Dipl.-Ing. Bettina Halwachs, BSc.

Novel Strategies for the In-depth Analysis of Complex Microbial Communities

DISSERTATION

to obtain the academic degree

Doktorin der technischen Wissenschaften

submitted at

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Ing. Dr.techn. Rudolf Stollberger

Institute for Medical Engineering

Second Supervisor

Dipl.-Ing. Dr.techn. Gerhard Thallinger

Bioinformatics Group

Institute for Knowledge Discovery

Graz, August 2014

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The text document uploaded to TUGRAZonline is identical to the present doctoral dissertation.

Graz, ______________________ ______________________

Datum / *Date* Unterschrift / *Signature*

Abstract

Microbiomics, the investigation of microbial communities at different stages of disease, at specific time-points, or at varying conditions in a particular habitat such as distinct areas of the human body, or environmental samples such as clean rooms, or extremophile ecosystems, is one of the most rapidly growing research areas nowadays. This is mainly facilitated by the development of novel molecular classification approaches, as well as the steadily decreasing sequencing costs during the last decade. As this research area is still evolving by new developments in sequencing techniques and constantly growing knowledge, there is a need for novel or adapted methods, approaches, and tools for all the analysis steps of the community characterization and classification workflow.

This thesis introduces new approaches, methods, and tools for important steps in the entire high-throughput characterization and classification process of complex microbial communities. At the experimental design level, the effects of sequencing library normalization on the final community profile and its diversity was investigated. Subsequently, the *Decontaminator*, an effective tool for the removal of contaminating sequences from the target data sets is introduced as a major improvement during sequence pre-processing. For the core step, the taxonomic classification, an internal transcribed spacer (ITS) reference database, for fungal sequences was created. Tests of the ITS amplicon classification, with a hand curated *in-silico* amplified and fully annotated ITS mock community, showed good results for *reference based* classification and *de-novo* OTU picking approaches based on the UNITE ITS reference sequences. Statistical analysis of determined community profiles was extended by methods for differentially abundant feature detection. Therefor, Metastats, edgeR, and limma+voom, were evaluated using simulated count data, revealing that the linear modeling approaches outperform Metastats for bigger library sizes and fold change values. Based on this evaluation result, real community profiles obtained from analyses conducted within this thesis were tested for differentially abundant features. Finally, with the transcriptome analysis of two *Campylobacter fetus* subspecies, the typical ϵ-proteobacterial promoter motif was also confirmed for *C. fetus sp.* Moreover, this kind of analysis introduces a future direction for more detailed investigation of specific members of a microbial community.

Keywords: High-throughput classification, Microbiome, Sequencing, DA feature detection, Transcriptome analysis, Library normalization

Zusammenfassung

Mikrobiomik, die Erforschung von mikrobiellen Gemeinschaften in verschiedenen Krankheitsstadien zu bestimmten Zeitpunkten, oder unter unterschiedlichen Bedingungen, in einem bestimmten Lebensraum (zB Körperregionen, spezielle Umgebungen wie Reinräume oder extremophile Ökosysteme), zählt zu dem am schnellsten wachsenden Forschungsgebieten. Diese Entwicklung wurde im letzten Jahrzehnt hauptsächlich durch Fortschritte im Bereich der neuen molekularen Klassifikationsansätze, und durch stetig sinkende Sequenzierungskosten unterstützt. Durch die Weiterentwicklung der Sequenzierungstechniken und dem stetigen Zuwachs an Wissen auf diesem jungen Forschungsgebiet besteht ein Bedarf an neuen oder verbesserten Verfahren, Methoden, und Werkzeugen für alle Ebenen des Auswertungsprozesses.

Diese Dissertation stellt neue Ansätze, Methoden, und Werkzeuge für die wichtigsten Schritte des gesamten Hochdurchsatz-Charakterisierungs- und Klassifizierungs-Prozesses von komplexen mikrobiellen Gemeinschaften vor. Auf der Ebene des experimentellen Designs wurden die Auswirkungen auf das mikrobielle Profil anhand von normalisierten Sequenz-Bibliotheken untersucht. Der Vorverarbeitungsschritt wurde um den entwickelten *Decontaminator*, einem effektiven Werkzeug zur Erkennung und Entfernung von verunreinigenden Sequenzen, erweitert. Für den Hauptanalyseschritt, der taxonomischen Klassifizierung, wurde eine Referenz-Datenbank für Pilzsequenzen basierend auf der Internal Transcribed Spacer (ITS) Markerregion erstellt. Um die Qualität und Zuverlässigkeit der ITS-Amplikon Klassifikation zu bewerten, wurde eine von Hand kuratierte und vollständig annotierte ITS Mock Gemeinschaft erzeugt, mit deren Hilfe UNITE als brauchbare Ressource für sowohl referenz als auch *de novo* basierte Klassifizierungsmethoden eignet. Die statistische Analyse der ermittelten mikrobiellen Profile wurde um Methoden zur Identifizierung von differenziell abundanten Gruppen erweitert. Dazu wurden die Methoden Metastats, edgeR, und limma+voom, mit simulierten Count-Daten getestet und evaluiert. Hier konnte gezeigt werden, dass die linearen Modellierungsansätze für größere Bibliotheks- und Effekt-Größen bessere Ergebnisse erzielen als Metastats. Schließlich konnte durch die Transkriptom-Analyse von zwei *Campylobacter fetus* Subspezies das ϵ-Proteobakterium Promotor Motiv auch für diese Subspezies bestätigt werden. Darüber hinaus wurde hier mit der durchgeführten Transkriptom-Analyse eine zukünftige Richtung für weiterführende detaillierte Untersuchungen von speziellen Mitgliedern der mikrobiellen Gemeinschaft vorgestellt.

Keywords: Hochdurchsatz Klassifizierung, Mikrobiom, Sequenzierung, DA Feature Identifizierung, Transkriptom Analyse, Library Normalisierung

List of Publications

This thesis is based on the following publications in peer-reviewed journals, book chapters, as well as upon unpublished work. The full text of already published articles is available via the supplementary information in Appendix on page 241 ff.

Journal Articles

Bragina A, Oberauner-Wappis L, Zachow C, **Halwachs B**, Thallinger GG, Müller H, Berg G: **The *Sphagnum* microbiome supports bog ecosystem functioning under extreme conditions** *Molecular Ecologyl* 2014, doi: 10.1111/mec.12885, *In press*.

Nilsson RH, Hyde KD, Pawlowska J, Ryberg M, Tedersoo L, Aas AB, Alias S, Alves A, Andersson CL, Antonelli A, Arnold AE, Bahnmann B, Bahram M, Bengtsson-Palme J, Berlin A, Branco S, Chomnunti P, Dissanayake A, Drenkhan R, Friberg H, Fröslev T, **Halwachs B**, Hartmann M, Henricot B, Jayawardena R, Kauserud H, Koskela S, Kulik T, Liimatainen K, Lindahl B, Lindner D, Liu JK, Maharachchikumbura S, Manamgoda D, Martinsson S, Neves MA, Niskanen T, Nylinder S, Pereira OL, Pinho DB, Porter TM, Queloz B, Riit T, Sanchez-Garcia M, Sousa de, F, Stefanczyk E, Tadych M, Takamatsu S, Tian Q, Udayanga D, Unterseher M, Wang Y, Wikee S, Yan Y, Larsson E, Larsson KH, Koljalg U, Abarenkov K: **Improving ITS sequence data for identification of plant pathogenic fungi** *Fungal Diversity* 2014, Special Issue:1-9.

Klymiuk I, Högenauer C, **Halwachs B**, Thallinger GG, Fricke FW, Steininger C: **A physicians' wish list for the clinical application of intestinal metagenomics** *PLOS Medicine* 2014, 11:e1001627

Kienesberger S, Sprenger H, Wolfgruber S, **Halwachs B**, Thallinger GG, Perez-Perez GI, Blaster MJ, Zechner EL, Gorkiewicz G: **Comparative Genome Analysis of Campylobacter fetus Subspecies Revealed Horizontally Acquired Genetic Elements Important for Virulence and Niche Specificity** *PLOS One* 2014, 9(1):e85491.

Halwachs B, Höftberger J, Stocker G, Snajder R, Gorkiewicz G, Thallinger GG: **High-Throughput Characterization and Comparison of Microbial Communities.** *Biomed Tech (Berl)* 2013, doi: 10.1515/bmt-2013-4312.

Braun A*, **Halwachs B***, Geier M, Weinhandl K, Guggemos M, Marienhagen J, Ruff AJ, Schwaneberg U, Rabin V, Torres-Pazmino DE, Thallinger GG, Glieder A: **MuteinDB: the mutein database linking substrates, products and enzymatic reactions directly with genetic variants of enzymes.** *Database (Oxford)* 2012:bas028

*Author contributed equally.

Book Chapters

Halwachs B, Gorkiewicz G, Thallinger GG: **High-Throughput Characterization and Comparison of Microbial Communities.** In *Computational Medicine, Tools and Challanges.* Edited by Trajanoski Z. Vienna, Austria: Springer; 2012:37-57[1]

[1] Computational Medicine by Springer. Reproduced with permission of Springer in the format Thesis/Dissertation via Copyright Clearance Center

List of Figures

List of Tables

Contents

1. Introduction

Continuous and rapid development of deoxyribonucleic acid (DNA) sequencing technologies and approaches, since the introduction of the *"first-generation sequencing"* protocols in 1977 by Sanger *et al.* [1] and by Maxam and Gilbert [2], have set the course for a new area of molecular diagnostics. Classical Sanger sequencing dominated the last three decades before it was gradually replaced by newer methods, the so-called *"next-generation sequencing"* (NGS) techniques [3]. Although this traditional sequencing method benefits from high accuracy (> 99.999 %) and long read length (> 800 bps) [4], these advantages are made to naught by long analysis times, costs, and throughput. The major advantage of NGS over Sanger sequencing is the ability to produce an enormous amount of data in a single run within a short period of time at a low cost. Fig. 1.1 shows sequencing costs associated with DNA sequencing, tracked by the National Human Genome Research Institute (NHGRI) over the last two decades. Compared with the hypothetical trend of Moore's Law [5], the drastic reduction of DNA sequencing costs is illustrated after the transition from traditional Sanger sequencing to NGS technologies in early 2008. Besides the advantages of time, throughput and costs, NGS allows sequence determination from amplified DNA fragments without cloning [6].

Especially in microbiology, the drastic reduction of sequencing costs is the main reason why molecular characterization approaches have almost replaced traditional cultivation based methods [7].

We live in a world which is dominated by microorganisms. This is supported by the fact that the number of microorganisms on earth exceeds the number of human beings by a factor of 10^{21} [8]. As the majority of these microorganisms can not be

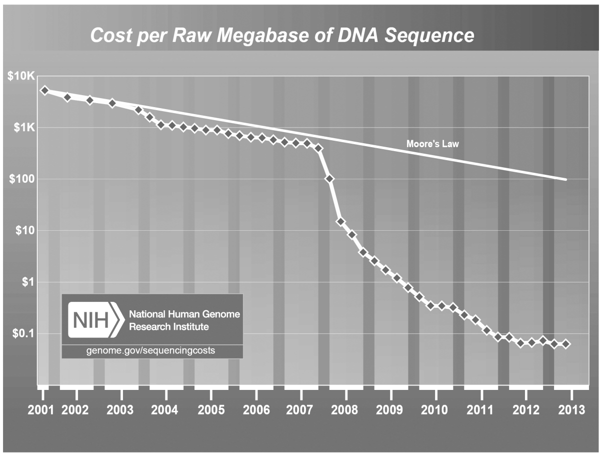


Figure 1.1.: Sequencing costs associated with DNA sequencing tracked by the National Human Genome Research Institute (NHGRI) over the last two decades. The hypothetical trend by Moore's Law [5], helps to illustrate the drastic reductions in DNA sequencing costs after the transition from traditional Sanger sequencing to NGS technologies in early 2008. (Source: National Human Genome Research Institute, 2014. Retrieved from http://www.genome.gov/sequencingcosts/)

cultivated in the laboratory, most of them have not been described or characterized yet [9]. Microorganisms are very diverse and include bacteria, archaea, fungi, protozoa, algae, lichens, and even viruses [10]. They have been found in almost all areas or environments of life, colonizing not only surfaces. Moreover, they live on human beings, animals, and plants. Microorganisms accomplish important functions in a variety of life cycles such as improvement of soil, water, digestive processes, or production of biofilms [11–13]. Additionally, the power of *beneficial* microbes has been used over the last century for the production of food, in agriculture, and presumably most importantly, in medicine [14, 15].

The human body is home to a wide range of microbial communities whose cells outnumber human cells 10 to 1 [8]. During the last years and the availability of NGS, it was possible to gain insights into microbial communities of different body sites such as the skin, intestines, oral cavity, or lungs in different states of health [16]. Although the relationship between the human host and its microbial coresidents is beneficial in many cases, sometimes it evolves to the contrary [17]. Changes in the microbial community composition have been related to digestive disorders and even to obesity [18, 19]. Furthermore, they are under suspicion to be responsible for skin or gum diseases [20, 21]. Reasons which lead to this mutualistic ("commensal") conversion of the relationship are still poorly understood and need further investigation [22].

In addition, to the thousands of beneficial microbes, host systems, such as the human body, are ordinarily inhabited by infectious microorganisms. As long as the immune system works properly, or the resistance of the host system is strong enough, these microbes are not able to overpopulate or move into areas where they do not occur normally. But when the balance of normal microbes is disrupted for some reasons, they can become the main cause for serious infections, diseases, and even lead to death. Such kinds of microorganisms are summarized as opportunistic pathogens [23].

A well-known representative are fungi from the *Candida* species. They belong to the normal human microbiota and have been found in the gastrointestinal and genital tract, the skin, and lower respiratory tract (LRT) of almost all healthy humans [24–27]. Although they are part of the "normal" microbiome of different human body sites, *Candida spp.* are considered as one of the most important human opportunistic pathogens [28]. Hence, distinguishing default fungal colonization from serious fungal infections is a critical point in medical diagnostics, especially in pulmonary samples [29]. Apart from the body's *Candida spp.*, any other environmental fungus is able to affect internal organs and cause substantial infections such as pneumonia. At special risks are immunocompromised patients, such as patients suffering from HIV or undergoing cancer treatment but also patients treated at intensive care units (ICUs) [30]. Nevertheless, *Candida spp.* are the most common pathogens causing serial severe fungal infections in humans worldwide [31].

The vast majority of microbes is found in the gastrointestinal (GI) tract [32]. Previous characterization approaches of the gut microbiome in humans and mice were able to establish a connection between the microbial composition of the gut microbiome and nutritional, as well as metabolic diseases or dysfunctions such as inflammatory bowel disease (IBD), obesity, and its related diseases (diabetes, nonalcoholic fatty liver disease, cardiovascular diseases, atherosclerosis, as well as certain cancers) [33–35]. Although these studies allowed deriving nutritional effects on the microbiome, as well as the impact of the microbiome on the host systems metabolism, little is known about the effect of certain molecules such as phospholipids on the gastrointestinal microbiome. Phospholipids are an important cell component and form a class of phosphoric and amphiphile lipids. The most important function of phospholipids is their ability of forming lipid bilayers in cell membranes [36] which play a crucial role in the communication and transportation of chemicals and ions [37].

Apart from nutritional factors which alter the intestinal and genital microbiome, the general population within these habitats is of special interest. For example, *Campylobacter* species have been recognized as emerging animal and human pathogens. Although the two major *Camplylobater fetus* subspecies *fetus* and *veneralis* (*Cff, Cfv*) are highly syntenic they differ strikingly in pathogenicity [38]. Whereas both subspecies are important factory farming pathogens, *Cfv* is adapted to bovines, causing infections which lead to abortion in cattle. In contrast, *Cff* is known to colonize the intestinal and genital-tract not only of bovines but also of sheep, birds and humans; causing diarrhea, serious invasive infections and even death [39, 40]. Genomic investigations on these pathogens allow insights into gene regulation, linkage of genes to particular pathotypes, as well as on their role in virulence and host tropism.

Fundamental ecological processes are carried out by diverse activities of complex microbial communities. Previous studies revealed that these communities do not only define a habitat moreover they interact with the host systems and are important indicators of responses to changed conditions [41–45]. Of special interest are environments inhabited by bacteria which are adapted to extreme conditions, such as drought, extreme temperatures, or low support of oxygen and nutrients. These habitats are putative sources for novel biocatalysts and enzymes. Industrial applications such as the production of biofuels, diverse drugs, fine chemicals and certain commodity products already benefit from the diverse enzymatic activities of microorganisms [46, 47].

