons have been undertaken, including the use of relative abundances or random feature sub-sampling (rarefaction). While these methodologies are the most frequently used, they are not fully capable to correct for these sample-wide differences. To increase comparability between samples the use of exogenous spike-in bacteria is proposed to correct for sample specific differences in microbial load. The methodology is tested on a dilution experiment with known differences between samples and successfully applied on a clinical microbiome data set. These experiments suggest that current analysis methods lack a pivotal angle on the data, that is comparability between samples differing in microbial load. Meanwhile, the proposed spike-in based calibration to microbial load (SCML) allows for accurate estimation of ratios of absolute endogenous bacteria abundances.

Furthermore, microbiome community analysis is heavily dependent on the resolution of the underlying read count data. While resolutions such as operational taxonomic units (OTUs)

generally overestimate diversity and create highly redundant and sparse data sets, agglomerations to common taxonomy can obfuscate distinct read count patterns of possible sub-populations inside the given taxonomy. Even though the ladder agglomeration strategy might be valid for taxonomy with low phenotypical divergence, plenty taxonomic lineages in fact contain highly diverse sub-species. Thus, a more appropriate taxonomic unit would adapt its resolution for those densely populated branches, allowing for different count resolutions inside the same community. Here the concept of adaptive taxonomic units (ATUs) is introduced and applied on a perturbation experiment including mice receiving antibiotics. For this data set the different classical count resolutions (i.e. collapsed to order, family or genus etc.) produce highly contradictory results.

Meanwhile, adaptive taxonomic units (ATUs) derived by hierarchical affinity merging (HAM) adapt the granularity of taxonomy to the underlying sequencing data. Branches of bacterial phylogeny that are highly covered in the data set receive a higher resolution than those that were infrequently observed. The algorithm hereby merges operational taxonomic units (OTUs) guided not only by sequence dissimilarity, but also by count distribution and OTU size. Due to the agglomeration the number of features is reduced significantly, lowering the complexity of the data, while preserving distributional patterns only observable at OTU level. Consequently, the sparsity of the count data is reduced significantly such that every ATU accumulates reasonable count number and can thus be reliably analysed. The algorithm is provided in the form of the R-Package *dOTUClust*.

Table of contents

List of Abbreviations

A. acidiphilus *Alicyclobacillus acidiphilus* 16, 18

ASCT allogeneic stem cell transplantation 17, *Glossary:* Allogeneic stem cell transplantation

ATU adaptive taxonomic unit xii

GI-GvHD gastrointestinal graft-versus-host disease 17

GvHD graft-versus-host disease 4

HAM hierarchical affinity merging xii, xiv, 41

HMP Human Microbiome Project 2

JSD Jensen-Shannon distance 44

NGS next-generation-sequencing 2, 5, 40

NIH National Institutes of Health 2

OTU operational taxonomic unit xi, xii, *Glossary:* Operational taxonomic unit

R. radiobacter *Rhizobium radiobacter* 16, 18

rRNA ribosomal RNA 5

S. ruber *Salinibacter ruber* 16, 18

SCML spike-in based calibration to microbial load xi, 17

WGSS whole-genome shotgun sequencing 2, 5

List of figures

List of tables

Chapter 1

Introduction

1.1 A brief history of microbiome research

It was when Anthonie van Leeuwenhoek observed the first unicellular microorganism in 1676 that his discovery initiated a long scientific endeavour finally leading to today's microbiology and microbiome research. The scientific community struggled nearly a century long to recreate the precision of Leeuwenhoek's microscopes, until with the invention of achromatic lenses the technology was advanced enough to finally enable a systematic description of the term bacterium by Christian Gottfried Ehrenberg [2]. Soon after, Louis Pasteur described the fermentation process and growth of yeast in 1860 [3]. With the classification of bacteria in four groups by Ferdinand Cohn, bacteriology and parts of today's bacterial taxonomy were founded. While Robert Koch and Pasteur initially focused their research on the role of germs in diseases [4, 5], Pierre-Joseph van Beneden concentrated his research on their positive role. With his studies on Animal Parasites and Messmates he initiated a shift in the perception of bacteria. He was the first in coining the terms commensalism and mutualism amidst the 1870s [6]. From this moment on bacteria were no longer only considered solely interacting with the host as pathogens, but also coexist without creating harm (commensalism) or even offer benefits to their host (mutualism). Around the dawn of the 20th century Ilya Ilyich Mechnikov authored a book on the age prolonging effect of *Lactobacilli*, based on his observations of the prolonged life of Bulgarians as consequence of their yoghurt consume [7]. Even though his hypothesis did not hold, his writings did start an enthusiasm for beneficial pro- and microbiota, which persists until today [8, 9].

Until the late 20th century the identification and characterisation of microbiota was mainly based on bacterial culture. Bacteria which happened to grow in culture, could be identified or

categorized by their phenotypic properties like growth medium, shape, size and biochemistry. Understandably, this approach was not applicable for slow growing or uncultivable bacteria, which was estimated to be true for at least 80% of microbiota found in faecal specimens [10]. In particular for anaerobic bacteria this posed a problem [11]. Additionally, phenotypic methodology was not necessarily decisive for certain bacteria [12]. Especially in the context of medical diagnostics this was of major concern, where the successful identification of a pathogen might decide on the fate of a patient. This changed with the availability of DNA sequencing, when Carl Woese and his colleagues offered a phylogenetic approach to distinguish bacteria based on the genetic sequence of specific genes in the late 1970's [13, 14]. Compared to bacterial culture, this marker gene approaches were culture-independent, allowing even the identification of uncultivable or slowly growing bacteria. On the back of this, the new approach allowed the extension of the classical eukaryote-prokaryote dichotomy, leading to today's system of three domains of life: Archaea, Bacteria and Eukaryotes [15]. This major leaps were soon followed by the first completely sequenced genome of a bacterium, *Haemophilus influenzae*, by Craig J. Venter and fellow researchers [16], which also happened to be the first study to perform whole-genome shotgun sequencing.

When the first results of the Human Genome Project were published in 2001 [17, 18], it became more and more imminent, that some functions, as well as many phenomenons of disease and health, could not solely be explained based on the human genome itself. This paved the way for a paradigm considering the human as a super-organism, consisting not only of its own genome, but a collection of foreign genes and functions offered by a plethora of endogenous microbiota, viruses and fungi. Such collective of micro-organisms and their functions inside a habitat was termed a microbiome [19, 20]. By outnumbering the human in terms of genes by a factor of 100, its role on health and disease in this host-microbe relationship was investigated ever since [11, 21–24]. A new scientific field was born: microbiome research.

In 2004 Venter *et al.* utilized whole-genome shotgun sequencing to analyse the microbial content of seawater samples from the Sargasso sea [25]. This metagenomic study was the first of its kind to analyse a environmental microbiome as a whole. On the basis of this study, further advances in sequencing technology and the emergence of next-generation-sequencing the Human Microbiome Project (HMP) was initiated in 2007, supported by the National Institutes of Health (NIH) Common Fund [26–29]. The aim of this project was to screen and analyse the composition and characteristics of the bacterial communities in all different human body habitats and assess their effects on health and disease of their host. Various large scale studies should follow, notably, Metagenomics of the Human Intestinal Tract (MetaHIT) [30], The Earth Microbiome project (EMP) [31] or The Flemish Gut Flora Project [32]. Each of them contributed to the revelation and characterisation of the relationships between host and microbe, but also among microbes themselves.

1.2 The gut microbiome

In humans, microbiomes are found on the skin and in all sorts of cavities like mouth, nose, lungs, vagina and gut. The latter is called the human gut microbiome and the analysis of its composition, as well as changes within it, will be the focus of this work. Throughout this thesis the term microbiome will be used as a description for the collection and community of bacteria in the gut.

Many studies in recent years reported links between the community composition or functions of the gut microbiome and the health and disease state of the host [22, 33]. In the following two sections examples for these relationships, positive and negative alike, are briefly introduced (see 1.2.1 and 1.2.2).

1.2.1 Beneficial effects of the gut microbiome

In a healthy state, the gut microbiome beneficially impacts our daily life. As soon as neonatal colonization takes place, the gut microbiota start to help in maturation, shaping and training of their hosts adaptive immune system [34–37]. They can also act in pro- and anti-inflammatory manner and directly induce certain immune responses [38]. These relationships, combined with their aid in the defence against pathogen invasion [39–42], highlight their importance for a functioning immune system. The gut microbiome further provides enzymatic reactions which complement our digestion and degradation capabilities of carbohydrates and other compounds in the gut [43–46], while being specialized on the host's diet and lifestyle [47–49]. Besides this, some studies suggest that gut bacteria also participate in the biosynthesis of essential amino acids and omega-6 fatty acids [50], signalling molecules like short chain fatty acids (SCFAs) [51, 52] or impact host biosynthesis [53].

1.2.2 Dysbiosis and disease association of the gut microbiome

All the aforementioned beneficial associations directly imply that a functional and healthy gut microbiome is essential for daily functions and the health of the host. As a consequence perturbations and dysbiosis of the gut microbiome are linked with several diseases and conditions in humans and mice. According to its essential role in digestion and nutrition, an impaired microbiome is highly associated with metabolic disorders like obesity, metabolic syndrome and type 2 diabetes [54, 55]. The microbiome composition and function is shown to be affected by the hosts diet [56], altered in obese compared to lean hosts [57, 58] and additionally contributing to increased energy harvest [59]. Comparably associations between gut microbial composition

and type 1 diabetes, an autoimmune disorder, are recently pointed out by several studies [60, 61].

As residents of the intestine and neighbours of the intestinal barrier, gut microbiota are also found to be involved in the development and severity of inflammatory bowel disease (i.e. ulcerative colitis and Crohn's disease) [62, 63]. Likewise, some pathobionts (i.e. non-pathological bacteria which can turn pathological) and pathogens can operate as risk factors, promoters and modulators of colorectal cancer [64–70].

Furthermore, the ability of the gut microbiota to modulate inflammation also impacts the well being of patients after undergoing allogeneic stem cell transplantation. The absence of certain residential bacteria can aggravate graft-versus-host disease (GvHD) in those patients [71]. Additionally, the use of systemic antibiotics increases the risk of GvHD by disrupting the protective microbiome [71–74].

1.2.3 Implications on medical treatment strategies

All the progress in understanding the interplay between gut microbiota and its host did not only unfold the vast diagnostic potential with regard to several diseases, but also paved the way for alternative treatment strategies targeting the microbiome. One successful example being faecal microbiota transfer (FMT) to treat patients with recurring *Clostridium difficile* infections [75]. This otherwise hard to fight infection is treated by transplanting gut bacteria from a healthy donor in the infected patient. After several repetitions some of the donor's microbiota become residents to the new host. These newly resident or replenished bacteria compete or even directly fight with the pathogen. Application of the same techniques to treat more complex conditions like Crohns disease or other inflammatory bowel diseases showed less optimistic results [76], indicating that the current FMT procedures in that field need further tuning and improvement.

Beyond that, a growing field of research is personalized or precision medicine, where insights on the patients gut microbiome could guide future therapeutic decisions and infer possible therapy outcome individually [77]. Examples for guided interventions could be the administration of precision drugs, pre- or pro-biotics to modulate the susceptibility of a patient towards a specific treatment [66, 74, 78–82]. Recently, some gut bacteria were shown to reduce the effectiveness of medications by its capability in the inactivation and degradation of drugs [83, 84].

In the following, I will give an overview about the rise of community profiling and its application. Further, I introduce state-of-the art protocols for microbiome community analysis and discuss their features and possible shortcomings.

1.3 Microbiome community profiling

Community profiling represents a first step in the investigation of links between the gut microbiome and certain phenotypes, disease and health alike. It is used to assess questions on presence, absence, functions or the dynamics of the gut microbiota by investigating the microbiome composition of a sample [85]. For this purpose for example faecal, luminal or biopsy samples can be screened by utilizing next-generation-sequencing (NGS) to perform either whole-genome shotgun sequencing (WGSS) or targeted amplicon sequencing. The choice of methodology depends on the scientific question. To answer questions on the functional capability of a microbiome, WGSS is used to analyse the gene content of the community residing in the sample [25, 86, 87].

Generally for WGSS, isolated DNA is sheered into fragments of random length, sequenced separately and the resulting fragments are assembled into contigs. These contigs can then be mapped against gene catalogues like KEGG GENES, enabling identification of functions and metabolic pathways involving these genes via KEGG Orthology [88, 89]. Because the contigs arrising from WGSS are longer than usually used marker genes, the resolution for identification of bacteria can be higher. Despite the possibilities of WGSS, confidently mapping from genes to organisms remains a very challenging task, which complicates investigations on composition and diversity based on such data. Additionally, WGSS is the more complex, time consuming and expensive of both methods and hence not available for every research group [85, 90].

A more widely used technique, especially, if the emphasis of the study lies on the microbiomes composition and diversity, is marker gene based amplicon sequencing. Protocols for this methodology involve DNA extraction followed by targeted polymerase chain reaction (PCR) amplification and subsequent sequencing. Dependent on the chosen target marker gene, this approach produces mainly sequences of bacterial origin compared to WGSS, which contains all sorts of meta-genomic content (i.e. DNA of human, viral, fungal and bacterial origin). The most used marker gene for community profiling of the microbiome is the 16S ribosomal RNA (rRNA) gene, a part of the 30S small subunit (SSU) of the prokaryotic ribosomal RNA. This gene has been shown to reliably reconstruct phylogenies and allows to distinguish bacteria from one another [13, 85, 91–93].

Compared to WGSS, 16S rRNA sequencing shows lower computational complexity and therefore is applicable to studies containing more samples (e.g. many patients or longitudinal study designs) [94]. Furthermore, because of superior curated reference databases, lower costs, as well as better detection of rare species 16S rRNA amplicon sequencing still is considered a viable and reasonable approach to profile a microbiome community structure up to genus level [95], especially if the amplicon size approximates full length of the 16S rRNA [96].

Being such a straight forward, fast and established technique, 16S rRNA amplicon sequencing is often used to address initial hypotheses and guide further validation experiments down the road. For example Goodrich *et al.* studied the gut microbiome of twins and unrelated persons for microbial taxa which are dependent on host genetics, when they found that a certain heritable taxon was enriched in individuals with low body mass index [58]. Based on this observation, they set up an experiment where they administered microbiome communities with and without this taxon to mice. Those which carried the bacterium showed reduced weight gain during the experiment (see [58] for details).

In the remainder of this introduction I will focus on community profiling based on 16S rRNA targeted amplicon sequencing and its analysis, as the data used in this thesis is solely based on this methodology.

1.3.1 Multiplexed 16S rRNA targeted amplicon sequencing

For the successful identification of bacteria by utilizing marker genes two things are mandatory: First, one needs to find a region in the genome (a gene or a part of a gene) which is conserved over many species. This means a gene with a conserved and important function which is mandatory for survival of the organism. Second, downstream to this target region there should be more variable regions, which are marked by evolutionary changes. The first part gives you the target position to start amplification. The second the sequence to investigate diversity. Ribosomes are an example of essential mechanisms and their rRNA shows high levels of genetic conservation while also containing some hypervariable regions.

Prokaryotic ribosomes consist of two molecular subunits: the small subunit (SSU) and the large subunit (LSU). Both subunits contain several ribosomal proteins and at least one rRNA species. For the LSU these are the 23S and 5S species and for the SSU it is the 16S rRNA species. The latter is generally not found in eukaryotes (except in mitochondria and chloroplasts). This specificity allows to target mainly for rRNA of bacterial and archaeal origin by targeting the 16S rRNA gene [97]. The roughly 1.540 base pair long 16S rRNA gene contains several conserved and nine hyper-variable regions (V1-V9) [98], allowing for taxonomic differentiation between bacteria [13, 91–93]. Due to its relatively short length (1.5kb) sequencing the 16S rRNA is fast and cheap compared to other marker genes. After two decades of usage as marker gene in phylogeny and microbial ecology, the 16S rRNA gene is also well characterized and many databases containing reference sequences like the Greengenes database [99, 100] or the SILVA ribosomal RNA gene database are available [101].

With 16S targeted amplicon sequencing the hyper-variable regions of the small subunit are selected, amplified and sequenced. Figure 1.1 illustrates the hyper-variable (red) and conserved

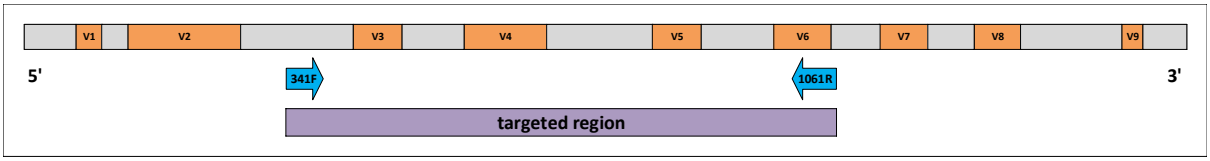

Figure 1.1: Schematic depiction of the distribution of hypervariable (red rectangles) and conserved (grey rectangles) regions on the 16S rRNA gene. Exemplary for amplicon sequencing of a targeted region (purple rectangle), forward and reverse primers are illustrated as blue arrows. This primer combination was chosen for all experiments associated with this thesis.

(grey) regions of the 16S rRNA gene schematically. Selection of specific regions can be controlled by the use of region specific primer pairs (e.g. blue arrows in Figure 1.1), which are chosen in the conserved parts upstream of the region of interest (e.g. purple rectangle in Figure 1.1) [102]. The hyper-variable regions on its own show different performance in the identification of certain bacteria [103]. If possible, read lengths spanning more than one hyper-variable region are chosen to increase the identification accuracy [104–106]. Generally, one to three consecutive variable regions can be selected per study, depending on the maximum read length determined by the sequencing technology. For each sample, a mixture of region specific primers is paired with a unique artificial sequence of length 12, the sample barcode, to be able to trace a sequences sample of origin [107, 108]. This multiplexing approach also allows to sequence multiple samples in the same sequencing run, without loosing information on the origin of a sequence.

Protocols for 16S amplicon sequencing differ depending on the used platform and technology, but follow a general procedure. First PCR amplification in combination with target-specific primers is performed to produce amplicons specific for the selected hyper-variable region (see above). Afterwards, these marker gene amplicons are purified and cleaned, before being finally sequenced. The resulting raw read sequences undergo quality control steps and are transformed to read count data, which lies the basis for studies on the microbiomes diversity and composition. Sequencing technologies to perform 16S amplicon sequencing include platforms by Roche, Illumina, Ion-Torrent (PGM), Oxford Nanopore and Pacific Biosciences (SMRT).

1.3.2 From raw reads to count data - Operational taxonomic units (OTUs)

Many microbiotas of the gut are still unknown or uncultivable, which complicates microbe classification based on phenotypes. Targeted amplicon sequencing of the 16S rRNA gene allows high throughput identification of these bacteria. Nevertheless, sequencing errors introduce non-biologic variance in the sequence, artificially increasing sequence dissimilarity [109]. Counting unique sequences would result in a highly complex count table with a huge number of features, many of which would have just arisen by error. Additionally, microbial ecology

is still missing a defined concept for clear demarcation of bacterial species, complicating the identification and quantification of microbiota in microbiome samples [110–112]. A pragmatic approach to address these issues is to cluster 16S rRNA sequences in groups or phylotypes based on their pairwise sequence similarity. In 2005 Blaxter *et al.* introduced this concept of molecular operational taxonomic units (MOTUs) [113]. These clusters should serve as a proxy for microbial "species", especially in settings where "robust taxonomic hypotheses are difficult to construct" (i.e. biosphere/microbiome) [113]. The naming of these clusters was soon simplified to OTUs.

The demarcation of OTUs generally is controlled by a threshold of percent sequence identity (usually 97%) [114], which allows sequences to cluster together which show divergence of up to 3% sequence identity. Mapping sequences against centroids of the different OTUs then creates count tables which reflect the abundance of each OTU (feature) per sample. Even though the concept of OTUs is criticized for its use of a general genetic similarity cut-off over all species [115, 116], it allows a general approximation on taxonomical composition and diversity and therefore is still the most widely used technique in the field to date. Especially, because of the variability of the 16S rRNA gene even inside the same genome, clustering of read counts seems a mandatory task before community analysis[117].

The last decade gave rise to many tools and algorithms offering clustering of 16S rRNA gene sequences into OTUs [118–120]. The most prominent workflow tools to pick OTUs and analyse the composition and diversity of microbiome samples are QIIME [119] and mothur [118]. While exhibiting smaller differences, both tools follow a general pattern of data processing: (i) quality control and demultiplexing, (ii) OTU demarcation (creation of OTU table), (iii) taxonomic assignment and finally, (iv) compositional and diversity analysis based on resulting OTU tables.

First, all raw sequences are checked for different quality criteria. These criteria can be thresholds for sequence length (minimum and maximum), base quality, maximum number of ambiguous bases, homopolymer length and primer mismatches. With QIIME this step also contains demultiplexing of sequences, which is the grouping of sequences based on their sample of origin by assessing the sample specific barcodes at the 5' end of each read. In the following, the filtered reads are checked for chimeric sequences. Because these hybrid sequences, which consist of two or more parent sequences, artificially skew diversity and species estimations, they are identified upfront and excluded from further analysis [121, 122]. In a next step OTUs are called (i.e. picked) based on a user defined sequence similarity threshold on the quality filtered sequences. There are three strategies to pick OTUs. First de novo OTU picking, which calls OTUs based on a clustering of all input sequences [123–125]. Second reference based OTU picking, guided by a reference database [104], and third a combination of both, called open-reference OTU picking.

While closed reference OTU picking is the fastest of all three, it is also restricted to the databases in use. This means that sequences which do not map against any reference sequence in the database are omitted from further analysis and therefore lost. De novo OTU picking on the other hand is independent of any database and therefore preserves and clusters each input sequence. Simultaneously, it is the more computationally expensive approach and therefore not applicable for larger datasets. Open reference OTU picking uses a reference database and all sequences, which would have been discarded due to not matching a reference in the database, are clustered de novo. This hybrid approach offers a faster execution compared to pure de novo OTU picking, while also retaining all input sequences. Nevertheless, for big datasets this approach can still be slow.

After successful OTU picking, each sequence is assigned the taxonomy of its parent OTU. This step again deviates between the strategies. In de novo OTU clustering for each OTU the most abundant sequence or representative sequence determines the taxonomy of the OTU and all its containing sequences. The closed-reference approach, on the other hand, uses the taxonomy of the already pre-clustered reference OTUs. The hybrid open-reference approach utilizes the principles of both other techniques to assign taxonomy to each OTU depending on its origin (de novo or closed). Finally, OTUs are assigned a taxonomy and a count table can be produced. At this point the user can choose at which resolution he wants to retrieve the count table. The options to choose from are either to get the OTU count table or a collapsed version of it at a user defined taxonomic rank (e.g. family). Even though the clustering into OTUs significantly reduces the number for features, OTU count tables still suffer from high sparsity, which means that many features are rarely observed throughout all samples and many zeros are present in the data. Generally, sparse read counts can occur due to insufficient sequencing depth (i.e. under-sampling), overpopulated reference databases (i.e. multiple hits on different reference genomes), or just because the species truly shows very low abundance [126].

More recently developed approaches count occurrences of unique sequences [127, 128]. This so called exact-sequence-variants (ESV, also referred to as amplicon-sequence-variants) offer a higher resolution than OTUs, but simultaneously suffer from highly increased sparsity and feature complexity.

1.3.3 Normalization and analysis of OTUs

Normalization strategies

To use the generated count tables (OTUs or taxonomy) as basis for further analysis it is necessary to make sure that the counts are comparable between samples. As library sizes (total read

counts per sample) vary, proper normalization is mandatory. Differences in library size can occur either due to technical or biological effects and it is hard to discern how much of the difference can be attributed to which of these sources [129]. These differences in library sizes complicate the identification of potential biomarkers by over- or underestimating effect sizes in differential abundance analysis. One often used approach to make species counts between samples comparable is to transform those into relative abundances. To do so each OTU count in a sample is divided by the total library size (i.e. total read counts) of the respective sample. Relating the abundance of an entity or species on the sample's library size offers a pragmatic approach to allow a certain degree of between sample comparability. However, library sizes are no fixed quantity, but rather a fraction of the originating environment. Hence, the library size of a sample is highly dependent on sequencing efficiency, as well as susceptible for under- and oversampling effects. Simultaneously, the compositional nature of relative abundances cannot capture absolute changes in microbiota abundance, if the microbial load between samples differs. Especially in disease context, patients are often subject of major disturbances (e.g. antibiotics exposure, diet or physical damage to the intestinal barrier) of the intestinal flora, which can result in different microbial load.

An alternative approach is to correct for different read depths by randomly sub-sampling to an even depth across samples (i.e. rarefaction) [130, 131]. For this purpose the smallest library size of all samples is chosen as count boundary for the random sampling. If for example the lowest library size is 5000, then all samples are randomly sub-sampled down to this number. Even though this approach eliminates heteroscedasticity for differing library sizes, it also reduces the information by throwing away features for every sample except one, which especially affects rare species. This again hampers differential abundance analysis by omitting probably important information [132].

Diversity analysis

Following quality filtering and normalization of the count data, ecological diversity of the community is the first object of investigations. Hereto three terms for measuring biodiversity can be examined, which were first described in community ecology by Robert Harding Whittaker in 1972: α-, β- and γ-diversity [133–135].

The first component, α-diversity, informs about the within-sample diversity and is a measure for the richness and evenness of each sample on its own. Richness measures the number of species in an ecological community, whereas evenness describes how homogeneous the abundances of these species are distributed in the community [133, 136]. There are several measures and diversity indices to inspect a samples richness or evenness [134]. The most popular diversity

indices in microbiome ecology to measure α-diversity are the Shannon index [137] and the Simpson index (including its transformations) [138].

The β-diversity on the other hand describes differences in diversity between different habitats or samples (between-sample diversity) [139]. The most common measure for β-diversity is the Bray-Curtis dissimilarity [140], which describes the absolute species overlap between two populations. While Bray-Curtis dissimilarity utilizes overlap, there are measures like UniFrac [141–143], which additionally take sequence similarity between species into account.

In recent years, several studies have linked decreased gut microbiome diversity or dysbiosis with a multitude of diseases and disorders like colorectal cancer [68], ulcerative colitis in children [144], Crohn's disease [145] or myalgic encephalytis/chronic fatigue syndrome [146]. However, because there is no true consensus in methodologies regarding the proper preprocessing and normalization of microbiome community data, these diversity measures can vary greatly between studies. This inter study variation can stem from the OTU-picking algorithm, the count resolution (e.g. 97%, 99% or taxonomic rank), the existence of chimeric sequences and technical noise (i.e. sequencing errors). For example screening the same population once based on 97% sequence identity OTUs and once based on 99% OTUs will produce different estimates of diversity, with the 99% being assessed as more diverse than the 97% OTUs.

Differential abundance and taxonomic biomarker discovery

In microbiome research, investigators are especially interested in the identification of microbiota, which are strongly associated with specific conditions. These key microbiota can be identified by searching for statistically significant differences in species abundance between for example healthy and disease populations or different host phenotypes [147, 148]. Several methods and tools have been proposed to assess differential abundance in microbiome studies, ranging from sample-wise comparisons (MEGAN [149, 150] and STAMP [151]), over simple statistical tests and principal components analysis as in mothur [152], UNIFRAC [141] and MG-RAST [153], up to more sophisticated approaches as Metastats [154], linear discriminant analysis (LDA) effect size (LEfSe) [147] or metagenomeSeq [148].

In general, differential abundance analysis is highly affected by the type of normalization strategy [129, 132].

1.4 Motivation

The ever increasing importance of the microbiome regarding the well being and health of humans makes it a popular investigation target for medical diagnostics and treatment planning. Just as knowledge about the patients microbiome can support the diagnostic process, so can it guide the decisions on possible treatment and medications. This impact makes reproducibility and comparability highly mandatory. Especially diagnostic settings demand proper standardization and normalization of the data. Absence of broadly accepted standards exacerbates this issue in microbiome studies, as highlighted by several critics [155–158]. Such normalization standard would allow to control for sample specific effects like differences in total meta-genomic concentrations or counts, sample preparation, as well as differences in cell lysis or sequencing efficiency.

Furthermore, read count data for microbiome studies tend to be sparse and exhibit many low read count entities (i.e. OTUs) [126]. Additionally, only a fraction of species inside the human gut for example are known, which makes counting read counts a cumbersome task. These issues further impede proper statistical analysis and researchers often reduce these sparse datasets by omitting OTUs with low read counts, based on arbitrarily chosen thresholds. Alternatively, all OTUs are collapsed by their assigned taxonomic rank (e.g. family or genus), reducing sequencing error and complexity. However, both methods tend to discard information which might be important. While the first does this directly, the second omits all entities missing an definite taxonomic assignment at the chosen level.

1.5 Thesis organization

Following this general introduction upcoming chapters of this thesis are organized as follows.

Chapter two (see 2) covers the standardisation of microbiome profiles by the use of exogenous spike-in bacteria and its impact on the interpretation of microbiome community data. This chapter is based on our publication, Staemmler et al. (2016) [1]. First I shine a light on state-of-the art analysis in microbiome research and point out resulting problems with these approaches. A chapter specific methods section follows. Finally, the results of this chapter are illustrated and discussed afterwards.

The optimization of microbiome community analysis with the help of feature binning is handled in chapter three (see 3). Motivation specific for that topic is given, uncertainties in current research are highlighted and concepts for tackling these are introduced. Build upon this, I exhibit

dOTUClust, an R package for feature binning in microbiome data. The rationale, as well as the working scheme of the incorporated algorithm are explained. As a proof-of-principle I apply *dOTUClust* on a small microbiome data set of mice under antibiotic exposure and present how this method strengthens microbiome analysis. Finally, I give a summary of the results of chapter three and point out implications of the findings of chapters two and three for future research in the field of microbiome community analysis in chapter four (see 4).

Chapter 2

Adjusting microbiome profiles for differences in microbial load by spike-in bacteria

This chapter has been adapted from Stämmler et al. [1]:
Stämmler F, Gläsner J, Hiergeist A, et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome*. 2016;4:28. doi:10.1186/s40168-016-0175-0.

All laboratory work for this chapter was performed by Joachim Gläsner and Andreas Hiergeist, with support by Nadja Reul, Claudia Deinzer, Christine Irtenkauf and Holger Melzl at the Institute of Clinical Microbiology and Hygiene at the University Clinic Regensburg. This included animal housing, harvesting, sample preparation, sequencing and quantification of 16S-rDNA copies. Human stool samples for this chapter were collected and provided by Prof. Dr. Holler and Dr. Weber and their staff at the Department of Haematology and Oncology at the University Clinic Regensburg. All experimental procedures were approved by the Ethics Committee of the University Medical Centre of Regensburg.

2.1 Abstract

Background Next-generation 16S ribosomal RNA gene sequencing is widely used to determine the relative composition of the mammalian gut microbiomes. However, in the absence of a reference, this does not reveal alterations in absolute abundance of specific operational taxonomic units if microbial loads vary across specimens.

Results Here we suggest the spiking of exogenous bacteria into crude specimens to quantify ratios of absolute bacterial abundances. We use the 16S rDNA read counts of the spike-in bacteria to adjust the read counts of endogenous bacteria for changes in total microbial loads. Using a series of dilutions of pooled faecal samples from mice containing defined amounts of the spike-in bacteria *Salinibacter ruber*, *Rhizobium radiobacter* and *Alicyclobacillus acidiphilus*, we demonstrate that spike-in-based calibration to microbial loads allows accurate estimation of ratios of absolute endogenous bacteria abundances. Applied to stool specimens of patients undergoing allogeneic stem cell transplantation, we were able to determine changes in both relative and absolute abundances of various phyla, especially the genus *Enterococcus*, in response to antibiotic treatment and radio-chemotherapeutic conditioning.

Conclusion Exogenous spike-in bacteria in gut microbiome studies enable estimation of ratios of absolute OTU abundances, providing novel insights into the structure and the dynamics of intestinal microbiomes.

2.2 Introduction

Current studies on community composition focus on the relative abundance or proportions of OTUs [56, 159]. As an example, a specific OTU may contribute 5 % to microbiome A and 10 % to microbiome B corresponding to a ratio of 1:2. If we further assume that the total number of bacteria or microbial load of A is four times larger than in B, the 5 % in A account for twice as many bacteria as the 10 % in B, thus bringing the actual ratio to 2:1.

Antibiotic treatment, diet, and/or disease affect both microbial loads and compositions. For example, Holler et al. [72] observed that the relative abundance of the genus *Enterococcus* in stool specimens collected from patients undergoing allogeneic stem cell transplantation (ASCT) can increase from undetectable levels prior to ASCT to up to 94 % after ASCT. More interestingly, this relative shift to *Enterococcus* was associated with an increased risk of acute gastrointestinal graft-versus-host disease (GI-GvHD). Without knowledge of total microbial load, however, it is impossible to infer whether this shift was the result of either an absolute increase in the number of *Enterococcus* or a decrease in the number of bacteria other than *Enterococcus*.