Of special interest within this context are *Sphagnum*-dominated bogs. Although they are seen as very unique, they represent a very wide spread type of terrestrial ecosystems. These bog ecosystems belong to the oldest and most constant vegetation forms on earth for thousands of years. Mosses of the genus *Sphagnum* comprise more than a hundred different species and belong to the most abundant type of bog vegetation in the Northern hemisphere. They greatly contribute not only to global carbon turnover, but also to global climate regulation [48]. *Sphagnum* mosses form a unique habitat which is characterized by abiotic factors such as high acidity and low temperature, extremely low concentrations of mineral nutrients and oxygen, and extremely varying water saturation levels [49]. As this phylogenetically old genus has no roots, important functions such as nutrient supply, protection, and defense by biofilm formation to ensure moss growth and health are implemented by phyllosphere bacteria [50]. Although microbial diversity of the *Sphagnum* microbiome is well-described, little is known about its function. Metagenomic analysis of *Sphagnum* mosses allow for the discovery of unique features and potential functional, as well as structural differences to already described metagenomes of plants, peat soils, or aquatic systems.

Technical advances of sequencing technologies have led to a new era of molecular medicine and biotechnology. Recently introduced and steadily improving molecular phylogenetic, culture independent analysis approaches allows answering different questions within the "*(meta)omics*" life cycle, (see Fig. 1.2). Targeted amplicon sequencing (Sec. 1.1.1) enables fast insights and an overview of microbial community composition. Additionally, whole genome shotgun sequencing (*metagenomics*) (Sec. 1.1.2) enables a structural and functional description of the complex microbial composition.

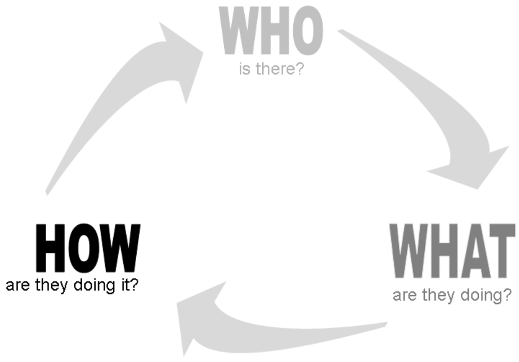


Figure 1.2.: Different questions within the *(meta)omics* life cycle can be answered by molecular phylogenetic, culture independent analysis approaches such as targeted amplicon sequencing (*Who is there?*), metagenomics (*What are they doing?*), and metatranscriptomics or -proteomics (*How are they doing it?*). Adopted from The Metagenomics Group at CBS[2]

Findings about metabolic activities and functions can be additionally enriched by *(meta)transcriptomic* (Sec. 1.1.3) approaches. They allow for deeper insights on how the communities are triggered and on what stimulates or inhibits particular activities. Hence, mechanisms which allow pathogenic taxa to proliferate and subsequently harm or alternate the host can be discovered and better understood. Additionally, our planet's microbial habitants and their capabilities can be further completed and described in more detail.

These methods result in an enormous amount of sequencing data which poses new and demanding challenges to the field of bioinformatics. High-throughput analysis and characterization methods are needed to process, analyze, and maintain this mass of data.

1.1. High-Throughput Characterization of Microbial Communities

Although NGS technologies have been principally used for whole genome and transcriptome sequencing [51], targeted sequencing of a specific gene region for solving questions in population genetics has become a very common technique. This technique, also called microbiome analysis, tries to reveal the microbial composition and diversity of a particular habitat. So, the microbiome is defined as the total number of microbial genomes in a defined environment at a particular state or time point or under predefined conditions [52]. Bacteria, archaea, lichens, fungi, or even viruses are summarized as microbes. Microbiomes of interest can be for example human body sites, soil, plants, clean rooms, foods, medical devices, or any other region of interest [16, 53, 54].

Rapid development of sequencing technologies and the introduction of a new sequencing strategy called *"whole genome shotgun sequencing"* (metagenomics), as well as refined analysis tools and methodologies, allowed also for structural, as well as for

[2] http://www.cbs.dtu.dk/researchgroups/metagenomics/metagenomics.php

functional investigations of complex microbial communities. Metagenomics is often used as a hypernym for both targeted amplicon sequencing, as well as for whole genome shotgun sequencing, which is based on the underlying analysis workflow, shown in Fig. 1.3.

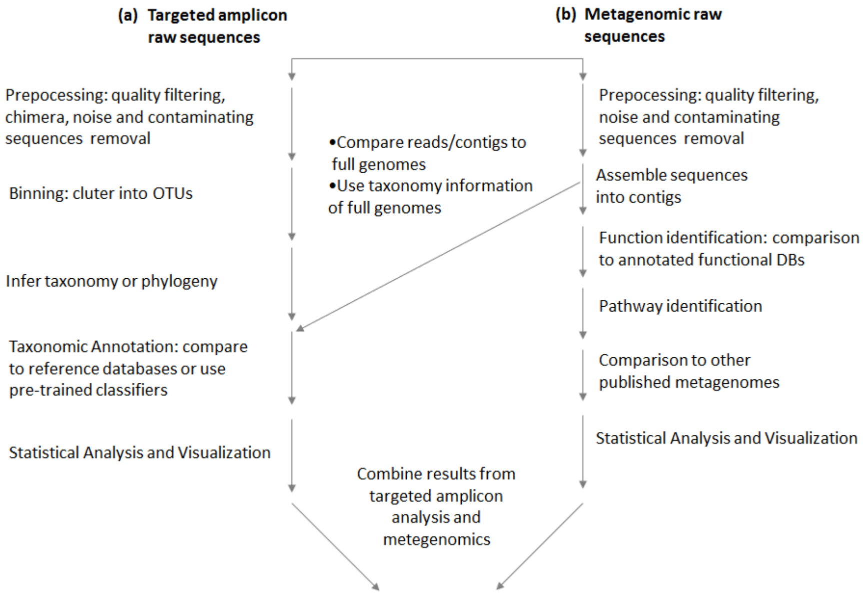

(a) Targeted amplicon raw sequences
(b) Metagenomic raw sequences
Prepocessing: quality filtering, chimera, noise and contaminating sequences removal
Binning: cluter into OTUs
Infer taxonomy or phylogeny
Taxonomic Annotation: compare to reference databases or use pre-trained classifiers
Statistical Analysis and Visualization
•Compare reads/contigs to full genomes
•Use taxonomy information of full genomes
Prepocessing: quality filtering, noise and contaminating sequences removal
Assemble sequences into contigs
Function identification: comparison to annotated functional DBs
Pathway identification
Comparison to other published metagenomes
Statistical Analysis and Visualization
Combine results from targeted amplicon analysis and metegenomics

Figure 1.3.: Summary of the bioinformatic analysis workflow of (a) targeted amplicon (left branch) and (b) metagenomics (right branch) data analysis. (*figure modified from [55]*)

1.1.1. Targeted Amplicon Sequencing (*"Who is there?"*)

To characterize and classify complex microbial communities, a marker gene which is shared amongst the whole domain is amplified by a set of universal primers from DNA, which is directly extracted from the environmental sample, followed by sequencing of the amplicons [56]. Typical marker genes are the ribosomal RNA of the small subunit (16S SSU rRNA) for bacteria and archaea [57], and the internal transcribed spacer (ITS) for fungi and lichens [58]. Both marker genes encode partially for ribosomal DNA. Ribosomes are shared amongst all organisms and are highly conserved within different species due to high evolutionary pressure and their role in protein biosynthesis [59]. Additionally, different ribosomal structures (Tab. 1.1) in prokaryotes and eukaryotes allow for distinct analysis of microbial communities of these two domains. Ribosomal subregions are relatively short (16S about 1.5 kbps, ITS about 800 kbps) making them faster and cheaper to sequence than many other unique microbial genes.

Table 1.1.: Description of ribosomal structure in prokaryotes and eucaryotes. Both ribosomes comprise a small and a large subunit, which can be distinguished according their sedimentation coefficient, molecular mass, and proteins.

Pro. ribosome	Large Subunit	Smal Subunit	# Protein
70S	50S	23S (2904 nt)	31
		5S (120 nt)	
	30S	16S (1542 nt)	21

Eu. Ribosome	Large Subunit	Smal Subunit	# Protein
80S	60S	26S (4718 nt)	49
		5.8S (160 nt)	
		5S (120 nt)	
	40S	18S (1874 nt)	33

Although the small ribosomal subunit (16S) is only about 1.5 kbps long, it exceeds the maximum read length of sequencing platforms of Illumina and Roche (∼400 bps (paired-end) and ∼470 bps, respectively). Therefore, only particular regions of the 16S gene, such as the hypervariable region 1 and 2 (V1 and V2) are amplified in typical community characterization studies. Van de Peer *et al.* discovered in 1996 [60] nine hypervariable regions (V1-V9) flanking highly conserved areas within the structure of the 16S genes [60], shown in Fig. 1.4. The length of the different regions ranges from about 200 to 470 bps which make them to an attractive loci for targeted amplicon studies [61].

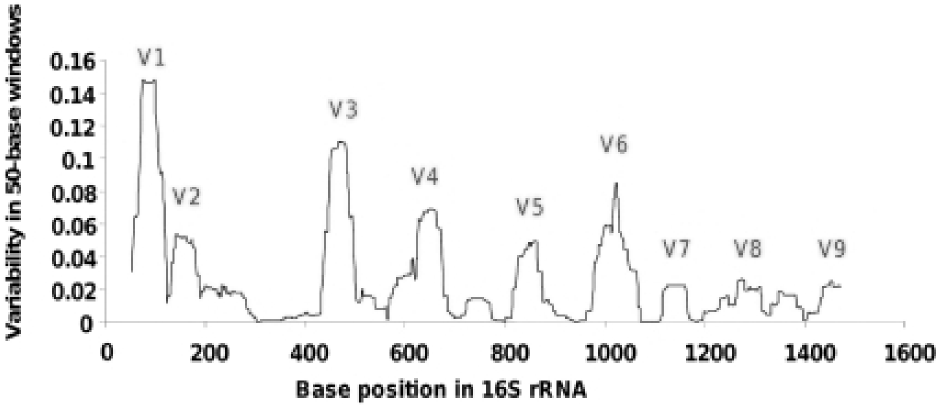


Figure 1.4.: The structure of the small ribosomal subunit is characterized by 9 hypervariable regions flanking highly conserved loci of the small ribosomal subunit gene. The entire SSU is about 1.5 kbps long and includes hypervariable regions (V1-V9) which range from about 200 to 470 bps. (*Figure taken from [62]*)

As a marker gene for the identification of fungi and lichens, a part of the eukaryotic ribosome is used. However, in this case, the internal transcribed spacer (ITS) one and two (ITS1, ITS2) have been introduced by Schoch *et al.* 2012 as universal loci for the characterization of fungi and lichens. The eukaryotic ribosome is organized in tandem repeats all over the genome which are separated by untranscribed spacer regions. Within the ribosome, the small subunit is separated by two internal spacers from the main large subunit [63], shown in Fig. 1.5. These internal spacers are transcribed into rRNA but are removed before the final ribosome is built. Compared to the prokaryotic ribosome, variable regions D1 and D2 have been identified within the LSU, [63] and have been used for fungal community characterization as well [64].

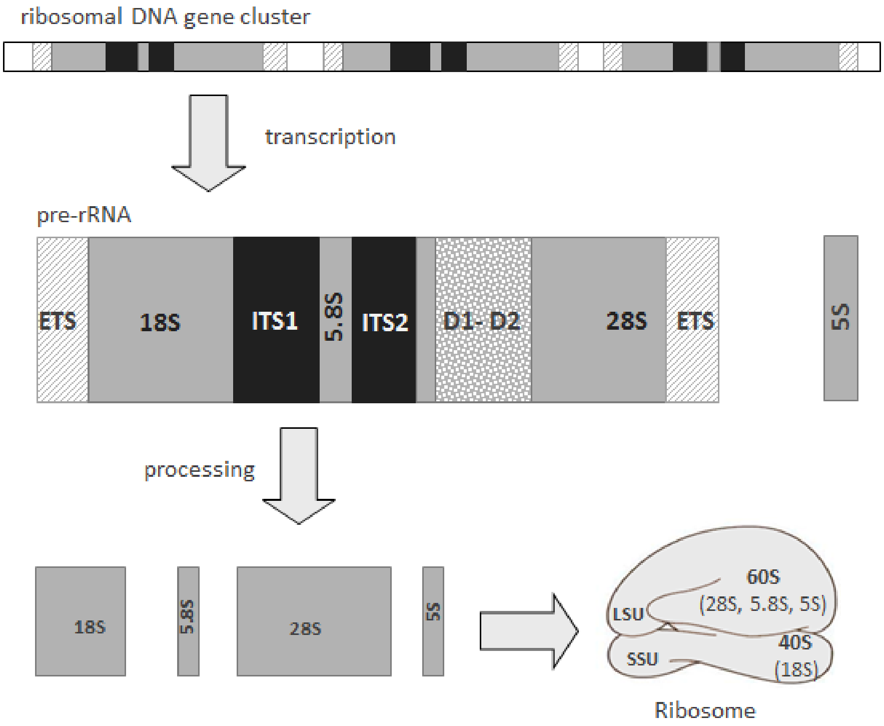


Figure 1.5.: The eukaryotic ribosome gene cluster is organized in tandem repeats along the genome which are separated by untranscribed spacers. It comprises the large subunit 60S (including 28S, 5.8S, 5S rRNA) and the small subunit 40S (18S rRNA), whereby the 5S rRNA can be encoded far apart from the main gene cassette. Variable regions D1/D2 have been identified within the beginning of LSU. SSU and LSU are separated by two internal transcribed spacer (ITS) regions. These regions are transcribed into rRNA but removed before the mature ribosome is finally formed.

In targeted amplicon studies a single sequencing run results in thousands of sequence reads for a given sample. To determine the phylogentic composition, individual sequences are assigned to operational taxonomic units (OTUs). Each OTU represents a specific taxonomic group at a particular phylogentic level (commonly 97 % similarity at the sequence level corresponds to distinct species [65], 95 % to distinct genus level groups). Finally, taxonomic classification and the quantitative number of reads which were assigned to a particular OTU completes its annotation. During the last decade, a variety of tools have been introduced to analyze amplicons of microbiome surveys. Basically, they can be divided into two main approaches based on the used OTU picking method: OTUs are either generated by ***de-novo* OTU picking**, which is based upon unsupervised clustering, or ***reference/taxonomy* OTU picking**, in which OTUs are formed by comparative classification using a reference database [55].

The basic workflow of both approaches is illustrated in Fig. 1.6. Briefly, *de novo* based OTU picking, shown in Fig. 1.6a, comprises the following core working steps: (1) pre-processing (sample splitting, trimming, removal of contaminating, or chimeric sequences, denoising, quality filtering); (2) aligning sequences using multiple sequence alignment (msa); (3) calculation of the distance between all sequences to allow for an accurate calculation; (4) *de novo* OTU picking, clustering of sequences according to their sequence similarity into distinct OTUs; (5) dereplication (for each OTU a representative sequence is selected); (6) classification (each OTU representative is classified either by a similarity search against a reference database or with an estimation approach); and (7) statistical analysis and visualization (PCA, heatmaps, DA analysis, phylogenetic distribution bar charts, calculation of diversity measures) [66].

Reference based OTU picking, shown in Fig. 1.6b, comprise the following working steps: (1) pre-processing (sample splitting, trimming, removal of contaminating, or chimeric sequences, denoising, quality filtering); (2) taxonomic classification by a similarity search against a reference database; (3) OTU generation by grouping sequences according to their taxonomic classification; and (4) statistical analysis and visualization (PCA, heatmaps, DA analysis, phylogenetic distribution bar charts, calculation of α- and β-diversity measures) [66].

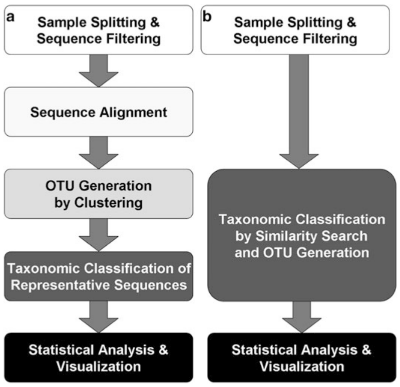

a
Sample Splitting & Sequence Filtering
Sequence Alignment
OTU Generation by Clustering
Taxonomic Classification of Representative Sequences
Statistical Analysis & Visualization
b
Sample Splitting & Sequence Filtering
Taxonomic Classification by Similarity Search and OTU Generation
Statistical Analysis & Visualization

Figure 1.6.: Typical workflow for **(a)** *de novo* based OTU comprising five main steps: (1) pre-processing, (2) sequence alignment, (3) sequence clustering, (4) taxonomic classification of cluster representative sequence, (5) statistical analysis and visualization. **(b)** *reference/taxonomy* based OTU picking comprising three main steps: (1) pre-processing, (2) taxonomic classification by a similarity search against a reference database; (3) OTU generation by grouping sequences according to their taxonomic classification; (4) statistical analysis and visualization. *Image taken from [66]*[3].

For the described steps, a variety of tools and pipelines have been developed. Most popular tools have been integrated in ready to use web-based pipelines, such as SnoWMAn [67], or CloVR [68]. Or summarized in command line packages such as mothur [69] or Quantitative Insights Into Microbial Ecology (QIIME) [70].