Application of synthetic spike-in standards allows for changing the profiles' reference points. The reference point of relative abundances is a fixed aliquot of 16S rDNA. These profiles are insensitive to the microbial load of a stool specimen. Adding controlled amounts of spike-in material allows for rescaling the profiles such that the measured concentrations of the standard are constant across samples, making the spike-in standard the new reference point of the profiles and the profiles sensitive to microbial loads. Spike-in strategies featuring different GC contents and covering a wide concentration range in combination with appropriate normalization strategies have already been proposed to correct for library preparation and nuisance technical effects in the inference of gene expression levels from RNA-Seq experiments [160]. This approach, as well as similar schemes employed in proteomics [161] and metabolomics [162], adds the spike-in standards to transcriptomes, proteomes and metabolomes only after cell lysis and extraction of mRNA, proteins and metabolites, respectively, and thus do not allow correction of variation originating from these critical experimental steps. Recently Jones et al. [157] suggested using whole cell spike-in controls for monitoring this technical variability in the field of microbiome research.

Extending their results, we here suggest the addition of exogenous viable spike-in bacteria to rescale the read counts of endogenous bacteria. We call this protocol spike-in based calibration to microbial load (SCML), and test it in a dilution experiment with defined absolute spike-in bacteria abundances against serially diluted background microbiomes. Moreover, we reconsider the emergence of *Enterococcus* as the predominant genus in ASCT using SCML.

2.3 Chapter Methods

2.3.1 Spike-in bacteria

In this study we used *Salinibacter ruber* (*S. ruber*) DSM 13855 T , an extreme halophilic bacterium found in hypersaline environments [163], *Rhizobium radiobacter* (*R. radiobacter*) DSM 30147 T, a non-phytopathogenic member of the Biovar I group of *Agrobacterium* found in soil and the plant rhizosphere [164], as well as the thermo-acidophilic, endospore forming soil bacterium *Alicyclobacillus acidiphilus* (*A. acidiphilus*) DSM 14558 T [165]. All bacteria were purchased from the DSMZ (German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany). These eubacteria belong to different phyla typically found in mammalian faecal microbiomes, contributing to *Bacteroidetes/Chlorobi* group, *Proteobacteria*, and *Firmicutes*, respectively. They do not exist in the gut microbiome under physiological conditions and are well distinguishable from bacteria commonly found in the gut using 16S rRNA gene sequencing. *S. ruber* and *R. radiobacter* are gram-negative bacteria, whereas *A. acidiphilus* is a spore-forming gram-positive bacterium. The difference in the chemical constitution of the cell wall accounts for a specific susceptibility to the cell lysis protocol used. Spike-in bacteria were harvested in the late logarithmic/early stationary growth phase by centrifugation and subsequently resuspended in 5 ml of sterile PBS buffer. Bacterial densities in suspensions were quantified by OD600 measurement using empirical conversion factors determined by direct microscopic cell counting. Accordingly, 1 OD600 unit corresponds to $4.6x10^9$ cells/ml for *S. ruber*, $1.4x10^9$ cells/ml for *R. radiobacter*, and $1.2x10^9$ cells/ml for *A. acidiphilus*, respectively. 16S rRNA gene copy numbers per genome for the spike-in bacteria were obtained from the rrnDB database [166]. Six pools of bacterial mock communities containing *S. ruber*, *R. radiobacter* and *A. acidiphilus* were generated according to the scheme provided in supplementary table A3.

2.3.2 Sample preparation and DNA extraction

Mouse specimens

For the validation experiment, cecum contents were collected from three 12-week-old male C57BL/6J mice (200 mg wet weight each), immediately suspended into 1 ml of PBS, homogenized by means of the TissueLyser II (QIAGEN, Hilden, Germany), pooled, adjusted with PBS to a total volume of 4 ml, and split into seven aliquots of 550 μl each. Six of these aliquots were diluted five times according to the scheme provided in supplementary table A2. Aliquot 7 was used as a non-spike control. Sixty microliters of the corresponding spike-bacteria pool (whole

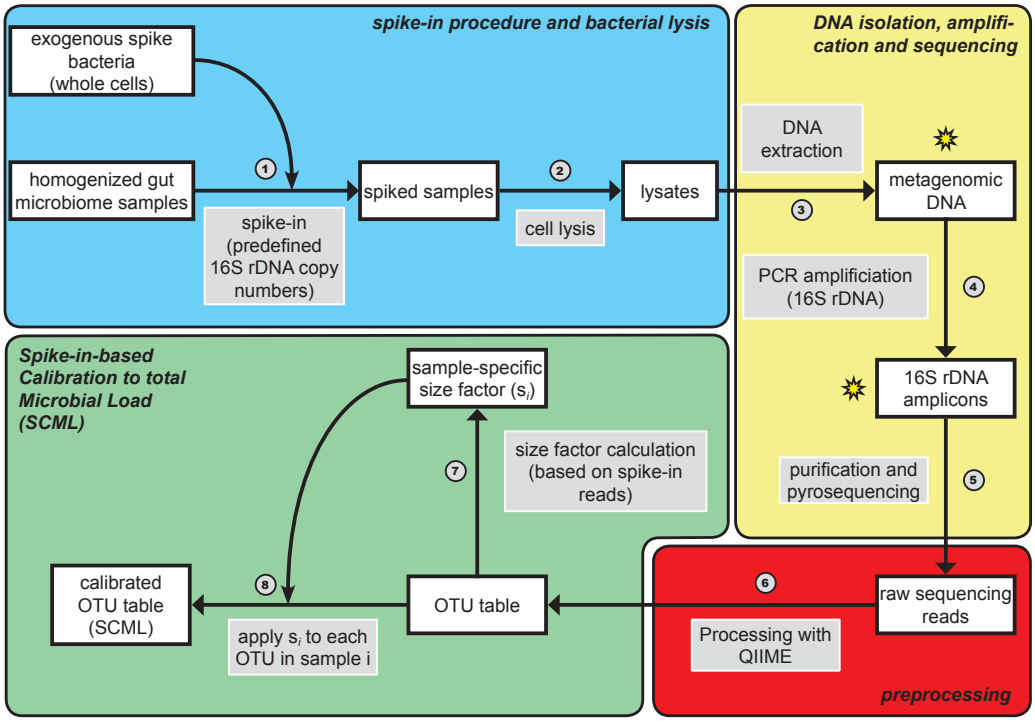

Figure 2.1: Procedural overview of proposed spike-in procedure and the spike-in-based calibration to total microbial load (SCML). The overview is divided into four sections: spike-in procedure and bacterial lysis (blue), DNA isolation, amplification and sequencing (yellow), pre-processing (red) and the actual spike-in-based calibration to microbial load (green). White-filled boxes depict procedural intermediates, while grey-filled boxes depict the different procedural steps. Each step is numbered. In the first step **(1)** whole cells of exogenous spike bacteria corresponding to a fixed number of 16S rDNA copies are added to homogenized microbiome samples. Bacterial lysis is performed on the resulting spiked samples **(2)**. Metagenomic DNA is extracted from the lysates **(3)** and PCR amplified using 16S rDNA specific primers **(4)**, creating 16S rDNA amplicons. These amplicons are purified and pyrosequencing is performed **(5)**. The resulting raw read counts are pre-processed with QIIME (quality filtering, demultiplexing and closed reference OTU picking) to generate OTU read count tables **(6)**. Based on the read counts associated with single or multiple reference spike-in bacteria, a size factor si for each sample i is calculated and applied to each OTU of this particular sample i (**8**, see methods section). This leads to an OTU read count table calibrated to differences in microbial load. These read counts can be utilized to more accurately assess changes between different samples. All depicted steps are described in detail in the methods section. Stars indicate points in the procedure at which qPCR is performed to identify possible errors in DNA isolation (metagenomic DNA) or PCR amplification (16S rDNA amplicons).

cells) containing the desired number of 16S rDNA copies (see supplementary table A3) were added to 250 μl of all prepared, unlysed stool dilutions (see Fig. 2.1, step 1) according to the scheme provided in supplementary table A2. Then, 180 μl of Bacterial Lysis Buffer (Roche, Mannheim, Germany) and 20 μl Proteinase K (Fermentas GmbH, Sankt Leon-Rot, Germany) were added. Samples were incubated at 65 °C for 10 min followed by five cycles of freezing in liquid nitrogen (1 min) and boiling in hot water (95 °C, 1 min). Following the addition of 400 μl of Bacterial Lysis Buffer and a mixture of 0.1-mm and 2.5-mm beads, samples were treated for 2 min at 30 Hz in the TissueLyser II. Subsequently, samples were heated at 95 °C for 15 min and centrifuged at 4 °C to pellet stool particles and beads (see Fig. 2.1, step 2). The final volume was adjusted to 1 ml and DNA was extracted (see Fig. 2.1, step 3) by means of the MagNA Pure 96 instrument employing the MagNA Pure 96 DNA and Viral NA Large Volume Kit (Roche). Nucleic acids were quantified using the NanoDrop ND-1000 (Thermo Scientific, Wilmington, DE, USA).

Human ASCT specimens

With approval of the Ethics Committee of the University Medical Centre of Regensburg and after receipt of signed informed consent forms, stool specimens were collected at four different time points: prior to administration of prophylactic antibiotics and radio-chemotherapeutic conditioning, on days 0, 7, and 14, respectively, after ASCT. Stool specimens were stored at -80 °C until analysis. Fifty mg (wet weight) of each stool specimen were suspended into 250 μl PBS and subsequently subjected to DNA extraction as described above. Spiking of *A. acidiphilus*, *S. ruber* and *R. radiobacter*, and 454-pyrosequencing were performed according to the validation protocol described above. For these experiments, bacterial cells of *S. ruber*, *R. radiobacter* and *A. acidiphilus* equal to $3.0x10^8$, $5.0x10^8$, $1.0x10^8$ 16S rDNA copies, respectively, were spiked into each crude sample.

2.3.3 Amplification of V3-V6 16S rDNA variable region and 454 pyrosequencing

Spike bacteria-specific qPCR was performed for all specimens (mice and human) to identify errors in DNA isolation before undergoing amplification and pyrosequencing (see Fig. 2.1). A total of 25 ng metagenomic DNA was used as a template to amplify the V3-V6 variable regions of the 16S rRNA gene. PCR was performed using primer pair 341 F-1061R containing Lib-L adaptors and Roche standard multiplex identifiers (MIDs) in a final volume of 40 µl containing 0.088 μM of each primer, 2 mM MgCl2, and 1 U Platinum Taq DNA Polymerase

(Life Technologies). The PCR amplification (see Fig. 2.1, step 4) was carried out over 30 cycles (30s at 95 °C, 45 s at 64 °C, 45 s at 72 °C) with an initial 5-min hot start at 95 °C and a final extension step (7 min at 72 °C). The resulting 790-bp amplicons were analysed by standard agarose gel electrophoresis on a 1.5 % (w/v) gel. The amplicons were extracted from agarose gels using the QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany) and purified with Agencourt AMPure XP beads (Beckman Coulter, Krefeld, Germany). Copy numbers of amplicons containing LibL-adaptors were determined using the KAPA Library Quant 454 Titanium/Lib-L Universal Kit (KAPA Biosystems, Wilmington, DE, USA) and pooled to a normalized library with a concentration of $1x10^7$ adaptor-labeled amplicon molecules/μl for each sample. This library was subjected to sequencing (see Fig. 2.1, step 5) using the GS FLX+ system (454/Roche) and the GS FLX Titanium LV emPCR Kit (Lib-L) applying 0.4 copies per bead. Sequencing was performed on a full PTP according to manufacturer's protocol using the GS-FLX Titanium Sequencing Kit XL+ and the acyclic flow pattern B. Sequencing raw data was processed with gsRunProcessor v2.9 (Roche) using quality filtering as defined by the default LongAmplicons 3 pipeline resulting in 895 Mb from 1,313,653 passed filter wells with a median read length of 706 bases.

2.3.4 Quantification of 16S rRNA gene copy number by qRT-PCR

Primer design and validation

Primers and probes for quantification of eubacterial 16S rDNA copies (supplementary table A4) were designed and evaluated in silico based on the RefNR sequence collection of the SILVA reference database release 119 [101] containing 534,968 16S rRNA sequences. The overall SILVA database coverage of universal 16S rDNA quantification primers 764 F and 907R allowing one primer mismatch was 86 %. Allowing no primer mismatches, specificity of primers and probes targeting , *R. radiobacter* and *A. acidiphilus* DNA exhibited specificities of 100 %. Specificity of primers and probes were further evaluated in silico using the blastn algorithm against the nucleotide collection (nt) database. Concentration of primers were optimized by titration in the range of the kit manufacturer's recommendations after PCR amplification of 16S rDNA targets from DNA extracts of human and murine faecal specimens. Samples were spiked prior to DNA extraction with defined cell counts of *S. ruber*, *R. radiobacter* and *A. acidiphilus*, which were quantified microscopically using a modified Neubauer counting chamber. PCR products were screened for nonspecific bands by agarose gel electrophoresis (probe based assays) or agarose gel and melting curve analysis (SYBR Green I based assays). Specificity was further evaluated by quantitative real-time PCR amplification of total 16S rDNA and 16S rDNA of spike-in bacteria from ten non-spiked murine and human DNA extracts.

Quantification of total 16S rDNA

To verify the experimental design, 16S rRNA gene copies of total and spike-in bacteria were determined by qRT-PCR on a LightCycler 480 II Instrument (Roche). Primers and probes used are shown in supplementary table A4. PCR reactions included 1 μM each of eubacterial 16S rRNA gene primers 764 F and 907R (quantification primers) and the LightCycler 480 SYBR Green I Master Kit (Roche). Quantification standards were generated by cloning complex PCR amplicon mixtures that were generated from a caecal microbiome DNA preparation of wild type C57BL/6J mice (using primers 341 F and 1061R) into the pGEM-T.Easy vector (Promega, Madison, WI, USA). Cloning of PCR amplicon mixtures was carried out to mimic a complex murine microbiota with respect to qPCR amplification efficiency in analyzed samples as far as possible. Quantification PCR was conducted over 40 cycles (95 °C for 10s, 60 °C for 15 s and 72 °C for 15 s) with an initial 10-min hot start at 95 °C.

Quantification of 16S rDNA of spike-in bacteria

16S rRNA gene copy numbers of the spike-in bacteria *S. ruber*, *R. radiobacter* and *A. acidiphilus* were determined with 16S rDNA-targeted species-specific primers and hydrolysis probes (see supplementary table A4). Quantification PCR was conducted using the LightCycler 480 Probes Master kit (Roche) in a 20-μl reaction volume containing 4 mM MgCl2, 0.25 μM of each primer, and 0.1 μM probes. Quantification standards were constructed by cloning full length 16S PCR amplicons of all spike-in bacteria (amplified using 27 F and 1492R primers) into pGEM-T.Easy. Quantification PCR was conducted over 40 cycles (95 °C for 30s, 60 °C for 30 s and 72 °C for 30s) with an initial 10-min hot start at 95 °C.

2.3.5 Computational analysis

We used a combination of QIIME [119] (v1.8.0) and R version 3.2.0 [167] with installed Bioconductor package [168] to process the read data. Reads were filtered for quality using QIIME's split_libraries.py script (see Fig. 2.1, step 6) with default parameters except minimum and maximum read length, which were set to 400 bp and 800 bp, respectively. This read length threshold covered 99.99 % of all sequencing reads. The filtered reads were mapped to OTUs built on the SILVA [101] database (release 111) using QIIME's pick_closed_reference_otus.py script (see Fig. 2.1, step 6) with default parameters. The reference database OTUs used here constituted computationally built clusters of the SILVA SSU (small subunit) ribosomal RNA database. The clustering (see Fig. 2.1, step 6) was achieved by UCLUST 1.2.20 [169] and provided by the

QIIME team (available at http://qiime.org/home_static/dataFiles.html). Since reads from the three spike-in bacteria mapped to multiple OTUs, due to multiple reference OTUs encoding for the same spike-in genus, we deleted all but one OTU encoding for each spike-in from the database before mapping, to accumulate all reads from the spike-in to just this one OTU. The used reference sequences for these three OTUs are available in Additional file 5. Raw sequencing data of the dilution experiment is deposited in the European Nucleotide Archive (ENA) under the study accession number PRJEB11953, at http://www.ebi.ac.uk/ena/data/view/PRJEB11953. Details of the sample design are shown in supplementary table A2. Relative abundances were calculated by dividing each OTU read count by total read count of the corresponding sample.

Ratios of absolute abundances were calculated by using the expectation that the counts of reference spike-ins are inversely correlated to total microbial load of the samples under investigation. Let $\bar{s}$ be the mean read count of the reference spike-in *S. ruber* over all samples (see Fig. 2.1, step 7). The read count of every OTU in a samplei is rescaled by a factor si that is calibrated such that the spike-in count is equal to $\bar{s}$ in every sample (see Fig. 2.1, step 8). SCML can be performed by the use of an individual spike-in bacterium or the sum of all reads obtained for multiple spike-in bacteria. For further analysis, the counts are $\log_2$ transformed.

To compare ratios derived from relative abundances and those derived by SCML, we calculate $\log_2$ ratios between every pair of samples for each method as a symmetrical measure of difference. Ratios of relative abundances are calculated by dividing the relative abundances of each OTU by its relative abundance in the compared sample, whereas ratios for SCML are calculated by means of the spike-in calibrated read counts (SCML data). If for example OTU A shows relative abundances of 20 % and 40 % in samples 1 and 2, respectively, the corresponding ratio for this comparison would be $\frac{0.4}{0.2} = 2$, i.e. the abundance of OTU A in sample 2 is two times higher than in sample 1. The corresponding $\log_2$ ratio would be $log_2(2) = 1$. Both ratios are calculated separately for each OTU.

For the combination approach of SCML, the read counts of *A. acidiphilus* and *R. radiobacter* were adjusted by their difference in the predefined spike-in concentration (supplementary table A2) towards *S. ruber*. If for example *A. acidiphilus* was added by design in half the concentration compared to *S. ruber*, then all reads by *A. acidiphilus* were multiplied by two. The adjusted read counts of *A. acidiphilus*, *R. radiobacter* and the raw read counts of *S. ruber* were summed up to one artificial entity. These summed reads were used in the same fashion as the *S. ruber* read counts in the single spike-in calculation. For the dilution experiment the adjustment of *A. acidiphilus* and *R. radiobacter* read counts was necessary, because both spike-ins were added in varying amounts in this experiment. In an application of our spike-in procedure (e.g. ASCT

specimens in this study) all spike-in bacteria cells are added at fixed amounts. Therefore, an adjustment of the spike-in read counts before the combination would be obsolete.

2.4 Chapter Results

2.4.1 Spike-in bacteria yield different read turnouts but correlate well with microbial loads

Figure 2.2a shows linear relationships between the spiked-in 16S rDNA copies (x-axis in $\log_2$ scale) of *A. acidiphilus* and *R. radiobacter*, respectively, and the resulting $\log_2$ read counts. The total number of spike-in reads increases with dilution of the background microbiome. Simultaneously, as a constant amount of *S. ruber* was added to each sample, the portion of the spike-in bacteria increases (Fig. 2.2b). As a result, the read count assigned to a spike-in OTU is expected to inversely correlate with the total microbial load.

Figure 2.2b shows box plots of the $\log_2$ transformed read counts of *S. ruber*, *R. radiobacter* and *A. acidiphilus* as a function of microbial loads across all 36 samples. The counts were adjusted for their varying spike-in concentrations by design. For example, if in an experiment the concentration of the *A. acidiphilus* spike-in was only 50 % of that of *S. ruber*, the *A. acidiphilus* counts were doubled. After adjustment of *A. acidiphilus* and *R. radiobacter*, we observe an inverse correlation of $\log_2$ spike-in counts with the microbial load (reciprocal dilution factor) for all three spike-in bacteria (Fig. 2.2b). In detail there is a correlation of $r = -0.834$ for *S. ruber*, $r = -0.795$ for *R. radiobacter* (adjusted) and $r = -0.725$ for *A. acidiphilus* (adjusted). Additionally, we observe that the three bacteria have notably different read yields, with *S. ruber* showing the highest counts.

2.4.2 SCML yields almost unbiased estimates of ratios of absolute abundances within taxonomic units

For comparing SCML to standard relative abundance analysis, we generated two data sets by scaling the read counts with respect to two different reference points: First, we scaled the observed read counts relative to the library sizes. This gives us the standard relative abundances (standard data). In a second data set we scaled the same counts relative to the spike-in reads of *S. ruber* (SCML data). We first compared the data for *A. acidiphilus* and *R. radiobacter* separately. By design the expected ratio for *A. acidiphilus* and *R. radiobacter* between every pair of samples is known. Figure 2.3 shows the observed inter-sample ratios for both data sets as a function of

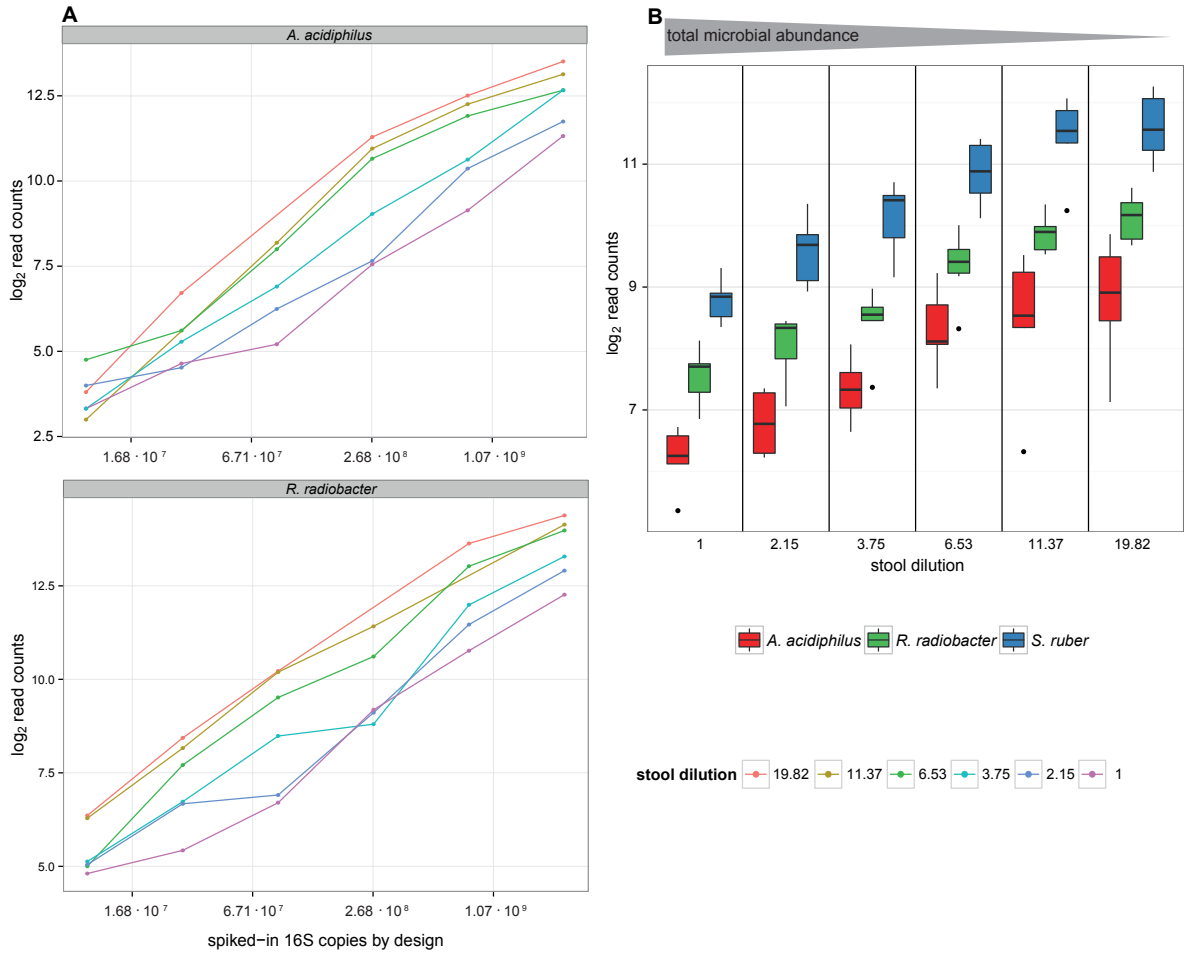

Figure 2.2: Log_2 transformed read counts of the three spike-in bacteria as a function of total microbial load. *S. ruber* was added at a constant number of 16S rDNA copies, while *A. acidiphilus* and *R. radiobacter* were spiked in variably (cf. supplementary table A2). **(a)** Resulting read counts of *A. acidiphilus* and *R. radiobacter* versus spiked-in 16S rDNA copies at different background stool microbiota dilutions. Each dot represents a caecal specimen, while the colour specifies its dilution. **(b)** Boxplots showing the read counts of all three spike-in bacteria as a function of total microbial load. The $\log_2$ read counts of *S. ruber* are coloured blue, while *A. acidiphilus* and *R. radiobacter* are coloured red and green, respectively. Read counts of *A. acidiphilus* and *R. radiobacter* were adjusted by a factor corresponding to their difference of the predefined spike-in concentration to *S. ruber*. The x-axis is discrete and represents increasing stool dilution (bottom), as well as decreasing microbial load from left to right (grey arrowhead on the top).

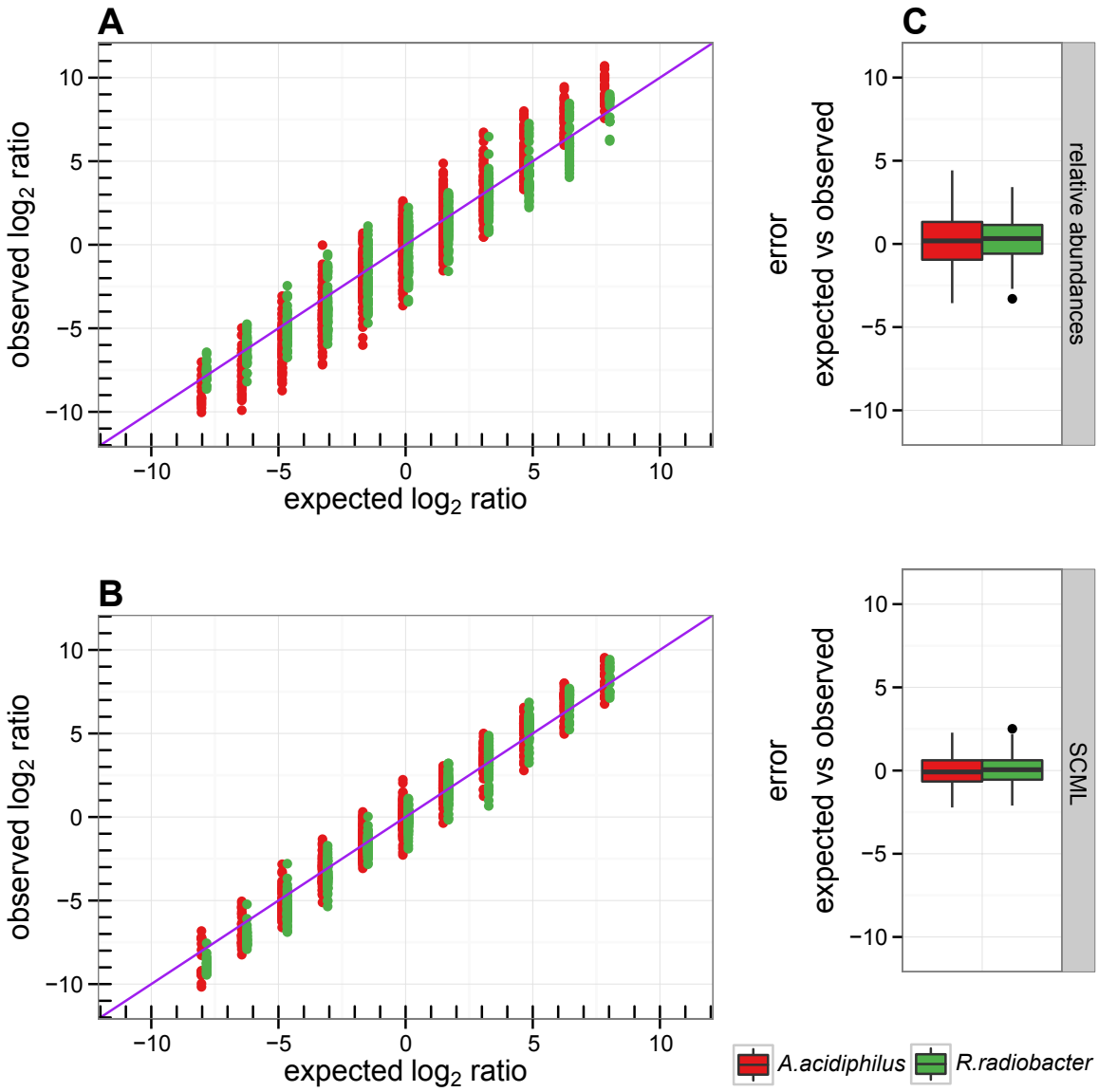

Figure 2.3: Comparison of $\log_2$ ratios derived from relative abundances and after applying SCML to *A. acidiphilus* and *R. radiobacter*. Observed $\log_2$ ratios versus expected $\log_2$ ratios of the spike-ins *A. acidiphilus* and *R. radiobacter* as derived from **(a)** relative abundances and **(b)** SCML by *S. ruber* for all pairwise sample comparisons. Both approaches were performed on the raw, not adjusted read counts of *A. acidiphilus* and *R. radiobacter*. The expected $\log_2$ ratios are calculated by the theoretical number of 16S rDNA copies predetermined in the design of the validation experiment (cf. supplementary table A2). The purple diagonal represents the identity, which represents the expected $\log_2$ ratios by design. The box plots in **(c)** show the error between the expected and observed $\log_2$ ratios for both approaches. The smaller this error, the better calibrated the ratios are.

expected ratios. Plot (a) was created using standard data, while plot (b) was created using SCML data. We observe a reduced systematic error in (b) when comparing the data trend to the identity line (purple). The standard data shows systematically overestimated ratios in both directions. SCML reduced this bias. Moreover, we observe a high variability of estimated ratios, which was almost cut in half by SCML (Fig. 2.3c). We next analysed the ratios for the background OTUs. By design, experimentally controlled ratios can be calculated from the dilution factor of the background microbiome. In contrast to *A. acidiphilus* and *R. radiobacter* the ratios derived from relative abundances (standard data) of these OTUs is zero by experimental design. Figure 2.4 shows the distribution of observed background ratios as a function of corresponding expected ratios. Plot (a) was created using standard data, while plot (b) was created using SCML data. As expected, ratios of relative abundances cannot capture shifts in microbial loads that do not affect the composition (Fig. 2.4a). In line with the previous observations, we observe a reduction of estimation variance when using SCML (Fig. 2.4c). Correlations between expected and observed $\log_2$ ratios were 0.359 and 0.833 for the standard data and the SCML data, respectively.

2.4.3 SCML allows more accurate estimation of ratios than calibrating for total 16S rRNA gene copies using qRT-PCR

Quantification of the total number of 16S rRNA gene copies by qRT-PCR may be used to determine microbial loads. To compare the practicability of the latter with SCML we used a SYBR Green-based qPCR assay to quantify 16S rDNA (supplementary table A1). Figure 2.5 shows observed and expected $\log_2$ ratios for background OTUs using either (a) SCML or (b) rescaling to constant total 16S rRNA gene copies. It is apparent that observed ratios derived from SCML show higher concordance with the expected ratios regarding estimation bias and variance. Correlations between expected and observed $\log_2$ ratios were 0.717 and 0.833 for the qPCR and the SCML approach, respectively. These findings are also supported by an overall lower error between the observed and expected $\log_2$ ratios when derived from SCML (Fig. 2.5c). However, it has to be acknowledged that the SYBR Green-based quantification method of the bacterial load has not been explicitly compared to probe-based formats, so any limitations/imprecisions possibly resulting from the use of this universal detection format were not taken into account.

2.4.4 Combining multiple spike-in bacteria reduces estimation errors

Figure 2.2b shows that the adjusted counts of all three spike-in bacteria reciprocally correlated with microbial loads. We next investigated whether taking the sum of all three spike-in read

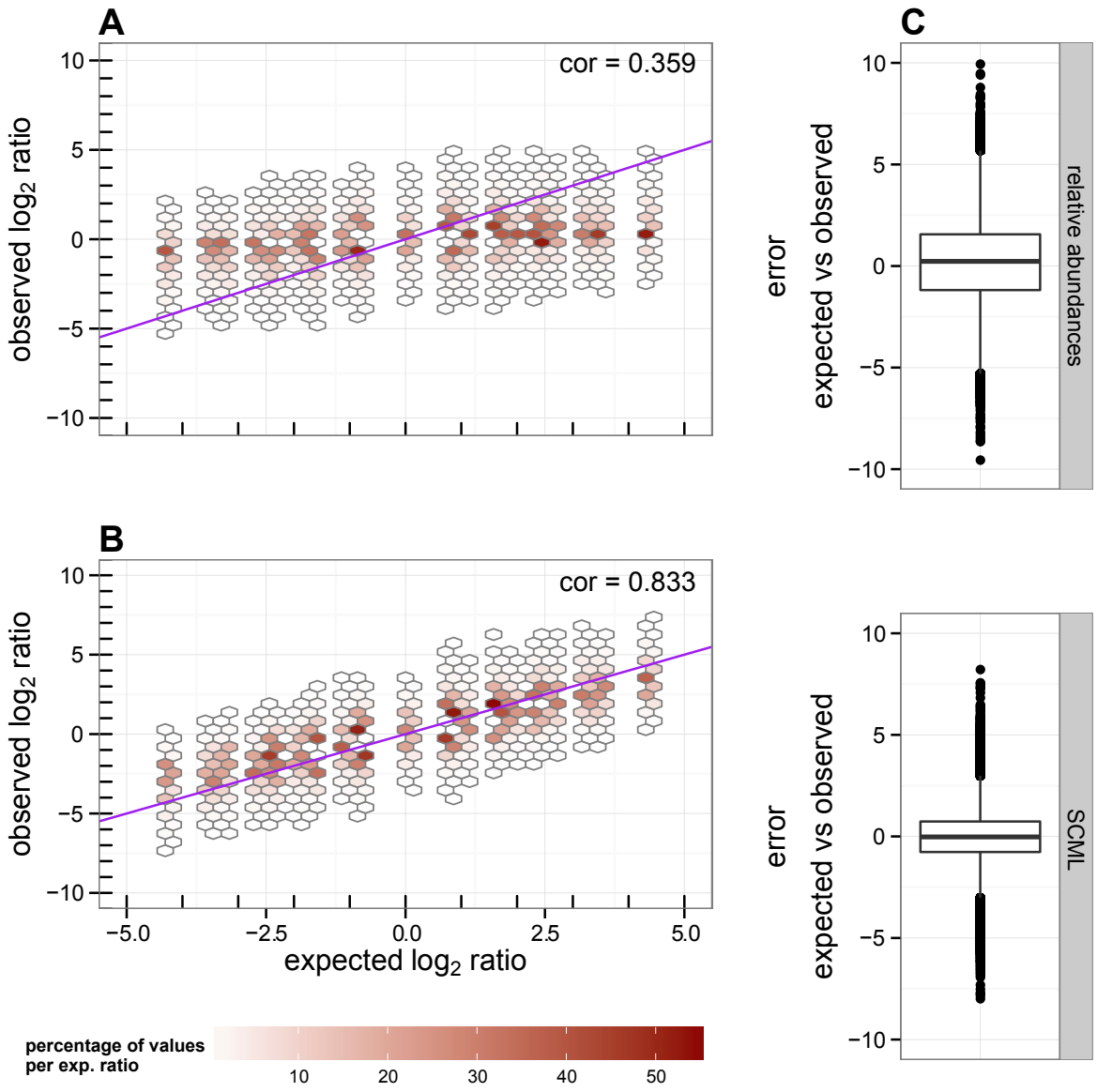

Figure 2.4: Comparison of $\log_2$ ratios derived from relative abundances and after applying SCML to all background OTUs. Observed $\log_2$ ratio versus expected $\log_2$ ratio of all background OTUs for all pairwise comparisons as derived from **(a)** relative abundances and **(b)** SCML by *S. ruber*. The data is binned to hexagons because of the high number of data points. The colour of each hexagon represents the percentage of counts at the corresponding level of expected $\log_2$ ratios contained in each bin. Bins that contributed to <0.05 % for each level of expected $\log_2$ ratio are omitted. The purple diagonal represents the identity, which represents the expected $\log_2$ ratios by design. The box-plots in **(c)** show the error between the expected and observed $\log_2$ ratios for both approaches. The smaller this error, the better calibrated the ratios are.

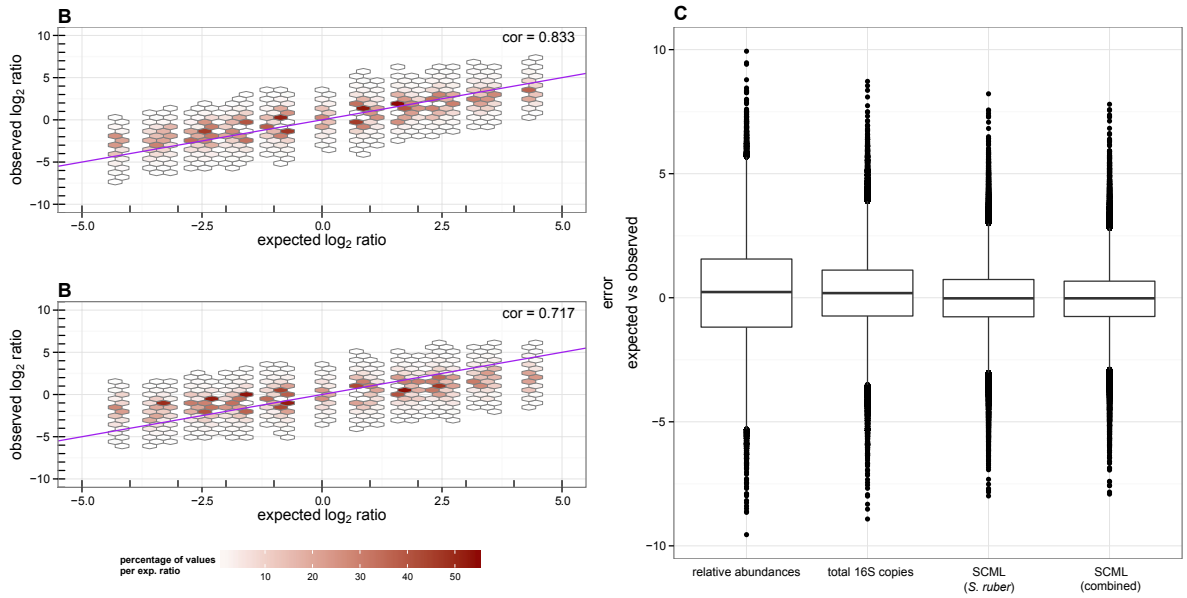

Figure 2.5: Comparison of SCML and normalization by qRT-PCR-derived total number of 16S rDNA copies to all background OTUs. Observed $\log_2$ ratio versus expected $\log_2$ ratio of all background bacteria OTUs for all pairwise sample comparisons after **(a)** SCML by *S. ruber* and **(b)** normalization by qRT-PCR derived total 16S rDNA copy number. The data is binned to hexagons because of the high number of data points. The colour of each hexagon represents the percentage of all counts at the corresponding level of expected $\log_2$ ratios contained in each bin. Bins that contributed to less than 0.05 percent for each level of expected $\log_2$ ratio are omitted. The purple diagonal represents the identity, which represents the expected $\log_2$ ratios by design. The box-plots in **(c)** summarize the error between the expected and observed $\log_2$ ratios for the four different approaches. The smaller this error, the better calibrated the ratios are. Variances of the $\log_2$ differences are 3.65, 2.01, 1.28 and 1.18 as derived from relative abundances, counts calibrated for differences in total number of 16S rRNA gene copies, SCML (by *S. ruber*) and combined SCML (by *S. ruber*, *A. acidiphilus* and *R. radiobacter*), respectively.

counts further improves the estimates. Since *A. acidiphilus* and *R. radiobacter* were spiked in variable amounts we had to adjust their counts prior to using them for calibration. For example, if in an experiment the concentration of the *A. acidiphilus* spike-in was only 50 % of that of *S. ruber*, the *A. acidiphilus* counts were doubled. We then used the sum of adjusted counts of all three spike-ins for calibration and repeated the analysis of the previous section. Figure 2.5c shows box-plots of the error between expected and observed $\log_2$ ratios for background OTUs based on relative abundances, read counts normalized by total 16S rDNA copies, as well as based on the SCML data with *S. ruber* only and the combined counts of all three spike-ins, respectively. The smaller this error, the better calibrated are the ratios of absolute abundances. Variances of these errors are 3.65, 2.01, 1.28 and 1.18, respectively. Thus, combined spike-ins yield a slightly increased precision compared to single spike-in usage. Correlations between expected and observed $\log_2$ ratios were 0.833 and 0.845 for the SCML and the combined SCML approach, respectively.

2.4.5 Calibration to microbial loads reveals absolute increase of *Enterococcus* in the intestine during allogeneic stem cell transplantation

Finally, we show that SCML expands our understanding of human microbiomes and their role in disease. Recently, a marked early loss of gastrointestinal microbiome diversity and an increase in relative abundance of members of the genus *Enterococcus* have been observed in the course of ASCT and found to increase the risk of developing acute GI-GvHD [71–73]. Since the data had been generated without spike-in bacteria, it had not been possible to conclude whether the observed increase in relative abundance of *Enterococcus* was the result of an increase in absolute abundance of *Enterococcus* or of a decrease in abundance of other bacterial species. Here we report on five patients, whose stool microbiomes were monitored prior to ASCT or on days 0 (d0), 7 (d7), and 14 (d14) after ASCT, respectively, using the proposed spike-in approach. Figure 2.6a shows the familiar diagram of relative microbiome composition without taking the spike-in bacteria into consideration. Reads contributing to the genus of *Enterococcus* are reported at genus resolution, while all other bacteria are shown on phyla resolution. In line with Holler et al. [72], we observe dramatic relative increases in *Enterococcus* abundance on days 7 and 14 after ASCT in three of the five patients. By scaling read counts to an even microbial load using the *S. ruber* counts, we observe marked changes in the microbial loads in the course of the treatment (Fig. 2.6b). Patient 5, for instance, shows an almost tenfold reduction of microbial load on day 14 after ASCT (*S. ruber* reads 4721) compared to pre-ASCT (*S. ruber* reads 515). In our study, specimens dominated by *Enterococcus* generally have low microbial loads (Fig. 2.6b). We also observe an absolute increase in abundance of the genus *Enterococcus* in these

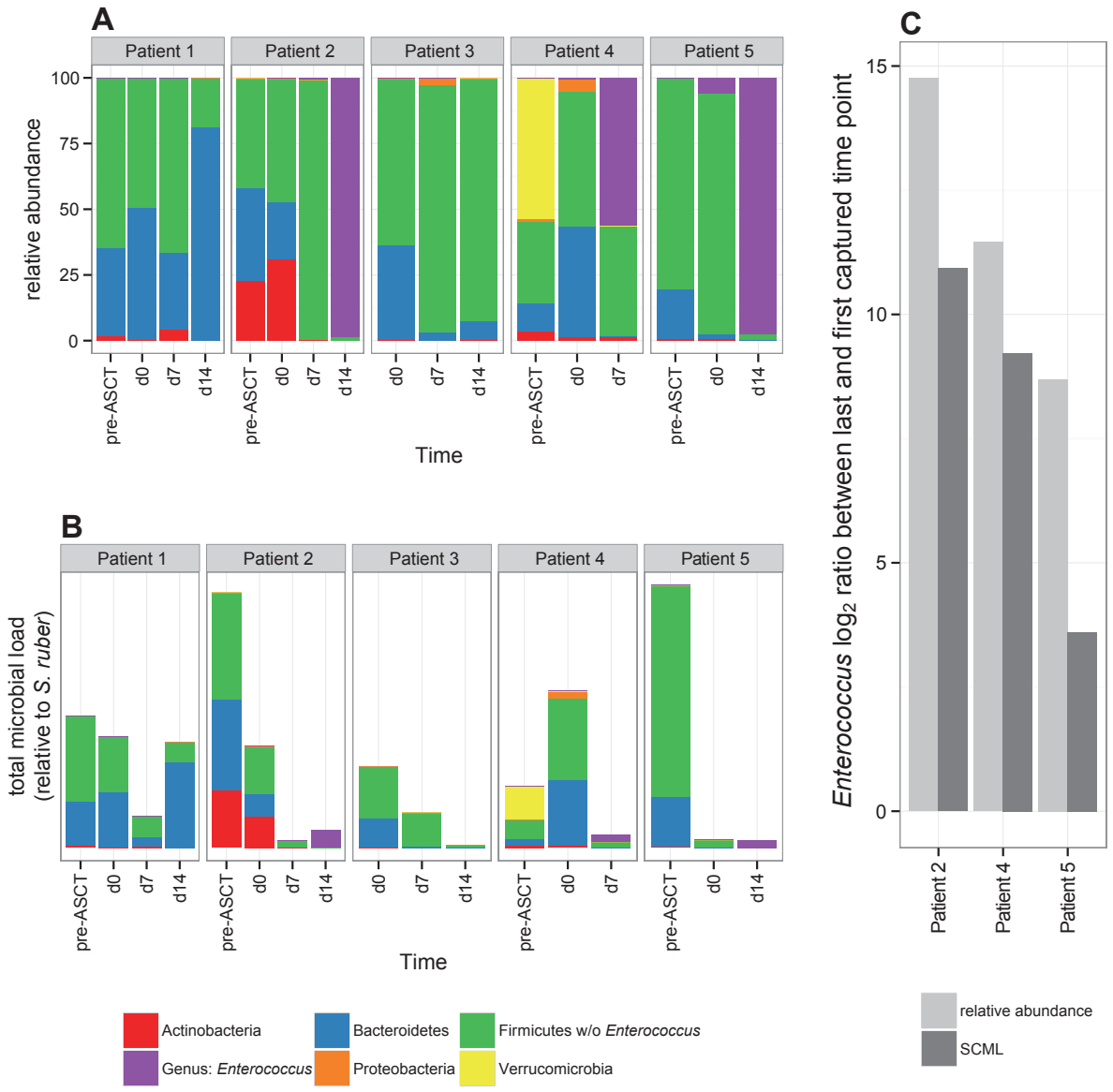

Figure 2.6: Bacterial abundances in stool specimens of ASCT patients. Specimens were collected prior to administration of prophylactic antibiotics and radio-chemotherapeutic conditioning (pre-ASCT) and on days 0, 7 and 14 after ASCT (d0, d7, d14). **(a)** Microbial composition given as in relative abundances; **(b)** read counts scaled to a uniform count of the spike-in S. ruber and **(c)** $\log_2$ ratios of *Enterococcus* of the last time point to pre-ASCT of patients 2, 4 and 5 as derived from relative abundances (light grey) and SCML (dark grey). In **(a)** and **(b)** the reads of the three spike-in bacteria are omitted. Additionally, the reads that contributed to the genus of *Enterococcus* are excluded from the *Firmicutes* phylum and coloured separately (purple).

microbiomes relative to the specimens collected before ASCT (Fig. 2.6c). Patients 2, 4 and 5 showed log_2 ratios of *Enterococcus* between the last and first time point of 10.93, 9.22 and 3.60, respectively, employing SCML, compared to log_2 ratios of 14.76, 11.46 and 8.69 based on standard data. This suggests that *Enterococcus* dominance is in fact associated with both a decrease in microbial load and a rise in absolute abundance of *Enterococcus*.

2.5 Chapter Discussion

Here we suggest the use of spike-in bacteria to calibrate multiple intestinal microbiome profiles to microbial loads (SCML). We employed *A. acidiphilus*, *R. radiobacter* and *S. ruber* as spike-in bacteria and demonstrated their excellent suitability for a comprehensive and informative profiling of gut microbiomes. Usually, these three bacteria are absent in the intestinal microbiomes of mammals, and their unique 16S rRNA gene sequences cannot be mistaken for those of bacteria found in the gastrointestinal tract. All three bacteria are valid reporters of the actual microbial load. Thus SCML adds a new perspective to gut microbiome profiling that expands the common relative microbiome composition analysis.

Variability in microbial loads of intestines is a genuine and potentially clinically relevant biological feature that remains underutilised in standard protocols. On a more technical side, adding whole cells prior to lysis enables control for DNA recovery and pyrosequencing errors as a side benefit. Following this, the addition of exogenous spike-ins could also enhance other studies like whole-genome sequencing, qPCR-based quantification of pathogens as well as approaches using alternative marker genes [157, 170, 171].

Bacterial species compete for nutrients and can mutually displace each other, while others can only live in symbiosis. These dynamics of the intestinal ecosystem shape the structure of microbiome profiles [172, 173]. Mutually displacing, e.g. concurrent or antagonistic, species display anti-correlated profiles, while those of symbiotic species are correlated [174]. This theoretical consideration holds true for absolute numbers of bacteria. Interpreting the correlation structure on the basis of relative numbers can be misleading. If one species grows in absolute number, this will lead (i) to an increase of its fraction within the microbiome and (ii) to a decrease of the fractions of all other species. Hence, every change of a single species affects relative counts for all other species generating notorious anti-correlation between profiles of different species due to compositionality [175, 176]. Importantly, this effect is independent of ecological processes like displacement and symbiosis. Thus, profiles calibrated to ratios of total microbial loads provide a less disturbed assessment of the dynamics of the intestinal ecosystem.

We observed different sequence read yields for the three spike-in bacteria even upon addition of identical numbers of 16S rDNA gene copies to mouse faeces sharing the same microbial load. Fortunately, this problem should only arise with comparisons of different species. As demonstrated, it does not affect the estimation of intra-species ratios between samples, where species-specific yields cancel.

There is a difference between absolute quantification and calibration of ratios of absolute abundances. The former needs calibration to a defined unit such as bacteria per volume. External spike-ins do not enable absolute quantification due to e.g. variable lysis efficiency across intestinal bacteria and variable 16S rDNA copy numbers. In ratios the unit cancels. Hence, if in a comparison of the same OTU in two samples SCML calibrated values show a ratio of 2, then the OTU is in fact represented (almost) twice as often. With standard relative data this is not the case, when microbial loads in these samples differ. Importantly, ratios between different OTUs are not calibrated by SCML. We can thus calibrate microbiome profiles to ratios of absolute abundance but not to absolute quantities of bacteria themselves.

A drawback of the spike-in approach is the propagation of PCR amplification errors from the spike-in bacterium to all other taxonomic units. Indeed, the spike-in counts can be affected by PCR amplification or sequencing errors. The earlier these errors occur, the more they could influence the final read tallies. By using these reads to calibrate microbial ratios, this error-derived variance propagates to all other taxonomic units. The calibration reduces bias, but inflates variance. One may attenuate this undesired effect by using multiple spike-in bacteria of fixed concentrations across samples and averaging or summing their counts as shown here.

Chapter 3

Dynamical refinement of operational taxonomic units with *dOTUClust*

In this chapter I highlight limitations in microbiome community analysis, which arise depending on the chosen count resolution and offer an algorithm to alleviate these discrepancies by dynamic adaptation of OTUs: hierarchical affinity merging (HAM). I implemented the algorithm in an R-package, *dOTUClust*. The chapter is organized as follows. First, an overview of state-of-the-art methods and analysis pipelines is given. Possible shortcomings of the different approaches are discussed. Further, basic principles needed for the algorithms in this package are introduced. In the following the R-Package itself and its general functions are described. After an in-depth description of the underlying algorithms I offer application of the package on a 16S rRNA gene amplicon sequencing data set as proof-of-principle and highlight the differences and concordances between different approaches. In the last two sections of this chapter results on this data set are presented and discussed.

3.1 Abstract

The most used technology to profile the bacterial compositions of complex microbiomes is 16S ribosomal RNA (rRNA) gene amplicon sequencing. To quantify micro-organisms 16S rRNA gene sequences must be assigned to taxonomic units, which are readily available on various levels of granularity, from fine-grained and computational determined operational taxonomic units (OTUs) to the coarse but time honoured taxonomic ranks. Choosing the optimal level of taxonomic resolution is not obvious when comparing microbiomes. With very high resolution the majority of taxonomic units have very few reads assigned to them, rendering the proper

identification of differential abundance near to impossible. On the other hand, with coarser units ecological differences that only affect sub-units are obfuscated. Depending on the studied microbiomes some branches of the microbial phylogeny are more populated than others. For densely populated branches we can use a fine grained taxonomic resolution, while for more sparsely populated branches we must settle for coarser resolutions. Thus, taxonomic units should adapt their resolution to specific microbiota distributions and different microbiome studies require varying layouts for taxonomic units.

Here I introduce adaptive taxonomic units (ATUs). For a given data set it calculates a hierarchy of computational taxonomies with decreasing granularity. Starting from highly resolved 99% sequence identity OTUs the algorithm successively merges taxonomic units that are both phylogenetically and ecologically similar until all sequences fall into a single unit. This adaptive taxonomic hierarchy merges units in sparsely populated branches long before it combines those in more dense areas. By determining the optimal resolution in the hierarchy one can balance biological or clinical requirements of taxonomic resolution with statistical needs for sufficiently high microbial counts. I show that the algorithm can distinguish similar taxonomic groups by their distributional pattern in a perturbation experiment and highlight the impact of this OTU agglomeration strategy on downstream analysis. The underlying algorithm to build ATUs is integrated into the R-package *dOTUClust*.

3.2 Chapter Motivation

In this section I will point out the motivation for a proper refinement of OTUs to improve microbiome community analysis. The general caveats of state-of-the-art analysis methods are highlighted and an alternative approach is proposed.

3.2.1 How to count - A compromise between resolution and power

Marker gene based microbiome community analysis relies on read count data. Raw sequencing reads are clustered into OTUs based on a defined sequence identity threshold (usually >97%). Based on representative sequences inside these OTUs taxonomy is assigned with the help of reference databases. The resulting OTU read counts are the basis for further analysis regarding the composition of the microbiome and the abundance of bacteria. In current studies there are several levels of count resolution to choose from. Either using a fine-grained resolution with plain OTU counts or a coarse-grained one by agglomeration of OTUs based on common taxonomy at a certain rank (collapsing taxonomies: mostly phylum, family or genus level) [155]. In the following, I will refer to these two granularities as OTU and taxonomy counts, respectively.

Often OTU count tables are sparse read count matrices, impeding statistical tests for differential abundance given the low read counts per OTU. As a consequence, features with low numbers of reads are filtered out upfront analysis, which leads to a loss of information on rare species. Agglomeration by taxonomic rank on the other hand can increase the read counts per entity, while additionally reducing the complexity of a data set. However, taxonomy counts might disguise actual diversity. Furthermore, taxonomy counts are dependent on the accuracy and completeness of the taxonomic assignment, as it defines how OTUs are collapsed by taxonomic rank. This is especially problematic in microbiome habitats like the gut, where a fair amount of residential bacteria remains unknown or uncultured and therefore proper reference sequences are missing [177–179]. If no unique taxonomic assignment can be made, OTUs can be only partially annotated (e.g. including family, but no genus annotation) [155]. This is the case if the inspected hypervariable region of the 16S rRNA gene is insufficient for proper taxonomic demarcation or if the taxonomy in the used database is incomplete.

To point out the strengths and shortcomings of each granularity, an example is illustrated in Figure 3.1. The underlying count data is shown in Figure 3.1 A. Counts by OTU and taxonomy (i.e. agglomerated by genus rank) on this data set are depicted in Figure 3.1 B and 3.1 C, respectively. As evident from the figure, the choice of granularity in this example affects the resulting conclusion:

A

OTU count table

OTU_ID	A	B	Genus
OTU_1	0	60	*Klebsiella*
OTU_2	58	0	*Klebsiella*
OTU_3	4	0	*Eggerthella*
OTU_4	5	0	*Eggerthella*
OTU_5	2	0	*Eggerthella*
OTU_6	4	0	*Eggerthella*
OTU_7	4	0	*Eggerthella*
OTU_8	5	0	*Eggerthella*

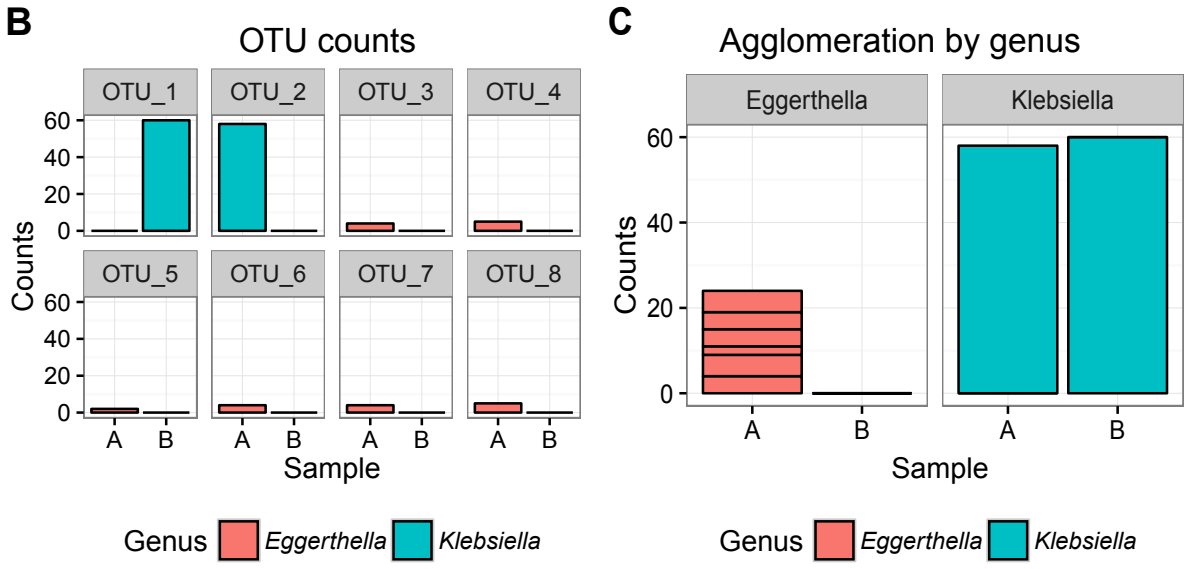

Figure 3.1: Read counts inspected at different levels of resolution in microbiome community analysis. **(A)** The table holds the read counts of eight OTU over two samples A and B, as well as the taxonomic assignment at genus level for each OTU. These read counts can either be investigated as OTUs **(B)**, or agglomerated read counts according to the assigned taxonomy (in this case genus) **(C)**. Both bar plots in **(B)** and **(C)** show read counts (y-axis) per sample (x-axis) for each OTU or genus level, respectively.

(i) Consider OTUs 1 and 2 in Figure 3.1 A, both encoding for the same bacterial genus: *Klebsiella*. Both show similar number of reads, but OTU 1 is only present in sample B and OTU 2 in sample A. Inspecting the OTU counts (Fig. 3.1 B) it can be concluded that there are two entities (i.e. sub-populations) of *Klebsiella*, each being mutually exclusive for one of the conditions. Given the underlying data, both OTUs can be considered differential abundant between condition A and B. (ii) On the contrary, if combining these OTUs by their shared taxonomy (Fig. 3.1 C) *Klebsiella* no longer shows differential abundance between condition A and B, hence information on this two opposing abundance patterns is lost. For OTUs 3-8 on the other hand an agglomeration by genus taxonomy would be beneficial to enrich low read counts. In this example either decision has its shortcoming: disguising differential abundance through agglomeration by taxonomy or loosing information in cause of sparse read OTU counts.

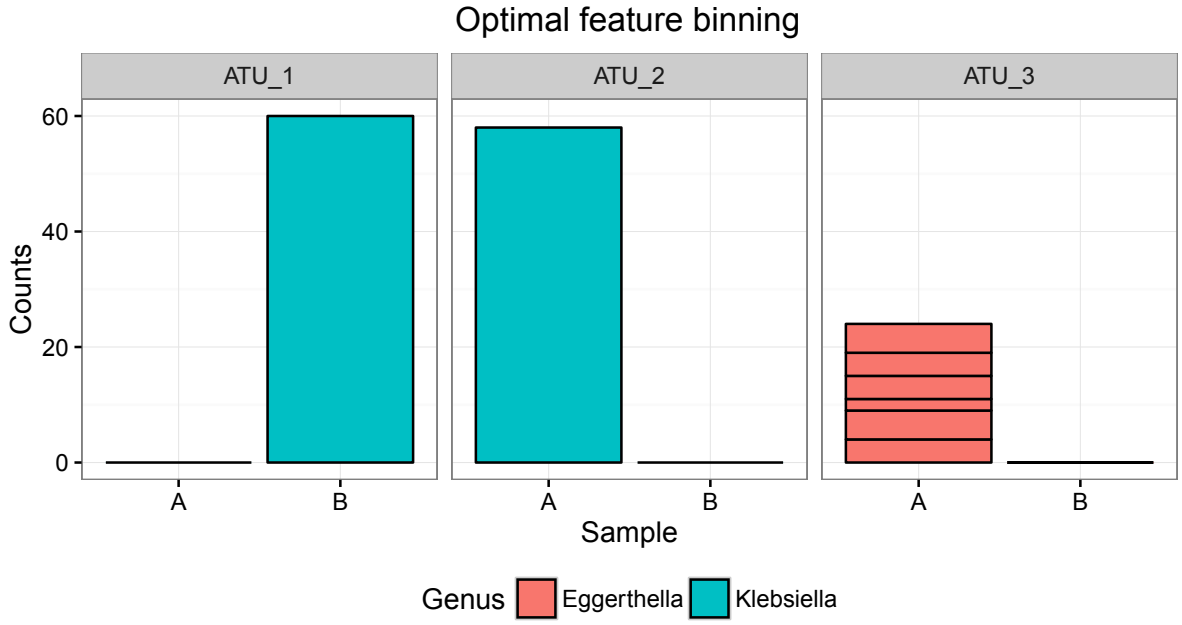

Figure 3.2: Example to illustrate ATU-binning on the toy example from Figure 3.1. Read counts are shown on the y-axis, while the x-axis represents two samples A and B. Bars are coloured according to the taxonomic assignment at genus level. Each facet represents an ATU.

We propose that a optimized agglomeration method would allow different levels of granularity given the underlying data. In the given example the use of coarse-grained taxonomy counts for the OTUs 3-8 and fine-grained OTU counts for OTUs 1-2 would be most appropriate. Because the resulting constructs are no longer pure OTUs, we hereby introduce ATUs and define this approach as ATU-binning. The ATUs for the example from Figure 3.1 are depicted in Figure 3.2. While OTU 1 and OTU 2 stay divided as ATU 1 and ATU 2, OTUs 3-8 are accumulated into ATU 3. Informations which could be utilized to control the degree of granularity of OTUs are already available in the data: distributional read count patterns and the size of each OTU.

3.2.2 Ecological similarity as guide for species demarcation

According to microbial ecology, bacterial species react on disturbances or changes of the environment [180, 181]. These varying environmental variables, which can act as evolutionary pressure on the residing species, can be for instance temperature, availability of nutrients or the introduction of a new species into the habitat. In the following, those species involved can show changes in their abundances. Youngblut *et al.* and others stated that species which react with divergent distributional abundance patterns on a given stimulus are most likely distinct species. Similar patterns of abundance on the other hand can indicate ecological coherence [182, 183]. Additionally, several sub-species have been shown to differ significantly in their ecological role, as well as geno- and phenotype, even if they share a common taxonomy [184].

The usage of ecological parameters to help in the demarcation of species was already proposed by Cohan *et al.* with their ecotype model [185, 186]. The authors further developed an algorithm, EcoSim, to refine the demarcation of sub-populations and identify eco-types by informing linkage based clustering with estimated ecological parameters like genetic drift and periodic selection [187]. Meanwhile Hunt *et al.* investigated the distributional patterns of isolated strains with their algorithm adaptML, which does not incorporate sequence information at all [188]. Both algorithms were used for specific demarcation tasks on small sets of well known sequences, but were not applicable for high dimensional next-generation-sequencing (NGS) data, spanning hundred thousands of read counts. Additionally, Koeppel *et al.* performed their clustering on sanger sequences, rather than 16S rRNA gene amplicons, which are generally shorter in length.

Preheim *et al.* recently utilized distributional information of read counts to inform OTU clustering [115]. The proposed algorithm, distribution-based clustering (dbOTU), aims to prevent clustering sequence reads into the same OTU, if they differ in distributional abundance. To achieve this the algorithm first picks a candidate sequence from the pool of unique sequences and searches for OTUs, whose representative sequences are close to the candidate in terms of genetic distance (Jukes-Cantor-corrected). If an OTU exists which fulfils this criterion, the algorithm checks, whether the candidate and the representative sequence of the OTU show similar distributional abundance patterns. To assess the similarity in distribution, distribution-based clustering counsels the chi-squared test. As long as the count distributions of candidate and reference sequence are not considered independent from each other (i.e. tested by chi-squared test), they are allowed to be merged. Otherwise, the candidate is evaluated against the next OTU, until a matching OTU is found. If no OTU could be found, which fulfils the aforementioned criteria, a new OTU is created with the candidate serving as representative sequence. This is repeated until all sequences are member of an OTU. Compared to other OTU calling algorithms, distribution-based clustering produces less ecological redundant OTUs, which increases power for downstream analysis

[115]. On top of these ecologically more consistent OTUs, Preheim *et al.* argue that compared to genetic distances, distributional patterns are more resilient towards sequencing errors, as well as variations in 16S rRNA sequence or copy numbers. Hence, sequences which originate from the same population show similar distributional patterns across samples, independent of possible errors or differences in 16S rRNA characteristics. According to the authors, the algorithm therefore also offers an alternative to classical de-noising. At the same time, counting unique reads and calculating distributional distances based on those is a very computational demanding task, especially for more complex communities (see the environmental sample in Preheim *et al.* [115]).