1.1.2. Shotgun Metagenomics (*"What are they doing?"*)

Metagenomics enables culture-independent studies based on the whole genetic information of complex microbial communities which are directly sampled from a particular environment. Sequencing of the whole genome provides information about structure, function, and interactions of the microbial community with its habitat [71]. This approach results in a much more complete community description than targeted amplicon studies. Moreover, with metagenomics it is possible to discover potentially novel biocatalysts or enzymes, and construct evolutionary profiles of community structure and function. Additionally, genomic linkages between function and phylogeny for uncultured organisms can be established [72]. The main steps involved in a typical metagenome project are: (1) filtering of raw reads prior to main downstream analysis, (2) assembling of reads into contigs, (3) comparing assembled contigs to

[3] Computational Medicine by Springer. Reproduced with permission of Springer in the format Thesis/Dissertation via Copyright Clearance Center

whole reference genomes, (4) taxonomic, as well as functional classification, and (5) identification of functions, corresponding pathways using pathway databases such as KEGG [73].

Furthermore, metagenomic profiles can be compared to other available metagenomes. Finally, targeted amplicon analysis and metagenomic results can be combined for final conclusions. Along with providing descriptive analysis about the composition and function of microbial communities, metatranscriptomics allow additionally for the investigation of underlying regulatory mechanisms. A variety of bioinformatic tools, methods, and algorithms have been developed for each single step of the metagenomic analysis workflow [55]. The most commonly and widespreadly used are combined in automated analysis servers and pipelines such as the MG-Rast server [74] and MEGAN [75].

1.1.3. (Meta)Transcriptomics (*"How are they doing it"*)

In contrast to a metagenome, a *(meta)transcriptome* comprises only sequence information of active - expressed genes at the time, place or state of investigation [76]. High-throughput sequencing of mRNAs obtained from natural microbial communities (*metatranscriptomics*) or from a single microbial genome *transcriptomics* are able to provide the first insights into their activities and regulatory mechanisms [77]. Especially, a technique called *differentially RNA-seq (dRNA-seq)* enables the selective analysis of primary transcripts in the genome [78]. Whether these genes in question are translated into proteins is triggered by a number of either enhancing or inhibiting factors. Investigations of the promoter region, 100-1000 bps 5′ upstream of the transcription start site (TSS), of a primary transcript allows identification of sequence motifs for regulatory elements and transcription factor binding sites. Therefore, conclusions about regulatory mechanisms are feasible [76].

1.2. Challenges of High-Throughput Sequencing Data Analysis

Sequencing technologies are rapidly improving, and due to their steadily decreasing costs, they have become more and more a standard procedure in research, as well as in clinical practice [79]. Unfortunately, bioinformatic tools, methods, and algorithms are struggling to keep pace with current developments, scientific findings, and newly arising requirements.

A ubiquitous and fundamental step in targeted amplicon studies is the pre-processing of the raw sequencing data. This is of great importance for two reasons. First, low quality sequences, artificially created, or contaminating sequences hamper and slow down the analysis process. Second and even more important, these kinds of data skew the analysis result (OTU inflation) and compromise correctness and quality of the final conclusions. Of special interest are misamplified fragments originating from the host environment, such as human DNA fragments in bacterial GI community studies. Commonly available tools such as *DeconSeq* [80] are based upon alignment against special sequence collections such as human or mouse. This is a problem in cases where the origin of the contamination is unclear, or misamplified sequences can originate from multiple sources.

The core step of high-throughput characterization studies is the taxonomic classification of the amplified fragments. Reference sequence databases are a key resource in the classification and characterization of complex microbial communities. Regardless of the classification approach used, a proper reference system is required for final taxonomic annotation. Although reference systems for prokaryotes (bacteria and archaea) are well-established [81], comparable systems for eukaryotes such as fungi and lichens are far away from being complete [81] even though they have taken on greater importance during the last years. These systems have to be continuously extended, curated, and maintained by experts to ensure high quality reference sequences, as well as complete and correct annotations. Collective annotation approaches such as initiated by Nilsson *et al.* [81] are important efforts towards reliable reference archives.

Curated reference databases such as UNITE's [82] formated versions for QIIME and mothur or classification systems such as SnoWMAn's fungal BLAT pipeline, or RDP's LSU [64] and ITS [83] classifier version have been recently released and are still under development.

The final step of each high-throughput characterization study is the visualization and statistical analysis of the classification results, shown in Fig. 1.3. Although statistical approaches such as PCA, rarefaction, as well as α- and β-diversity measures are well-established for data description, mechanisms for testing significant changes of OTU abundance between different groups, places, states, or time points are lagging behind. The classification result of microbial community surveys is represented as a so-called *feature matrix* containing the number of reads observed (counts) for each OTU (feature) for every single sample. This kind of representation is similar to the final result of RNA-seq experiments. The major purpose of this type of experiment is the detection of differentially expressed (DE) genes. Based on the similar nature of these two kinds of problems and outputs, the evaluation of this already well-established tools are needed on community data, to enrich final statistical analysis of the final community profile.

1.3. Objectives

The aim of this thesis is to introduce and evaluate new, as well as existing approaches, methods, and tools for single steps of the entire high-throughput characterization and classification process of complex microbial communities. Furthermore, these new or extended approaches, methods, and tools are used for the analysis of real datasets to solve different biological, medical, or ecological questions.

The **major aims** which are specified within this thesis are listed below.

- **Investigation of the influence of sequencing library normalization on the final community profile and its diversity**
- **Development of an application for the identification and removal of contaminating sequences**
- **Integration and evaluation of resources for fungal community analysis**
- **Evaluation and adaption of methods for differentially abundant feature detection**
- **Transcriptome analysis of the *Campylobacter fetus* subspecies *fetus* and *veneralis***

The following sections link the specified objectives with the projects and data, which are analyzed within this thesis and the respective results, which are used to implement, evaluate and adopt approaches, methods, and tools.

1.3.1. Investigation of how sequencing library normalization affects community profile and diversity

The effect of sequence library normalization should be investigated by the analysis and comparison of a standard and a normalized sequencing library, which originates from the *metagenome moss project* (Sec. 2.1.4). This sequencing approach was planned by members of the Bioinformatics Group of Dr. Gerhard Thallinger[4] together with

[4] Bioinformatics Group, Institute for Knowledge Discovery, Graz University of Technology, Graz, Austria

the team of Prof. Dr. Gabriele Berg[5]. Aside from the comparison of untreated and normalized sequencing libraries, the main aim of this Illumina-based metagenomic approach is to facilitate deeper insights into specific biochemical pathways and adaptive strategies through the analysis of significantly changing functional subsystems.

1.3.2. Development of an application for identification and removal of contaminating sequences

An application for automated detection and removal of contaminating sequences should be developed within the scope of this thesis. Firstly, the tool should be evaluated using a well-described sequence set. Subsequently, the new pre-processing approach should be applied on the amplicon sequence set generated within the *diarrhea study* (Sec. 2.1.5) to demonstrate the effects of contaminating sequences on community diversity.

1.3.3. Integration and evaluation of resources for fungal community analysis

The fungal amplicon set, originating from a bronchoalveolar lavage survey should be analyzed using the targeted amplicon sequencing pipeline SnoWMAn. Therefore, a reference database for ITS amplicons has to be generated, tested, and incorporated into the analysis pipeline. Additionally, the bacterial community profile of this survey should be determined. The study was planned and realized by the team of Prof. Robert Krause[6], MD, to investigate the relationship of risk factors for *Candida* colonization (Sec. 3.8). Therefore, the obtained community profile has to be tested for DA abundant features, in the different experimental conditions of the ITS community profiles, as well as in the bacterial communities. Furthermore, the results of the high-throughput classification should be compared to the results of the traditional BAL culture analysis.

[5] Institute of Environmental Biotechnology, Graz University of Technology,Graz, Austria

[6] Section of Infectious Diseases and Tropical Medicine, Medical University of Graz, Graz, Austria

1.3.4. Evaluation and adaption of methods for differentially abundant feature detection

To evaluate different methods for differential feature detection, simulated count data (according to [84]), with a known number of truly DA features should be tested with different methods. According to the result of this evaluation approach, community profiles of data sets, which are evaluated within this thesis, should be tested for differentially abundant features. In detail, the count data obtained from the datasets created withing the *metagenome moss project* (Sec. 2.1.4), as well as from the *Candida* (Sec. 2.1.1) and GI mouse amplicon studies (Sec. 2.1.3) are within the scope of the differentially abundant feature detection.

1.3.5. Transcriptome analysis of *Campylobacter fetus* subspecies fetus and veneralis

The expression data which is analyzed within this thesis was generated by the team of Dr. Sabine Kienesberger[7] and Ass.-Prof. Gregor Gorkiewicz[8], MD, in the course of the comprehensive study on two *Campylobacter fetus* subspecies (Sec. 2.1.2). Besides, the comparative analysis effort of the two *Campylobacter fetus* subspecies, regulatory elements which might influence metabolism and virulence of the subspecies have been of special interest. Therefore, differentially RNA-sequencing (dRNA-seq) was performed. This data should be subjected to automated TSS identification and categorization, as well as for the subsequent motif analysis in the determined promoter regions.

[7] Institute of Molecular Biosciences, University of Graz, Graz, Austria

[8] Institute of Pathology, Medical University of Graz, Graz, Austria

2. Methods

The following sections describe the datasets, databases, approaches, applications, methods, tools, and algorithms which were used to solve the discussed results of this thesis. Different applications and resources are grouped according to their main topic or overall characteristics.

2.1. Datasets

Five distinct datasets have been analyzed at different steps, of the analysis workflow, using different high-throughput analysis approaches to investigate and answer various questions. The following sections introduce and describe the experimental design, as well as how the data was sampled, prepared, and analyzed prior to bioinformatic analysis.

2.1.1. Bronchoalveolar Lavage (BAL) Study

The sampling effort for the BAL study comprise 55 adult patients (age $>$ 18 years) who were assigned according their health state and medical treatment into three main groups (1,2,3). Group 1 and 2 were additionally split into two more subgroups according to antibiotic treatment (A = no antibiotic treatment, B = with antibiotic treatment).

Group 1 (control group) includes fifteen healthy patients who did not show any clinical, radiological or laboratory evidence for an infectious diseases at sampling time point. *Group 2* comprise thirteen non-neutrophenic intubated and mechanically ventilated

patients who where treated at the intensive care unit (ICU). Non of them showed indications for a community acquired (CAP) or ventilator-associated (VAP) pneumonia according to common case definitions [85, 86].
Group 3 includes twenty-seven patients who showed indications of VAP, CAP, or aspiration associated pneumonia (ASP) according to common case definitions [85, 86], as well as by X-rays of the lungs which were examined by blinded and independent investigators. All patients in group 3 were treated with antibiotics because of their disease state.
Patients who received antifungal therapy within 8 weeks before study start, as well as patients with pulmonary diseases such as chronic obstructive pulmonary disease (COPD), sarcoidosis, asthma bronchiale, malignant diseases of the lung, or intestinal lung disease were excluded. Further exclusion criteria comprise all kinds of immunosupressive therapy, or HIV.

A total of 59 samples were obtained by deep tracheal aspiration (all samples group 1), bronchoscopic bronchoalveolar lavage of the right lung (all samples of group 2), or directed pulmonary infiltrates suggestive of VAP. Samples were brought immediately after extraction to the in-house microbiology laboratory of the Medical University of Graz, aliquoted and stored at -70 °C until further analysis. Bacterial, as well as fungal DNA was extracted from both tracheal secretion and BAL samples by using the MagNA Pure LC DNA Isolation Kit III[9]. The variable region 4 (V4) of the 16S small-subunit (SSU) ribosomal gene was amplified from the obtained DNA isolated by PCR, using the forward and revers primers given in Tab. 2.1 in combination with 30 6-mer multiplexing identifiers (MID).

Table 2.1.: Forward and reverse sequencing primers used for amplification of the variable region four (V4) of the 16S gene within the BAL study.

Name	Dir.	Sequence
V4_RDP_FWD	FWD	AYTGGGYDTAAAGNG
V4_RDP_REV1	REV	TACCRGGGTHTCTAATCC
V4_RDP_REV2	REV	TACCAGAGTATCTAATTC
V4_RDP_REV3	REV	CTACDSRGGTMTCTAATC
V4_RDP_REV4	REV	TACNVGGGTATCTAATCC

[9] MagNA Pure LC, Roche Diagnostics, Vienna; http://www.roche.at

The fungal ITS1 region was amplified in triplicate from DNA sample extracts using the forward and reverse primers, given in Tab. 2.2 in combination with 30 6-mer MIDs.

Table 2.2.: Forward and reverse sequencing primers used for amplification of the internal transcribed spacer region 1 (ITS1) of the fungal ribosomal gene within the BAL study.

Name	Dir.	Sequence
ITS1F	FWD	CTTGGTCATTTAGAGGAAGTAA
ITS2	REV	GCTGCGTTCTTCATCGATGC

After PCR an amplicon library was generated using equimolar amounts of PCR products derived from the individual samples and bound to sequencing beads. Final sequencing was performed on a Roch 454 GS FLX system at the Center of Medical Research[10] (ZMF) according to the manufacturers protocol.

Ethic statement

The study was approved by the institutional review board of the Medical University of Graz (protocol no. 19-322 ex 07/08). From all subjects written informed consent was obtained.

2.1.2. *Campylobacter fetus* Study

Campylobacter (C.) strains were grown on Columbia blood agar (CBA) plates containing 5 % sheep blood[11] at 37 °C in a microaerobic atmosphere[12] for 24 h [87]. A total number of 102 *C. fetus* strains were characterized distinctly to subspecies level and subsequently tested in polymerase chain reaction (PCR) screens. Biochemical identification of subspecies *C. fetus fetus* and *C. fetus veneralis* were performed according to growth in the presence of 1 % (wt/vol) glycine and the reduction of 0.1 % sodium selenite in liquid culture [88]. For all isolates a subspecies-specific PCR assay [89] was applied. For special cases, amplified fragment length polymorphism analysis

[10] Graz, Austria; http://www.medunigraz.at/zmf/

[11] bioMerieux, Marcy l'Etoile, France; http://www.biomerieux.fr/

[12] GENbag/GENbox MicroAir; bioMerieux, Marcy l'Etoile, France; http://www.biomerieux.fr/

[90] and pulsed-field gel electrophoresis [91] was performed, to clarify equivocal results [88]. For one representative of each subspecies (*Campylobacter fetus fetus, Cff* and *Campylobacter fetus veneralis, Cfv*) a library was prepared for differentially RNA sequencing (dRNA-seq) according to the protocol described by Sharma *et al.* 2010 [92]. Briefly: extracted RNA was split into aliquots for cDNA library pairs construction. One aliquote was treated with Terminator-5-phospate-dependent exonuclease[13] (TEX) to deplete processed RNAs (denoted TEX+); untreated library (denoted TEX-) [92]. cDNA libraries were constructed by *vertis* Biotechnology AG[14] prior to TEX treatment. Cluster amplification was performed with Illumina's TruSeq PE Cluster Kit v.5[15] on a Cluster Station. Libraries (TEX+, TEX-) were sequenced on two distinct lanes on Illuminas HiSeq 2000 platform according to the TruSeq SBS 36 Cycle Kits v.5[15] and a 91 bps single-end protocol at the Institute for Molecular Infection Biology[16] (IMIB). Final sequencing image files were processed with Illumina's Sequencing Control Software (SCS), Real Time Analysis (RTA) v2.6, and CASAVA v.1.7[6] [38].

2.1.3. Gastrointestinal Mouse Study

The project was divided into two main sampling efforts. The first experimental setting comprise seven wildtpye (WT), Mdr2 knockout (Mdr2-KO), and bile-duct ligated (BDL) mice, each treated under normal diet (chow) for a period of eight weeks. Mdr2-KO mice lack the liver specific P-glycoprotein which triggers the phosphatidylcholine transport across the canalicular membrane [93]. As a consequence, secretion of phospholipids into the bile is not possible within this type of mice. In BDL mice linkage between bile and liver and so any kind of secretion is interrupted [94]. Fecal (F), as well as mucosal (M) samples from different colonic locations, ileum (Ile), jejunum (Jej), and caecum (Cae) had been collected. The experimental design of the first sampling effort is summarized with the supplementary information in Appendix Tab. A.1.

The second sampling effort comprise two phenotypes, wildtype, and Mdr2-KO

[13] Epicentre (an Illumina company); Madison, WI, USA; http://www.epibio.com/

[14] Munich, Germany; http://www.vertis-biotech.com/

[15] Illumina Inc. San Diego, CA, USA, http://www.illumina.com/

[16] Wuerzburg, Germany; http://imib-wuerzburg.de/

mice which were treated under a normal and a phospholipid enriched diet (phosphatidylcholines (PC); 5 % Phosphatidylcholin-enriched chow) over a period of eight weeks. Fecal (F), as well as mucosal (M) samples from different colonic locations, ileum (Ile), colon (Col), and caecum (Cae) had been collected. The experimental design of the second sampling effort is summarized with the supplementary information in Appendix Tab. A.2 and Tab. A.3.