More recently Frøslev *et al.* presented a post-clustering curation method called LULU [189]. Instead of tackling the initial demarcation of OTUs, like some of the aforementioned methods, the authors wanted to curate already existing OTU count tables. Their hypothesis was, that taxonomic redundancy in OTUs often is a result of smaller differences in sequence variants and therefore an artefact of general OTU clustering strategy. The corresponding R-Package *lulu* utilizes patterns of co-occurrence together with sequence dissimilarity measures to identify such artefact OTUs which might have arisen from a more abundant OTU, either due to sequencing error or phylogenetic variability. By merging possibly related OTUs, this approach reduces the artificial redundancy which otherwise would result in overestimating diversity.

Both, dbOTU (successor dbOTU3) and LULU focus on the reduction of erroneous OTUs and are bound by user provided thresholds of sequence similarity [115, 189]. Yet, as microbiome research often focuses on disease associations, it might be of larger interest to identify hubs of ecologically similar performing entities, rather than create or curate plain species proxies. For the construction of optimal feature binning after successful OTU picking, as shown in Figure 3.2 with ATUs, I hereby present hierarchical affinity merging (HAM). While being guided by OTU size, as well as sequence- and distributional similarity, the algorithm identifies a optimal set of mixed granularities, covering the middle ground between taxonomic and OTU resolution. We apply the algorithm, which is shipped in the R-package *dOTUClust*, on a data set of mice receiving antibiotics and investigate its implications on downstream diversity analysis. Being the most related method, we also compare our results with those achieved by post-clustering curation method LULU. A plethora of alternative methods were proposed to overcome the uncertainty in species demarcation by utilizing other proxies for bacterial species (e.g. minimum-entropy-decomposition, oligotyping or exact sequence variants) [128, 190, 191]. However, since these methods abandon the concept of OTUs altogether, they pose a totally different angle point on species demarcation and are therefore not discussed in detail.

3.3 Chapter Methods

The sequencing data used in this chapter was kindly provided as a proof-of-principle for my methodology by Prof. Dr. Dr. Gessner and the Institute of Medical Microbiology and Hygiene at the University Clinic Regensburg. Animal housing, treatment, sample preparation, DNA extraction and 16S rRNA gene amplicon sequencing were also performed by members of the institute. Therefore all preliminary steps before data retrieval are only described briefly in the following two sub-sections to offer context of the data set and allow reproducibility (see 3.3.1 and 3.3.2).

3.3.1 Sample preparation and DNA extraction

Two male mice were housed at the animal facility of the University Clinic Regensburg and treated with a cocktail of antibiotics over a course of 56 days. The treatment was applied orally via fixed antibiotic concentrations in the daily drinking water. This antibiotic cocktail contained four different antibiotics, namely Neomycin, Metronidazole, Ampicillin and Vancomycin (for concentration levels see supplementary table B2). Faeces of the mice were collected right before application of the antibiotic cocktail (preABX), at days 28 and 56 (10wksABX) of the treatment, as well as four weeks after end of treatment (postABX). Faecal DNA extraction was performed for a total of eight specimens (2 mice x 4 time points) by a slightly modified version of the standard extraction protocol from the QIAamp DNA Stool Mini Kit (QIAGEN, Hilden, Germany) handbook. For each specimen 20 mg of faeces was weighed, accompanied by ASL Lysis buffer (Qiagen Kit) and subjected to cell lysis under repeated bead beating, subsequent heating and centrifugation. In the following, DNA isolation and purification was performed in close concordance to the standard protocol for pathogen detection with the QIAamp DNA Stool Mini Kit under the use of Inhibitex tablets, Proteinase K and silica-membrane-containing columns. Elution of DNA from the columns was performed under the addition of 80 µl AE buffer.

3.3.2 16S rRNA gene amplicon sequencing and total 16S qPCR

The isolated DNA of each specimen was spiked with already isolated DNA of the spike-ins used in part one of this thesis (see 2.3.1) and subsequently subjected to qPCR to quantify the total amount of 16S rRNA gene copy numbers. Specimens of the time points before treatment, at day 56 into antibiotics and after treatment were additionally subjected to 16S rRNA gene amplicon sequencing to investigate the microbial composition of the corresponding specimens. Protocols,

as well as used technology, primers and hydrolysis probes for qPCR and 16S rRNA gene amplicon sequencing were concordant to those used in the spike-in experiment (see subsection 2.3.3 for details). Primers and probes for Sequencing and qPCR can be found in supplementary table A4. Results of total 16S rRNA gene copy quantification via qPCR can be found in supplementary table B3.

3.3.3 Preprocessing and OTU clustering

Raw sequences were de-noised with FlowClus 1.0 [192], to reduce sequencing and PCR artefacts prior to further analysis. Minimum and maximum sequence length for elimination were chosen to be 400 and 800bp (expected read length based on the chosen 16S primers) with a maximal flow value of 6.49. The maximum numbers of homopolymers and of ambiguous bases accepted in a read were set to be 8 and 6, respectively. Furthermore, reads with an overall average quality score below 25 and a sliding window average quality score below 20 (window size 50bp) were eliminated. An overview of all chosen parameters for FlowClus are provided in supplementary table B4. The de-noised sequencing reads were demultiplexed and quality filtered by QIIME [119] (v1.8.0) with the split_libraries.py command using default parameters, except the minimum and maximum read length set to 400 and 800 bp according to the expected sequence length and to be concordant with the used FlowClus parameters. In order to allow higher differentiation, a sequence identity threshold of 99% was chosen to pick OTUs against the Greengenes database (version 13.8). This step was performed by using QIIME's pick_closed_reference_otus.py script with a sequence identity of 99%, while also providing the corresponding 99% reference OTUs from Greengenes. All other parameters were left on default. The resulting OTUs count matrices are loaded into R version 3.2.0 [167] with installed Bioconductor package [168] to be used as input for HAM. As additional input for HAM, also the used reference sequences from the OTU-picking step, as well as the otu-taxonomy map of the greengenes database are imported into R.

3.3.4 Features to assess OTU similarity

In order to decide which OTUs should be merged, the algorithm has to assess the relatedness between OTUs in the form of similarity or dissimilarity measures. It hereby distinguishes between two major aspects of similarity. First of all, it should be able to describe how phylogenetically similar two OTUs are. Hereby, the 16S rRNA reference sequences of each OTU can be used to compare sequence identity and therefore phylogenetic relatedness.

However, as already introduced, phylogenetic similarity does not necessarily imply similar ecological behaviour. Sub-species, sharing high sequence similarity in their 16S rRNA might adapt differently well in the same ecological niche. If for example one sub-species does be affected by an certain environmental factor, the other one might be immune to it. Therefore, to assess the ecological relatedness between two OTUs it can be evaluated how similar the read counts of both distribute over all samples. In the following I describe how both features of relatedness are calculated and how they are combined to guide HAM.

3.3.4.1 Phylogenetic similarity - The Levenshtein Distance

To achieve a measure of sequence similarity I decided to use the Levenshtein distance. This distance was introduced by Vladimir Levenshtein and is also known as the minimum edit distance [193]. It is defined as the minimum of insertions, deletions and substitutions needed to transform a string A into string B. This definition makes it a essential measure in computational biology for example to capture the distance between DNA sequences. The most prominent application is in Hirschberg's algorithm to determine the optimal alignment between two strings [194]. The distance itself fulfils all metric axioms and therefore is a true distance metric. For its application in HAM the Levenshtein distance is calculated in the initial step of the algorithm with help of the *stringdist* package in R [195].

3.3.4.2 Ecological similarity - The Jensen-Shannon Distance

To measure the difference in count distribution between two OTUs I decided to use the Jensen-Shannon distance (JSD). It is defined as the square root transformation of the Jensen-Shannon divergence

$$JSD(\vec{x},\vec{y}) = \sqrt{div_{JS}(p(\vec{x})||q(\vec{y}))} \tag{3.1}$$

where $\vec{x}$ and $\vec{y}$ are read count vectors of two different OTUs over all samples and p and q the probability distributions of these vectors, respectively. Compared to the Jensen-Shannon-Divergence, which does not fulfil triangle inequality, JSD fulfils all axioms for a distance metric [196, 197]. The Jensen-Shannon-Divergence was introduced by Jianhua Lin [198] and is defined as follows:

$$div_{JS}(P||Q) = \frac{1}{2}div_{KL}(P||\bar{m}) + \frac{1}{2}div_{KL}(Q||\bar{m}), \tag{3.2}$$

whereas $\bar{m} = (p+q)/2$, that is the mean probability distribution of p and q, and div_{KL} the Kullback-Leibler divergence. The latter was first introduced in information theory by Salomon

Kullback and Richard Leibler in 1951 [199] and is defined as:

$$div_{KL}(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}. \tag{3.3}$$

Jensen-shannon divergence was successfully applied in the field of alignment-free genome comparison via feature frequency profiles [200]. The distance variant, JSD, was used by Arumugam *et al.* for sample-wise clustering of microbiome profiles [201]. In this application read counts over all OTUs of each sample was used as distribution to assess inter-sample similarity. For optimized calculation of JSD in terms of speed, a parallelized C++ version of the algorithm (written by J.J. Allaire and Jim Bullard) was adapted and integrated into *dOTUClust* via Rcpp [202, 203].

3.3.5 Dissimilarity score as merging guidance

To guide the merging process in HAM a dissimilarity score is calculated in each step, which evaluates the best merging pair of OTUs. This dissimilarity score counsels three measures to assess the similarity for each pair of OTUs. First the Levenshtein distance, which is the minimum edit distance between the corresponding reference sequences of these OTUs. Second the JSD, which measures the similarity in read count distribution between the OTUs. I define *minN* as the minimum between the sums of total read counts of two OTUs. Given an OTU read count table M, *minN* between OTUs i and j is defined as

$$minN_{i,j} = \min(\sum_{k=1}^{m} M_{i,k}, \sum_{k=1}^{m} M_{j,k}), \tag{3.4}$$

whereas m is the number of samples. This measure was created to ensure that small OTUs are preferably merged over larger OTUs.

All measures are either updated (Levenshtein) or recalculated (*minN* and *JSD*) each step for every possible pairwise combination of OTUs and the dissimilarity score is recalculated. In order to prevent that ne of the incorporated measures (e.g. Levenshtein) dominates the others just by its numerical range, all measures are transformed to a common scale before the dissimilarity score is calculated.

The dissimilarity score between OTUs i and j is defined as the maximum between the Levenshtein distance (LV) and the minimum of *JSD* and *minN*:

$$d_{i,j} = \max(LV'_{i,j}, \min(JSD'_{i,j}, minN'_{i,j})), \tag{3.5}$$

whereas LV', JSD' and $minN'$ denote the transformations of LV, JSD and $minN$ to a common scale.

The rationale behind equation 3.5 is the following: First the minimum between JSD and $minN$ controls that JSD is only considered if there are enough counts to safely investigate the distribution by JSD. Second the maximum term ensures that the dissimilarity score between two OTUs can only be as low as the maximum between Levenshtein or JSD. Given two OTUs which are highly similar in terms of both measures, this leads to a small dissimilarity score and therefore a higher chance to be merged. On the other hand, if one of the measures shows a low similarity for the pair of OTUs in question, this leads to a high dissimilarity score, which decreases the chances of being merged at an early step of the algorithm. For example, two OTUs, which share the highest similarity in terms of JSD compared to all other OTU pairs, will receive a high dissimilarity score, if they show the lowest similarity regarding Levenshtein distance. If, however, one of both OTUs is small in terms of read counts, mostly Levenshtein is driving the dissimilarity score.

Because all three measures show different numerical scales, each score is brought to a common scale. To achieve this, a sequence A, which follows an exponential distribution, is created. It is defined as follows: Let A be a strictly monotonically increasing sequence with m elements

$$A = (a_j)_{j=1}^{m} \; or \; (a_1, a_2, \cdots, a_m), \tag{3.6}$$

whereas m is the number of elements in the lower triangle of the square distance matrix (i.e. $m = (n - \sqrt{(n)})/2$, with n being the total number of elements in the square distance matrix (i.e. number of all possible OTU-pairs) and $\sqrt{n}$ the number of diagonal elements). Then all elements in A are defined by an exponential function $f(x)$ divided by its integral between 0 and 1:

$$a(x) = \frac{f(x)}{\int_0^1 f(x)dx} \; x \in \mathbb{R} \mid 0 \leq x \leq 1, \tag{3.7}$$

whereas $f(x)$ is defined as follows:

$$f(x) = 1 - e^{-x}. \tag{3.8}$$

The resulting sequence, as well as the elements in JSD, $minN$ and $Levenshtein$ are ordered by their specific ranks. Finally, the distance measures values are set to the value of A according to their respective rank order. Therefore, the lowest value of each distance matrix gets assigned the lowest value of A:

$$JSD'_{rank(JSD)} = A_{rank(A)}. \tag{3.9}$$

Because the distance matrices of all three measures are symmetric, all calculations are only performed on the lower triangle of all matrices to reduce complexity and computation time. After the transformation the dissimilarity score will be assigned according to equation 3.5. Due to the nature of the beforehand transformation the dissimilarity score is bounded between 0 (very similar) and 1.718282 (very dissimilar, see eq. 3.8). The dissimilarity score then can be used to guide HAM.

3.3.6 Hierarchical Affinity Merging (HAM)

HAM combines OTUs based on the calculated dissimilarity score, define in equation 3.5, in a stepwise manner. This score counsels sequence similarity, distributional information and the size of both OTUs. Similar to hierarchical agglomerative clustering, HAM combines each step the best scoring pair of clusters. The algorithm is illustrated in Figure 3.3 and will be described in the following paragraph.

HAM takes an OTU count table, a set of corresponding reference sequences for each OTU, as well as a taxonomic mapping as input, with the latter being optional. Based on these inputs, HAM identifies in each step the minimum dissimilarity score of all pairwise comparisons and merges the corresponding OTUs. In very rare occasions taking the minimum dissimilarity to identify the next merge does create ambiguous results. This can happen, as soon as at least two sets of candidate OTUs share the same dissimilarity score (i.e. a tie). In this case HAM falls back to inspect each of the three distance measures on its own until it finds an optimum (i.e. minimum distance) between the candidates in question.

After successful merging the dissimilarity score needs to be updated. Whereas *JSD* and *minN* have to be recalculated because read counts change due to the accumulation of OTUs, Levenshtein distance can be updated through linkage. Which linkage shall be used can be chosen by the user from the following methods: complete, single or average linkage (default is complete). The updating saves a lot of computational time and is possible because the Levenshtein distance fulfils all criteria of a true distance metric [193]. After updating the dissimilarity score the algorithm again identifies the minimum score and merges the corresponding OTUs. Again, *JSD* and *minN* are recalculated and the Levenshtein distance of both merging OTUs is updated by linkage methods. These steps are repeated until only one cluster is left (c.f. see Figure 3.3). To identify the optimal k to cut the hierarchy, HAM checks during each clustering step if a convergence criterion is fulfilled. Currently, there are two convergence criteria implemented in

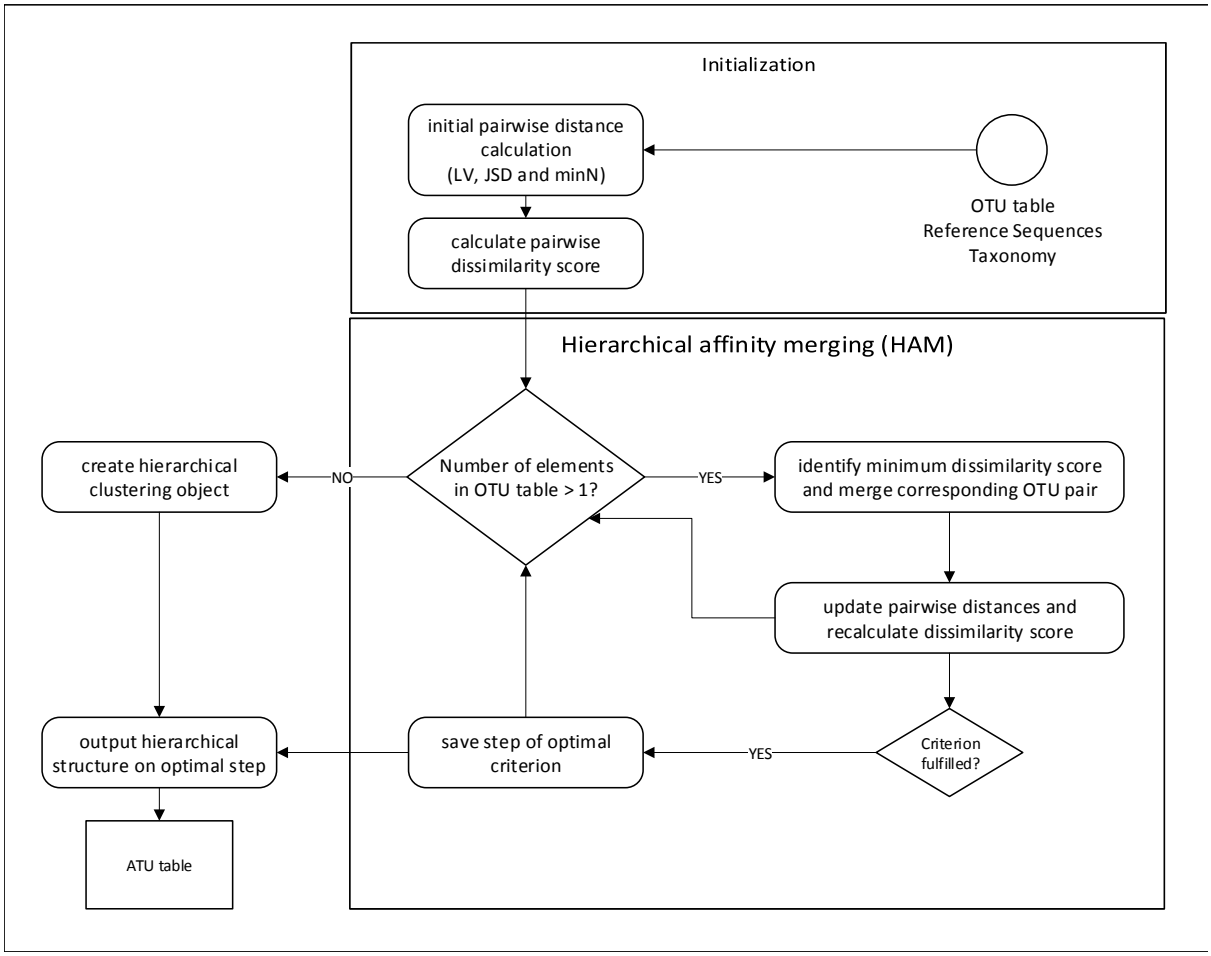

Figure 3.3: Flowchart illustrating the main algorithm of *dOTUClust*: hierarchical affinity merging (HAM). The algorithm starts with an OTU table and its corresponding reference sequences for each OTU (e.g. as provided by QIIME) and initializes by calculating all pairwise distance measures (i.e. Levenshtein distance between reference sequences, Jensen-Shannon distance (JSD) and *minN*). All measures are transformed and combined into the dissimilarity score, as described in section 3.3.5. In a next step the algorithm then identifies the OTU pair with the lowest dissimilarity score assigned and merges those two OTUs. After successful merging, the distance measures are updated and the dissimilarity score recalculated. In each step the convergence criteria are calculated and kept for later usage. The merging and update steps are repeated until all OTUs are merged into one single entity. When the full hierarchy is build, the algorithm identifies based on the calculated convergence criteria the optimal k to cut the hierarchy. This value of k is then used to build the final ATU count table and mapping.

HAM to choose from: (i) at least 95% of all read counts are part of an ATU. (ii) At least 95% of all clusters present contain at least n reads (default: n = 40). Alternatively, *k* can be provided by the user. After HAM is completed, the hierarchy is cut first step *k* which fulfilled the chosen criterion and the resulting clustering is used as template for the agglomeration of OTUs into final ATUs.

Box 3.1: Adaptive Taxonomic Units (ATUs)

Adaptive Taxonomic Units (ATUs) are defined as an agglomeration of OTUs, guided by sequence identity, distributional similarity and cluster size. ATUs are build in a stepwise hierarchical manner. The building of ATUs is based on Hierarchical Affinity Merging (HAM). Compared to other methods, ATUs do not aim to reduce sequencing error/bias, but try to increase statistical power by merging groups of interest with highly similar distributions and therefore decreasing complexity. Hence, OTUs with highly similar distribution can end up in the same ATU, even if they are of different taxonomic origin, as long as their sequence identity is not too far away. In general ATUs can be considered a proxy for functional and ecological clusters inside a microbiome profile.

3.3.7 R-Packages used

Data manipulation tasks were performed with the help of *data.table* [204], as well as instances of the *tidyverse* [205]. For visualization in this chapter the packages *ggplot2* [206], *ggpubr* [207], *gridExtra* [208], *cowplot* [209], *dendextend* [210] and *ComplexHeatmap* [211] were used. Calculation and testing of beta-diversity was performed with *vegan* [212]. Parallel C++ code was incorporated via *Rcpp* and *RcppParallel* [213–215]. Levenshtein distances were calculated via *stringdist* [195]. For comparison reasons *lulu* [189] was used to retrieve curated read counts. For the means of reproduction source code, data files, documentation as well as further information on all methods is provided in the electronic supplement of this thesis.

3.4 Chapter Results

In this section I first offer a proof of principle that results on OTU and taxonomy count data differ on a data set of mice screened before, whilst and after antibiotic treatment. In the following I apply *dOTUClust* on this data set to offer more robust results, independent of taxonomic assignment and highlight the differences to both previous methods.

3.4.1 Proof of principle - mice receiving antibiotics

3.4.1.1 Assumptions based on read count data are biased by the level of granularity

To assess whether assumptions based on OTU and taxonomy counts are consistent, I applied HAM to a subset of OTUs of the in house antibiotics mice dataset. Abundances of 263 OTUs (rows) are shown in a heatmap in Figure 3.4. Samples (columns) are annotated according to the treatment time point. The time points are before (preAbx), whilst (10wksAbx) or after antibiotic treatment (postAbx). The cell colouring indicates the abundance for each OTU. While lighter tones of blue indicate a higher abundance, dark blue indicates lower abundances and read counts of zero are coloured white. In the heatmap most OTUs are solely present at a specific time point of antibiotic treatment, either before or after. Only a few of these entities reconstitute after treatment (top of the heatmap), while another fraction tends to be present only whilst and after treatment. Based on this illustration a researcher might conclude, that there seems to be a shift in OTUs between preAbx and postAbx time points.

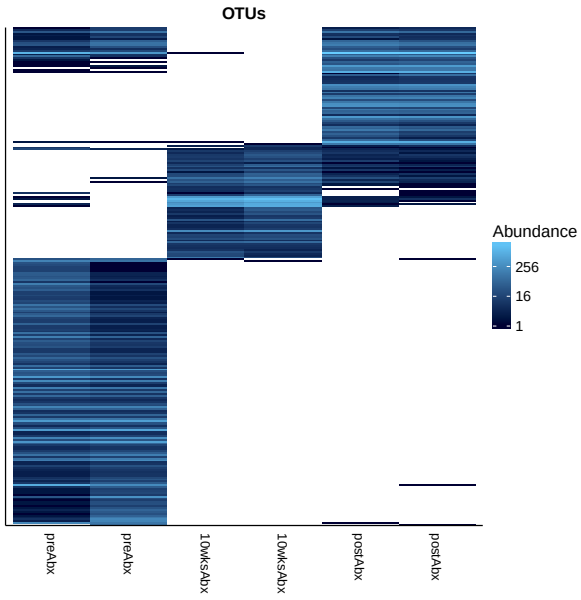

Figure 3.4: Heatmap illustrating the read count abundance of all OTUs (y-axis) for six caecal mice samples (x-axis). The samples were taken at different time points of the treatment: before antibiotics administration (preAbx), ten weeks after onset of antibiotic treatment (10wksAbx) and after antibiotic treatment and washout (postAbx).

However this mutual exclusivity can be disguised if the OTUs are agglomerated by their corresponding taxonomic rank. A common agglomeration by the family rank is shown in Figure 3.5 A. In this heatmap the rows correspond to the family bins according to the assigned taxonomy. In this resolution, compared to the OTU counts in Figure 3.4, we observe no family that is only present after antibiotic treatment. All families displayed were present before treatment.

While for instance *Lactobacillaceae*, *Ruminococcaceae*, *Erysipelotrichaceae* and *Peptostreptococcaceae* get lost during treatment, but reoccur after antibiotic treatment, some families (i.e. *Anaeroplasmataceae*, *Prevotellaceae*, *Bacteroidaceae* and *Porphyromonadaceae*) were totally lost due to antibiotic treatment in the observed time span. Other families either increased (*Enterobacteriaceae*) or decreased (*S24-7*, *Lachnospiraceae*) during antibiotic treatment (10wksAbx). These three families are also the highest abundant ones in this data set, comprising for 26%, 35% and 20% of total read counts, respectively.

As an example for the disguise of differential patterns I inspected the OTUs which are part of the *Lachnospiraceae* family, as displayed in Figure 3.5 B and which are comprised of 9938 reads (20% of total read counts). Figure 3.5 B shows a shift between OTUs only present before (highlighted in light green) and only present after antibiotic treatment (highlighted in blue). This assumption for *Lachnospiraceae* can not be made if read counts are agglomerated at family rank (see Fig. 3.5 A, highlighted in red).

3.4.1.2 OTU and taxonomy granularity are both prone to loss of information

Agglomeration of read counts by taxonomic rank is dependent on the quality and completeness of the taxonomy assignment after OTU calling. Depending on the rank chosen for agglomeration, read counts that are not annotated at that particular rank are lost in the resulting taxonomy count tables. Table 3.2 illustrates how many OTUs are not annotated on the ranks of family, genus and species in the antibiotics dataset. On the order rank still all OTUs, and therefore all reads, are annotated. This changes for the antibiotics data set for the family rank, where already 17% of a all OTUs (see left table in tab. 3.2) are not annotated. Even though this loss only corresponds to 7% of total read counts at this level, the information loss gets worse for each deeper taxonomic rank. On genus level already 69% of OTUs (66% of total reads) are lost due to this agglomeration strategy. By agglomerating at species level on this data set, one looses 90% OTUs (96% of total reads). We further investigated, whether this effect was introduced by the read count filtering (OTUs with less than 10 total read counts) which we performed beforehand and show that, despite minor fluctuations, the results are concordant on the full antibiotics data set (i.e. without prior read count filtering, see right table in tab. 3.2).

Considering the sparsity of OTU count tables, we also observe a loss of information induced by read count filtering. For our subset of the antibiotics data set we loose 150 of 263 OTUs (57%) if we filter for at least 50 read counts over all samples (see table 3.2). Increasing the count threshold leads to a loss of 186 OTUs (70.7%) for a minimum of 100 reads, climbing up to 212 (80.6%) OTUs for a minimum of 200 reads over all samples. This loss of features corresponds to 3473 (6.9%), 5947 (11.8%) and 9403 (18.7%) read counts for thresholds of 50,

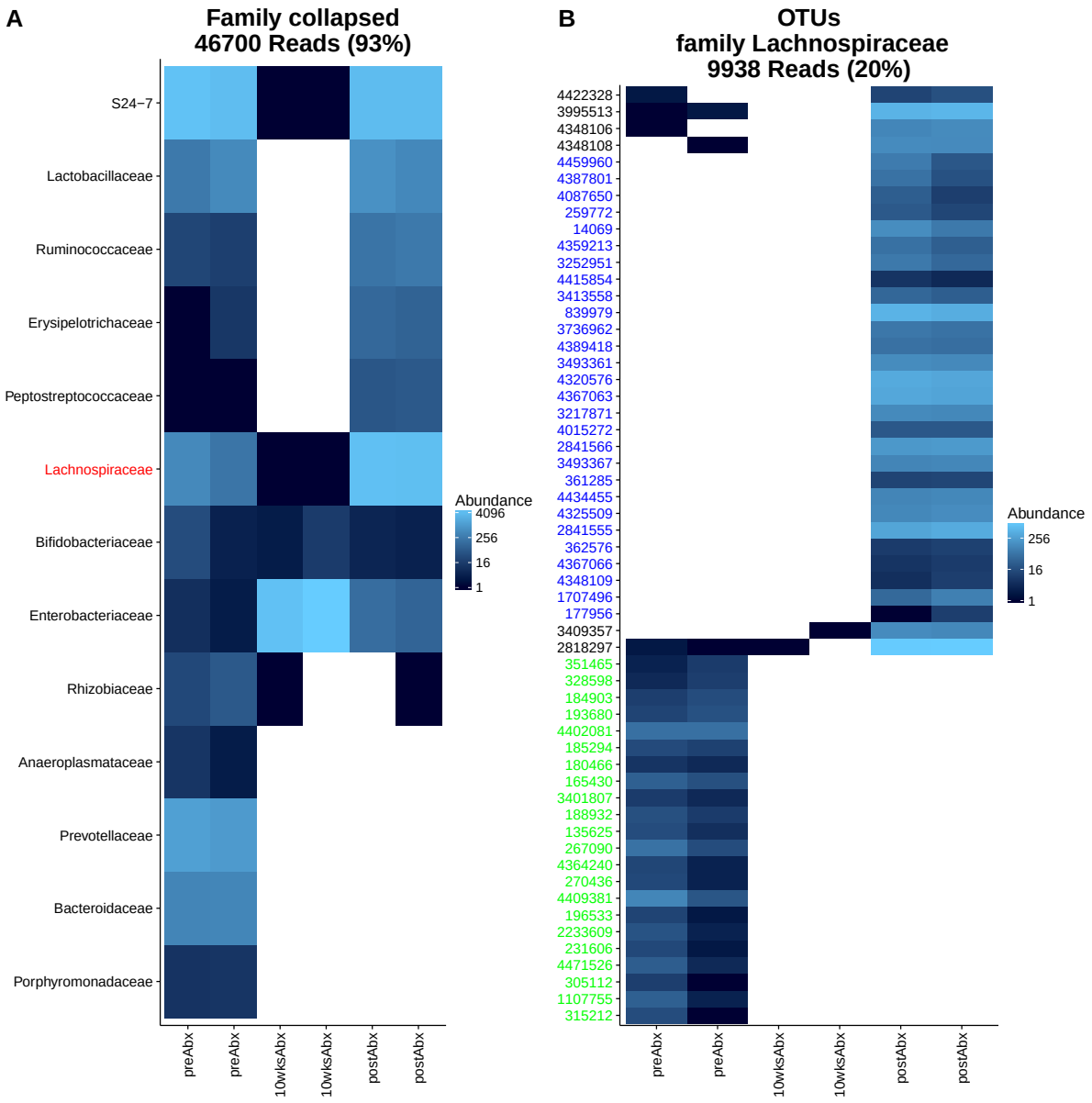

Figure 3.5: Heatmap showing read count abundances of all family levels **(A)** and all OTUs encoding for family of *Lachnospiraceae* **(B)** for each sample (x-axis) of the antibiotics dataset subset. Samples are annotated by time point of extraction. Lighter tones of blue correspond to higher read count abundance of each entity. Read abundances of *Lachnospiraceae* are marked red in **(A)**. In **(B)** OTUs which are only present after antibiotic treatment (postAbx) are highlighted in blue and those only present before antibiotic treatment are highlighted in green. Beneath each plot title the number of read counts represented in each heatmap is presented, as well as its percentage of total read counts.

subset (263 OTUs)			full data set (1095 OTUs)		
chosen taxonomic rank		**not annotated**	**chosen taxonomic rank**		**not annotated**
Order	OTUs	0 (0 %)	Order	OTUs	0 (0 %)
	Reads	0 (0 %)		Reads	0 (0 %)
Family	OTUs	45 (17 %)	Family	OTUs	176 (16 %)
	Reads	3628 (7 %)		Reads	3992 (8 %)
Genus	OTUs	182 (69 %)	Genus	OTUs	684 (62 %)
	Reads	33368 (66 %)		Reads	34949 (66 %)
Species	OTUs	237 (90 %)	Species	OTUs	950 (87 %)
	Reads	48294 (96 %)		Reads	50307 (96 %)

Table 3.1: Annotation status of OTUs from the unfiltered antibiotics dataset (1095 OTUs, right table) and a filtered subset of it (263 OTUs, left table) depending on the taxonomic rank chosen for agglomeration and the corresponding sequence reads affected by this. For the left table OTUs were omitted if they had less than 10 read counts over all samples. Each number is followed by its percentage of total OTUs or total sequence reads in parenthesis.

	count granularity					
	OTUs			**ATUs**		
Count treshold	50	100	200	50	100	200
Features filtered	150	186	212	5	10	15
Reads filtered	3473	5937	9403	219	572	1371

Table 3.2: Overview of lost information due to minimum read count filtering for OTU and ATU granularity. For different thresholds the number of features are shown (i.e. OTUs and ATUs), which fall below the minimum read counts, as well as the sum of the corresponding read counts involved.

100 and 200, respectively. As illustrated by table 3.2 this loss of features can be substantially attenuated by building ATUs with *dOTUClust* on the antibiotics subset. For the same thresholds used for OTUs, only 5 (14.3%), 10 (28.6%) and 15 (42.9%) ATUs are lost, which corresponds to 219 (0.4%), 572 (1.1%) and 1371 (2.7%) read counts, respectively.

3.4.1.3 HAM allows for dynamical enrichment of OTU count data by utilizing ecological and phylogenetic properties in microbiome community analysis

We applied HAM on the 263 OTU subset of the antibiotics data set shown in Figure 3.4. The algorithm converged after 228 steps, leaving with 35 features left, consisting of 31 ATUs and 4 OTUs (i.e. singletons). The chosen convergence criterion was to reach 95% of reads being part of an ATU.

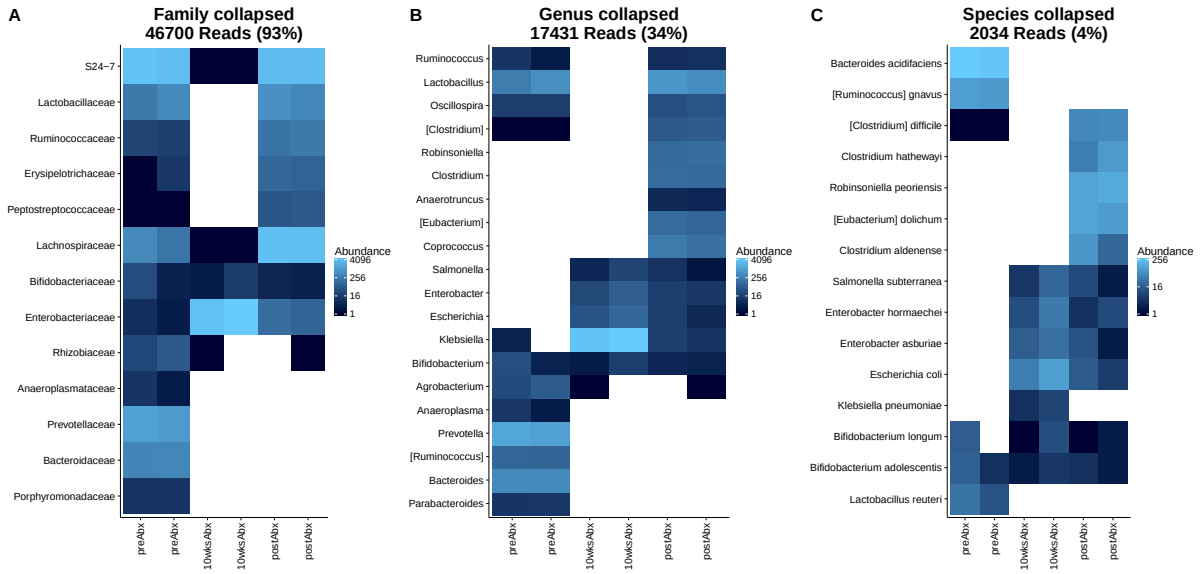

Figure 3.6: Read count abundances of the antibiotics subset (263 OTUs) summarized by taxonomic annotation on different taxonomy ranks: **(A)** family, **(B)** genus **(C)** species. The y-axis corresponds to the taxonomic entities and the x-axis to the samples. Samples are annotated by the corresponding time point of extraction. All OTUs missing annotation at the specific rank are considered not abundant and omitted from these illustrations. The value underneath each plot title shows how many reads are represented (i.e. annotated at the specific rank) by the corresponding taxonomic agglomeration, followed by the percentage of total reads in parenthesis (for detail see 3.1).

The resulting hierarchy of OTUs is illustrated by a dendrogram in Figure 3.7. The coloured OTU labels indicate different family ranks. Height in the dendrogram corresponds to the merging step, counting from inside to the outside. The underlying ATU mapping is shown in table B5 (see appendix). This table holds for each OTU in the subset the corresponding taxonomy, the resulting ATU assignment by HAM and information about the outlier status based on Levenshtein and Jensen-Shannon distance. A OTU inside an ATU is considered as possible outlier if its distances are greater than those of the 75% quantile of all the members of that given ATU. Interestingly, we observe only a small fraction of OTUs holding different taxonomic annotations inside of ATUs, even though taxonomic relation is only modelled in parts via the Levenshtein distance.

We next investigated the aforementioned example of *Lachnospiraceae* OTUs and their assignment by HAM. Figure 3.8 illustrates patterns of 56 OTUs which are annotated as *Lachnospiraceae* (A), the corresponding agglomeration by taxonomy (C) and 10 ATUs found by HAM containing *Lachnospiraceae* assigned OTUs (B). The latter can therefore contain also OTUs, which are not assigned as *Lachnospiraceae*. For all panels the y-axis denotes $\log_{10}$ counts and the x-axis lists samples as previously. Each facet in panel A represents an OTU and in panel B an ATU. The background colours in A encode the ATU membership of each OTU, corresponding to the colouring in panel B. Dashed borders surrounding the facets in both panels

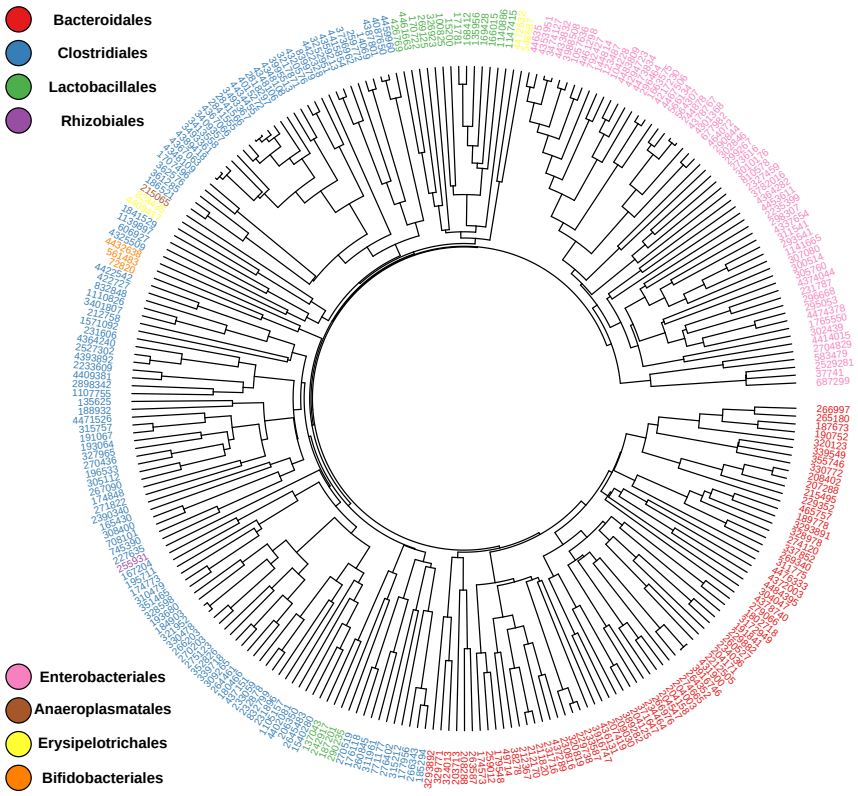

Figure 3.7: Dendrogram of the full hierarchy of OTUs as derived by HAM on the subset of the antibiotics data set. Each tip denotes a OTU and is coloured by its corresponding family rank taxonomy. The chronological order of merging steps is encoded in the height of the dendrogram and starts from the outside of the circle.

A and B indicate total read counts below 50, which would be omitted by read count filtering in this scenario. Colouring of the bars corresponds to the genotype of each mice.

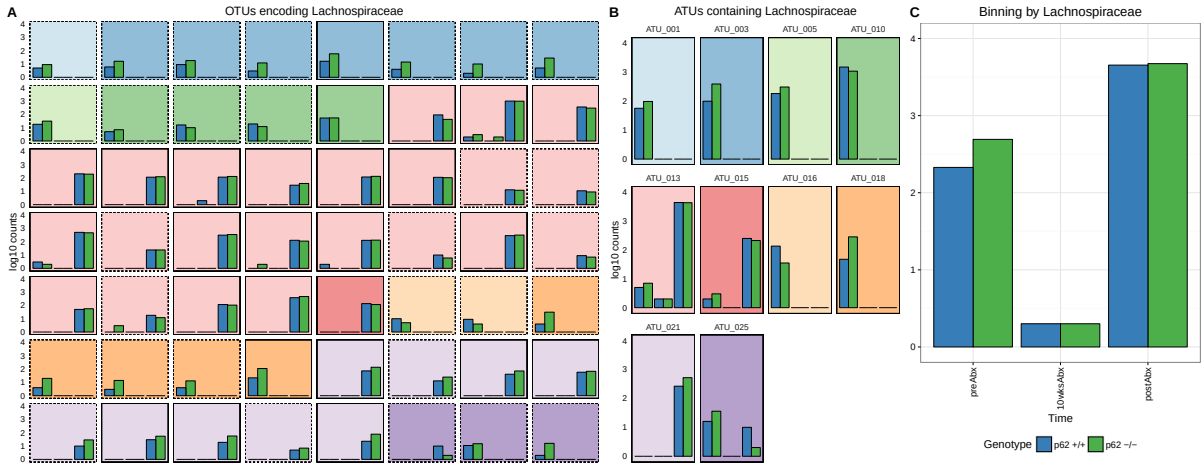

Figure 3.8: Comparison of three levels of granularity: OTU counts **(A)**, ATU counts **(B)** and counts on family level **(C)**. Log_{10} counts (y-axis) on a OTU subset of the antibiotics mice data set. Each bar in a facet corresponds to a sample (x-axis). The samples were taken at different time points of the treatment: before antibiotics administration (preAbx), ten weeks after onset of antibiotic treatment (10wksAbx) and after antibiotic treatment and washout (postAbx). Samples are grouped by time point. The colour of the bar encodes for the genotype. This plot illustrates all OTUs encoding for *Lachnospiraceae* family **(A)**, the corresponding ATUs containing at least on of these OTUs, as well as an agglomeration of all OTUs at the family level of *Lachnospiraceae*. The background colour in **(A)** encodes the ATU membership of each OTU and relates to the background colour in **(B)**. Dashed lines arround facets point out OTUs with less than 50 overall counts. Composition of the ATUs in (**B**) is illustrated in more detail in Figure 3.9. Differences in total counts between panel **(B)** and the others arise because ATUs also contain OTUs which are annotated differently or not at all.

While taxonomy agglomeration in figure 3.8 suggests a decrease during antibiotics treatment, followed by a reconstitution hereafter, ATU-binning with HAM conserves the different patterns which were already observed in the OTU count data for *Lachnospiraceae* (see also Fig. 3.5 B). In summary two distinct patterns are found: On one hand ATUs which are only present before treatment (ATUs 1, 3, 6 , 10, 16 and 18), and on the other hand ATUs which mainly occur after antibiotic usage (ATUs 13, 15 and 21). Two patterns, which are clearly no longer distinguishable in family agglomeration (see Fig. 3.5 C).

Similar effects can be observed for the family of *Enterobacteriaceae* and the genus of *Lactobacillus*, which are shown in panels B and C in Figure 3.9 alongside with *Lachnospiraceae* (panel A). These heatmaps visualize $\log_{10}$ read counts of OTUs, which either are encoding for the taxonomic rank of interest or are part of an ATU which contains at least one named entity of this rank. Colouring of the heatmap cells fades from dark blue to yellow with increasing read counts. For each of these OTUs one coloured bar at the right side of each heatmap indicates the ATU membership (ATU ID) and another on the left side of each heatmap whether the OTU

itself would sustain a read count filtering for 50 overall read counts (Threshold). Additionally, each panel holds information about assigned taxonomy for each OTU on different taxonomic ranks (A: Order and Family, B: Family and Genus, C: Genus, right side of each heatmap). If the annotation on the specific rank is missing the corresponding bar is coloured light grey.

Figure 3.9 A offers a more detailed depiction of the composition of the ATUs from Figure 3.8 B. Here we observe that these ATUs not only contain OTUs encoding for *Lachnospiraceae* (green coloured, first segment), but also for other families like *Peptostreptococcaceae* (lavendel coloured, second segment) or *Ruminococcaceae* (orange coloured, third segment). Additionally, there are OTUs which have no taxonomic assignment at family level (light grey, fourth segment). These would have been omitted in analysis based on agglomeration by taxonomy. Strikingly there are no contradictory patterns accumulated within the different ATUs, even though HAM produces more ATUs than expected by the distributional patterns present for this subset. This would not hold for collapsing taxonomies at family rank, as already shown in Figure 3.5 (as indicated by the family colour bar).

For the family of *Enterobacteriaceae* in 3.9 B the algorithm identifies six ATUs, consisting of 57 OTUs. These ATUs depict two distinct patterns: First, OTUs are present only whilst the administration of antibiotics, and second, OTUs arise at administration and stay present hereafter. Again patterns within ATUs are consistent with each other for all ATUs and these patterns would have been misleading by family counts. For this example all OTUs in the heatmap do encode for the same family, *Enterobacteriaceae*, but differ in their genus assignment. While most OTUs are members of *Klebsiella* (green) or not annotated at genus level (grey), a scattered fraction (6 OTUs) encodes for the genera of *Escherichia* (orange, 3 OTUs), *Enterobacter* (blue, 2 OTUs) and *Salmonella* (rose, 1 OTU).

As a third example Figure 3.9 C shows the members of five different ATUs, which comprise OTUs of the genus *Lactobacillus* (18 OTUs). Two of these ATUs are singletons, containing only one member each. These are named by the corresponding OTU instead of receiving an ATU identifier (orange and darker green at the top of the figure). This example once more shows the possible disguising effect of taxonomy based counts. While most OTUs are present before and after antibiotic treatment, there are five, which do not reconstitute afterwards. This subset is successfully separated by HAM from the aforementioned pattern (see singleton 326923 and ATU 26) and could not have been identified on read counts collapsed to genus level (see genus annotation bar on the right side of Fig. 3.9 C). Additionally the members of ATU 26 would be lost during read count filtering on OTU counts, but are conserved by being merged as ATU. A similar case can be observed for the candidate family *S24-7* from the order of *Bacteroidales*. Being a candidate clade, this family does not offer deeper annotation (i.e. genus) and would

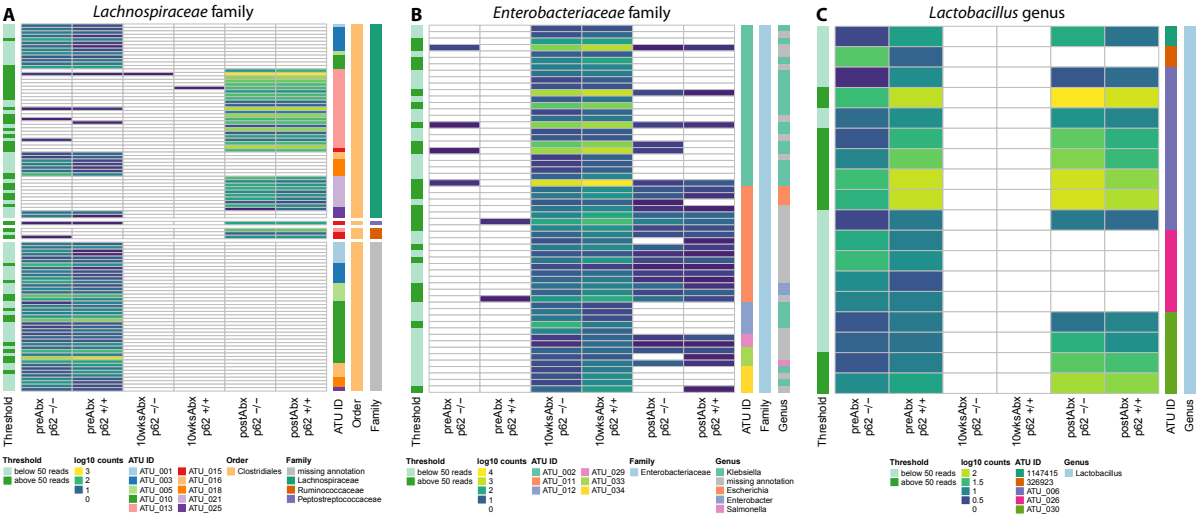

Figure 3.9: Heatmap displaying $\log_{10}$ read counts of different subsets of OTUs (y-axis) for each sample (x-axis) of the antibiotics dataset. Samples are annotated on the bottom according to the time of extraction and their respective genotype. An OTU was included if it is a member of any ATU containing *Lachnospiraceae* **(A)**, *Enterobacteriaceae* **(B)** or *Lactobacillus* **(C)**, respectively (e.g. for **(B)** ATUs 2, 11, 12, 29, 33, 34). ATUs can contain non merged OTUs, which are annotated by their respective OTU ID (e.g. 1147415 and 326923 in **(C)**). The column threshold on the left highlights if a OTU on its own would pass (dark green) a read count filtering of 50 overall read counts or not (light green). Additionally, each OTU is annotated with its ATU membership (ATU ID) and one or more taxonomic ranks, ranging from Order to Genus, indicating how these OTUs would be agglomerated by taxonomy. The heatmap cells are splitted by a gap according to the family annotation of each OTU in **(A)**. Due to the nature of the algorithm ATUs can contain OTUs with different taxonomic annotations (see subsections 3.3.6 and 3.3.5 for details on that manner).

be counted as one entity by taxonomic agglomeration while ATUs show at least two distinct distributional patterns (see figure 3.10).

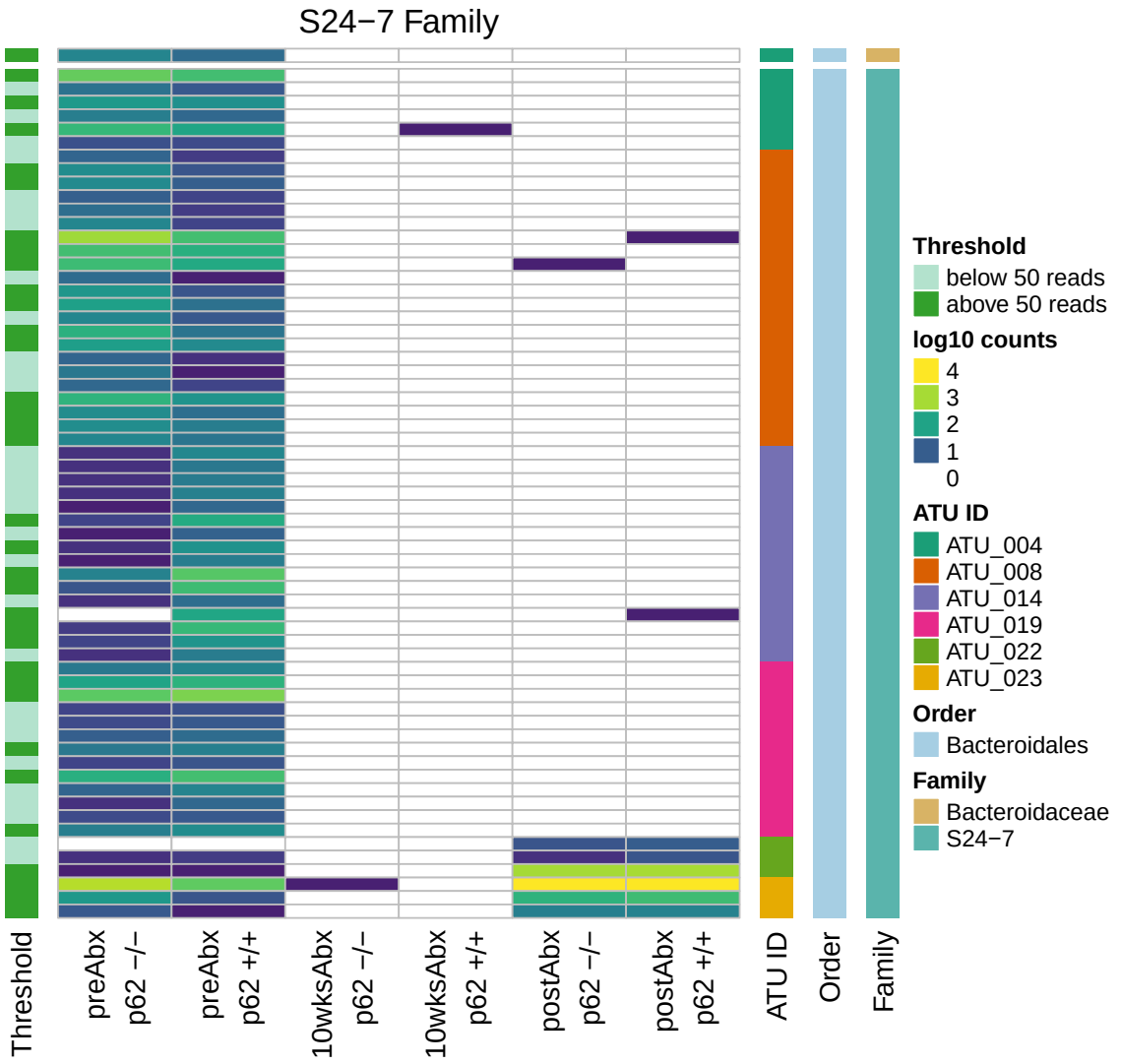

Figure 3.10: Heatmap displaying $\log_{10}$ read counts of different subsets of OTUs (y-axis) for each sample (x-axis) of the antibiotics dataset. Samples are annotated on the bottom according to the time of extraction and their respective genotype. Here only OTUs are shown which are assigned to the *S24-7* candidate family. The column threshold on the left highlights if a OTU on its own would pass (dark green) a read count filtering of 50 overall read counts or not (light green). Additionally, each OTU is annotated with its ATU membership (ATU ID) and by its assigned taxonomy at order and family rank, indicating how these OTUs would be agglomerated by taxonomy. Due to the nature of the algorithm ATUs can contain OTUs with different taxonomic annotations (see subsections 3.3.6 and 3.3.5 for details on that manner).

All three examples of Figure 3.9 highlight the inconsistencies arising from the use of taxonomy counts like incomplete taxonomic assignment or disguised patterns of occurrence. The

use of ATUs rather than taxonomy counts allows for the assessment of these otherwise lost patterns, while preserving information by not relying on incomplete taxonomy. Strikingly, the distributional patterns within the different ATUs are consistent.

Furthermore, due to the agglomeration of similar distributed OTUs less features and reads are lost as consequence of read count filtering (see also Fig. 3.8 A and B - as indicated by dashed lines around the facets) and therefore more information is preserved. For the antibiotics subset (263 OTUs) this effect is also illustrated in the right column of table 3.2. After utilizing HAM only 219 read counts (0.44%) would be lost in total due read count filtering (minimum of 50 reads overall), compared to 3473 reads (6.9%) on OTU count level.

Most strikingly, the amalgamation of OTUs into ATUs by HAM preserved several low abundant entities, which otherwise would have been ignored (see Figures 3.8, 3.9 and 3.10). These entities were too small by its own to be considered differentially abundant between the different time points, but accumulate huge evidence when considered as part of ATUs.

3.4.1.4 Diversity estimates based on ATUs perform closer to expectancy by experimental design

In typical microbiome studies α- and β-diversity are estimated after successful OTU building/picking. However, as many of the used measures are dependent from the number of species in a sample, merging OTUs together and therefore reducing the complexity of microbiome count data would also impact several α- and β-diversity measures. We therefore decided to investigate the effect of ATU-binning on the three most common diversity measures: the number of observed species, the Shannon Index and the Simpson's Index of Diversity. Furthermore, we compared the resulting measures with those derived from plain OTUs and OTUs after post-clustering curation with LULU [189]. This particular method was chosen for comparison, as it is the most related methodology to HAM that we are currently aware of.

All calculated α-diversity measures for the antibiotics subset are shown for each sample and underlying count resolution in figure 3.11. The different sub-figures A, B C and D illustrate observed species, Simpson's Index of Diversity, Shannon index and Shannon equitability (evenness), respectively, on the y-axis for each sample on the x-axis. Diversity measures calculated on plain OTUs are indicated by light green, ATUs by light blue or on OTUs curated by LULU by dark blue bars.

Observed species based on OTUs show significantly higher values for each sample compared against those based on ATUs and LULU curated OTUs, which are both roughly six times lower. However, the estimates based on HAM and LULU are very similar for this comparison

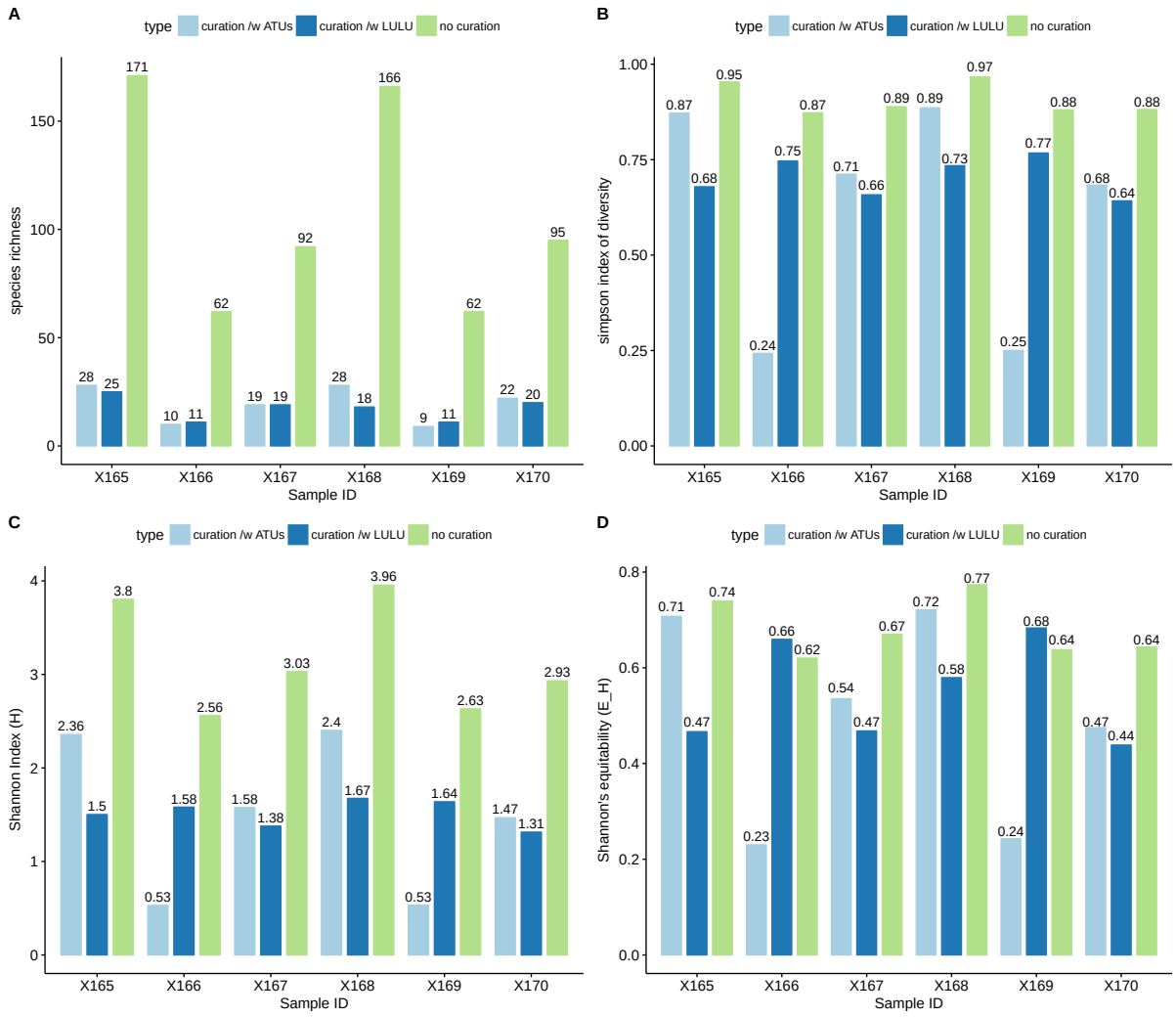

Figure 3.11: Comparison of species richness (y-axis, A), Simpson's Index of Diversity (y-axis, B) and Shannon Index (y-axis, C) per sample (x-axis) on the 263 OTUs subset of the antibiotics dataset. All diversity measures are shown after curation by HAM (light blue) and LULU (dark blue) versus those calculated based on OTUs (light green). Values of the bars are printed above each bar. For each shown measure higher values indicate higher diversity. While A and C are bound by positive infinity, B and D are bound by 0-1.

(see figure 3.11 A). Ignoring the differences in magnitude, all three counting methods show a similar pattern over all samples regarding observed species. Decrease of observed species whilst antibiotic treatment and increase hereafter (four weeks after antibiotics).

Simpson's Index of diversity stays nearly stable if calculated on plain OTUs or LULU curated ones (see figure 3.11 B). However, both methods show fluctuations in the measure facing different directions regarding the treatment samples: a marginal decrease in diversity after onset of treatment for OTUs and a marginal increase in diversity for LULU curated OTUs. Calculations based on ATUs still show the expected antibiotics effect accompanied by a harsh decrease in diversity during treatment, followed by a reconstitution hereafter.

In figure 3.11 C both OTUs and ATUs show a similar pattern in Shannon Index compared to Simpson's Index of diversity based on ATUs in figure 3.11 B. However, for LULU curated OTUs no significant decline in Shannon Index can be observed during antibiosis.

The forth plot in figure 3.11 (see D) shows the Shannon equitability as a measure of evenness. Values reaching 1 mean that abundances are evenly distributed over all residing species and values reaching 0 indicate the opposite. Both OTU and ATU counts seem to be more evenly distributed in samples before and after treatment, compared to the samples during antibiotic treatment. However, ATU counts being less evenly distributed in comparison to OTU counts, especially during antibiosis (see samples X166 and X169). Comparably, this effect seems to be the other way around with LULU curated OTU counts, being more evenly distributed whilst treatment, compared to before and after it.

The diversity calculations were repeated on the full data set (1095 OTUs) and are illustrated in figure 3.12. The only observable difference compared to calculations on the subset was that ATUs based observed species are significantly higher than those derived from LULU curated OTUs. All other effects observed in the subset (see figure 3.11) can also be observed in the full data set (see figure 3.12). A comparison of membership between both methods is illustrated as alluvial diagram in supplemental figure B1.

Regarding β-diversity (i.e. bray curtis), no significant differences between methods could be observed via ordination analysis (stratified by time point of treatment, shown in supplemental Figure B2). However, this effect was expected based on other studies which showed that collapsing of closely related OTUs does not effect bray-curtis distance [216]. The only exception for this data set was agglomeration to species rank, which again was obvious as only a small portion of OTUs have proper species annotation in this data set.