Details about mice treatment, surgical procedures, and sample collection are available as supplementary information, Appendix (pp 178, 176). Briefly: the entire colon was ligated and removed. To ensure purity of the extracted sample, distinct colonic regions were ligated too. Fecal samples were obtained by incision of the particular colonic regions and extraction of its entire content. Subsequently, the colonic section was washed twice in 10 ml of sterile 0.9 % NaCl solution to remove all traces of feces. Obtained samples were stored at -20 °C until further analysis.

Community DNA was extracted using the Magna Pure LC DNA III Isolation Kit[17], according to the manufacturers protocol[18]. Hypervariable region one and two (V1-V2) of the 16S small ribosomal subunit was amplified from the obtained DNA isolated by PCR, using the forward and revers primers given in Tab. 2.3, in combination with 30 10-mer MIDs. After PCR an amplicon library was generated using equimolar amounts of PCR products derived from the individual samples and bound to sequencing beads. Final sequencing was performed on a Roche 454 GS FLX instrument at the Center of Medical Research[19] (ZMF), according to the manufacturer's recommendations.

Table 2.3.: Forward and reverse sequencing primers used for amplification of the variable region one and two (V1-V2) within the GI mouse study.

Name	Dir.	Sequence
V12_RDP_FWD	FWD	AGAGTTTGATCCTGGCTCAG
V12_RDP_REV	REV	CTGCTGCCTYCCGTA
V12_RDP_REV1	REV	ATTACCGCGGCTGCTGG

[17] MagNA Pure LC, Roche Diagnostics, Vienna; http://www.roche.at/

[18] Magna Pure LC DNA Isolation Kit III (Bacteria, Fungi), Version 13, November 2012

[19] Graz, Austria; http://www.medunigraz.at/zmf/

Ethic statement

The study was approved by the Federal Ministry of Science and Research[20], Austria, according to (TVG, BGBI. Nr. 501/1989 i.d.F. BGBI. l Nr. 162/2005; TVN, GZ: BMWF-66.010/0046-II/3b/2012).

2.1.4. *Sphagnum* Moss Study

Peat moss samples of type *Sphagnum magellanicum* Brid. (section Sphagnum) were collected from the Austrian Alpine bog Pirker Waldhochmoor[21] in December 2011. Four replicates represented by gametophyte[22] living moss plants were collected from four independent sampling points separated by 15 m each. The collected samples were stored in sterile plastic bags at 4 to 8 °C, during transportation to the laboratory, for further processing. Community DNA isolation of the *S. magellanicum* microbiome was performed by hybridization of 200 g of each sample in Stomacher bags (20 g/bag) blended with 0.85 % NaCl solution (50 ml/bag); followed by subsequent shaking of the diluted samples in a Stomacher laboratory blender[23] for 3 min. Plant residuals were removed from the suspension by a two-stage filtering process (500 μm and 63 μm). After discarding the supernatant[24] the remaining pellets were resuspended in 1.5 ml of 0.85 % NaCl and centrifuged at high speed (10,000 g, 4 °C) for 20 min. Finally, the obtained pellets were stored at -70 °C before DNA isolation. Community DNA was extracted using the FastDNA Spin Kit for Soil[25] according to the manufacturer's standard operating procedure (SOP). Before final sequencing, DNA aliquots from all samples were pooled together. Paired-end whole genome shotgun sequencing was performed by Eurofines MWG Operon[26] on the Illumina HiSeq 2000[27] platform (2x 100 bp). To allow for deeper ecological analysis most dominant sequences were removed by applying a normalization treatment (Sec. 2.9.1) on one aliquot of the total

[20] Vienna, Austria; http://www.bmwfw.gv.at/

[21] N46°37′38.66″ E14°26′5.66″

[22] The haploid (sexual) state of adult plants and fungi.

[23] BagMixer, Interscience, Saint Nom, France; http://www.interscience.fr/

[24] Material that floats on the surface of a liquid.

[25] BIO 101, Qbiogene Inc., Carlsbad, CA, USA; http://www.qbiogene.com

[26] Ebersberg, Germany; http://www.eurofinsgenomics.eu/

[27] Illumina Inc. San Diego, CA, USA; http://www.illumina.com/systems/

community DNA prior to sequencing. The other aliquot was sequenced untreated using the standard protocol. Library normalization, as well as the final sequencing was performed by Eurofines MWG Operon[28].

2.1.5. Diarrhea Study

To investigate alternations of the colonic microbiota in response to osmotic diarrhea four voluntary healthy Caucasian male adults were tested. Age of the subjects ranged from 26 to 47 at a body mass index (BMI) range of 24 to 26.6. None of the four subjects (A-D) suffered either from diarrhea or had been treated with antibiotics for at least twelve months prior to study start. During the study stool frequency and consistency were daily monitored and recorded according to the Bristol stool chart [95]. Fig. 2.1 illustrates the study design comprising the four main treatment periods and the four sample collection time points.

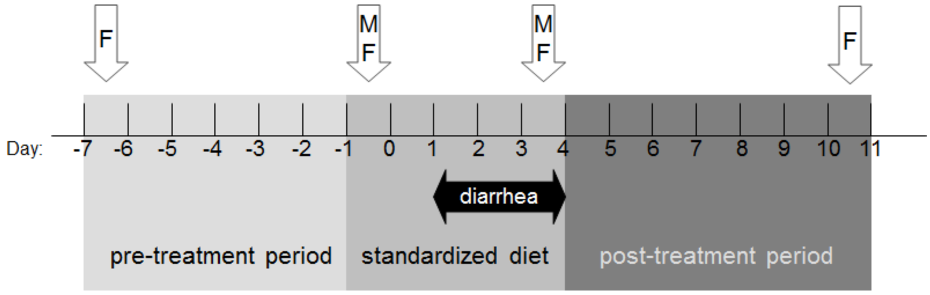


Figure 2.1.: Experimental design of the study "Alterations in the colonic microbiota in response to osmotic diarrhea" [96]. From day -7 to day -2 all subjects were set on a free diet, followed by a controlled standard diet from day -1 to day 0. Stool samples (F) were collected one week before and one week after the induction of diarrhea (remission). Additionally, two more fecal samples were taken together with mucosal biopsys (M) before the first dose of PEG (day 0) and when diarrhea was maximally pronounced (day 3) [96]. *The figure is modified from [96]*

.

Pre-treatment period lasted for six intervention free days which was followed by five days of *standard diet* (total calorie intake of 2150 kcal/d comprising 85 g protein, 77 g fat, 250 g carbohydrates, and 25 g fiber). On day three of the standard diet period, osmotic diarrhea was induced by a dose of 50 g tid (150 g/d) of osmotic

[28] Ebersberg, Germany; http://www.www.eurofinsgenomics.eu/

laxative polyethylene gycol (PEG) 4000[29]. This treatment was applied on the next three consecutive days. Finally, seven days of stool observation and free diet, *post-treatment period* completed the study. Fecal samples (F), directly taken from stool, were collected at four different time points by all subjects (-7, 0, 4, 7). Additionally, colonic mucosa (M) samples were obtained by biopsy from three subjects (B-D), on day zero and day four. Therefore, the targeted region was properly prepared before two biopsies were taken. Both sample types were immediately frozen and stored at $-20\,^{\circ}\mathrm{C}$ for further analysis. Subjects, time points, and tissue types are summarized with the Appendix Tab. D.1.

DNA was extracted using the QIAamp DNA Stool Mini Kit[30] (stool) and the QIAmp DNA Mini Kit[30] (mucosal tissue) according to the manufacturers protocol. To increase bacterial DNA yield the stool homogenate was incubated in boiling water for 5 minutes prior to DNA extraction. Hypervariable region V1-V2 had been amplified with a set of universal primers, given in Tab. 2.4 in combination with 6-mer MIDs. Sequencing was performed on a Roch 454 GS FLX Sequencer at the Center of Medical Research[31] (ZMF).

Table 2.4.: Forward and reverse sequencing primers used for amplification of the variable region one and two (V1-V2) within the diarrhea study.

Name	Dir.	Sequence
BSF8	FWD	AGAGTTTGATCCTGGCTCAG
BSR357	REV	CTGCTGCCTYCCGTA

Ethic statement

The study was approved by the institutional review board of the Medical University of Graz (protocol no. 20-090 ex 08/09). From all subjects written informed consent was obtained.

[29] Forlax, Merck, Vienna, Austria; http://www.merck.at

[30] Qiagen, Hilden, Germany; http://www.qiagen.com/

[31] Graz, Austria; http://www.medunigraz.at/zmf/

2.2. High-Throughput Characterization of Microbial Communities using SnoWMAn

The Straightforward Novel Webinterface for Microbiome Analysis[32] (SnoWMAn) [67] was developed as a user-friendly and straightforward web application for pre-processing, taxonomic classification, visualization, and statistical analysis of sequences obtained from targeted amplicon sequencing experiments. It consists of five different analysis pipelines, BLAT [97] and JGAST [98] adhering the reference/taxonomy based OTU picking approach. Additionally, pipelines based on UCLUST [99], RDP [100], and mothur [69] are available, which support OTU picking by cluster formation (according the *de novo* OTU picking approach). A typical SnoWMAn community analysis includes three main steps. First, obtained sequence data, quality files, primer sequences, and descriptive metadata information has to be uploaded to SnoWMAn's data repository. In the second step, one out of five different pipelines and respective parameters have to be chosen for community data analysis. Finally, the third step provides common visualization and statistical analysis capabilities on the classification result [66].

Once the sequence data (FASTA [101] formated, and optional quality information), as well as the mandatory description file (as plain text file) is uploaded, it can be selected for taxonomic classification with one of the available pipelines, as described in Sec. 2.2.1 and Sec. 2.2.2.

Pre-processing within all pipelines includes sample splitting, optional quality-, chimera-, and duplicate filtering. In addition, sequences can be removed from downstream analysis according to their minimum length or the maximum number of ambiguous bases (N's). Primers added during the amplification step can be also removed prior to classification to improve overall quality, as well as to speed up the analysis.

[32] SnoWMAn; https://snowman.genome.tugraz.at/

2.2.1. BLAT Pipeline

The BLAT pipeline is based upon the reference/taxonomic OTU picking approach, see Fig. 1.6b. Each sequence of the microbial community is compared using the alignment tool BLAT described by Kent 2002 [97] to an indexed and FASTA formated reference database which is linked to a taxonomic annotation file. In a next step the reference database for the BLAT classification step can be selected by the user. Reference databases for bacteria and archaea 16S, as well as for fungal ITS amplicons are available. The final job submission procedure, and options for visualization and statistical analysis are generally discussed for all pipelines in Sec. 2.2.3.

2.2.2. RDP Pipeline

The RDP pipeline is based upon the Ribosomal Database Project (RDP) classifier [100] and adheres to the *de novo* OTU picking approach, see 1.6a. The classifier is implemented according a Naïve Bayes algorithm and can be trained on any kind of sequence data. Currently, classification models for 16S rRNA (bacteria and archaea) and fungal large ribosomal subunit (LSU) [64] sequences are available. Additionally, training data and a classifier version for fungal internal transcribed spacer (ITS) amplicons are announced [83]. Major RDP pipeline parameters are the structural alignment model and the classifier version which have to be selected based on the data within the penultimate step. Furthermore, cluster similarity and clustering steps can be adapted for each analysis run. For analysis the pre-processed sequences are aligned based on the selected secondary structure aligned model (such as *Infernal* [102]) to identify the overall shared "core region". Sequences are then clustered according to their sequence similarities into OTUs. For each cluster (OTU) a representative sequence (proxy) is selected for final taxonomic annotation of the entire cluster. The selected pre-trained RDP-classifier version is used to estimate taxonomic classification for each proxy. In the de-replication step classification of each OTU specific proxy is assigned to all members of the OTU. Based on the classification result several statistical analysis and visualization capabilities are available via the web-application. They are generally summarized and discussed for all pipelines in Sec. 2.2.3.

2.2.3. Statistical Analysis and Visualization

Before the analysis is finally submitted all selected parameters according to the chosen pipeline are summarized to give the user the chance for a last check. Once the analysis is started its status and a rough time estimation can be monitored via the web interface. Calculation time varies considerably according to the total number of sequences, ranging from hours to days. Therefore, users can choose e-mail notification optionally on analysis completion.

Final classification result can be visualized using different chart types such as bar charts, pie charts, or line plots in absolute or relative scale. In addition, α-, as well as β-diversity, rarefaction curves, and principal component analysis (PCA) [103] can be calculated to compare microbial composition across samples or groups. Phylogenetic overlap between samples can be easily visualized using integrated Venn diagrams [104]. Figures and underlying data generated by the web-application can be easily exported either as PNG, SVG, or as Microsoft Excel file. Furthermore, results of all intermediate steps such as filtered sequence data, results of distance calculation, clustering, and taxonomic classification can be exported and downloaded for further analysis [67].

2.3. Reference Sequence Databases

The following sections describe two databases and one database collection which were used for classification of the datasets described in Sec. 2.1.

2.3.1. Greengenes

The *Greengenes*[33] [105] 16S reference database includes well-curated, non-chimeric, and complete sequences of the 16S small-subunit ribosomal gene for bacteria and

[33] Greengenes; http://greengenes.secondgenome.com/

archaea [105]. The main features of Greengenes are (1) a standardized set of descriptive fields, (2) identification of potential chimeric sequences, and (3) taxonomic assignment using multiple sequence alignment (MSA). Additionally, taxonomic annotation for each sequence is available from independent curators, including Norman Pace [106], Wolfgang Ludwig [107], Phil Hugenholtz [108], as well as from National Center for Biotechnology Information (NCBI) [109], and the Ribosomal Database Project (RDP) [110].

2.3.2. M5nr - The M5 Non-Redundant Protein Database

The *M5nr*[34] [111] represents an indexed and searchable protein database which combines certain resources of many popular sequence databases such as NCBI GenBank[35] [109] and RefSeq[36] [112], KEGG[37] [73], Gene Ontology[38] (GO) [113], the Integrated Microbial Genomes at the Joint Genome Institute [114], the SEED project[39] [115], VBI's[40] PATRIC[41] [116], the evolutionary genealogy of genes: Non-supervised Orthologous Groups[42] (eggNOG) [117], and UniProt[43] [118] in one single database [111]. Similarity searches across this database collection is performed either by using BLAST [119] or BLAT [97].

2.3.3. UNITE

The *UNITE* system for DNA based fungal species circumscriptions[44] [82] is a reference system for molecular identification of fungi based on their ITS sequence. Each fungal species contained in the reference system is represented by at least two ITS sequences

[34] M5nr http://tools.metagenomics.anl.gov/m5nr/
[35] NCBI GenBank; http://www.ncbi.nlm.nih.gov/genbank/
[36] NCBI Reference Sequence Databse; http://www.ncbi.nlm.nih.gov/refseq/
[37] Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg/
[38] Gene Ontology; http://www.geneontology.org/
[39] SEED project http://theseed.org/
[40] Virginaia Bioinformatics Institute, Blacksburg, VA, USA; http://www.vbi.vt.edu/
[41] Pathosystems Resource Integration Center, Blacksburg, VA, USA; http://patric.vbi.vt.edu/
[42] eggNOG; http://eggnog.embl.de/
[43] UniProt; http://www.uniprot.org/
[44] UNITE; http://unite.ut.ee/

of the International Nucleotide Sequence Database Collaboration (INSDC)[45] [120]. Unite entries are given a unique and stable name based on the accession number type. Taxonomic and ecological annotations are regular uniformed, updated, and corrected as far as possible [81]. In addition, a subsystem of representative sequences is released, mainly for local sequence similarity searches. Therefore, sequences are clustered on different sequence similarity thresholds (97-99 %) into so-called *"species hypothesis" (SH)*. A random or manual selected representative is chosen for each of the hypothesis to represent the cluster within the subsystem. The web-based identification system is open to public and, in addition, different versions or formats of the UNITE database are offered for download[46] [82].

2.4. Sequencing Technologies and Platforms

Sequencing is the technique used to determine the primary structure, the series of base pairs in fragments of nucleotide sequences such as DNA or RNA. Platforms described in Sec. 2.4.1 and 2.4.2 are based upon *"sequencing by synthesis"* (SBS). This method uses single stranded DNA fragments which are sequenced by synthesizing the complementary strand along it [121]. Basically, the described NGS techniques can be distinguished either by template preparation protocols, chemistry, detection approaches, or their underlying base calling methods [122]. The following discussions will briefly introduce the two different sequencing technologies and commercial platforms that have been used for sequencing data generation of the processed datasets within this thesis.

2.4.1. Roche Genome Sequencer (GS) FLX Instrument

The 454 GS FLX instrument [123] is based on the detection of luminescence created during conversion of pyrophosphate (*pyrosequencing*) [124–126]) rather than chain termination with dideoxynucleotides (traditional Sanger principle [1]). It comprises

[45] INSDC; http://www.insdc.org/

[46] http://unite.ut.ee/repository.php

four major working steps: (1) ligation of adapters to DNA fragments; (2) emulsion polymerase chain reaction (PCR, amplification); (3) distribution of beads among a picotiter plate; and (4) pyrosequencing [127]. During sequencing for each base the four nucleotides are subsequently incorporated into the single strand of DNA. Successful incorporation of a nucleotide releases pyrophosphate (PPi) stoichiometrically. The following reaction leads to emission of light which can be detected by a camera and facilitates the sequence identification by the detected flowgram. The instrument is able to create 400-600 kbps per run with 400-500 bps read lengths in about 10 days. The vast amount within this analysis cycle is needed for the sample preparation. The actual sequencing step of the 454 GS FLX instrument takes only about 8 hours.

2.4.2. Illumina MiSeq and HiSeq Platforms

The *Illumina* sequencing technology [128] relies on SBS and is also known as short-read sequencing because of its maximum read length of $\sim$150-300 bps (Read length can be almost doubled by using paired-end mode, yielding in a maximum read length of $\sim$300-500 bps for each end of the template). Different instrument types such as the Genome Analyzer (GA) II, the MiSeq or HiSeq platform use more or less the same chemistry and are all based on the same sequencing principle which consists of three main working steps: (1) library preparation; (2) cluster generation (bridge-amplification); and (3) sequencing. During the sequencing step single stranded DNA is synthesized by adding the four types of fluorescently labeled bases at one time in each single step. Not incorporated molecules are washed away in each step which allows for subsequent sequence identification by the fluorescent signal [128].

Main difference of the currently available platforms are throughput and time. The *HiSeq* platform for instance, is able to produce up to 180 gbps (in Rapid Run Mode) per sample with maximum read length of 150 bps (paired-end). Eight samples can be processed without multiplexing within a single run in about 7 days. In contrast the *MiSeq* which processes only one sample per run (without multiplexing) yields to much less data per sample (540 mbps - 15 gbps) but benefits from lower run times, of 24-55 hours [129].

2.5. Development Tools

Basic software development tools, frameworks, and libraries have been used within this thesis. The following sections give a short introduction and overview on the used tools.

2.5.1. Java

The *Java* programming language [130] is one of the most popular programming languages in use, especially for client-server web applications [131]. The language is object-oriented and adheres class-based design patterns, and benefits from almost no implementation dependencies. Java applications are compiled to bytecode (class-files) which enables execution, in a platform independent manner, on any computer architecture providing a Java virtual machine (JVM) [130].

2.5.2. IGB-BioJava

IGB-BioJava is based upon the BioJava project [132], which is a freely available, open-source Java software project. It presents a Java framework which was primarily developed for processing of different kinds of biological data using the Java programming language. The major goal of the BioJava project is simplifying bioinformatics data analysis, processing, as well as application development. It comprises analytical and statistical routines and algorithms, parsers and converters, for various common file formats. The Institute for Genomics and Bioinformatics (IGB)[47] is hosting an extended and customized version of the general BioJava library, the so-called IGB-BioJava.

2.5.3. JFreeChart

JFreeChart [133] is a freely available, open-source Java chart library which simplifies integration of professional high quality charts in Java applications. JFreeChart supports

[47] Graz, Austria; http://genome.tugraz.at/

a wide rage of different chart and output types such as Swing components, image files (special noteworthy PNG and JPEG are supported), and common vector graphics file formats (such as SVG, PDF, and EPS). The flexible software design facilitates easy extension and is applicable for client-side, as well as for server-side application [133].

2.5.4. The R Project for Statistical Computing

The R project [134] is a powerful programming language and environment for statistical computing and graphics. R can be described as framework for data manipulation, calculation, and graphical visualization. The major strength of R is its suite of operators for array and matrix manipulations. As programming language R, is well-developed and supports common concepts such as conditionals, loops, recursions, user defined functions, as well as input and output facilities. The R environment represents a coherent system which is very flexible and easy to adapt and extend. For example it can be linked at runtime to C or C++ routines for computationally-intensive tasks or extended by packages originating for example from the Bioconductor project (Sec. 2.5.5) [134].

2.5.5. Bioconductor

Bioconductor[48] [135] is an open source, open development, software project to provide and develop bioinformatic, as well as computational biology tools for analysis, visualization, and comprehension of high-throughput genomic data generated by wet lab experiments or moleculare biology. As it is based on the R programming language, Biodonductor components are released and distributed as R packages. Moreover, the analysis framework offers a large number of meta-data packages which provide additional information about metabolic pathways, microarrays, genomes, organisms, and other annotations for data enrichment. A wide range of analysis facilities such

[48] Bioconductor; http://www.bioconductor.org/

as basic sequence analysis, statistical testing, DNA microarray, RNA-seq, ChIP-seq, annotation, flow cytometry, and other data analysis mechanisms are available [135].

2.6. Decontamination and Chimera Filtering

2.6.1. Blast Like Alignment Tool - BLAT

The *Blast Like Alignment Tool (BLAT)* [97] is a very effective and fast tool for rapid detection of sequence homology in highly similar sequences (nucleotide identity $\geq$ 95 %, translated protein identity $\geq$ 95 %) [136]. It was developed for measuring sequence homology of biological sequences such as DNA, RNA, or protein sequences in order to get information about their biological function. Although BLAT is based upon the *Basic logical alignment search tool* BLAST [119] heuristics, it does not calculate the optimal alignment of two sequences. It can use arbitrary sequence database and input sequence file to create the final BLAT result list, in decreasing order according to the calculated score. For each sequence of the input file the corresponding hit in the reference database and its qualitative parameters, such as *percentage of identity*, *number of gaps*, *number of mismatches*, *alignment length, and positions* are given [97]. Typical applications of BLAT include cross-species protein or mRNA alignments in order to determine homology, as well as detection of gene family members or protein-coding sequences of a specific gene [136].

2.6.2. DeconSeq

DeconSeq[49] [80] is a publicly available tool (web-based and stand alone version) for rapid, automated identification, and removal of contaminating sequences in metagenomic as as well in targeted amplicon datasets (minimum read length 150 bps) by alignment based comparison against reference genomes. The tool offers pre-processed reference databases for complete genomes, such as human, mouse, bacterial, and viral genomes. DeconSeq is based upon a modified version of the BWA-SW [137] aligner

[49] DeconSeq; http://deconseq.sourceforge.net/

which developed for mapping low-divergent sequences against a large reference genome [80].

2.6.3. UCHIME

UCHIME[50] [138] was developed as fast and efficient algorithm for detecting chimeric sequences which were formed out of two or more different fragments during PCR in targeted amplicon studies. The core algorithm is based on a 3-way alignment approch for each query sequence against two potential "parent" sequences from a external reference database (*reference mode*). Additionally, UCHIME offers chimera detection in *de novo* mode which uses the input database in combination with the abundance information for each input sequence as reference database for calculating the 3-way-alignment [138].

2.6.4. Acacia

Acacia[51] [139] is a publicly available Java program for rapid and conservative error correction of homopolymer over- and under-calls in pyrosequencing data. In contrast to other tools, Acacia does not use all-against-all alignments. Homopolymer regions are identified by using the quicker but less sensitive approach of empirical-derived models [139].

2.7. Internal Transcribed Spacer (ITS) Mock Community

The following sections describe fundamental methods, tools, and databases (Sec. 2.3.3) which were used to create the first mock community based on ITS amplicons.

[50] UCHIME; http://www.drive5.com/uchime/

[51] Acacia; http://acaciaerrorcorr.sourceforge.net/

2.7.1. ITSx

ITSx[52] [140] is a software utility to identify and extract ITS subregions ITS1, ITS2 and other ribosomal parts (small subregion (SSU), large subregion (LSU), 5.8S) from large Sanger, as well as from high-throughput sequencing datasets. Subregional sequences are extracted based on the predicted positions of the ribosomal genes. Therefore, position predictions based on Hidden Markov Models (HMM) [141] which are computed from large alignments comprising twenty eukaryotic groups are used. ITSx is written in Perl for Unix-based systems and is publicly available [140].

2.7.2. ecoPCR

ecoPCR[53] [142] is an electronic (*in silico*) polymerase chain reaction (PCR) running tool which relies on the very efficient pattern matching algorithm Agrep [143]. It allows PCR amplification simulation of a set of given input sequences using forward and revers primer pairs. The software is developed for Unix platforms and is freely available for download.

2.8. Differentially Abundant (DA) Feature Analysis

The next sections introduce briefly different statistical methods for detecting differentially abundant (DA) features in microbiome samples. The methods are based on *count data* obtained from high-throughput sequencing experiments. Counts are represented as discrete number of reads which have been observed for a particular feature in a selected sample. In microbiome samples features are represented by operational taxonomic units (OTUs) which can be seen as distinct species when clustered at a sequence similarity of at least 97 %. The input to all methods is presented as so called *Feature Abundance Matrix*. Within this matrix rows correspond to specific features, and columns to a single sample.

[52] ITSx; http://microbiology.se/software/itsx/

[53] ecoPCR; http://www.grenoble.prabi.fr/trac/ecoPCR

2.8.1. Metastats

Metastats[54] [144] facilitates pair-wise comparisons of multiple samples from two different groups. It is applicable not only to 16S rRNA surveys but also to high-throughput metagenomics data (using the extended Metastas approach, metagenomeSeq [145]). The Metastats approach is based on two main assumptions. First, input data can be grouped according to a certain criteria such as treatment, disease state, or gender into two distinct groups. Each group comprises multiple individuals (samples). Second, for each feature of a particular sample count data, representing the relative abundance of the feature is available. To overcome sampling depths bias across multiple samples, the raw abundance counts are simply normalized to the total contribution of each feature per sample. The major strength of Metastats is handling sparsly-sampled features using Fishers's exact test [144, 146]. Differential abundance is tested according to a two-sided t-test. Whether the detected DA feature is statistically significant, is evaluated using a nonparametric t-test [147]. To control large false positive numbers of the t-statistics the metastats approach employs the false discovery rate (FDR) [144]. Therefore the significance of the test is evaluated by a q-value, which is calculated as described in Storey and Tibshirani [148].

2.8.2. edgeR - Empirical Analysis of digital gene expression (DGE) in R

edgeR[55] [149] is an R Bioconductor [134, 135] package designed for the analysis of replicated count-based expression data such as obtained from RNA-seq [150, 151], ChIP-Seq [152], proteomics, or metagenomics experiments. The implementation is originally based on a methodology for serial analysis of gene expression (SAGE [146]) of microarray experiments introduced by Smyth 2004 [153]. Testing for differential abundance is applicable for two or more groups, with replicate measurements in at least one of the groups. Statistical methods covered by edgeR are based on the negative binomial (NB) distribution as a model for dispersion estimation, as well as for exact tests, empirical Bayes methods, and generalized linear models, when

[54] Metastats; http://metastats.cbcb.umd.edu/

[55] edgeR, Bioconductor; http://www.bioconductor.org/packages/release/bioc/html/edgeR.html

working with more complex experiments. These and similar methods such as multiple testing procedures that share information across all observations help to improve final inference [149].

2.8.3. Limma: Linear Models for Microarray Data

Limma[56] [154] is an R Bioconductor [134, 135] package for differential expression analysis using linear modeling features for microarray experiments. It is designed for simple replicated study designs, as well as for experiments with two or more groups, direct or factorial designs, and time course experiments. The package is based on normally distributed, continuous log-ratios, or -intensities obtained from microarray experiments. The basic principle is to fit a linear model to the gene expression data. For stable analysis, even for experiments with a small number of samples variance shrinkage based on an empirical Bayes approach is used to borrow information across samples and finally to estimate the biological variance.

2.8.4. voom

The *voom* [155] method *"transforms"* discrete count values as obtained from RNA-seq experiments into normalized log-counts per million ($\log_2$-cpm) and associated precision weights. Subsequently, the transformed values are ready for linear modeling and can be entered into the limma [154] analysis pipeline or any other microarray analysis pipline operating on precision weights. Therefore, the mean-variance relationship of the log-counts, as well as the precision weight for each observation is estimated [155].

2.9. Metagenome Treatment and Analysis

The following sections describe the normalization protocol, as well as the used metagenomic analysis server for analysis and comparison of the *Sphagnum* moss sequencing

[56] Limma, Bioconductor; http://www.bioconductor.org/packages/2.12/bioc/html/limma.html

libraries.

2.9.1. Library Normalization Protocol

Normalization of sequencing libraries is part of the library preparation step before sequencing. Normalization and sequencing of the *Sphagnum magellanicum* moss communities was performed by Eurofins MWG Operon [57] according to their protocol, see Appendix page 188ff. Briefly: (1) One denaturation and reassociation cycle of the DNA followed by (2) separation from reassociated ds-DNAs from remaining ss-DNAs (normalized DNA) by passing the mixture over a hydroxylapatite column, and (3) finally, PCR amplification of ss-DNAs after hydroxylapatite chromatography.

2.9.2. MG-RAST - The Metagenomics Rast Server

The Metagenomics Rast (MG-RAST) server[58] [74] is a web-based phylogenetic and functional annotation and analysis platform for metagenomic datasets. Furthermore, amplicon (16S, 18S, LSU, ITS) and metatranscritpome (RNA-seq) sequence datasets are supported. The pipeline offers the capacity to analyze large shotgun metagenomic data sets up to terabases. It combines numerous bioinformatic tools and databases for quality control, clustering, and taxonomic classification, as well as protein prediction based on nucleic acid sequence datasets generated by next-generation sequencing platforms. In addition, results can be visualized using principal component analysis (PCA), hierarchical clustering (HC) [156], or heat maps. Furthermore the MG-Rast sever supports comparisons between or to the 15,105[59] publicly available metagenomes.

[57] Ebersberg, Germany; http://www.eurofinsgenomics.eu/

[58] MG-RAST; http://metagenomics.anl.gov/

[59] As from: January 2014.

2.10. Primary Transcript Analysis and Motif Identification in two *C. fetus* subspecies

The transcription start site (TSS) identification and subsequent promoter region analysis in two *Campylobacter fetus* subspecies was realized combining numerous bioinformatic tools which are introduced in the following sections. Intermediate data produced by these tools was modified, processed, combined, and evaluated by using R, see Sec. 2.5.4.

2.10.1. MEME - Multiple Em for Motif Elicitation

MEME[60] [157] analyzes a set of given DNA or protein sequences for similarities and produce a motif for each pattern it discovers amongst them. Within MEME, motifs do not contain gaps and are presented as position-dependent letter-probability matrices. For each position in the pattern, these matrices describe the probability of each possible letter. Using statistical modeling techniques, MEME determines automatically the best pattern and returns the number of occurrences, as well as the description of each found motif in common formats such as HTML, XML, and plain TEXT [157].

2.10.2. Sequence Logo

Sequence logos[61] [158] are a frequently used technique to investigate and visualize conserved regions, as well as the frequencies and the total conservation among aligned sequences. It represents the degree of conservation of nucleotides for each position by a stack of letters, with the relative size of the letters presenting their frequency. The information content of each position is directly proportional to the total height of the letters in bits [158].

[60] MEME; http://meme.nbcr.net/meme/

[61] http://weblogo.berkeley.edu/logo.cgi

2.10.3. CLC Genomics Workbench

The *CLC Genomics Workbench*[62] [159] is designed as a user-friendly graphical cross-platform desktop application supporting and integrating typical NGS analysis and visualization tools, algorithms, and workflows. In addition to all features of the CLC Main Workbench [160], it includes capabilities for classical genomics, epigenomics, transcriptomics, read mapping, as well as for *de novo* assembly [159].

2.10.4. The Sequence Alignment/Map format and SAMtools

The *Sequence Alignment/Map (SAM)* format [161] was developed as simple and generic alignment format for read alignments against reference sequences. It supports common sequencing platforms, as well as read aligners and read lengths up to 128 mbps. Furthermore, the format offers a well-defined interface for downstream analyses, such as genotyping, variant detection, and assembly. The major strengths of the format are its flexibility, the compact size, and its efficiency in random access of the contained mapping information. Even more compact in size is the binary equivalent to SAM, the so-called BAM format. Due to indexing and positional sorting, specific genomic regions can be processed without loading the entire alignment. In addition to the SAM format, the software package *SAMtools* offers various utilities for parsing, processing, and conversion of alignments in SAM/BAM format. Most notable functions of the software package are removal of PCR duplicates, sort and merge alignments, conversion from and to different alignment formats, generate per-position information in the pileup format, SNP and Indel variant calling, as well as illustration in a text-based viewer [161].

[62] CLC Bio AS, Aarhus, Denmark; http://www.clcbio.com

3. Results

3.1. Investigation of How Sequencing Library Normalization Affects the Community Profile and its Diversity

The main goal of library normalization is to remove the most dominant sequence patterns to some kind of equilibrium between different abundant species.