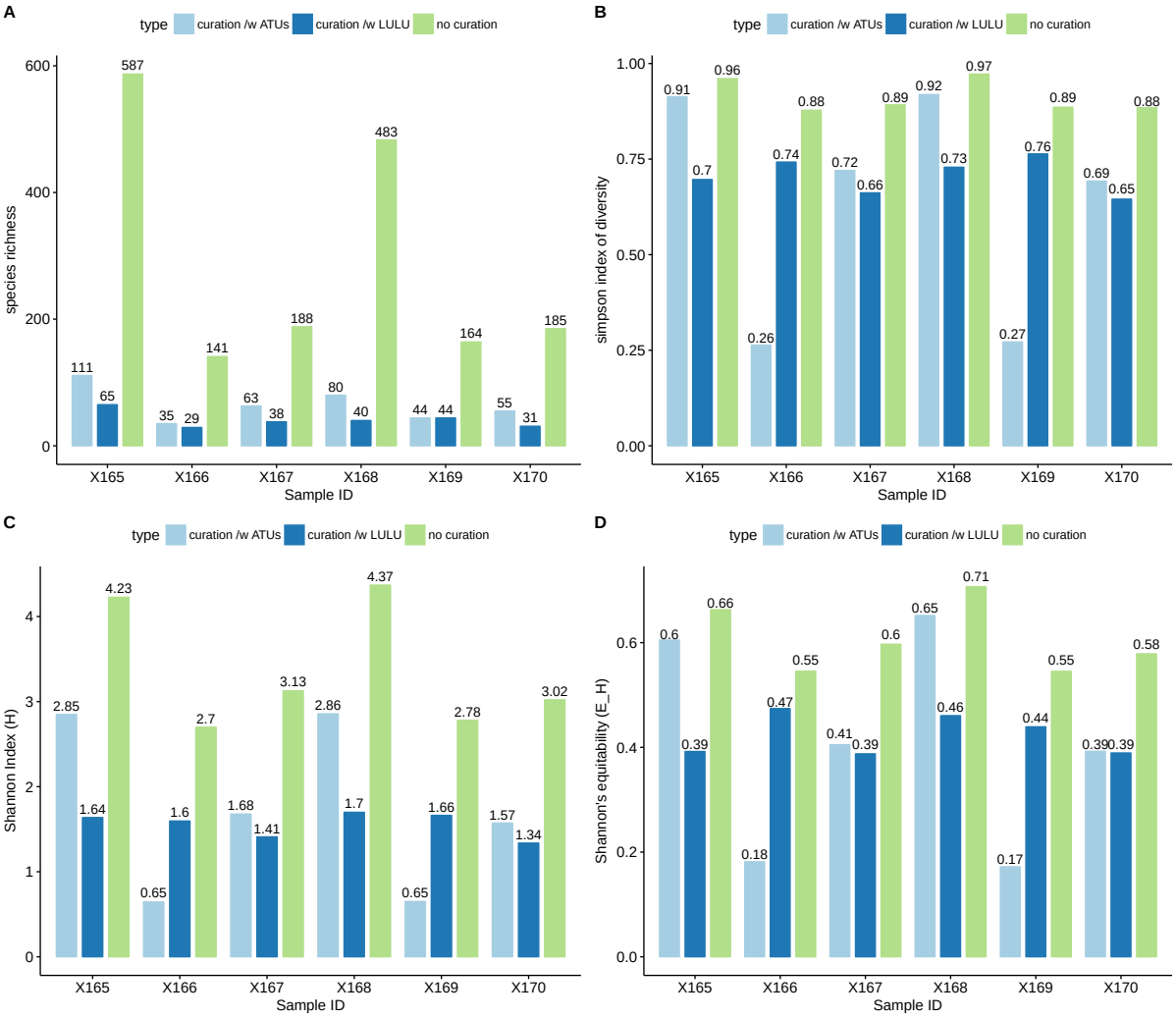

Figure 3.12: Comparison of species richness (observed species, y-axis) per sample (x-axis) on 1095 OTUs of the antibiotics dataset (full dataset). Compared are observed species after curation by HAM (light blue) and LULU (dark blue) versus those calculated based on OTUs (light green). Values of the bars are printed above each bar. For each shown measure higher values indicate higher diversity. While A and C are bound by positive infinity, B and D are bound by 0-1.

3.5 Chapter Discussion

In recent years researchers interest in the gut microbiome rapidly increased, especially for its involvement in diseases which are linked to metabolic disorders. Several studies analysed the composition and dynamics of the gut microbiome and found associations with a plethora of diseases and disorders, but also links to health and well being. However, a missing consensus on the definition of bacterial species in the domain of high-throughput 16S marker gene sequencing [217], as well as proper handling of the data leave room for improvement.

Sparse read counts in microbiome studies for instance aggravate meaningful statistics [148]. The high complexity of microbiome related count data has been shown to distort statistical analysis like biomarker discovery and differential abundance analysis. While many techniques have been proposed to handle this, including less error-prone OTU calling algorithms [115, 217–219] or methodologies designed for sparse count data [148, 220, 221], no clear consensus is found in methodology. The probably most naïve approach includes the elimination of OTUs with small read counts. For this approach an arbitrarily chosen threshold defines what count is considered too low to be included in further analysis. Even though this method removes possible erroneous OTUs, unsurprisingly, it also excludes possible important species just because they show low read counts [126]. Anyhow, the information was measured as part of the study and therefore should not be thrown away that easily just because the low abundances render statistics far more complicated.

Another commonly used approach to overcome sparsity is the agglomeration of OTUs to mutual taxonomic ranks, which is also incorporated in current state-of-the-art microbiome analysis tools [119, 222]. Even though this approach significantly reduces complexity, it also carries three major shortcomings with it. First, taxonomic abundance is highly dependent on existing knowledge about phylogenetic content of the microbiome in question. Depending on the investigated cavity even state-of-the-art reference databases do not sufficiently cover all possibly residing organisms. Especially, missing reference genomes for uncultivable or unknown bacteria complicate proper identification. At latest when it comes to taxonomic annotation in order to give meaning to the clusters in question, even de-novo clustering approaches, which are favoured over reference based methods, are somewhat dependent on high coverage taxonomic databases. Despite joint efforts like the human microbiome project and other consortia gradually increased the coverage of several reference databases in recent years, they still remain far from being complete. Besides this, it has been shown, that species delineation based on 16S rRNA sequences might be inconclusive [223].

Secondly, taxonomy is often assigned based on a fixed similarity threshold towards the taxonomic reference, while an appropriate cut off might differ between families or genera [217]. Third and

more strikingly, read count agglomeration by taxonomy can disguise true effects of otherwise ecologically different sub-populations. Hence, even though this methodology reduces sparseness and complexity in the data, it only exchanges this statistical problem for another by lumping possible distinct entities together, which disguises underlying biology.

In order to provide a proof-of-principle for above mentioned shortcomings, I chose an experimental setting which offers great differences between the inspected samples in regards of diversity and distribution: a perturbation experiment including antibiotic treatment. For this purpose mice were screened before, whilst and after receiving a ten week course of antibiotic treatment. For this data set I was able to show, that both OTU- and taxonomy-counts suffer from the aforementioned shortcomings. While OTU-counts showed high complexity, sparseness and redundancy among OTUs, taxonomy-wise agglomeration disguised ecological structures in the community. Especially the latter created conflicting ecological patterns compared to inspecting plain OTUs, as shown for the families of *Lachnospiraceae*, *Enterobacteriaceae* and the genus of *Lactobacillus*. Choosing one of these methods therefore posed a trade-off between complexity and resolution, offering opposing interpretations of ecology.

Application of the newly proposed concept of ATUs on the same data set, resulted in less sparse count tables compared to OTUs. Because of the enrichment far fewer read counts would be lost during read count filtering (e.g. 4.8% versus 22.19% of reads for ATUs and OTUs, respectively, for a threshold of 200 reads overall). Instead of ignoring this large portion of reads the algorithm agglomerated the information into hubs of ecologically similar performing entities. Most notably, many OTUs were too small to be considered insightful, but as soon as these OTUs were amalgamated into ATUs they revealed impressive evidence. Evidence which would have been un-detected when only investigating OTUs.

Unsurprisingly, the merging into ATUs also resulted in a substantial reduction of complexity from initially 1095 OTUs down to 202 entities, consisting of 116 ATUs and 86 singletons (i.e. unmerged OTUs). For a pre-filtered OTU-table (at least 10 read counts over all samples) containing 263 OTUs, HAM identified 31 ATUs and 4 singletons.

Additionally, taxonomic independence of ATUs ensured the conservation of several insufficiently annotated OTUs, while agglomeration by taxonomy would be restricted to the annotation status of the underlying OTUs. For the antibiotics data set an agglomeration by genus rank already rendered 62% of 1093 OTUs unidentifiable by the means of taxonomy. Depending on how these insufficient labelled OTUs are handled, this could correspond to a loss of 66% (34949) of total read counts.

More strikingly, we showed that while agglomeration by taxonomy disguised opposing distributional patterns, which could be clearly observed on OTU level, ATUs conserved most

of these patterns. This effect was especially apparent for the families of *Lachnospiraceae*, *Enterobacteriaceae* and the genus of *Lactobacillus* in the given data set. Similarly, the candidate family of *S24-7* in the order of *Bacteroidales* exhibited at least two opposing patterns by ATUs, while taxonomic agglomeration would have resulted in one single entity, suggesting no difference between the sub-populations. Notably, according to literature, the aforementioned clades contain many ecologically distinct species, which have been associated with a highly diverse phenotype, affecting both, gut health and disease [145, 224–229]. Therefore agglomeration by taxonomy would not be the appropriate resolution for microbiome analysis on these clades. On the other hand, ATUs derived by HAM provided an increased resolution for bacterial families which are hard to be separated by taxonomy alone, as been shown by this proof-of-principle experiment.

Several studies have shown that diversity estimates based on OTUs tend to overestimate [189, 219, 230]. Considering the antibiotics data set we observed the same effect when comparing species richness calculated on raw OTUs. Both, ATUs and curated OTUs suggested just a sixth of the observed species. Albeit, all three counting methods captured the expected trend in observed species. Regarding species diversity, one might intuitively expect, that a longer course of antibiotic treatment would drastically decrease the microbial diversity of the gut microbiome. ATU counts showed this general trend in both diversity indices, Shannon Index and Simpson's Index of diversity. Meanwhile, OTUs mostly showed this trend for the Shannon Index. However, throughout all measures, ATUs result in lower diversity estimates than suggested by OTUs. Species evenness for both, raw OTU and ATU count data, drops during antibiotic treatment and slightly recovers after washout. However, the effect seems more prominent based on ATUs. In contrast, evenness based on curated OTUs suggests more evenly distributed species during antibiosis, compared to before and after treatment. The opposing results observed for LULU curated OTUs might be explained by differences the agglomeration method between both algorithms. LULU uses co-occurrence patters based purely on presence and absence. This measure is far less effective on small data sets (i.e. low number of samples), which after all might create more erroneous merges than expected. In comparison, HAM tries to cluster OTUs which are similar in regards of distribution and sequence similarity (Levenshtein) to identify hubs of ecological similar performing OTUs. These differences could lead to distinct distributional agglomerations, which again would impede evenness and therefore also Simpson's Index of Diversity. Only regarding observed species, both HAM and LULU yield concordant results. Overall, we showed that diversity estimation based on ATUs does not create opposing results compared to diversity estimation based on OTUs. However, ATUs do result in lower diversity estimates and suggest a more drastic decrease in species evenness.

The general motivation for microbiome community analysis is to find association between bacterial species and a medical condition. This leads to the notion, that distributional patterns

might be more informative than taxonomy. As many species are interchangeable by other species performing the same task, its reasonable to assume, that these bacteria will share similar traits in an ecological manner. At this point it has to be clarified, that ATUs are no proxies for bacterial species in the classical sense of OTUs, rather than proxies for ecologically similar performing entities or patterns. By determining the optimal resolution in the hierarchy ATUs balance biological or clinical requirements of taxonomic resolution with statistical needs for sufficiently high microbial counts. Generally, HAM can be applied besides standard OTU and taxonomy based microbiome community analysis without the need to change experimental designs or standard protocols. The resulting additional insights could guide hypothesis finding and the design of validation experiments, especially in perturbation studies or clinical settings.

Chapter 4

Conclusion and outlook

Over the past decade a plethora of studies have pushed boundaries and expectations by identifying a vast number of associations of the human gut microbiome in health and disease. In that context marker based microbiome community analysis has proven itself as a formidable tool for hypothesis building. While several of the proposed associations have been proven by well planned follow up experiments, for others controversial results have been reported. Especially, the compositional nature of microbiome profiles conflicts with many of the currently available analysis strategies, introducing spurious correlations, over- or wrongly estimated fold changes and as a consequence ill defined hypothesis. These effects are further amplified by differences in microbial load (i.e. the total amount of bacterial material) between different samples. These differences often arise as a consequence of treatment, disease or technical variations. They are especially apparent in patients which show a compromised immune system or receive antibiotics, which complicates comparison with control groups, showing a higher microbial load. In classical protocols for 16S rRNA amplicon sequencing these differences are often levelled by saturation effects during PCR, which is performed to increase the meta-genomical content of a sample prior to amplicon sequencing. Because many of the utilized normalization strategies in 16S rRNA gene based microbiome community analysis are performed in-silico, they are often not able to fully capture differences in microbial load between samples. The presented methodology SCML offers control for differences in microbial load by using fixed concentrations of exogenous whole-cell spike-ins. This was shown with the reconstruction of expected fold changes in a thoroughly designed dilution experiment and furthermore utilized in a clinical setting involving disturbance of the gut microbiome. Fold changes derived on raw counts or relative abundances were significantly overestimated compared to those calculated on SCML calibrated read counts, which further highlights the importance to correct for these kind of sample differences to enhance microbiome community analysis.

Similarly, alternative strategies have been proposed to account for differences in microbial load. These strategies utilize purified exogenous genomic DNA [231] or synthetic DNA [232] as spike-ins. While both methods generally allow more control over the number of spike-templates added to each sample, in-silico designed synthetic DNA spike-ins additionally enable improved traceability of the resulting sequences. However, as both spike-ins are added after cell lysis, they do not offer additional control for differences in lysis efficiency compared to the use of exogenous whole-cell spike-ins as used in SCML. Microbiome community analysis needs to control for differences in microbial load, especially in clinical settings where disturbances of the microbial community would be expected and directly impact decisions on treatment.

Besides technological or physiological differences between samples, also different experimental handling have been found to affect the results of 16S rRNA based microbiome community analysis [233]. Furthermore, the level of taxonomic resolution at which counts are agglomerated or counted does impact the outcome of such analysis. Generally, to identify bacteria, sequencing reads are clustered into taxonomic units to account for sequencing errors and biological variability in the 16S rRNA gene. The resulting OTUs are the highest available resolution above pure read counts and serve as a proxy for bacterial species. While the fine-grained resolution of OTUs results in the possibility to observe more subtle changes in patterns, this increased resolution comes with the cost of higher complexity. Consequently, the found OTU-count tables exhibit a high degree of redundancy and are often sparsely populated, which significantly impedes proper statistical analysis. Alternatively, OTUs are often agglomerated to a shared taxonomy. While this results in more densely populated count tables (i.e. higher counts and reduced number of entities) it also obscures ecological differences that only affect sub-units. Choosing one of these resolutions therefore always comes with a trade-off which could produce opposing results.

With the introduction of the concept of ATUs an additional perspective on microbiome profiles is offered. The underlying algorithm HAM allows for the identification of ecological and phylogenetic consistent hubs in microbiome profiles by enabling varying count resolutions in the same community. For this purpose, the algorithm hierarchically merges existing OTUs guided by a pairwise affinity-score. This score counsels sequence identity, count distributional similarity and entity size to guide its decision which OTUs shall be merged. Like other hierarchical algorithms, HAM first defines a full hierarchy on all entities (i.e. OTUs) and subsequently identifies the optimal cutting point for the given hierarchy. The resulting ATUs are more resilient against loss of information due to read count filtering and therefore preserve more information from the experiment than OTU or taxonomy-based strategies. Moreover, ATUs conserve treatment specific ecology related patterns, while drastically reducing complexity and sparseness of the count table instance. Meanwhile, ATUs are independent of underlying taxonomy and can therefore also preserve ecological patterns even if a proper taxonomy is missing.

Compared to other methods which incorporate distributional patterns, HAM counsels distributional similarity, sequence similarity and entity size coequally, not serially [115, 189, 234]. Additionally, HAM finds a complete hierarchy, offering a informative structure unlike a definite clustering. By determining the optimal granularity for the given hierarchy, a balance between experimental requirements on taxonomic resolution with statistical needs for sufficiently enriched microbial counts. Besides this, it has to be clarified, that ATUs derived from HAM cannot be considered as proxies of bacterial species in the classical sense of OTUs, rather than proxies for ecological-consistent performing species.

While OTUs still offer the most pragmatic approach to investigate the community composition of microbiomes, recent discussions suggested to abandon the concept of OTUs in the favour of counting exact sequence variants (ESVs) [128, 191, 235, 236]. However, a shift towards ESVs would drastically increase the already overwhelming complexity and sparsity of microbiome related count data. Independently from the outcome of this discussion, I suggest that approaches like the here proposed HAM should be applied to microbiome count data to balance the clinical need for taxonomic distinction and phenotypical or ecological abstraction, offering an excellent starting point to form association-based hypothesis and guide future experiments.

Glossary

Allogeneic stem cell transplantation

"A procedure in which a person receives blood-forming stem cells (cells from which all blood cells develop) from a genetically similar, but not identical, donor. This is often a sister or brother, but could be an unrelated donor." [237] 17, I

Operational taxonomic unit

Cluster of read counts which share at least a defined sequence identity (usually 97%). The most abundant sequence inside this cluster determines the representative sequence and therefore also the taxonomical assignment. xi, xii, 35, I

Appendices

Appendix A - Supplemental files for chapter 2

Table A1: Species-specific and total 16S rDNA copies as measured by quantitative realtime-PCR for the dilution experiment. For the original table see Additional File 1 of Stämmler *et al.* (2016) [1].

Sample /	*A. acidiphilus* 16S rDNA copies per sample	*S. ruber* 16S rDNA copies per sample	*R. radiobacter* 16S rDNA copies per sample	Total 16S rDNA copies per sample
65	1.05E+07	7.75E+08	8.35E+09	5.50E+10
66	2.59E+07	8.60E+08	2.53E+09	3.47E+10
67	9.85E+07	7.35E+08	6.75E+08	1.82E+10
68	3.80E+08	8.10E+08	3.34E+08	1.45E+10
69	5.50E+08	4.85E+08	9.05E+07	6.40E+09
70	3.17E+09	8.40E+08	2.61E+07	1.59E+10
71	2.88E+07	1.02E+09	2.63E+09	6.05E+10
72	8.45E+07	6.10E+08	5.45E+08	2.82E+10
73	2.44E+08	6.65E+08	1.59E+08	1.80E+10
74	1.12E+09	8.70E+08	1.55E+08	1.61E+10
75	1.85E+09	6.30E+08	2.41E+07	1.05E+10
76	4.40E+06	3.79E+08	4.27E+09	1.52E+10
77	2.93E+07	2.97E+08	3.74E+08	1.55E+10
78	4.26E+08	1.14E+09	2.01E+08	2.83E+10
79	1.08E+09	1.10E+09	8.85E+07	2.12E+10
80	3.12E+09	9.50E+08	2.77E+07	1.92E+10
81	8.90E+06	7.40E+08	7.30E+09	2.29E+10
82	2.64E+07	8.50E+08	4.37E+09	1.36E+10
83	2.55E+08	9.30E+08	1.55E+08	3.48E+10
84	1.32E+09	1.01E+09	8.85E+07	3.57E+10
85	3.14E+09	9.60E+08	2.77E+07	2.32E+10
86	8.10E+06	6.65E+08	7.25E+09	2.66E+10
87	2.93E+07	8.45E+08	3.07E+09	1.64E+10
88	1.03E+08	8.15E+08	7.15E+08	8.90E+09
89	1.03E+09	1.04E+09	8.25E+07	4.32E+10
90	3.08E+09	9.05E+08	3.52E+07	3.13E+10
91	9.85E+06	7.65E+08	9.65E+09	3.47E+10
92	2.69E+07	8.40E+08	2.99E+09	2.24E+10
93	6.75E+07	6.60E+08	7.30E+08	7.60E+09
94	3.85E+08	9.10E+08	2.62E+08	9.35E+09
95	3.44E+09	1.09E+09	2.98E+07	5.80E+10
96	5.60E+06	5.50E+08	6.00E+09	2.88E+10
97	2.62E+07	7.90E+08	3.51E+09	2.17E+10
98	1.05E+08	9.70E+08	8.85E+08	1.59E+10
99	3.82E+08	8.30E+08	3.91E+08	1.19E+10
100	1.22E+09	1.00E+09	1.14E+08	1.21E+10
101	neg	neg	neg	3.76E+10
102	3.05E+07	6.90E+08	3.10E+09	9.50E+09

Table A2: Design of dilution experiment. 2 x 100 mg cecum contents were collected from three C57BL/6J mice, immediately suspended in 1 ml of PBS, pooled, adjusted with PBS to a total volume of 4 ml and divided in 7 aliquots of 550 µl each. Spike ref., *Salinibacter ruber*; spike 1, *Alicyclobacillus acidiphilus*; spike 2, *Rhizobium radiobacter*. For the original table see Additional File 2 of Stämmler *et al.* (2016) [1].

Aliquot 1	**sample 65**	**sample 66**	**sample 67**	**sample 68**	**sample 69**	**sample 70**
stool dilution (mass)	1:1,00 (39,45 mg)	1:2,15 (18,34 mg)	1:3,75 (10,53 mg)	1:6,53 (6,04 mg)	1:11,37 (3,47 mg)	1:19,82 (1,99 mg)
16S rRNA copies spike ref	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
16S rRNA copies spike 1	1.00E+07	3.00E+07	9.00E+07	2.70E+08	8.10E+08	2.43E+09
16S rRNA copies spike 2	2.43E+09	8.10E+08	2.70E+08	9.00E+07	3.00E+07	1.00E+07
total 16S rRNA copies spiked	2.54E+09	9.40E+08	4.60E+08	4.60E+08	9.40E+08	2.54E+09
total 16S copies (qRT-PCR)	5.50E+10	3.47E+10	1.82E+10	1.45E+10	6.40E+09	1.59E+10

Aliquot 2	**sample 71**	**sample 72**	**sample 73**	**sample 74**	**sample 75**	**sample 76**
stool dilution (mass)	1:1,00 (39,45 mg)	1:2,15 (18,34 mg)	1:3,75 (10,53 mg)	1:6,53 (6,04 mg)	1:11,37 (3,47 mg)	1:19,82 (1,99 mg)
16S rRNA copies spike ref	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
16S rRNA copies spike 1	3.00E+07	9.00E+07	2.70E+08	8.10E+08	2.43E+09	1.00E+07
16S rRNA copies spike 2	8.10E+08	2.70E+08	9.00E+07	3.00E+07	1.00E+07	2.43E+09
total 16S rRNA copies spiked	9.40E+08	4.60E+08	4.60E+08	9.40E+08	2.54E+09	2.54E+09
total 16S copies (qRT-PCR)	6.05E+10	2.82E+10	1.80E+10	1.61E+10	1.05E+10	1.52E+10

Aliquot 3	**sample 77**	**sample 78**	**sample 79**	**sample 80**	**sample 81**	**sample 82**
stool dilution (mass)	1:1,00 (39,45 mg)	1:2,15 (18,34 mg)	1:3,75 (10,53 mg)	1:6,53 (6,04 mg)	1:11,37 (3,47 mg)	1:19,82 (1,99 mg)
16S rRNA copies spike ref	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
16S rRNA copies spike 1	9.00E+07	2.70E+08	8.10E+08	2.43E+09	1.00E+07	3.00E+07
16S rRNA copies spike 2	2.70E+08	9.00E+07	3.00E+07	1.00E+07	2.43E+09	8.10E+08
total 16S rRNA copies spiked	4.60E+08	4.60E+08	9.40E+08	2.54E+09	2.54E+09	9.40E+08
total 16S copies (qRT-PCR)	1.55E+10	2.83E+10	2.12E+10	1.92E+10	2.29E+10	1.36E+10

Aliquot 4	**sample 83**	**sample 84**	**sample 85**	**sample 86**	**sample 87**	**sample 88**
stool dilution (mass)	1:1,00 (39,45 mg)	1:2,15 (18,34 mg)	1:3,75 (10,53 mg)	1:6,53 (6,04 mg)	1:11,37 (3,47 mg)	1:19,82 (1,99 mg)
16S rRNA copies spike ref	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
16S rRNA copies spike 1	2.70E+08	8.10E+08	2.43E+09	1.00E+07	3.00E+07	9.00E+07
16S rRNA copies spike 2	9.00E+07	3.00E+07	1.00E+07	2.43E+09	8.10E+08	2.70E+08
total 16S rRNA copies spiked	4.60E+08	9.40E+08	2.54E+09	2.54E+09	9.40E+08	4.60E+08
total 16S copies (qRT-PCR)	3.48E+10	3.57E+10	2.32E+10	2.66E+10	1.64E+10	8.90E+09

Aliquot 5	**sample 89**	**sample 90**	**sample 91**	**sample 92**	**sample 93**	**sample 94**
stool dilution (mass)	1:1,00 (39,45 mg)	1:2,15 (18,34 mg)	1:3,75 (10,53 mg)	1:6,53 (6,04 mg)	1:11,37 (3,47 mg)	1:19,82 (1,99 mg)
16S rRNA copies spike ref	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
16S rRNA copies spike 1	8.10E+08	2.43E+09	1.00E+07	3.00E+07	9.00E+07	2.70E+08
16S rRNA copies spike 2	3.00E+07	1.00E+07	2.43E+09	8.10E+08	2.70E+08	9.00E+07
total 16S rRNA copies spiked	9.40E+08	2.54E+09	2.54E+09	9.40E+08	4.60E+08	4.60E+08
total 16S copies (qRT-PCR)	4.32E+10	3.13E+10	3.47E+10	2.24E+10	7.60E+09	9.35E+09

Aliquot 6	**sample 95**	**sample 96**	**sample 97**	**sample 98**	**sample 99**	**sample 100**
stool dilution (mass)	1:1,00 (39,45 mg)	1:2,15 (18,34 mg)	1:3,75 (10,53 mg)	1:6,53 (6,04 mg)	1:11,37 (3,47 mg)	1:19,82 (1,99 mg)
16S rRNA copies spike ref	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
16S rRNA copies spike 1	2.43E+09	1.00E+07	3.00E+07	9.00E+07	2.70E+08	8.10E+08
16S rRNA copies spike 2	1.00E+07	2.43E+09	8.10E+08	2.70E+08	9.00E+07	3.00E+07
total 16S rRNA copies spiked	2.54E+09	2.54E+09	9.40E+08	4.60E+08	4.60E+08	9.40E+08
total 16S copies (qRT-PCR)	5.80E+10	2.88E+10	2.17E+10	1.59E+10	1.19E+10	1.21E+10

Aliquot 7	**sample 101**
stool dilution (mass)	1:1,00 (39,45 mg)
16S rRNA copies spike ref	∅
16S rRNA copies spike 1	∅
16S rRNA copies spike 2	∅
total 16S copies (qRT-PCR)	3.76E+10

	sample 102
stool dilution (mass)	∅
16S rRNA copies spike ref	1.00E+08
16S rRNA copies spike 1	3.00E+07
16S rRNA copies spike 2	8.10E+08
total 16S rRNA copies spiked	9.40E+08
total 16S copies (qRT-PCR)	9.50E+09

	Pool 1	**Pool 2**	**Pool 3**	**Pool 4**	**Pool 5**	**Pool 6**	**no spike-in**
S. ruber	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	∅
A. acidiphilus	1.00E+07	3.00E+07	9.00E+07	2.70E+08	8.10E+08	2.43E+09	∅
R. radiobacter	2.43E+09	8.10E+08	2.70E+08	9.00E+07	3.00E+07	1.00E+07	∅

Sample 65	Sample 66	Sample 67	Sample 68	Sample 69	Sample 70	Sample 101
Sample 76	Sample 71	Sample 72	Sample 73	Sample 74	Sample 75	
Sample 81	Sample 82	Sample 77	Sample 78	Sample 79	Sample 80	
Sample 86	Sample 87	Sample 88	Sample 83	Sample 84	Sample 85	
Sample 91	Sample 92	Sample 93	Sample 94	Sample 89	Sample 90	
Sample 96	Sample 97	Sample 98	Sample 99	Sample 100	Sample 95	
	Sample 102					

Table A3: Six pools of bacterial mock communities containing *Alicyclobacillus acidiphilus*, *Rhizobium radiobacter* and *Salinibacter ruber* as used for spike-in in the dilution experiment. For the original table see Additional File 3 of Stämmler *et al.* (2016) [1].

Name	DNA-Sequence (5'-3')	E. coli 16S rDNA Nucleotide Position	Purpose of use	Reference
GM3F	agagtttgatcmtggc	8	Spike-in Quantification standards	Klindworth et al., 2013 [102]
1492R	taccttgttacgactt	1492	Spike-in Quantification standards	Klindworth et al., 2013 [102]
764F	caaacaggattagataccc	764	Quantification of total 16S rDNA copies	Imase et al, 2008 [238]
907R	ccgtcaattcctttragttt	907	Quantification of total 16S rDNA copies	Lane et al., 1991 [239]
341F	CCATCTCATCCCTGCGTGTCTCCGACTCAG<MID>cctacgggaggcagcag	341	454/Pyrosequencing and Total 16S quantification standards	Klindworth et al., 2013 [102]
1061R	CCTATCCCCTGTGTGCCTTGGCAGTCTCAGcrrcacgagctgacgac	1061	454/Pyrosequencing and Total 16S quantification standards	Klindworth et al., 2013 [102]
Aacidi-238TM	6FAM-agctagttggtgaggtaacggcccacc-BBQ	238	Quantification of 16S rDNA copies A. acidiphilus	This study [1]
Aacidi-193F	gaggaaagttgcaaatgcaaca	193	Quantification of 16S rDNA copies A. acidiphilus	This study [1]
Aacidi-453R	aggagctttccactctccttat	453	Quantification of 16S rDNA copies A. acidiphilus	This study [1]
Rradio-166TM	LC670-aattaataccgcatacgccctacg-BBQ	166	Quantification of 16S rDNA copies R. radiobacter	This study [1]
Rradio-126-F2	ggaacatacccttttcctgcgg	126	Quantification of 16S rDNA copies R. radiobacter	This study [1]
Rradio-197-R2	gccaatccttccccgataaatc	197	Quantification of 16S rDNA copies R. radiobacter	This study [1]
Salini-180TM	LC640-cacgtcgtctggatcccgcatg-BBQ	180	Quantification of 16S rDNA copies S. ruber	This study [1]
Salini-7F	agagtttgatcatggctcag	7	Quantification of 16S rDNA copies S. ruber	This study [1]
Salini-413R	tacgccccataggggtgt	413	Quantification of 16S rDNA copies S. ruber	Antón et al. [163]

Table A4: Primers and hydrolysis probes, their purpose of use, as well as their references, used in this study. For the original table see Additional File 4 of Stämmler *et al.* (2016) [1].

Table A5: Metadata mapping file for the dilution experiment. Linker primer sequences (CCTACGGGNGGCWGCAG) and reverse primers (crrcacgagctgacgac) were the same for all samples and are therefore omitted from this table. The following meta information was available for each sample: the sample barcode (BarcodeSequence), the pool used for spike-in (Treatment), the corresponding dilution factor (Dilution), a running sample number (Description), as well as the background mass in *mg* (Background). Numbers in he columns Dilution and Background are rounded to two decimal places. For the full meta file as used in QIIME (without rounded values) see Additional File 7 of Stämmler *et al.* (2016) [1].

SampleID	BarcodeSequence	Treatment	Dilution	Description	Background [*mg*]
MID01	ACGAGTGCGT	Pool1	1.00	65	39.45
MID02	ACGCTCGACA	Pool2	2.15	66	18.34
MID03	AGACGCACTC	Pool3	3.75	67	10.53
MID05	ATCAGACACG	Pool4	6.53	68	6.04
MID07	CGTGTCTCTA	Pool5	11.37	69	3.47
MID08	CTCGCGTGTC	Pool6	19.82	70	1.99
MID09	TAGTATCAGC	Pool2	1.00	71	39.45
MID10	TCTCTATGCG	Pool3	2.15	72	18.34
MID11	TGATACGTCT	Pool4	3.75	73	10.53
MID12	TACTGAGCTA	Pool5	6.53	74	6.04
MID13	CATAGTAGTG	Pool6	11.37	75	3.47
MID14	CGAGAGATAC	Pool1	19.82	76	1.99
MID15	ATACGACGTA	Pool3	1.00	77	39.45
MID16	TCACGTACTA	Pool4	2.15	78	18.34
MID17	CGTCTAGTAC	Pool5	3.75	79	10.53
MID31	AGCGTCGTCT	Pool6	6.53	80	6.04
MID19	TGTACTACTC	Pool1	11.37	81	3.47
MID20	ACGACTACAG	Pool2	19.82	82	1.99
MID18	TCTACGTAGC	Pool4	1.00	83	39.45
MID21	CGTAGACTAG	Pool5	2.15	84	18.34
MID22	TACGAGTATG	Pool6	3.75	85	10.53
MID23	TACTCTCGTG	Pool1	6.53	86	6.04
#MID25	TCGTCGCTCG	Pool2	11.37	87	NA
#MID26	ACATACGCGT	Pool3	19.82	88	NA
MID27	ACGCGAGTAT	Pool5	1.00	89	39.45
MID28	ACTACTATGT	Pool6	2.15	90	18.34
MID29	ACTGTACAGT	Pool1	3.75	91	10.53
MID30	AGACTATACT	Pool2	6.53	92	6.04
MID26	ACATACGCGT	Pool3	11.37	93	3.47
MID32	AGTACGCTAT	Pool4	19.82	94	1.99
MID33	ATAGAGTACT	Pool6	1.00	95	39.45
MID34	CACGCTACGT	Pool1	2.15	96	18.34
MID35	CAGTAGACGT	Pool2	3.75	97	10.53
MID36	CGACGTGACT	Pool3	6.53	98	6.04
MID37	TACACACACT	Pool4	11.37	99	3.47
MID38	TACACGTGAT	Pool5	19.82	100	1.99
MID39	TACAGATCGT	noSpike	NA	101	NA
MID40	TACGCTGTCT	SpikeKO	NA	102	NA

Table A6: Metadata mapping file for the ASCT experiment. Linker primer sequences (CCTACGGGNGGCWGCAG) and reverse primers (crrcacgagctgacgac) were the same for all samples and are therefore omitted from this table. The following meta information was available for each sample: the sample barcode (BarcodeSequence), the patient it belongs to (Patient), the time of extraction (Time), as well as a anonymised specimen code (Description). For the full meta file as used in QIIME see Additional File 8 of Stämmler *et al.* (2016) [1].