3.1.1. Metagenome analysis of sequencing libraries

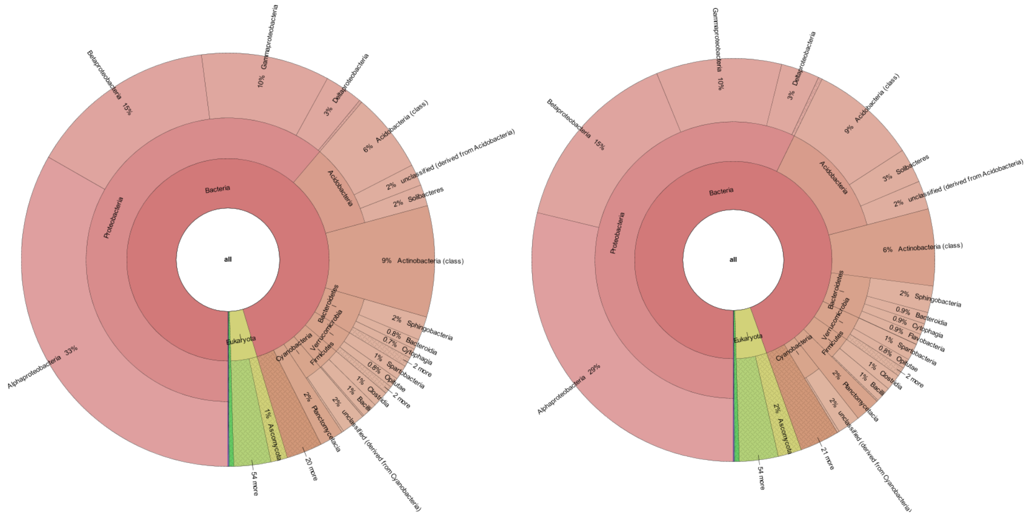
To investigate the effects of library normalization on the final taxonomic composition paired-end, 2 x 100 bps, sequencing was performed, multiplexed, on one lane of the HiSeq 2000 for the standard and the normalized library, see Sec. 2.9.1. Subsequently, both libraries were analyzed using the metagenomic annotation pipeline MG-Rast [74]. ∼172 Mio. and ∼141 Mio. sequences were analyzed after merging paired-end reads (in retain-mode[63]) and default quality-based filtering. The data was classified using the M5nr+ database [111] as annotation reference, using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15, measured in aa for protein and bps for RNA databases. Community composition analysis of the remaining ∼80 Mio. and ∼67 Mio. sequences, down to species level, is given in Tab. 3.1 and illustrated by Fig. 3.1a and Fig. 3.1b.

The survey was targeted towards functional systems carried out by bacteria. Although the majority of the sequences were assigned to the Bacteria kingdom (Tab. 3.1),

[63] non-overlapping paired-ends will be retained in the output file as individual (non-joined) sequences.

Table 3.1.: Absolute and relative domain distribution calculated from the standard (a) library and the normalized (b) library. OTU counts are given in brackets next to the corresponding absolute value. The data was classified using the M5nr+ database as annotation reference, using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15, measured in aa for protein and bps for RNA databases.

Library	Bacteria		Eukaryota		Other		Unassigned /-unclassified	
Standard	63,674,687 (7,976)	79.39 %	2,835,478 (10,704)	3.54 %	242,091 (330)	0.30 %	13,448,219 (1,152)	16.77 %
Normalized	51,655,401 (7,628)	76.73 %	2,767,353 (10,346)	4.11 %	212,289 (422)	0.32 %	12,690,154 (1,226)	18.85 %



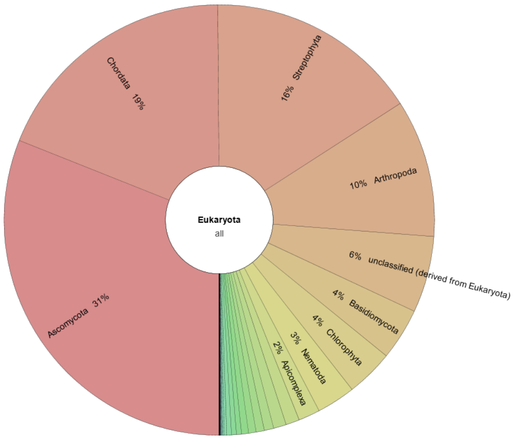
(a) standard

(b) normalized

Figure 3.1.: Taxonomic distribution for the standard (a) and the normalized (b) library, illustrated by a Krona plot [162]. The data was classified using the M5nr+ database as annotation reference, using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15, measured in aa for protein and bps for RNA databases.

and apart from a proportion of *unassigned* sequences, about 4 % of total sequences originate from an eukaryotic host - mainly fungal and animal material (see Fig. 3.1b).

Furthermore, the effect of library normalization on community diversity was investigated by rarefaction analysis (see Sec. 3.3). Apart from the fact, that sampling is still not complete, as more sampling or deeper sequencing would still increase the final number of OTUs. Richness was shown to be higher within the normalized sequencing libraries, although comprising less sequencing reads compared to the standard library.



(a) standard

(b) normalized

Figure 3.2.: Eukaryotic, contaminating community amount, of the moss metagenome, for the standard (a) and the normalized (b) sequencing library illustrated by a Krona plot [162]. The data was classified using the M5nr+ database as annotation reference, using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15, measured in aa for protein and bps for RNA databases.

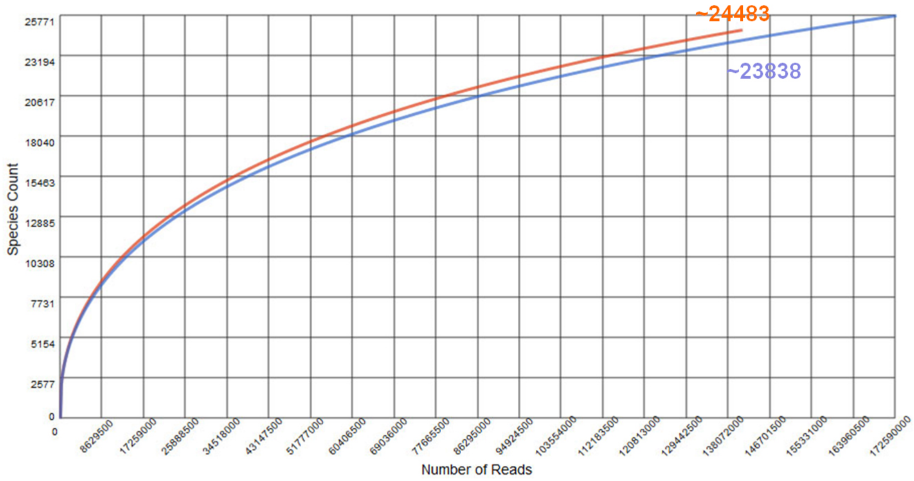


Figure 3.3.: Rarefaction curves calculated from the metagenomes of the standard (blue) and normalized (red) library. α-Diversity (species count) at $\sim$138 Mio. reads is given next to the corresponding rarefaction curve. In addition, rarefaction analysis visualizes, that sampling is still not complete, as more sampling or deeper sequencing would still increase the final number of OTUs. Furthermore, Richness was shown to be higher within the normalized sequencing libraries, although comprising less sequencing reads compared to the standard library.

To test for statistical significance of changes between the standard and the normalized library the Pearson $\mathcal{X}^2$-test [163] was performed using the `chiq.test` R function based on the taxonomic domain distribution obtained by MG-Rast, Tab. 3.1. The Pearson's Chi-squared test with Yates' continuity correction [163] (X-squared = 45,770.61, df = 1, p-value < 2.2e-16) confirmed that cDNA sequencing library normalization affects overall domain distribution.

3.1.2. Identification of functional subsystems

The normalized sequencing data was assembled by a *de novo* approach into contigs using the CLC Genomics Workbench [159] (version 4) and the recommended default settings. 1,115,029 scaffolded contigs were obtained by this approach with an average length of 501 bps. The assembled contigs were exported to FASTA [101] format using the CLC Genomics Workbench export utils and provided for further analysis to the team of Prof. Berg[64], who performed functional subsystem analysis based on the revealed contigs.

Within this inter-environmental comparison of the *S. magellanicum* with publicly available metagenomes (summarized as *higher plant* and *peat soil* metagenomes), 198 functional subsystems were manually selected and subsequently tested for statistical significant changes, within this thesis, as described in Sec. 3.4. The distribution of the count data was checked, prior to analysis, by application of the Kolomogorov-Smirnov-Test [164] on the raw abundances of selected metagenomes (Appendix Tab. B.1). In a next pre-processing step, scale normalization factors were calculated to account for the different library sizes of the raw data samples, prior to significance analysis. To make the count data ready for linear modeling, raw counts were transformed using the voom [155] function. Additionally, the probability distribution of each group was visualized before and after data transformation using density plots, Fig. 3.4.

Finally, changes of subsystems between the different groups were assessed by statistical analysis using the linear modeling approach implemented by the R Bioconductor

[64] Institute of Environmental Biotechnology, Graz University of Technology, Graz, Austria

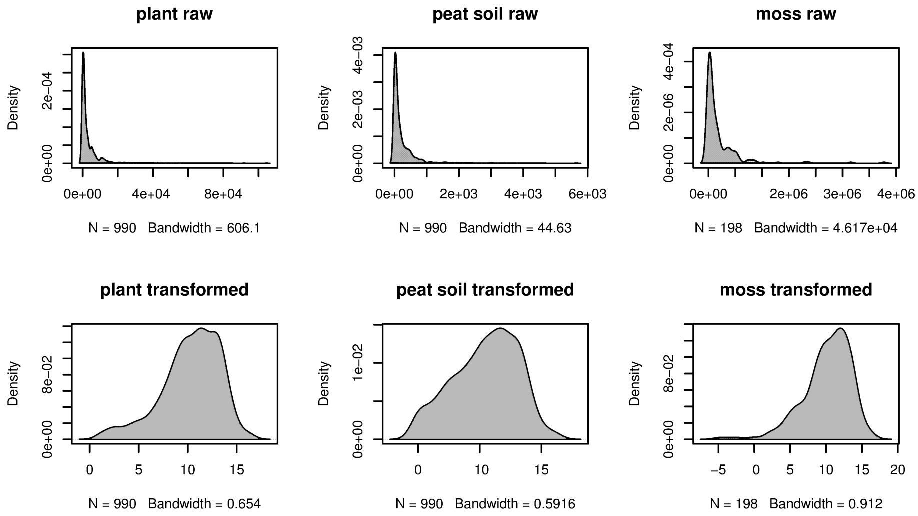


Figure 3.4.: Density plots of the statistically analyzed metagenomes illustrating the distribution of the raw count data before and after data transformation using voom for the *S. magellanicum*, as well as for the publicly available higher plant and peat soil metagenomes. The figure nicely illustrates the successful transformation of the raw data distribution using the voom function towards the normal distribution.

package limma+voom (version 3.16.8) [154]. From the 198 tested features, 106 and 37 functional subsystems were detected as differentially abundant within *S. magellanicum* and higher plants or *S. magellanicum* and peat soil, respectively. 26 functional subsystems were found to be differentially abundant in both habitats.

Tab. 3.2 and Tab. 3.3 list the top 10 differential abundant subsystems between *S. magellanicum* and higher plants and peat soils, respectively. A full list of differentially abundant subsystems of *S. magelanicum* compared to *higher plant metagenomes* and *S. magelanicum* compared to *peat soil metagenomes*, is available as supplementary information, in Appendix Tab. B.2 and Tab. B.4, respectively.

Table 3.2.: Top ten differentially abundant functional subsystems detected between *S. magellanicum* and higher plant metagenomes detected by limma+voom. Subsystems identified as differentially abundant in both comparisons are highlighted by bold font type.

Subsystmes level 1	Subsystems level 2	LogFC	AveExpr	t-val	p-val	Adj. p-val
S. magellanicum/higher plants metagenomes						
Stress response	**Dessication stress**	**-8.52**	**3.88**	**-5.73**	**0.00**	**0.0007**
Dormancy and sporulation	**Spore DNA protection**	**-5.14**	**3.24**	**-4.67**	**0.00**	**0.0022**
Phages, prophages, plasmids, transposable elements	Gene Transfer Agent (GTA)	-4.53	8.11	-11.83	0.00	0.0000
Membrane transport	Protein secretion system. type IV	-3.04	6.74	-9.68	0.00	0.0000
Membrane transport	Protein secretion system. type VII (chaperone/usher pathway. CU)	-2.30	9.94	-11.28	0.00	0.0000
Cofactors, vitamins, pigments, prostetic groups	**Coenzyme B**	**-2.25**	**3.49**	**-2.60**	**0.02**	**0.0473**
Clustering-based subsystems	Putative GGDEF domain protein related to agglutinin secretion	-1.99	6.56	-6.40	0.00	0.0003
Iron acquisition and metabolism	Siderophores	-1.80	9.24	-6.03	0.00	0.0005
Clustering-based subsystems	**Hypothetical associated with RecF**	**-1.74**	**7.76**	**-12.82**	**0.00**	**0.0000**
Clustering-based subsystems	**Related to menaquinone-cytochrome C reductase**	**-2.14**	**4.46**	**-5.70**	**0.00**	**0.0033**

Table 3.3.: Top ten differentially abundant functional subsystems detected between *S. magellanicum* and peat soil metagenomes detected by the limma+voom function. Subsystems identified as differentially abundant in both comparisons are highlighted by bold font type.

Subsystmes level 1	Subsystems level 2	LogFC	AveExpr	t-val	p-val	Adj. p-val
S. magellanicum/peat soils metagenomes						
Stress response	**Dessication stress**	**-10.57**	**3.88**	**-6.88**	**0.00**	**0.0013**
Dormancy and sporulation	**Spore DNA protection**	**-7.68**	**3.24**	**-6.84**	**0.00**	**0.0013**
Cofactor. vitamins, prostetic groups, pigments	**Coenzyme B**	**-5.70**	**3.49**	**-6.47**	**0.00**	**0.0018**
Respiration	Reverse electron transport	-4.32	4.51	-4.47	0.00	0.0114
Phages. prophages. transposable elements. plasmids	-	-3.54	5.71	-5.76	0.00	0.0033
Respiration	Sodium ion-coupled energetics	-3.31	6.35	-4.37	0.00	0.0114
Secondary metabolism	Plant octadecanoids	-2.88	3.52	-5.40	0.00	0.0042
Clustering-based subsystems	Proteasome related clusters	-2.84	4.25	-4.55	0.00	0.0114
Clustering-based subsystems	**Tricarboxylate transporter**	**-2.44**	**10.17**	**-7.80**	**0.00**	**0.0013**
Clustering-based subsystems	**Related to menaquinone-cytochrome C reductase**	**-2.14**	**4.46**	**-5.70**	**0.00**	**0.0033**

3.2. Development of an Application for Identification and Removal of Contaminating Sequences

The Decontaminator is a platform independent JAVA [130] command line application which enables detection and removal of randomly amplified sequence fragments originating for example from the host system, or from other non-marker gene DNA. It allows usage either as command line application or as a part of an analysis platform such as SnoWMAn [67], or QIIME [70]. The application requires BLAT [97] output, `blast8` formatted, as well as the original target amplicon sequence file, in FASTA format, as the initial input. Due to its generic design it can be easily extended to further input, as well as other output formats, on demand. Combining the best BLAT hits with the targeted amplicons during decontamination, the Decontaminator separates true amplicons from contaminations according to the user specified thresholds. Finally, statistic charts and tables are provided additionally to the filtered sequences to ensure a comprehensive decontamination procedure. The basic Decontaminator IO workflow is illustrated in Fig. 3.5.

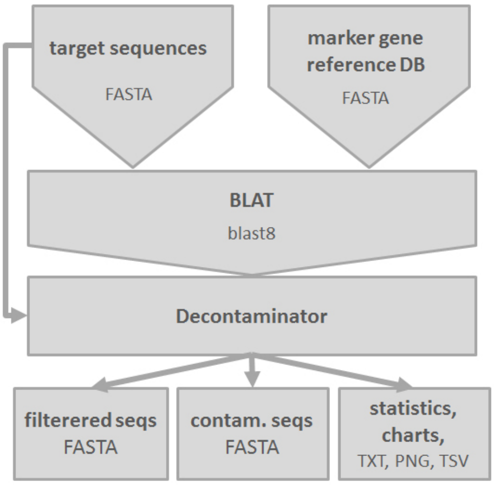


Figure 3.5.: Decontaminator basic IO workflow. The Decontaminator is based upon a BLAT homology search, using the targeted amplicons and an appropriate marker gene reference DB in FASTA format as input. Combining the best BLAT hits with the targeted amplicons during decontamination, the Decontaminator separates true amplicons from contaminations, according to the user specified thresholds. Finally, statistic charts and tables are provided additionally to the filtered sequences to ensure a comprehensive decontamination procedure.

To account for technical sequences, fragments such as barcodes (MIDs), or primers, at the beginning of the amplicons, the optional parameters barcode length (-bcl) and primer length (-pl) can be specified. Both parameters are used to calculate the true sequence length for the amplicon, which is needed for calculating the coverage percentage, between the query and the subject sequence (query coverage, QC). For

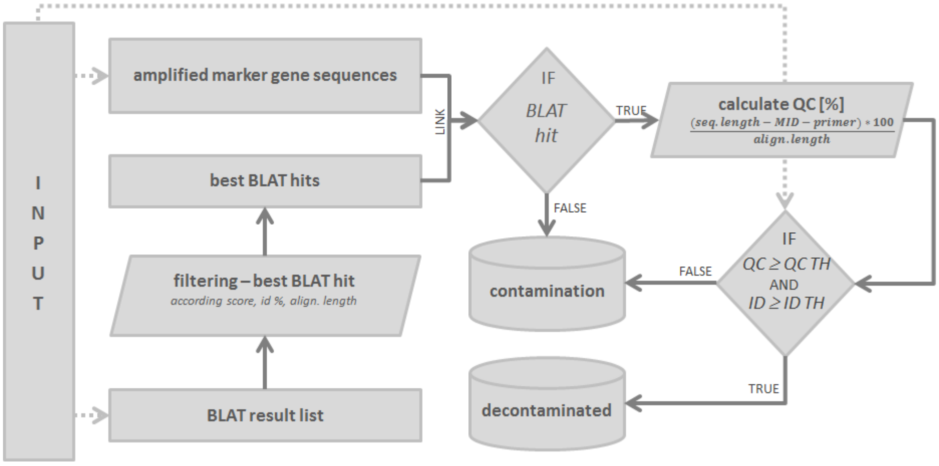


Figure 3.6.: The Decontaminator filtering workflow. First, targeted amplicon sequences are imported from the input *FASTA* file. In parallel the BLAT hit list is filtered according to score, alignment length (align. length), and percentage of identity (ID %), to extract the best BLAT hit. Second, for each amplicon, the corresponding BLAT hit is evaluated. Entries with no BLAT hit are excluded. In the other case, amplicon length is used to calculate the query coverage (QC), in respect to MID and primer length. Finally, reads which do not satisfy specified thresholds (TH) for identity [%] (ID) and query coverage (QC) are excluded as well.

each input sequence the corresponding best BLAT hit, according to percentage of identity (ID), alignment length, and bit-score is selected. In combination with the calculated query coverage, these are the main parameters which are used for sequence evaluation. Sequences with no BLAT hit at all are discarded as they do not show any similarity to the marker gene structure. Additionally, sequences below the thresholds for min. QC and for min. percentage identity are discarded as well. Both parameters can be specified by the user, selecting the `-c` and `-i` option, respectively. Thus, also chimeric fragments are likely to be removed by the Decontaminator because of low query coverage, as shown in Fig 3.6. A full list of the Decontaminator parameters and usage is included within the Appendix C on page 192.

Firstly, the Decontaminator had been evaluated using a small 16S test data set, containing 295 true 16S fragments (region V1-V2), which were tagged with 6 bps MIDs and amplified with the forward primers `AGAGTTTGATCCTGGCTCAG` and `AYTGGGYDTAAAGNG` [165]. In a second step, 25 randomly amplified human fragments were added to the initial test set. These fragments were created by *in silico* amplification using ecoPCR

[142], with the same primers as used for the 16S fragments, within the human genome (Homo_sapiens.GRCh37 release 72). Finally, 57 manually created chimeric sequences, based on 16S fragments of the test set, were added. Artificial chimeric sequences, were formed by a custom Java program, which combines fragments of two different reads randomly. For each of these two sets a separate Decontaminator, as well as a DeconSeq run was performed.

Results of the Decontaminator evaluation are summarized in Tab. 3.4. Parameters and thresholds for different Decontaminator test cases are given as supplementary information, Appendix Tab. C.1.

Table 3.4.: Decontaminator result summary of the first evaluation with a small 16S sequence set manually contaminated with (a) 25 randomly amplified human sequence fragments, as well as with (b) 57 manually created chimeras. In addition, the same dataset was filtered by DeconSeq.

	(a) incl. 25 human seq		**(b) incl. 57 chimeras**	
	(320 sequences)		(377 sequences)	
	Decontaminator	DeconSeq	Decontaminator	DeconSeq
Seqs. Usable	295		340	352
Seqs. totally filtered	25		37	25
Seqs. low query coverage	0		12	0
Seqs. Low identity	0		0	0
Seqs. no BLAT hit (Decontaminator)	25		25	NA

3.3. Integration and Evaluation of Resources for Fungal Community Analysis

To facilitate high-throughput classification and characterization of fungal communities, an appropriate reference sequence set based on the marker gene for fungi, the internal transcribed spacer region, was needed within the analysis pipeline SnoWMAn. Therefore, the UNITE [82] reference set (release 15.10.2013), corresponding to the species hypothesis (SH) resulting from clustering at 97 % sequence similarity was processed and incorporated into SnoWMAn.

UNITE provides sequence and annotation information in separate files which are linked with a UNITE specific identifier. Hence, 21,984 sequences and corresponding annotations, down to species level, were combined to a (FASTA) database file and a corresponding Greengenes [105] formatted annotation lookup file, using a custom Java program. Finally, these newly created resources were incorporated into SnoWMAn's BLAT analysis pipeline.

3.3.1. Validation set for ITS classification resources

For evaluation of the BLAT ITS reference sequence database based on UNITE, as well as for the evaluation of the most recently introduced RDP classifier [100] for ITS amplicons (beta version), or any other classification system for ITS fragments, *in silico* mock communities for the ITS1/2 and ITS1 region were created. These mock communities are based on 2,248 fungal sequences which cover the entire ITS1 and ITS2 region, including 5.8S, as well as parts of 18S and LSU. Flanking regions of 18S and LSU are necessary for providing primer binding sites in the *in silico* amplification step. Sequences were selected manually, quality checked, and provided by Henrik R. Nilsson[65] and Kessy Abarenkov[66]. The selected mock targets cover all major fungal phyla: (1) Basidiomycota (BAS), (2) Ascomycota (ASC), (3) Chytridiomycota (CHY), (4)

[65] Department of Biological and Environmental Sciences, University of Gothenburg, Sweden

[66] Natural History Museum, University of Tartu, Estonia

Early Diverging Linages (EAR, former Zygomycota), and (5) Glomeromycota (GLO), see also Tab. 3.5.

Table 3.5.: Absolute and relative sequence distribution of the ITS mock communities at the phylum level. Counts are presented for the manually selected fungal raw sequences (2,248), as well as for the *in silico* amplified ITS1/ITS2 (1,363) and the ITS1 (1,965) region.

	raw (2,248)		amplified ITS1/2 (1,363)		amplified ITS1 (1,965)	
	counts	[%]	counts	[%]	counts	[%]
Ascomycota	952	42	639	47	922	46.92
Basidiomycota	640	28	403	30	984	50.08
Glomeromycota	370	16	262	19	23	1.83
Early Diverging Linages (Zygomycota)	187	8	38	3	0	0
Chytridiomycota	99	4	21	2	36	1.17

The true sequence distribution of the ITS1/2 mock community for taxonomic levels from the *phylum* to the *species* is illustrated within Fig. 3.7a-e. For taxonomic levels lower than phylum, phylogenetic groups which cover less than 2 % of the total sequences abundance, have been summarized within *Other*.

Data tables for Fig. 3.16a-e are provided as supplementary information in Appendix Tab. E.1-E.6.

The true sequence distribution of the ITS1 mock community for taxonomic levels from the *phylum* to the *species* is illustrated within Fig. 3.8a-e. For taxonomic levels lower than phylum, phylogenetic groups which cover less than 2 % of the total sequences abundance, have been summarized within *Other*.

Data tables for Fig. 3.8a-e are provided as supplementary information in Appendix Tab. E.7-E.9.

To ensure that sequences cover the same genetic region, a multiple sequence alignment (MSA) was done, as well as an inspection with ITSx [140], prior to *in silico* amplification. Full GenBank [109] records were retrieved for all selected sequences using the given Accession numbers via using the Entrez eUtils [166] querying system. All obtained records were combined into one file which was pre-processed for *in silico* PCR by the ecoPCRFormater [142] and the NCBI taxonomy database [167] dump (release 24.10.2013). For the ITS1/2 mock community, the universal primers ITS1-F

(CTTGGTCATTTAGAGGAAGTAA) and ITS4 (TCCTCCGCTTATTGATATGC) [168] were used to amplify fragments between 150 bps and 1,000 bps, allowing maximal 3 bps mismatches within the primer binding site. Finally, a total set of 1,363 sequences were amplified and remain for the first *in silico* ITS1/ITS2 mock community. Successfully amplified fragments were enriched with their corresponding taxonomic annotation, down to the species level, according to INSDC [120] using the given Accession number.

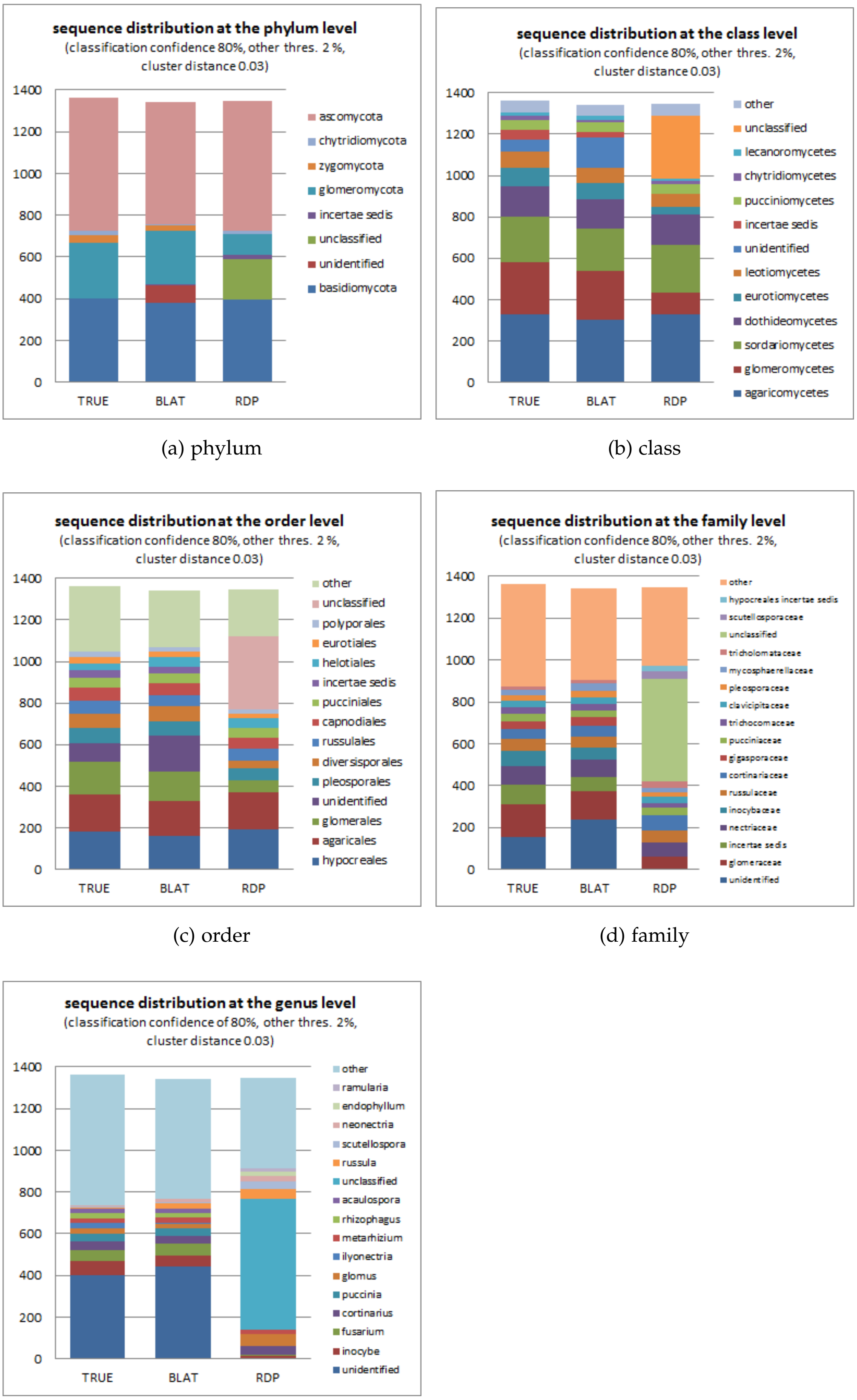
The ITS1 region was amplified with a maximal mismatch of 3 bps within the primer binder site using the universal primers ITS1 (TCCGTAGGTGAACCTGCGG) and ITS2 (GCTGCGTTCTTCATCGATGC) [169]. ecoPCR was again used to amplify fragments between 150 bps and 700 bps. Finally, 2,017 were sucessfully amplified, whereby 1,965 were fully annotated to at least a fungal phylum, according to INSDC [120] using the given Accession number.

3.3.2. Validation of ITS classification resources

The created *in silico* ITS1/2 and ITS1 mock communities were used to evaluate SnoWMAn's BLAT pipeline based on the UNITE reference sequences (version 15.10.2013) and the RDP ITS classifier (beta), which is part of SnoWMAn's mothur [69] pipeline. Apart from removal of sequences containing ambiguous bases, no special pre-processing, such as chimera filtering, or denoising, was performed on the raw sequences. Hence, 17 and 24 sequences were removed due to these criteria by default pre-processing within the analysis of the ITS1/2 and ITS1 mock, respectively, in both cases. Tab. 3.6 summarizes the sample overview of the taxonomic mock data analysis using SnoWMAn's mothur [69] pipeline with the RDP ITS classifier (beta) for taxonomic classification.

The sample overview of the mock data analyzed by SnoWMAn's BLAT pipeline using the UNITE reference DB (version 15.10.2013) is summarized in Tab. 3.6.

To compare the classification result of both approaches, sequence distribution amongst available taxonomic levels, was exported from SnoWMAn. Fig. 3.7a-e illustrates sequence distribution from the phylum to genus level of the classification result



(a) phylum

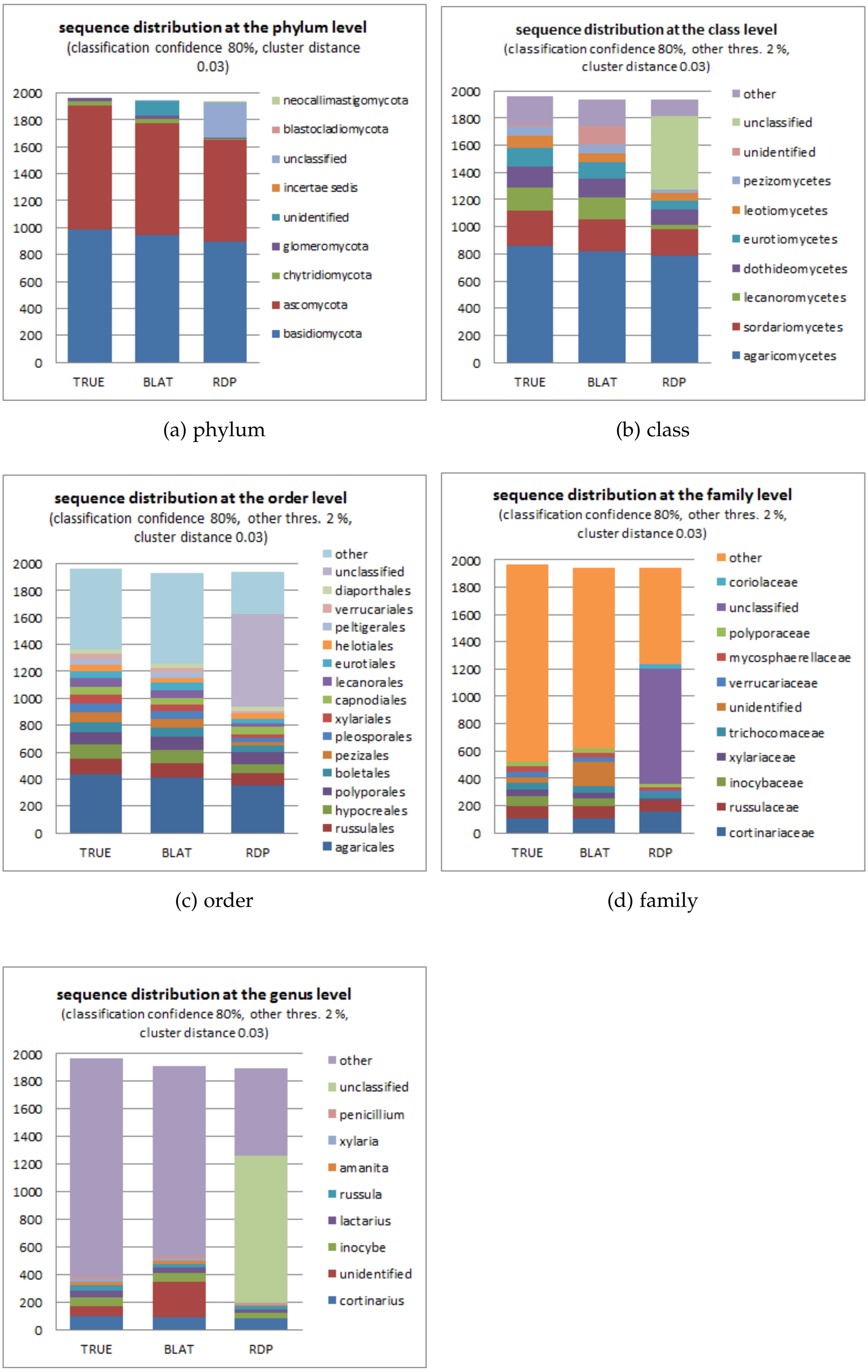
(b) class

(c) order

(d) family

(e) genus

Figure 3.7.: Sequence distribution of the ITS1/2 mock community, for the true composition, as well as obtained by classification with BLAT and the RDP ITS classifier (beta) at the class level, at a classification confidence of 80 %, a cluster distance of 0.03 and for taxa covering more than 2 % of total sequence abundance.