SampleID	BarcodeSequence	Treatment	Time	Description
MID26	ACATACGCGT	Patient2	preASCT	TT1
MID27	ACGCGAGTAT	Patient1	preASCT	TO1
MID28	ACTACTATGT	Patient1	d0	TO2
MID29	ACTGTACAGT	Patient1	d7	TO3
MID30	AGACTATACT	Patient1	d14	TO4
MID32	AGTACGCTAT	Patient2	d0	TT2
MID33	ATAGAGTACT	Patient2	d7	TT3
MID34	CACGCTACGT	Patient2	d14	TT4
MID35	CAGTAGACGT	Patient3	d0	TX2
MID36	CGACGTGACT	Patient3	d7	TX3
MID37	TACACACACT	Patient3	d14	TX4
MID38	TACACGTGAT	Patient4	preASCT	TZ1
MID39	TACAGATCGT	Patient4	d0	TZ2
MID40	TACGCTGTCT	Patient4	d7	TZ3
MID41	TAGTGTAGAT	Patient5	preASCT	UN1
MID42	TCGATCACGT	Patient5	d0	UN2
MID43	TCGCACTAGT	Patient5	d14	UN4

	Sample 65	**Sample 66**	**Sample 67**	**Sample 68**	**Sample 69**	**Sample 70**
Alicyclobacillus	1.00E+07	3.00E+07	9.00E+07	2.70E+08	8.10E+08	2.43E+09
Salinibacter	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
Rhizobium	2.43E+09	8.10E+08	2.70E+08	9.00E+07	3.00E+07	1.00E+07

	Sample 71	**Sample 72**	**Sample 73**	**Sample 74**	**Sample 75**	**Sample 76**
Alicyclobacillus	3.00E+07	9.00E+07	2.70E+08	8.10E+08	2.43E+09	1.00E+07
Salinibacter	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
Rhizobium	8.10E+08	2.70E+08	9.00E+07	3.00E+07	1.00E+07	2.43E+09

	Sample 77	**Sample 78**	**Sample 79**	**Sample 80**	**Sample 81**	**Sample 82**
Alicyclobacillus	9.00E+07	2.70E+08	8.10E+08	2.43E+09	1.00E+07	3.00E+07
Salinibacter	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
Rhizobium	2.70E+08	9.00E+07	3.00E+07	1.00E+07	2.43E+09	8.10E+08

	Sample 83	**Sample 84**	**Sample 85**	**Sample 86**	**Sample 87**	**Sample 88**
Alicyclobacillus	2.70E+08	8.10E+08	2.43E+09	1.00E+07	3.00E+07	9.00E+07
Salinibacter	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
Rhizobium	9.00E+07	3.00E+07	1.00E+07	2.43E+09	8.10E+08	2.70E+08

	Sample 89	**Sample 90**	**Sample 91**	**Sample 92**	**Sample 93**	**Sample 94**
Alicyclobacillus	8.10E+08	2.43E+09	1.00E+07	3.00E+07	9.00E+07	2.70E+08
Salinibacter	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
Rhizobium	3.00E+07	1.00E+07	2.43E+09	8.10E+08	2.70E+08	9.00E+07

	Sample 95	**Sample 96**	**Sample 97**	**Sample 98**	**Sample 99**	**Sample 100**
Alicyclobacillus	2.43E+09	1.00E+07	3.00E+07	9.00E+07	2.70E+08	8.10E+08
Salinibacter	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08	1.00E+08
Rhizobium	1.00E+07	2.43E+09	8.10E+08	2.70E+08	9.00E+07	3.00E+07

Table A7: Spike-in concentrations by design as used for reproduction. For the original table see Additional File 12 of Stämmler *et al.* (2016) [1].

Appendix B - Supplemental files for chapter 3

Table B1: Experimental meta data of the antibiotics mice dataset used in chapter 3. Shown are the sample identifier (SampleID), the used multiplex barcode (BarcodeSequence), the used reverse primer (ReversePrimer), the genotype of the mice (Genotype), the time point of extraction in the experiment (Time), the dilution factor of uses spike-in controls (SpikeDil), the internal number of the mice, an alternative genotype nomenclature (Type) and the internal mouse identifier (Description) for each of the six samples.

SampleID	BarcodeSequence	ReversePrimer	Genotype	Time	SpikeDil	Mice	Type	Description
165	ACAGTATATA	crrcacgagctgacgac	p62 -/-	preAbx	1:1	34714	KO	MID48
166	ACGCGATCGA	crrcacgagctgacgac	p62 -/-	10wksAbx	1:100	34714	KO	MID49
167	ACTAGCAGTA	crrcacgagctgacgac	p62 -/-	postAbx	1:10	34714	KO	MID50
168	AGCTCACGTA	crrcacgagctgacgac	p62 +/+	preAbx	1:1	34712	WT	MID51
169	AGTATACATA	crrcacgagctgacgac	p62 +/+	10wksAbx	1:100	34712	WT	MID52
170	AGTCGAGAGA	crrcacgagctgacgac	p62 +/+	postAbx	1:10	34712	WT	MID53

Table B2: Antibiotic compounds and their concentration in the drinking water fed to the mice while treatment period. Concentrations are given in milligram per millilitre.

Antibiotic compound	Concentration
Neomycin	0.25 mg/ml
Metronidazole	0.5 mg/ml
Ampicillin	0.5 mg/ml
Vancomycin	0.25 mg/ml

Table B3: Total 16S rRNA copies as measured by qPCR on diluted and undiluted samples. Compared to 16S rRNA amplicon sequencing, qPCR measurements were performed at an additional time point of 28 days into antibiotic treatment (ABT).

	16S-copies in 2 microlitre by qPCR (diluted 1:100)				16S-copies total by qPCR (20 milligram feces)			
Mice number	before ABT	28 days into ABT	56 days into ABT	4 weeks after ABT	before ABT	28 days into ABT	56 days into ABT	4 weeks after ABT
34712	4.3E+07	2.1E+06	1.33E+05	4.28E+06	1.7E+11	8.3E+09	5.31E+08	1.71E+10
34714	7.7E+07	2.9E+06	1.37E+05	6.34E+06	3.1E+11	1.1E+10	5.47E+08	2.53E+10

Table B4: Overview of parameter values chosen for raw sequence denoising by FlowClus. This denoising-pipeline was used to prepare raw sequences from the antibiotics experiment (see chapter 3).

FlowClus parameter settings	
Parameter	**Value**
Min. sequence length	400
Max. sequence length for elimination	800
Max. ambiguous bases allowed	6
Max. homopolymer length allowed	8
Min. average quality score	25
Min. window quality score	length = 50
Min. window quality score	qual = 20
Noisy flow interval	0.50-0.70
Max. flow value	6.49

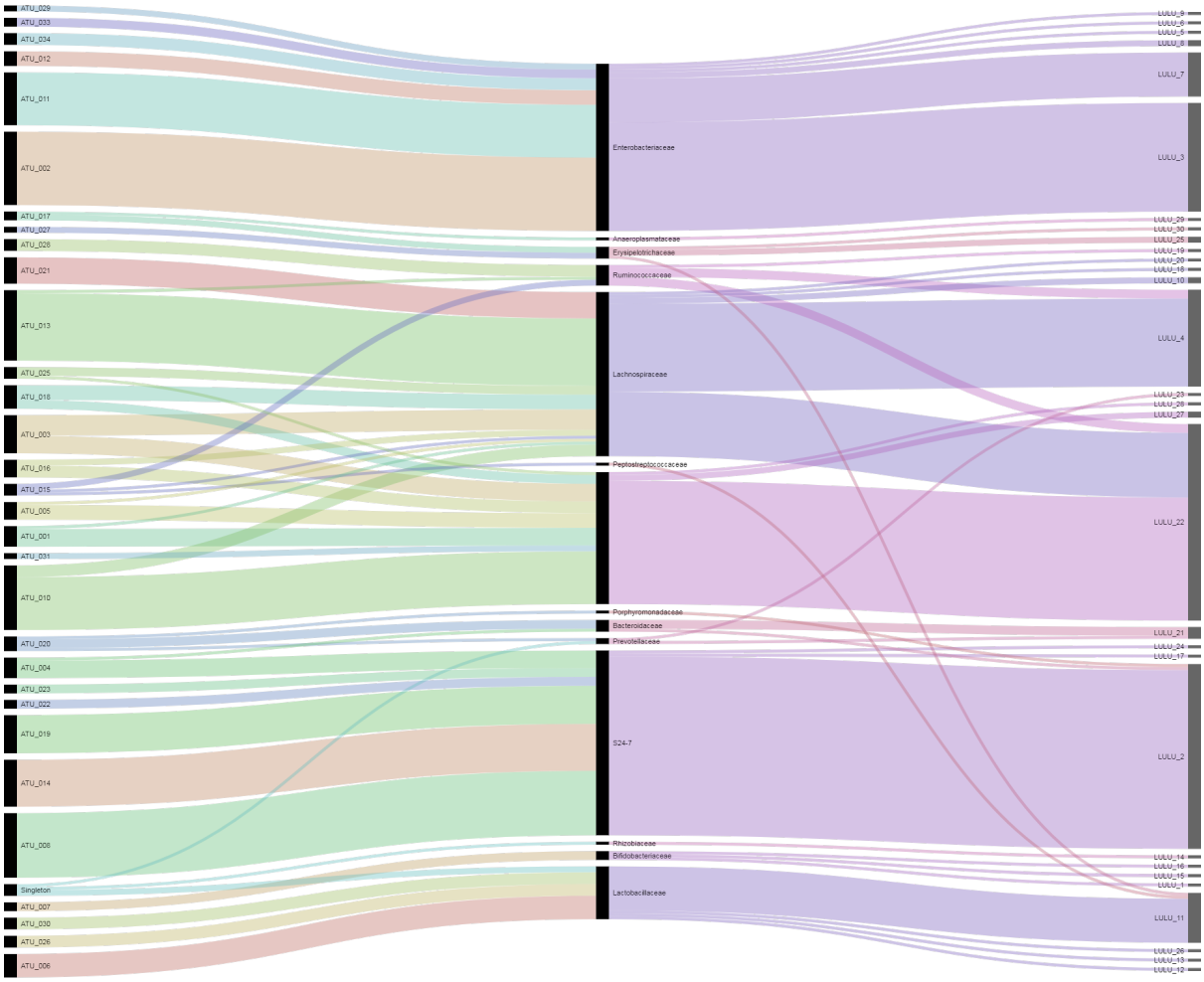

Figure B1: Alluvial diagram visualizing the mapping between ATUs and OTUs curated by LULU connected via the family rank assignment of the underlying OTUs. While the left column represents ATU membership, the right column indicates OTU membership after curation and the middle column shows the corresponding family assignment. The black nodes for each method are scaled according to their number of reads in the corresponding binning. Colours are non-unique and only used for visual traceability. Each node consists of all membership OTUs, therefore each node has as many connections as there are members inside it. Connections are drawn for each OTU over all columns.

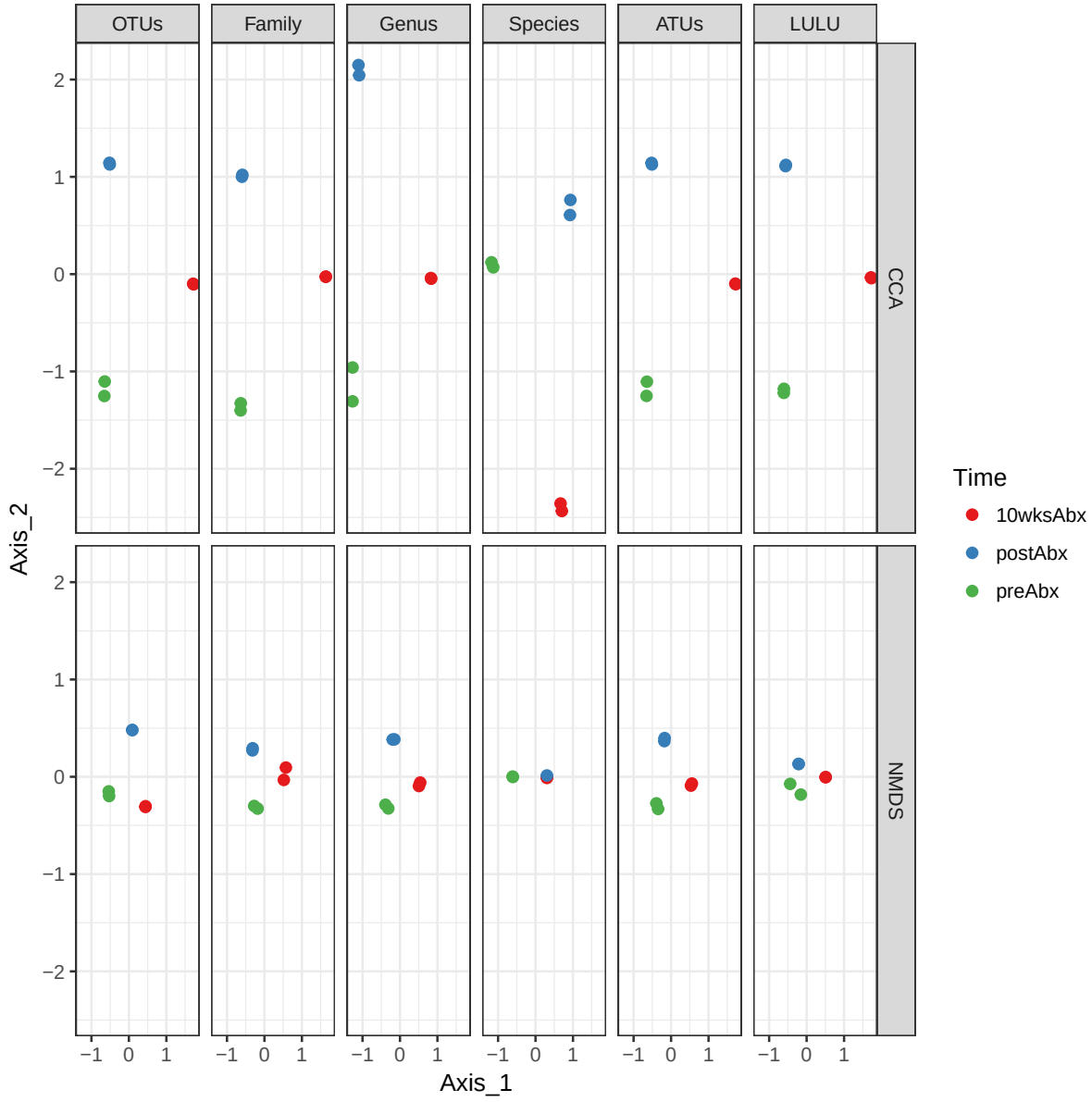

Figure B2: Comparison of separability by bray-curtis distance calculated on OTUs, family-, genus-, species taxonomy, ATUs and OTUs curated by LULU. Each dot represents a sample, while colour indicates the treatment time point. Ordination analysis was performed on canonical correspondence analysis (CCA) and non-metric multidimensional scaling (NMDS) and plotted above each other. Values of the axis are on the same scale, but of different interpretation.

OTU_ID	ATU	Phylum	Class	Order	Family	Genus	Species	Outlier_LV	Outlier_JSD
4422542	ATU_001	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
1110826	ATU_001	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
212758	ATU_001	Firmicutes	Clostridia	Clostridiales				TRUE	TRUE
3401807	ATU_001	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	[Ruminococcus]	gnavus	FALSE	FALSE
832848	ATU_001	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
422727	ATU_001	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
1571092	ATU_001	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
792427	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
296464	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
4378767	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	TRUE
4217230	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
1057636	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
4294723	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
464072	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	TRUE
3988508	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella	pneumoniae	FALSE	FALSE
4473834	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
144814	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
4301368	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
4483809	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
4433832	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
3474127	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
4461298	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
4353951	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
564307	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
4467506	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
104228	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
44635	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
1663575	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
244878	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	TRUE
566134	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
677982	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	TRUE
123487	ATU_002	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE

Table B5: Mapping for each of the 263 OTUs of the antibiotics subset. For each OTU the taxonomy is given, as well as the corresponding ATU membership. The last two columns indicate whether a OTU inside an ATU is considered to be a outlier in terms of JSD or Levensthein-distance (75th quantile).

174848	ATU_003	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
315757	ATU_003	Firmicutes	Clostridia	Clostridiales				FALSE	TRUE
135625	ATU_003	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	[Ruminococcus]	gnavus	TRUE	FALSE
305112	ATU_003	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	TRUE
267090	ATU_003	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4471526	ATU_003	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
270436	ATU_003	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
193064	ATU_003	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
191067	ATU_003	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
188932	ATU_003	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
196533	ATU_003	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
271822	ATU_003	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
327965	ATU_003	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
269340	ATU_004	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
328978	ATU_004	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
3293891	ATU_004	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
274120	ATU_004	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
189778	ATU_004	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
337852	ATU_004	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
311775	ATU_004	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	acidifaciens	TRUE	FALSE
308400	ATU_005	Firmicutes	Clostridia	Clostridiales				FALSE	TRUE
2390340	ATU_005	Firmicutes	Clostridia	Clostridiales				TRUE	FALSE
165430	ATU_005	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
708101	ATU_005	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
745390	ATU_005	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
227635	ATU_005	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
15200	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
166015	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
168412	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
100825	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	TRUE
171781	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
135956	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		TRUE	FALSE
1140886	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
169428	ATU_006	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE

Table B5: (continued)

561483	ATU_007	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	longum	FALSE	FALSE
4432638	ATU_007	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	adolescentis	TRUE	FALSE
72820	ATU_007	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	longum	FALSE	FALSE
260527	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
234464	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
204125	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
234036	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
204547	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
264352	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
207419	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
389282	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
204158	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
3916746	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
431900	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
191841	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			TRUE	FALSE
3172949	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
1802718	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
229882	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
274665	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
371647	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
266976	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
209030	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
204003	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
2212505	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
204171	ATU_008	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE

Table B5: (continued)

193680	ATU_010	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
275123	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
832799	ATU_010	Firmicutes	Clostridia	Clostridiales				TRUE	FALSE
206350	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
339718	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
437151	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
264461	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
321952	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
184903	ATU_010	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	[Ruminococcus]	gnavus	FALSE	FALSE
180466	ATU_010	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
2645483	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
352826	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
309265	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
233059	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
266203	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
4402081	ATU_010	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	[Ruminococcus]	gnavus	TRUE	FALSE
231896	ATU_010	Firmicutes	Clostridia	Clostridiales				TRUE	FALSE
270203	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
1540280	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	TRUE
1105157	ATU_010	Firmicutes	Clostridia	Clostridiales				TRUE	TRUE
330478	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
232878	ATU_010	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE

Table B5: (continued)

1765550	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia	coli	FALSE	FALSE
4337654	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
293541	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
305760	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
300514	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	TRUE
307080	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
231787	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
4374044	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
4474378	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
2235399	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia	coli	FALSE	FALSE
311541	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
296668	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
298307	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
1141665	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia	coli	FALSE	FALSE
4453611	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	hormaechei	TRUE	FALSE
295053	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
4414015	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	asburiae	FALSE	FALSE
302439	ATU_011	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
583479	ATU_012	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
37741	ATU_012	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		TRUE	FALSE
687299	ATU_012	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
2529281	ATU_012	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
2704829	ATU_012	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE

Table B5: (continued)

3493367	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
3493361	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
3995513	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4320576	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
3217871	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4389418	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4434455	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4348109	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	TRUE
4348108	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4348106	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
839979	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			TRUE	FALSE
362576	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			TRUE	FALSE
1707496	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Clostridium	hathewayi	FALSE	TRUE
4015272	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
2841555	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
3409357	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
2818297	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
361285	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			TRUE	FALSE
4422328	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	TRUE
4367066	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4367063	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
2841566	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
3413558	ATU_013	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
186521	ATU_013	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae			TRUE	FALSE

Table B5: (continued)

212170	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
437289	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
259012	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
3916747	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
320819	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			TRUE	FALSE
211820	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
229738	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
179548	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
230816	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
212367	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
231716	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
49714	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
233587	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
38278	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
174573	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
436131	ATU_014	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
1841529	ATU_015	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira		FALSE	FALSE
606927	ATU_015	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	[Clostridium]	difficile	FALSE	TRUE
1139897	ATU_015	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus		FALSE	FALSE
4325509	ATU_015	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Robinsoniella	peoriensis	FALSE	FALSE
351465	ATU_016	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
310479	ATU_016	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
174773	ATU_016	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
167204	ATU_016	Firmicutes	Clostridia	Clostridiales				TRUE	FALSE
195711	ATU_016	Firmicutes	Clostridia	Clostridiales				FALSE	TRUE
328598	ATU_016	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
828435	ATU_017	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae			FALSE	FALSE
4409417	ATU_017	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	[Eubacterium]	dolichum	FALSE	FALSE
215065	ATU_017	Tenericutes	Mollicutes	Anaeroplasmatales	Anaeroplasmataceae	Anaeroplasma		FALSE	FALSE

Table B5: (continued)

2898342	ATU_018	Firmicutes	Clostridia	Clostridiales				FALSE	TRUE
2527302	ATU_018	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
231606	ATU_018	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4393892	ATU_018	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
4364240	ATU_018	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
2233609	ATU_018	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4409381	ATU_018	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
1107755	ATU_018	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
207288	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
229352	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			TRUE	FALSE
355746	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
190752	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
187673	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
330772	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
208402	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
215495	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
265180	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
465757	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			TRUE	FALSE
266997	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
339549	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
320123	ATU_019	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
304047	ATU_020	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	acidifaciens	FALSE	FALSE
4372003	ATU_020	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Parabacteroides		FALSE	FALSE
4484395	ATU_020	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	acidifaciens	FALSE	FALSE
4378740	ATU_020	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella		FALSE	FALSE
4476333	ATU_020	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides		FALSE	TRUE
4415854	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
14069	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus		TRUE	FALSE
4359213	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
3252951	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4087650	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
4387801	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
259772	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus		TRUE	FALSE
4459960	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Clostridium	aldenense	FALSE	TRUE
3736962	ATU_021	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus		FALSE	FALSE

Table B5: (continued)

329771	ATU_022	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
3293892	ATU_022	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
324013	ATU_022	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			TRUE	FALSE
208288	ATU_023	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
263587	ATU_023	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	FALSE
203713	ATU_023	Bacteroidetes	Bacteroidia	Bacteroidales	S24-7			FALSE	TRUE
177956	ATU_025	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			FALSE	FALSE
266343	ATU_025	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
185294	ATU_025	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	[Ruminococcus]	gnavus	FALSE	FALSE
315212	ATU_025	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae			TRUE	FALSE
137043	ATU_026	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
242917	ATU_026	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		TRUE	FALSE
187201	ATU_026	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	reuteri	FALSE	FALSE
290235	ATU_026	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	TRUE
548587	ATU_027	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	[Eubacterium]	dolichum	FALSE	FALSE
4472632	ATU_027	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	[Eubacterium]	dolichum	FALSE	FALSE
260845	ATU_028	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus		TRUE	FALSE
311961	ATU_028	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus		FALSE	FALSE
270519	ATU_028	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira		FALSE	FALSE
176118	ATU_028	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira		FALSE	FALSE
302846	ATU_029	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
290844	ATU_029	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE
4461663	ATU_030	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
170722	ATU_030	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
426769	ATU_030	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
269125	ATU_030	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		FALSE	FALSE
771177	ATU_031	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
276402	ATU_031	Firmicutes	Clostridia	Clostridiales				FALSE	FALSE
3782016	ATU_033	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	TRUE
4364282	ATU_033	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Salmonella	subterranea	FALSE	FALSE
2327459	ATU_033	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	FALSE

Table B5: (continued)

273616	ATU_034	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
810578	ATU_034	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			FALSE	TRUE
3010176	ATU_034	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Klebsiella		FALSE	FALSE
299267	ATU_034	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae			TRUE	FALSE
326923	Singleton	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		NA	NA
1147415	Singleton	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus		NA	NA
279066	Singleton	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella		NA	NA
255931	Singleton	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Agrobacterium		NA	NA

Table B5: (continued)

Note on electronic supplement

This thesis is accompanied by an electronic supplement in form of a DVD. It contains all supplemental files already shown in this appendix, as well as additional files whose format did not allow to be attached to the written form of the thesis. These additional files are needed for reproduction of the results and include

- FASTA files,
- OTU, taxonomy and read count tables,
- code for reproduction of chapters 2 and 3,
- code for the recreation of the figures
- and source code of the R-Package *dOTUClust*.

References

[1] Frank Stämmler et al. "Adjusting microbiome profiles for differences in microbial load by spike-in bacteria". In: *Microbiome* 4.1 (2016), p. 28. ISSN: 2049-2618. DOI: 10.1186/s40168-016-0175-0. URL: http://dx.doi.org/10.1186/s40168-016-0175-0.

[2] CG Ehrenberg. "Symbolae physicae animalia evertebrata exclusis insectis". In: *Symbolae physicae, seu Incones adhue ineditae corporum naturalium novorum aut minus cognitorum, quae ex itinerebus per Libyam, Aegyptum, Nubiam, Dengalum, Syriam, Arabiam et Habessiniam. Pars Zoologica* 4 (1828).

[3] James A. Barnett. "A history of research on yeasts 2: Louis Pasteur and his contemporaries, 1850–1880". In: *Yeast* 16.8 (2000), pp. 755–771. ISSN: 1097-0061. DOI: 10.1002/1097-0061(20000615)16:8<755::AID-YEA587>3.0.CO;2-4. URL: http://dx.doi.org/10.1002/1097-0061(20000615)16:8%3C755::AID-YEA587%3E3.0.CO;2-4.

[4] R Koch. "Untersuchungen ueber Bakterien V. Die Aetiologie der Milzbrand-Krankheit, begruendent auf die Entwicklungsgeschichte des Bacillus Anthracis. Beitr. z. Biol. D. Pflanzen 2: 277-310". In: *Milestones in Microbiology* 1556 (1876).

[5] Louis Pasteur, J Joubert, and C Chamberland. "The germ theory of disease". In: *CR Hebdom Seances Acad Sci* 86 (1878), pp. 1037–1052.

[6] PJ Van Beneden. "Animal Parasites and Messmates." In: *New York. D. Appleton & Co* (1876).

[7] Elie Metchnikoff. *The prolongation of life*. Putnam, 1908.

[8] Scott H Podolsky. "Metchnikoff and the microbiome". In: *The Lancet* 380.9856 (2012), pp. 1810–1811.

[9] Philip Arthur Mackowiak. "Recycling Metchnikoff: probiotics, the intestinal microbiome and the quest for long life". In: *Frontiers in public health* 1 (2013), p. 52.

[10] Antonia Suau et al. "Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut". In: *Applied and environmental microbiology* 65.11 (1999), pp. 4799–4807.

[11] SH Duncan, P Louis, and HJ Flint. "Cultivable bacterial diversity from the human colon". In: *Letters in applied microbiology* 44.4 (2007), pp. 343–350.

[12] PCY Woo et al. "Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories". In: *Clinical Microbiology and Infection* 14.10 (2008), pp. 908–934.

[13] George E Fox et al. "Classification of methanogenic bacteria by 16S ribosomal RNA characterization". In: *Proceedings of the National Academy of Sciences* 74.10 (1977), pp. 4537–4541.

[14] Ramesh Gupta, Jan M Lanter, and Carl R Woese. “Sequence of the 16S ribosomal RNA from Halobacterium volcanii, an archaebacterium”. In: *Science* 221.4611 (1983), pp. 656–659.

[15] Carl R Woese, Otto Kandler, and Mark L Wheelis. “Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.” In: *Proceedings of the National Academy of Sciences* 87.12 (1990), pp. 4576–4579.

[16] Robert D Fleischmann et al. “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd”. In: *Science* 269.5223 (1995), p. 496.

[17] Eric S Lander et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (2001), pp. 860–921.

[18] J Craig Venter et al. “The sequence of the human genome”. In: *science* 291.5507 (2001), pp. 1304–1351.

[19] MN Burge. *Fungi in biological control systems*. Manchester University Press, 1988.

[20] Joshua Lederberg and Alexa T McCray. “Ome SweetOmics–A Genealogical Treasury of Words”. In: *The Scientist* 15.7 (2001), pp. 8–8.

[21] Dwayne C Savage. “Microbial ecology of the gastrointestinal tract”. In: *Annual Reviews in Microbiology* 31.1 (1977), pp. 107–133.

[22] Lora V Hooper and Jeffrey I Gordon. “Commensal host-bacterial relationships in the gut”. In: *Science* 292.5519 (2001), pp. 1115–1118.

[23] Paul B Eckburg et al. “Diversity of the human intestinal microbial flora”. In: *science* 308.5728 (2005), pp. 1635–1638.

[24] Steven R Gill et al. “Metagenomic analysis of the human distal gut microbiome”. In: *science* 312.5778 (2006), pp. 1355–1359.

[25] J Craig Venter et al. “Environmental genome shotgun sequencing of the Sargasso Sea”. In: *science* 304.5667 (2004), pp. 66–74.

[26] Peter J Turnbaugh et al. “The human microbiome project: exploring the microbial part of ourselves in a changing world”. In: *Nature* 449.7164 (2007), p. 804.

[27] Human Microbiome Project Consortium. *Data Analysis and Coordination Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP)*. URL: http://www.hmpdacc.org/ (visited on 02/26/2017).

[28] Human Microbiome Project Consortium et al. “A framework for human microbiome research”. In: *Nature* 486.7402 (2012), pp. 215–221.

[29] Human Microbiome Project Consortium et al. “Structure, function and diversity of the healthy human microbiome”. In: *Nature* 486.7402 (2012), pp. 207–214.

[30] Junjie Qin et al. “A human gut microbial gene catalogue established by metagenomic sequencing”. In: *nature* 464.7285 (2010), pp. 59–65.

[31] Jack A Gilbert, Janet K Jansson, and Rob Knight. “The Earth Microbiome project: successes and aspirations”. In: *BMC biology* 12.1 (2014), p. 69.

[32] Alexandra Zhernakova et al. “Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity”. In: *Science* 352.6285 (2016), pp. 565–569.

[33] Jack A Gilbert et al. “Microbiome-wide association studies link dynamic microbial consortia to disease”. In: *Nature* 535.7610 (2016), pp. 94–103.

[34] Roderick I Mackie, Abdelghani Sghir, and H Rex Gaskins. “Developmental microbial ecology of the neonatal gastrointestinal tract”. In: *The American journal of clinical nutrition* 69.5 (1999), 1035s–1045s.

[35] June L Round and Sarkis K Mazmanian. “The gut microbiota shapes intestinal immune responses during health and disease”. In: *Nature Reviews Immunology* 9.5 (2009), pp. 313–323.

[36] Craig L Maynard et al. “Reciprocal interactions of the intestinal microbiota and immune system”. In: *Nature* 489.7415 (2012), pp. 231–241.

[37] H. Chung et al. “Gut immune maturation depends on colonization with a host-specific microbiota”. In: *Cell* 149 (2012). DOI: 10.1016/j.cell.2012.04.037. URL: http://dx.doi.org/10.1016/j.cell.2012.04.037.

[38] Denise Kelly et al. “Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR-γ and RelA”. In: *Nature immunology* 5.1 (2004), pp. 104–112.

[39] Mary C Rea et al. “Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against Clostridium difficile”. In: *Proceedings of the National Academy of Sciences* 107.20 (2010), pp. 9352–9357.

[40] Nobuhiko Kamada et al. “Control of pathogens and pathobionts by the gut microbiota”. In: *Nature immunology* 14.7 (2013), pp. 685–690.

[41] C. G. Buffie and E. G. Pamer. “Microbiota-mediated colonization resistance against intestinal pathogens”. In: *Nat Rev Immunol* 13 (2013). DOI: 10.1038/nri3535. URL: http://dx.doi.org/10.1038/nri3535.

[42] Charlie G Buffie et al. “Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile”. In: *Nature* 517.7533 (2015), pp. 205–208.

[43] NI McNeil. “The contribution of the large intestine to energy supplies in man.” In: *The American journal of clinical nutrition* 39.2 (1984), pp. 338–342.

[44] EN Bergman. “Energy contributions of volatile fatty acids from the gastrointestinal tract in various species”. In: *Physiological reviews* 70.2 (1990), pp. 567–590.

[45] JH Cummings and GT Macfarlane. “Role of intestinal bacteria in nutrient metabolism”. In: *Clinical nutrition* 16.1 (1997), pp. 3–11.

[46] Harry J Flint et al. “Microbial degradation of complex carbohydrates in the gut”. In: *Gut microbes* 3.4 (2012), pp. 289–306.

[47] Carlotta De Filippo et al. “Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa”. In: *Proceedings of the National Academy of Sciences* 107.33 (2010), pp. 14691–14696.

[48] Alan W Walker et al. “Dominant and diet-responsive groups of bacteria within the human colonic microbiota”. In: *The ISME journal* 5.2 (2011), pp. 220–230.

[49] Tanya Yatsunenko et al. “Human gut microbiome viewed across age and geography”. In: *Nature* 486.7402 (2012), pp. 222–227.

[50] Estelle Devillard et al. “Metabolism of linoleic acid by human gut bacteria: different routes for biosynthesis of conjugated linoleic acid”. In: *Journal of bacteriology* 189.6 (2007), pp. 2566–2570.

[51] Gwen Tolhurst et al. “Short-chain fatty acids stimulate glucagon-like peptide-1 secretion via the G-protein–coupled receptor FFAR2”. In: *Diabetes* 61.2 (2012), pp. 364–371.

[52] Brian T Layden et al. “Short chain fatty acids and their receptors: new metabolic targets”. In: *Translational Research* 161.3 (2013), pp. 131–140.

[53] Jessica M Yano et al. “Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis”. In: *Cell* 161.2 (2015), pp. 264–276.

[54] Patrice D Cani et al. “Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet–induced obesity and diabetes in mice”. In: *Diabetes* 57.6 (2008), pp. 1470–1481.

[55] Sridevi Devaraj, Peera Hemarajata, and James Versalovic. “The human gut microbiome and body metabolism: implications for obesity and diabetes”. In: *Clinical chemistry* 59.4 (2013), pp. 617–628.

[56] P. J. Turnbaugh et al. “The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice”. In: *Sci Transl Med* 1 (2009). DOI: 10.1126/scitranslmed.3000322. URL: http://dx.doi.org/10.1126/scitranslmed.3000322.

[57] Ruth E Ley et al. “Obesity alters gut microbial ecology”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.31 (2005), pp. 11070–11075.

[58] Julia K Goodrich et al. “Human genetics shape the gut microbiome”. In: *Cell* 159.4 (2014), pp. 789–799.

[59] Peter J Turnbaugh et al. “An obesity-associated gut microbiome with increased capacity for energy harvest”. In: *nature* 444.7122 (2006), pp. 1027–131.

[60] Christopher T Brown et al. “Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes”. In: *PloS one* 6.10 (2011), e25792.

[61] Aleksandar D Kostic et al. “The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes”. In: *Cell host & microbe* 17.2 (2015), pp. 260–273.

[62] A. D. Kostic, R. J. Xavier, and D. Gevers. “The microbiome in inflammatory bowel disease: current status and the future ahead”. In: *Gastroenterology* 146 (2014). DOI: 10.1053/j.gastro.2014.02.009. URL: http://dx.doi.org/10.1053/j.gastro.2014.02.009.

[63] Y. Haberman et al. “Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature”. In: *J Clin Invest* 124 (2014). DOI: 10.1172/JCI75436. URL: http://dx.doi.org/10.1172/JCI75436.

[64] Aleksandar D Kostic et al. “Genomic analysis identifies association of Fusobacterium with colorectal carcinoma”. In: *Genome research* 22.2 (2012), pp. 292–298.

[65] Weiguang Chen et al. “Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer”. In: *PloS one* 7.6 (2012), e39743.

[66] N. Iida et al. “Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment”. In: *Science* 342 (2013). DOI: 10.1126/science.1240527. URL: http://dx.doi.org/10.1126/science.1240527.

[67] K. Klimesova et al. “Altered gut microbiota promotes colitis-associated cancer in IL-1 receptor-associated kinase M-deficient mice”. In: *Inflamm Bowel Dis* 19 (2013). DOI: 10.1097/MIB.0b013e318281330a. URL: http://dx.doi.org/10.1097/MIB.0b013e318281330a.

[68] Jiyoung Ahn et al. “Human gut microbiome and risk of colorectal cancer”. In: *Journal of the National Cancer Institute* (2013), djt300.

[69] Joseph P Zackular et al. “The gut microbiome modulates colon tumorigenesis”. In: *MBio* 4.6 (2013), e00692–13.

[70] Joseph P Zackular et al. “The human gut microbiome as a screening tool for colorectal cancer”. In: *Cancer prevention research* 7.11 (2014), pp. 1112–1121.

[71] R. R. Jenq et al. “Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation”. In: *J Exp Med* 209 (2012). DOI: 10.1084/jem.20112408. URL: http://dx.doi.org/10.1084/jem.20112408.

[72] E. Holler et al. “Metagenomic analysis of the stool microbiome in patients receiving allogeneic stem cell transplantation: loss of diversity is associated with use of systemic antibiotics and more pronounced in gastrointestinal graft-versus-host disease”. In: *Biol Blood Marrow Transplant* 20 (2014). DOI: 10.1016/j.bbmt.2014.01.030. URL: http://dx.doi.org/10.1016/j.bbmt.2014.01.030.

[73] D. Weber et al. “Low urinary indoxyl sulfate levels early after transplantation reflect a disrupted microbiome and are associated with poor outcome”. In: *Blood* 126 (2015). DOI: 10.1182/blood-2015-04-638858. URL: http://dx.doi.org/10.1182/blood-2015-04-638858.

[74] Daniela Weber et al. “Rifaximin preserves intestinal microbiota balance in patients undergoing allogeneic stem cell transplantation”. In: *Bone marrow transplantation* (2016).

[75] Els van Nood et al. “Duodenal infusion of donor feces for recurrent Clostridium difficile”. In: *New England Journal of Medicine* 368.5 (2013), pp. 407–415.

[76] Loek P Smits et al. “Therapeutic potential of fecal microbiota transplantation”. In: *Gastroenterology* 145.5 (2013), pp. 946–953.

[77] Thomas M Kuntz and Jack A Gilbert. “Introducing the Microbiome into Precision Medicine”. In: *Trends in Pharmacological Sciences* 38.1 (2017), pp. 81–91.

[78] David Zeevi et al. “Personalized nutrition by prediction of glycemic responses”. In: *Cell* 163.5 (2015), pp. 1079–1094.

[79] Sophie Viaud et al. “The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide”. In: *science* 342.6161 (2013), pp. 971–976.

[80] Kathleen Machiels et al. “Specific members of the predominant gut microbiota predict pouchitis following colectomy and IPAA in UC”. In: *Gut* (2015), gutjnl–2015.

[81] Marie Vétizou et al. “Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota”. In: *Science* 350.6264 (2015), pp. 1079–1084.

[82] Ying Taur et al. “The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation”. In: *Blood* 124.7 (2014), pp. 1174–1182.

[83] Henry J Haiser et al. “Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta”. In: *Science* 341.6143 (2013), pp. 295–298.

[84] Corinne Ferrier Maurice, Henry Joseph Haiser, and Peter James Turnbaugh. “Xenobiotics shape the physiology and gene expression of the active human gut microbiome”. In: *Cell* 152.1 (2013), pp. 39–50.

[85] Justin Kuczynski et al. “Experimental and analytical tools for studying the human microbiome”. In: *Nature Reviews Genetics* 13.1 (2012), pp. 47–58.

[86] Berend Snel, Peer Bork, and Martijn A Huynen. “Genome phylogeny based on gene content”. In: *Nature genetics* 21.1 (1999), pp. 108–110.

[87] Ravi Ranjan et al. “Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing”. In: *Biochemical and biophysical research communications* 469.4 (2016), pp. 967–977.

[88] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[89] Minoru Kanehisa et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D353–D361.

[90] Ohad Manor and Elhanan Borenstein. “Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome”. In: *Microbiome* 5.1 (2017), p. 19.

[91] William G Weisburg et al. “16S ribosomal DNA amplification for phylogenetic study.” In: *Journal of bacteriology* 173.2 (1991), pp. 697–703.

[92] Christopher P Kolbert and David H Persing. “Ribosomal DNA sequencing as a tool for identification of bacterial pathogens”. In: *Current opinion in microbiology* 2.3 (1999), pp. 299–305.

[93] Ian Kroes, Paul W Lepp, and David A Relman. “Bacterial diversity within the human subgingival crevice”. In: *Proceedings of the National Academy of Sciences* 96.25 (1999), pp. 14547–14552.

[94] Juan Jovel et al. “Characterization of the gut microbiome using 16S or shotgun metagenomics”. In: *Frontiers in microbiology* 7 (2016).

[95] Neethu Shah et al. “Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics.” In: *Pacific Symposium on Biocomputing*. Vol. 16. 2011, pp. 165–176.

[96] Pablo Yarza et al. “Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences”. In: *Nature Reviews. Microbiology* 12.9 (2014), p. 635.

[97] J Michael Janda and Sharon L Abbott. “16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls”. In: *Journal of clinical microbiology* 45.9 (2007), pp. 2761–2764.

[98] Jurgen Brosius et al. “Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli”. In: *Proceedings of the National Academy of Sciences* 75.10 (1978), pp. 4801–4805.

[99] Todd Z DeSantis et al. “Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB”. In: *Applied and environmental microbiology* 72.7 (2006), pp. 5069–5072.

[100] Daniel McDonald et al. “An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea”. In: *The ISME journal* 6.3 (2012), pp. 610–618.

[101] C. Quast et al. “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. In: *Nucleic Acids Res* 41 (2013). DOI: 10.1093/nar/gks1219. URL: http://dx.doi.org/10.1093/nar/gks1219.

[102] Anna Klindworth et al. “Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies”. In: *Nucleic acids research* (2012), gks808.

[103] Minseok Kim, Mark Morrison, and Zhongtang Yu. "Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes". In: *Journal of microbiological methods* 84.1 (2011), pp. 81–87.

[104] Zongzhi Liu et al. "Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers". In: *Nucleic acids research* 36.18 (2008), e120–e120.

[105] Timothy J Hamp, W Joe Jones, and Anthony A Fodor. "Effects of experimental choices and analysis noise on surveys of the "rare biosphere"". In: *Applied and environmental microbiology* 75.10 (2009), pp. 3263–3270.

[106] Anna Engelbrektson et al. "Experimental factors affecting PCR-based estimates of microbial species richness and evenness". In: *The ISME journal* 4.5 (2010), pp. 642–647.

[107] Jonas Binladen et al. "The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing". In: *PloS one* 2.2 (2007), e197.

[108] Christian Hoffmann et al. "DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations". In: *Nucleic acids research* 35.13 (2007), e91.

[109] Victor Kunin et al. "Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates". In: *Environmental microbiology* 12.1 (2010), pp. 118–123.

[110] W Ford Doolittle and R Thane Papke. "Genomics and the bacterial species problem". In: *Genome biology* 7.9 (2006), p. 116.

[111] Mark Achtman and Michael Wagner. "Microbial diversity and the genetic nature of microbial species". In: *Nature Reviews Microbiology* 6.6 (2008), p. 431.

[112] W Ford Doolittle and Olga Zhaxybayeva. "On the origin of prokaryotic species". In: *Genome research* 19.5 (2009), pp. 744–756.

[113] M. Blaxter et al. "Defining operational taxonomic units using DNA barcode data". In: *Philos Trans R Soc Lond Ser B Biol Sci* 360 (2005). DOI: 10.1098/rstb.2005.1725. URL: http://dx.doi.org/10.1098/rstb.2005.1725.

[114] Konstantinos T Konstantinidis and James M Tiedje. "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.7 (2005), pp. 2567–2572.

[115] Sarah P Preheim et al. "Distribution-based clustering: using ecology to refine the operational taxonomic unit". In: *Applied and environmental microbiology* 79.21 (2013), pp. 6593–6603.

[116] Thomas S. B. Schmidt, João F. Matias Rodrigues, and Christian von Mering. "Limits to robustness and reproducibility in the demarcation of operational taxonomic units". In: *Environmental Microbiology* 17.5 (2015), pp. 1689–1706. ISSN: 1462-2920. DOI: 10.1111/1462-2920.12610. URL: http://dx.doi.org/10.1111/1462-2920.12610.

[117] Tomáš Větrovsky and Petr Baldrian. "The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses". In: *PloS one* 8.2 (2013), e57923.

[118] P. D. Schloss et al. "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities". In: *Appl Environ Microbiol* 75 (2009). DOI: 10.1128/AEM.01541-09. URL: http://dx.doi.org/10.1128/AEM.01541-09.

[119] J. G. Caporaso et al. “QIIME allows analysis of high-throughput community sequencing data”. In: *Nat Methods* 7 (2010). DOI: 10.1038/nmeth.f.303. URL: http://dx.doi.org/10.1038/nmeth.f.303.

[120] Robert C Edgar. “UPARSE: highly accurate OTU sequences from microbial amplicon reads”. In: *Nature methods* 10.10 (2013), pp. 996–998.

[121] Brian J Haas et al. “Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons”. In: *Genome research* 21.3 (2011), pp. 494–504.

[122] Robert C Edgar et al. “UCHIME improves sensitivity and speed of chimera detection”. In: *Bioinformatics* 27.16 (2011), pp. 2194–2200.

[123] Patrick D Schloss and Jo Handelsman. “Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness”. In: *Applied and environmental microbiology* 71.3 (2005), pp. 1501–1506.

[124] Yijun Sun et al. “A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis”. In: *Briefings in bioinformatics* (2011), bbr009.

[125] Patrick D Schloss and Sarah L Westcott. “Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis”. In: *Applied and environmental microbiology* 77.10 (2011), pp. 3219–3226.

[126] Sophie J Weiss et al. *Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data*. Tech. rep. PeerJ PrePrints, 2015.

[127] Benjamin J Callahan et al. “DADA2: high-resolution sample inference from Illumina amplicon data”. In: *Nature methods* 13.7 (2016), p. 581.

[128] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis”. In: *The ISME journal* 11.12 (2017), p. 2639.

[129] Sophie Weiss et al. “Normalization and microbial differential abundance strategies depend upon data characteristics”. In: *Microbiome* 5.1 (2017), p. 27.

[130] Andrew Brewer and Mark Williamson. “A new relationship for rarefaction”. In: *Biodiversity and Conservation* 3.4 (1994), pp. 373–379.

[131] Nicholas J Gotelli and Robert K Colwell. “Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness”. In: *Ecology letters* 4.4 (2001), pp. 379–391.

[132] Paul J McMurdie and Susan Holmes. “Waste not, want not: why rarefying microbiome data is inadmissible”. In: *PLoS computational biology* 10.4 (2014), e1003531.

[133] Robert H Whittaker. “Evolution and measurement of species diversity”. In: *Taxon* (1972), pp. 213–251.

[134] Tom CJ Hill et al. “Using ecological diversity measures with bacterial communities”. In: *FEMS microbiology ecology* 43.1 (2003), pp. 1–11.

[135] Lou Jost. “Partitioning diversity into independent alpha and beta components”. In: *Ecology* 88.10 (2007), pp. 2427–2439.

[136] Robert K Colwell. “Biodiversity: concepts, patterns, and measurement”. In: *The Princeton guide to ecology* (2009), pp. 257–263.

[137] Claude Elwood Shannon. “Communication in the presence of noise”. In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[138] Edward H Simpson. “Measurement of diversity.” In: *Nature* (1949).

[139] AE Marrugan. “Ecological diversity and its measurement”. In: *Ecological diversity and its measurement* (1988).

[140] J Roger Bray and John T Curtis. “An ordination of the upland forest communities of southern Wisconsin”. In: *Ecological monographs* 27.4 (1957), pp. 325–349.

[141] Catherine Lozupone and Rob Knight. “UniFrac: a new phylogenetic method for comparing microbial communities”. In: *Applied and environmental microbiology* 71.12 (2005), pp. 8228–8235.

[142] Catherine A Lozupone et al. “Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities”. In: *Applied and environmental microbiology* 73.5 (2007), pp. 1576–1585.

[143] Micah Hamady, Catherine Lozupone, and Rob Knight. “Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data”. In: *The ISME journal* 4.1 (2010), p. 17.

[144] Sonia Michail et al. “Alterations in the gut microbiome of children with severe ulcerative colitis”. In: *Inflammatory bowel diseases* 18.10 (2012), pp. 1799–1808.

[145] Dirk Gevers et al. “The treatment-naive microbiome in new-onset Crohn’s disease”. In: *Cell host & microbe* 15.3 (2014), pp. 382–392.

[146] Ludovic Giloteaux et al. “Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome”. In: *Microbiome* 4.1 (2016), p. 30.

[147] Nicola Segata et al. “Metagenomic biomarker discovery and explanation”. In: *Genome biology* 12.6 (2011), R60.

[148] Joseph N Paulson et al. “Differential abundance analysis for microbial marker-gene surveys”. In: *Nature methods* 10.12 (2013), pp. 1200–1202.

[149] Daniel H Huson et al. “MEGAN analysis of metagenomic data”. In: *Genome research* 17.3 (2007), pp. 377–386.

[150] Suparna Mitra, Bernhard Klar, and Daniel H Huson. “Visual and statistical comparison of metagenomes”. In: *Bioinformatics* 25.15 (2009), pp. 1849–1855.

[151] Donovan H Parks and Robert G Beiko. “Identifying biologically relevant differences between metagenomic communities”. In: *Bioinformatics* 26.6 (2010), pp. 715–721.

[152] Patrick D Schloss et al. “Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities”. In: *Applied and environmental microbiology* 75.23 (2009), pp. 7537–7541.

[153] Folker Meyer et al. “The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes”. In: *BMC bioinformatics* 9.1 (2008), p. 386.

[154] James Robert White, Niranjan Nagarajan, and Mihai Pop. “Statistical methods for detecting differentially abundant features in clinical metagenomic samples”. In: *PLoS computational biology* 5.4 (2009), e1000352.

[155] Julia K Goodrich et al. “Conducting a microbiome study”. In: *Cell* 158.2 (2014), pp. 250–262.

[156] Rashmi Sinha et al. “The microbiome quality control project: baseline study design and future directions”. In: *Genome biology* 16.1 (2015), p. 276.

[157] M. B. Jones et al. “Library preparation methodology can influence genomic and functional predictions in human microbiome research”. In: *Proc Natl Acad Sci U S A* 112 (2015). DOI: 10.1073/pnas.1519288112. URL: http://dx.doi.org/10.1073/pnas.1519288112.

[158] Marc A Sze and Patrick D Schloss. “Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome”. In: *MBio* 7.4 (2016), e01018–16.

[159] O. Koren et al. “A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets”. In: *PLoS Comput Biol* 9 (2013). DOI: 10.1371/journal.pcbi.1002863. URL: http://dx.doi.org/10.1371/journal.pcbi.1002863.

[160] D. Risso et al. “Normalization of RNA-seq data using factor analysis of control genes or samples”. In: *Nat Biotechnol* 32 (2014). DOI: 10.1038/nbt.2931. URL: http://dx.doi.org/10.1038/nbt.2931.

[161] T. Geiger et al. “Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics”. In: *Nat Protoc* 6 (2011). DOI: 10.1038/nprot.2010.192. URL: http://dx.doi.org/10.1038/nprot.2010.192.

[162] L. Wu et al. “Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly 13C-labeled cell extracts as internal standards”. In: *Anal Biochem* 336 (2005). DOI: 10.1016/j.ab.2004.09.001. URL: http://dx.doi.org/10.1016/j.ab.2004.09.001.

[163] J. Anton et al. “Salinibacter ruber gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds”. In: *Int J Syst Evol Microbiol* 52 (2002). DOI: 10.1099/00207713-52-2-485. URL: http://dx.doi.org/10.1099/00207713-52-2-485.

[164] L. Zhang et al. “Genomic analysis of Agrobacterium radiobacter DSM 30147(T) and emended description of A. radiobacter (Beijerinck and van Delden 1902) Conn 1942 (Approved Lists 1980) emend. Sawada et al. 1993.” In: *Stand Genomic Sci* 9 (2014). DOI: 10.4056/sigs.4688352. URL: http://dx.doi.org/10.4056/sigs.4688352.

[165] H. Matsubara et al. “Alicyclobacillus acidiphilus sp nov., a novel thermo-acidophilic, omega-alicyclic fatty acid-containing bacterium isolated from acidic beverages”. In: *Int J Syst Evol Microbiol* 52 (2002).

[166] S. F. Stoddard et al. “rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development”. In: *Nucleic Acids Res* 43 (2015). DOI: 10.1093/nar/gku1201. URL: http://dx.doi.org/10.1093/nar/gku1201.

[167] *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2015.

[168] W. Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. In: *Nat Methods* 12 (2015). DOI: 10.1038/nmeth.3252. URL: http://dx.doi.org/10.1038/nmeth.3252.

[169] R. C. Edgar. “Search and clustering orders of magnitude faster than BLAST”. In: *Bioinformatics* 26 (2010). DOI: 10.1093/bioinformatics/btq461. URL: http://dx.doi.org/10.1093/bioinformatics/btq461.

[170] I. Brukner et al. “Assay for estimating total bacterial load: relative qPCR normalisation of bacterial load with associated clinical implications”. In: *Diagn Microbiol Infect Dis* 83 (2015). DOI: 10.1016/j.diagmicrobio.2015.04.005. URL: http://dx.doi.org/10.1016/j.diagmicrobio.2015.04.005.

[171] A. S. Amend, K. A. Seifert, and T. D. Bruns. “Quantifying microbial communities with 454 pyrosequencing: does read abundance count?” In: *Mol Ecol* 19 (2010). DOI: 10.1111/j.1365-294X.2010.04898.x. URL: http://dx.doi.org/10.1111/j.1365-294X.2010.04898.x.

[172] E. K. Costello et al. “The application of ecological theory toward an understanding of the human microbiome”. In: *Science* 336 (2012). DOI: 10.1126/science.1224203. URL: http://dx.doi.org/10.1126/science.1224203.

[173] I. Cho and M. J. Blaser. “The human microbiome: at the interface of health and disease”. In: *Nat Rev Genet* 13 (2012).

[174] J. Walter and R. Ley. “The human gut microbiome: ecology and recent evolutionary changes”. In: *Annu Rev Microbiol* 65 (2011). DOI: 10.1146/annurev-micro-090110-102830. URL: http://dx.doi.org/10.1146/annurev-micro-090110-102830.

[175] J. Aitchison. “A New Approach to Null Correlations of Proportions”. In: *J Int Ass Math Geol* 13 (1981). DOI: 10.1007/BF01031393. URL: http://dx.doi.org/10.1007/BF01031393.

[176] J. Aitchison. “The Statistical-Analysis of Compositional Data”. In: *J Roy Stat Soc B Met* 44 (1982).

[177] EG Zoetendal, M Rajilić-Stojanović, and WM De Vos. “High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota”. In: *Gut* 57.11 (2008), pp. 1605–1615.

[178] Mirjana Rajilić-Stojanović and Willem M de Vos. “The first 1000 cultured species of the human gastrointestinal microbiota”. In: *FEMS microbiology reviews* 38.5 (2014), pp. 996–1047.

[179] Konstantinos T Konstantinidis, Ramon Rosselló-Móra, and Rudolf Amann. “Uncultivated microbes in need of their own taxonomy”. In: *The ISME Journal* 11.11 (2017), p. 2399.

[180] Steven D Allison and Jennifer BH Martiny. “Resistance, resilience, and redundancy in microbial communities”. In: *Proceedings of the National Academy of Sciences* 105.Supplement 1 (2008), pp. 11512–11519.

[181] Diana R Nemergut et al. “Global patterns in the biogeography of bacterial taxa”. In: *Environmental microbiology* 13.1 (2011), pp. 135–144.

[182] Nicholas D Youngblut et al. “Lineage-specific responses of microbial communities to environmental change”. In: *Applied and environmental microbiology* 79.1 (2013), pp. 39–47.

[183] Georg K. Gerber, Andrew B. Onderdonk, and Lynn Bry. “Inferring Dynamic Signatures of Microbes in Complex Host Ecosystems”. In: *PLOS Computational Biology* 8.8 (Aug. 2012), pp. 1–14. DOI: 10.1371/journal.pcbi.1002624. URL: http://dx.doi.org/10.1371%2Fjournal.pcbi.1002624.

[184] Paul I Costea et al. “Subspecies in the global human gut microbiome”. In: *Molecular systems biology* 13.12 (2017), p. 960.

[185] Frederick M Cohan. “What are bacterial species?” In: *Annual Reviews in Microbiology* 56.1 (2002), pp. 457–487.

[186] Frederick M Cohan and Elizabeth B Perry. “A systematics for discovering the fundamental units of bacterial diversity”. In: *Current Biology* 17.10 (2007), R373–R386.

[187] Alexander Koeppel et al. “Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics”. In: *Proceedings of the National Academy of Sciences* 105.7 (2008), pp. 2504–2509.

[188] Dana E Hunt et al. “Resource partitioning and sympatric differentiation among closely related bacterioplankton”. In: *Science* 320.5879 (2008), pp. 1081–1085.

[189] Tobias Guldberg Frøslev et al. “Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates”. In: *Nature communications* 8.1 (2017), p. 1188.

[190] A Murat Eren et al. “Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences”. In: *The ISME journal* 9.4 (2015), pp. 968–979.

[191] Mikhail Tikhonov, Robert W Leach, and Ned S Wingreen. “Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution”. In: *The ISME journal* 9.1 (2015), pp. 68–80.

[192] John M. Gaspar and W. Kelley Thomas. “FlowClus: efficiently filtering and denoising pyrosequenced amplicons”. In: *BMC Bioinformatics* 16.1 (Mar. 2015), p. 105. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0532-1. URL: https://doi.org/10.1186/s12859-015-0532-1.

[193] Vladimir I Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[194] Daniel S. Hirschberg. “A linear space algorithm for computing maximal common subsequences”. In: *Communications of the ACM* 18.6 (1975), pp. 341–343.

[195] M.P.J. van der Loo. “The stringdist package for approximate string matching”. In: *The R Journal* 6 (1 2014), pp. 111–122. URL: https://CRAN.R-project.org/package=stringdist.

[196] Dominik Maria Endres and Johannes E Schindelin. “A new metric for probability distributions”. In: *IEEE Transactions on Information theory* 49.7 (2003), pp. 1858–1860.

[197] Ferdinand Österreicher and Igor Vajda. “A new class of metric divergences on probability spaces and its applicability in statistics”. In: *Annals of the Institute of Statistical Mathematics* 55.3 (2003), pp. 639–653.

[198] Jianhua Lin. “Divergence measures based on the Shannon entropy”. In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145–151.

[199] Solomon Kullback and Richard A Leibler. “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[200] Gregory E Sims et al. “Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions”. In: *Proceedings of the National Academy of Sciences* 106.8 (2009), pp. 2677–2682.

[201] Manimozhiyan Arumugam et al. “Enterotypes of the human gut microbiome”. In: *nature* 473.7346 (2011), p. 174.

[202] JJ Allaire and Jim Bullard. *Rcpp-Gallery: Parallel Distance Matrix Calculation with RcppParallel*. https://github.com/RcppCore/rcpp-gallery/blob/gh-pages/src/2014-07-15-parallel-distance-matrix.cpp. 2015.

[203] Dirk Eddelbuettel and Romain François. “Rcpp: Seamless R and C++ Integration”. In: *Journal of Statistical Software* 40.8 (2011), pp. 1–18. URL: http://www.jstatsoft.org/v40/i08/.

[204] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*. R package version 1.10.4-3. 2017. URL: https://CRAN.R-project.org/package=data.table.

[205] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. 2017. URL: https://CRAN.R-project.org/package=tidyverse.

[206] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

[207] Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.1.6. 2017. URL: https://CRAN.R-project.org/package=ggpubr.

[208] Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. 2017. URL: https://CRAN.R-project.org/package=gridExtra.

[209] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 0.9.2. 2017. URL: https://CRAN.R-project.org/package=cowplot.

[210] Tal Galili. "dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering". In: *Bioinformatics* (2015). DOI: 10.1093/bioinformatics/btv428. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/22/3718/17122682/btv428.pdf. URL: https://academic.oup.com/bioinformatics/article/31/22/3718/240978/dendextend-an-R-package-for-visualizing-adjusting.

[211] Zuguang Gu, Roland Eils, and Matthias Schlesner. "Complex heatmaps reveal patterns and correlations in multidimensional genomic data". In: *Bioinformatics* (2016).

[212] Jari Oksanen et al. *vegan: Community Ecology Package*. R package version 2.4-6. 2018. URL: https://CRAN.R-project.org/package=vegan.

[213] Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. ISBN 978-1-4614-6867-7. New York: Springer, 2013. DOI: 10.1007/978-1-4614-6868-4.

[214] Dirk Eddelbuettel and James Joseph Balamuta. "Extending extitR with extitC++: A Brief Introduction to extitRcpp". In: *PeerJ Preprints* 5 (Aug. 2017), e3188v1. ISSN: 2167-9843. DOI: 10.7287/peerj.preprints.3188v1. URL: https://doi.org/10.7287/peerj.preprints.3188v1.

[215] JJ Allaire et al. *RcppParallel: Parallel Programming Tools for 'Rcpp'*. R package version 4.3.20. 2016. URL: https://CRAN.R-project.org/package=RcppParallel.

[216] Jonathan Thorsen et al. "Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies". In: *Microbiome* 4.1 (2016), p. 62.

[217] Mohamed Mysara et al. "Reconciliation between operational taxonomic units and species boundaries". In: *FEMS microbiology ecology* 93.4 (2017), fix029.

[218] Susan M Huse et al. "Ironing out the wrinkles in the rare biosphere through improved OTU clustering". In: *Environmental microbiology* 12.7 (2010), pp. 1889–1898.

[219] Yan He et al. "Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity". In: *Microbiome* 3.1 (2015), p. 20.

[220] Jun Chen and Hongzhe Li. "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis". In: *The annals of applied statistics* 7.1 (2013).

[221] Zachary D Kurtz et al. "Sparse and compositionally robust inference of microbial ecological networks". In: *PLoS computational biology* 11.5 (2015), e1004226.

[222] Paul J McMurdie and Susan Holmes. “phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data”. In: *PloS one* 8.4 (2013), e61217.

[223] Jacqueline ZM Chan et al. “Defining bacterial species in the genomic era: insights from the genus Acinetobacter”. In: *BMC microbiology* 12.1 (2012), p. 302.

[224] Chin Wen Png et al. “Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria”. In: *The American journal of gastroenterology* 105.11 (2010), p. 2420.

[225] Indrani Mukhopadhya et al. “IBD—what role do Proteobacteria play?” In: *Nature Reviews Gastroenterology and Hepatology* 9.4 (2012), p. 219.

[226] Conor J Meehan and Robert G Beiko. “A phylogenomic view of ecological specialization in the Lachnospiraceae, a family of digestive tract-associated bacteria”. In: *Genome biology and evolution* 6.3 (2014), pp. 703–713.

[227] Keishi Kameyama and Kikuji Itoh. “Intestinal colonization by a Lachnospiraceae bacterium contributes to the development of diabetes in obese mice”. In: *Microbes and environments* 29.4 (2014), pp. 427–430.

[228] Kate L Ormerod et al. “Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals”. In: *Microbiome* 4.1 (2016), p. 36.

[229] Huey-Huey Chua et al. “Intestinal Dysbiosis Featuring Abundance of Ruminococcus gnavus Associates With Allergic Diseases in Infants”. In: *Gastroenterology* 154.1 (2018), pp. 154–167.

[230] Nam-Phuong Nguyen et al. “A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity”. In: *NPJ biofilms and microbiomes* 2 (2016), p. 16004.

[231] Wenke Smets et al. “A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing”. In: *Soil Biology and Biochemistry* 96 (2016), pp. 145–151.

[232] Dieter M Tourlousse et al. “Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing”. In: *Nucleic acids research* 45.4 (2017), e23–e23.

[233] Jolinda Pollock et al. “The madness of microbiome: Attempting to find consensus “best practice” for 16S microbiome studies”. In: *Applied and Environmental Microbiology* (2018), AEM–02627.

[234] Scott W Olesen, Claire Duvallet, and Eric J Alm. “dbOTU3: A new implementation of distribution-based OTU calling”. In: *bioRxiv* (2016), p. 076927.

[235] A Murat Eren et al. “Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data”. In: *Methods in Ecology and Evolution* 4.12 (2013), pp. 1111–1119.

[236] Amnon Amir et al. “Deblur rapidly resolves single-nucleotide community sequence patterns”. In: *MSystems* 2.2 (2017), e00191–16.

[237] NIH National Cancer Institute. *NCI Dictionary of Cancer Terms: Allogeneic stem cell transplantation*. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=270732 (visited on 02/03/2017).

[238] Masato Imase et al. “Construction of an artificial symbiotic community using a Chlorella–symbiont association as a model”. In: *FEMS microbiology ecology* 63.3 (2008), pp. 273–282.

[239] DJ Lane. “16S/23S rRNA sequencing”. In: *Nucleic acid techniques in bacterial systematices* (1991), pp. 115–175.