(a) phylum

(b) class

(c) order

(d) family

(e) genus

Figure 3.8.: Sequence distribution of the ITS1 mock community, for the true composition, as well as obtained by classification with BLAT and the RDP ITS classifier (beta) at the class level, at a classification confidence of 80 %, a cluster distance of 0.03 and for taxa covering more than 2 % of total sequence abundance.

Table 3.6.: Sample overview ITS mock communities analyzed by SnoWMAns mothur pipline and the RDP ITS classifier (beta). The table presents the number of determined OTUs for different cluster distances. In addition, the number of raw, filtered, not classified, unique, and the final number of classified sequences are given for both approaches.

		raw seqs.	filtered seqs.	not seqs.	classified seqs.	unique seqs.	OTUs						
							0.00	0.01	0.02	0.03	0.04	0.05	0.06
ITS1/2	RDP(beta)	1,363	17	0	1,346	1,141	1,045	937	878	821	775	743	715
	BLAT	1,363	17	2	1,344	1,139				615			
ITS1	RDP(beta)	1,965	24	0	1,941	1,938	1,938	1,937	1,932	1,922	1,905	1,885	1,865
	BLAT	1,965	24	0	1,941	1,938				1,707			

obtained by BLAT and the RDP ITS classifier (beta) for both mock communities, at a classification confidence of 80 %. For taxonomic levels lower than phylum, sequences not covering more than 2 % of overall abundance are summarized by the group *Other*.

The taxonomic classification results have been compared against the true sequence distribution at the phylum down to the genus level. Tab. 3.7 presents the absolute and relative numbers of the correct classified taxons for each phylogenetic level for BLAT and the RDP ITS classifier (beta).

Table 3.7.: Summary of taxonomic classification of BLAT and the RDP ITS classifier (beta) on the introduced ITS mock communities. True mock sequence distribution have been compared to the obtained taxonomic classification of both approaches. The table presents absolute and relative numbers of correctly classified taxons for each phylogenetic level.

	ITS1/2 mock				ITS1 mock			
	BLAT		RDP(beta)		BLAT		RDP(beta)	
	# corr. class.	[%]	# corr. class.	[%]	# corr. class.	[%]	# corr. class.	[%]
phylum	1,026	76.0	1,062	78.0	1,830	94.0	1,664	85.0
class	972	72.0	919	68.0	1,813	93.0	1,398	72.0
order	853	63.0	719	53.0	1,794	92.0	1,201	61.0
family	691	51.0	500	37.0	1,752	90.0	896	46.0
genus	514	38.0	273	20.0	1,630	83.0	749	38.0

3.4. Evaluation and Adaption of Methods for Differentially Abundant Feature Detection

Three different approaches, two well-established methods for differentially expressed gene detection in microarray and RNA-seq experiments (edgeR [149] and limma+voom [154, 155]) and Metastats [144] have been evaluated using simulated count data representing the result of a typical targeted amplicon sequencing experiment comprising two different groups. Additionally, efficiency of true positive detection was tested using different effect sizes.

3.4.1. Simulation of count data

The simulated count data, which was used for the evaluation of the different methods, was created according to the approach provided by McMurdie and Holmes [84]. In particular a set of 2,000 features with 30 truly differentially abundant features, in 2 distinct conditions, for four different effect sizes (fold change) are created for subsequent evaluation. The full list of settings for community profile simulation is provided in Tab. 3.8. Differentially abundant features within the two groups, are created by duplication of a given community profile and randomly modifying features according to the specified fold change values to create the second condition. With the specified settings, seen in Tab. 3.8, all possible combinations, based on known community profile patters from feces, were simulated. Finally, 72 community profiles with a known number of truly differentially abundant features (# true DA features = 30) were created.

Table 3.8.: Settings used for community profile simulation, according to McMurdie and Holmes [84], for subsequent evaluation of three DA feature detection approaches.

settings	values
# conditions	2
# min. seques per OTU	15
# max. OTUs	2,000
# samples in each condition	3; 5
# of reads per sample	2,000; 7,000; 10,000
# of replicates	1:3
effect sizes	1.25; 2.5; 5.0; 10.0
# truly DA features	30
# sample type	Feces
# total simulations	72

3.4.2. Evaluation of DA feature detection using Metastats (R version)

DA detection with Metastats was performed by adapting the Metastats R script provided by James R. White[67], version April 2009, and the simulated count data created in Sec. 3.4.1.

Prior to the evaluation analysis, results generated by Metastats' web-service were compared to the results of the R reimplementation. Apart from slight differences of p- and q-values, which are very likely due to rounding errors, methods can be treated as identical.

Table 3.9.: Analysis settings used for DA detection with metastats.

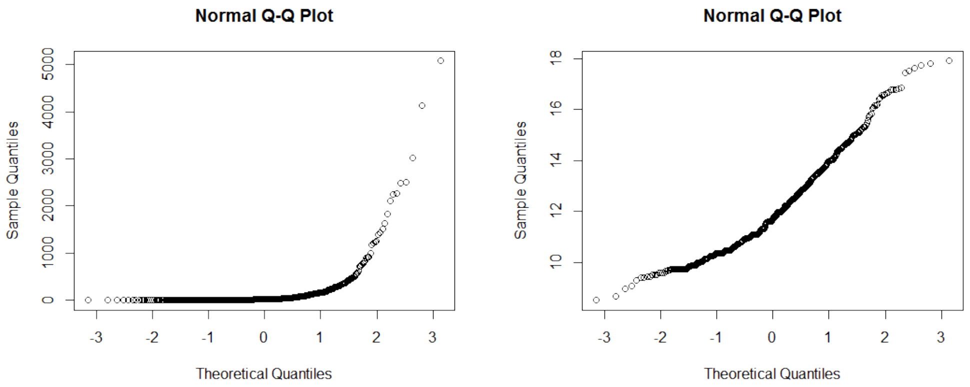
parameter	value
significance threshold	0.05
significance by	p values
# bootstrapping permutations	1,000

Tab. 3.9 summarizes the main settings which were used for DA detection in the 72 simulated datasets. As the simulated count data comprise in any case more than 2 samples, a two sample t-test is computed for each feature. Subsequently, distribution of the null t-statistics is estimated by the specified number of bootstrapping permutations (default 1000 permutations). Finally, q-values, for the FDR control are calculated using previously determined p-values according to the Fisher's exact test [144, 146]. Results of the DA detection, for the different simulation conditions, are summarized in Tab. 3.10-3.12.

3.4.3. Evaluation of DA feature detection using limma+voom

The R code of the limma+voom vignette [170], see Chapter 9.2 Sec. *Two Groups*, was adapted for analysis of the simulated count data. To provide continuous and normally distributed data, the simulated count data was transformed by *voom*, (Sec. 2.8.3) prior to linear modeling. To avoid overflow because of taking the log of zero, all values were increased by 1. Count data before and after data transformation by voom is illustrated by Fig. 3.9(a) (before) and 3.9(b) (after), at the example of the simulated count data *2000_Feces_3_10.00_10*.

[67] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA, https://github.com/icj/Metagenomics/blob/master/metastats.R



(a) raw

(b) transformed by voom

Figure 3.9.: Distribution of the count data, 2000_Feces_3_10.00_10, (a) before and (b) after data transformation using the voom function.

According to the description of the limma user guide, the matrix containing simulated count data was converted into a `DGEList` data object for further processing. To account for different sequencing library sizes within the different samples, scale normalization was applied on the simulated data. After specifying a design matrix based on the experimental conditions, data was transformed by voom. In a next step, it was fit to a linear model, which was subsequently used for calculating the empirical Bayes statistics [153]. Finally, p-values were adjusted by using the method described by Benjamini and Hochberg [171]. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant. Results of the DA detection are summarized in Tab. 3.10-3.12.

3.4.4. Evaluation of DA feature detection using edgeR

The examples of the edgeR user guide [172] for differential expression analysis for analyzing two or more groups, comprising replicated data, was adapted and used for detection of DA features within the simulated data. Firstly, count data was increased by one to prevent taking the log of zero, and turned into a `DGEList` data object. edgeR was used according to the glm approach, which permits for more general comparisons. Therefore, a model matrix, which describes the treatment conditions, was created based on the experimental groups. Before data was fit to a linear model (`glmFit`), count data was normalized. In addition common, as well as tagwise dispersion was estimated. Subsequently, likelihood ratio tests were conducted on the two coefficients in the linear model using the `glmLRT` function. Finally, p-values were adjusted by using the method by Benjamini and Hochberg [171]. Only features with an adjusted p-value (FDR) less than 0.05 were considered as differentially abundant. Results of the DA detection are summarized in Tab. 3.10-3.12.

3.4.5. Result summary of DA feature detection

The result of the DA detection for the different sets, of simulated count data, was grouped by the number of maximum reads per sample, and summarized in Tab. 3.10-3.12 for all three evaluated approaches.

Table 3.10.: Result summary of DA feature detection for evaluation of simulated count data, with a maximum library size of 2,000 sequences, which was tested with metastats, edgeR, and limma+voom. For the maximum number of 30 truly differentially abundant features, the number of correctly identified (true positives, TP), incorrectly identified (false positives, FP), not detected (false negatives, FN), and the number of "called" features is given. In addition, for each condition the FDR is calculated.

sample type	metastats					edgeR					voom+limma				
2000_Feces	# sig	# TP	# FP	# FN	FDR [%]	# sig	# TP	# FP	# FN	FDR [%]	# sig	# TP	# FP	# FN	FDR [%]
1_ 1.25_3	1	1	0	29	0.00	0	0	0	0		0	0	0	0	
1_ 1.25_5	3	1	2	29	66.67	3	0	3	30	100.00	0	0	0	0	
1_ 2.50_3	6	5	1	25	16.67	9	2	7	28	77.78	41	29	12	1	29.27
1_ 2.50_5	15	8	7	22	46.67	20	10	10	20	50.00	36	30	6	0	16.67
1_ 5.00_3	1	1	0	29	0.00	5	1	4	29	80.00	31	30	1	0	3.23
1_ 5.00_5	18	9	9	21	50.00	25	10	15	20	60.00	37	29	8	1	21.62
1_10.00_3	11	7	4	23	36.36	10	2	8	28	80.00	34	30	4	0	11.76
1_10.00_5	37	21	16	9	43.24	30	14	16	16	53.33	54	30	24	0	44.44
2_ 1.25_3	0	0	0	0	0.00	0	0	0	0		0	0	0	0	
2_ 1.25_5	0	0	0	0		0	0	0	0		7	7	0	23	0.00
2_ 2.50_3	17	9	8	21	47.06	15	5	10	25	66.67	32	23	9	7	28.13
2_ 2.50_5	22	13	9	17	40.91	22	8	14	22	63.64	30	30	0	0	0.00
2_ 5.00_3	8	5	3	25	37.50	10	2	8	28	80.00	49	30	19	0	38.78
2_ 5.00_5	29	18	11	12	37.93	20	6	14	24	70.00	34	29	5	1	14.71
2_10.00_3	20	13	7	17	35.00	25	8	17	22	68.00	32	29	3	1	9.38
2_10.00_5	43	26	17	4	39.53	34	17	17	13	50.00	33	30	3	0	9.09
3_ 1.25_3	0	0	0	0		0	0	0	0		5	4	1	26	20.00
3_ 1.25_5	6	4	2	26	33.33	0	0	0	0		3	3	0	27	0.00
3_ 2.50_3	4	2	2	28	50.00	5	0	5	30	100.00	30	28	2	2	6.67
3_ 2.50_5	19	11	8	19	42.11	26	12	14	18	53.85	45	29	16	1	35.56
3_ 5.00_3	20	11	9	19	45.00	25	9	16	21	64.00	34	30	4	0	11.76
3_ 5.00_5	26	17	9	13	34.62	26	11	15	19	57.69	37	30	7	0	18.92
3_10.00_3	17	9	8	21	47.06	16	2	14	28	87.50	25	25	0	5	0.00
3_10.00_5	31	21	10	9	32.26	27	14	13	16	48.15	30	30	0	0	0.00

Table 3.11.: Result summary of DA feature detection for evaluation of simulated count data, with a maximum library size of 7,000 sequences, which was tested with metastats, edgeR, and limma+voom. For the maximum number of 30 truly differentially abundant features, the number of correctly identified (true positives, TP), incorrectly identified (false positives, FP), not detected (false negatives, FN), and the number of "called" features is given. In addition, for each condition the FDR is calculated.

sample type	metastats					edgeR					voom+limma				
7000_Feces	# sig	# TP	# FP	# FN	FDR [%]	# sig	# TP	# FP	# FN	FDR [%]	# sig	# TP	# FP	# FN	FDR [%]
1_ 1.25_3	0	0	0	0		0	0	0	0		0	0	0	0	
1_ 1.25_5	2	0	2	30	100.00	3	2	1	28	33.33	0	0	0	0	
1_ 2.50_3	44	21	23	9	52.27	44	27	17	3	38.64	41	29	12	1	29.27
1_ 2.50_5	67	21	46	9	68.66	37	30	7	0	18.92	36	30	6	0	16.67
1_ 5.00_3	59	16	43	14	72.88	34	30	4	0	11.76	31	30	1	0	3.23
1_ 5.00_5	66	22	44	8	66.67	46	30	16	0	34.78	37	29	8	1	21.62
1_10.00_3	51	15	36	15	70.59	37	30	7	0	18.92	34	30	4	0	11.76
1_10.00_5	83	27	56	3	67.47	43	30	13	0	30.23	54	30	24	0	44.44
2_ 1.25_3	1	1	0	29	0.00	0	0	0	0		0	0	0	0	
2_ 1.25_5	3	2	1	28	33.33	8	7	1	23	12.50	7	7	0	23	0.00
2_ 2.50_3	27	12	15	18	55.56	38	24	14	6	36.84	32	23	9	7	28.13
2_ 2.50_5	60	17	43	13	71.67	44	29	15	1	34.09	30	30	0	0	0.00
2_ 5.00_3	41	15	26	15	63.41	46	30	16	0	34.78	49	30	19	0	38.78
2_ 5.00_5	71	20	51	10	71.83	37	30	7	0	18.92	34	29	5	1	14.71
2_10.00_3	48	16	32	14	66.67	31	30	1	0	3.23	32	29	3	1	9.38
2_10.00_5	8	5	3	25	37.50	42	30	12	0	28.57	33	30	3	0	9.09
3_ 1.25_3	0	0	0	0		4	4	0	26	0.00	5	4	1	26	20.00
3_ 1.25_5	1	0	1	30	100.00	3	3	0	27	0.00	3	3	0	27	0.00
3_ 2.50_3	54	18	36	12	66.67	36	28	8	2	22.22	30	28	2	2	6.67
3_ 2.50_5	53	16	37	14	69.81	57	29	28	1	49.12	45	29	16	1	35.56
3_ 5.00_3	56	18	38	12	67.86	42	30	12	0	28.57	34	30	4	0	11.76
3_ 5.00_5	89	27	62	3	69.66	48	30	18	0	37.50	37	30	7	0	18.92
3_10.00_3	28	8	20	22	71.43	25	25	0	5	0.00	25	25	0	5	0.00
3_10.00_5	76	25	51	5	67.11	34	30	4	0	11.76	30	30	0	0	0.00

Table 3.12.: Result summary of DA feature detection for evaluation of simulated count data, with a maximum library size of 10,000, which was tested with metastats, edgeR, and limma+voom. For the maximum number of 30 truly differentially abundant features, the number of correctly identified (true positives, TP), incorrectly identified (false positives, FP), not detected (false negatives, FN), and the number of "called" features is given. In addition, for each condition the FDR is calculated.

sample type	metastats					edgeR					voom+limma				
10000_Feces	# sig	# TP	# FP	# FN	FDR [%]	# sig	# TP	# FP	# FN	FDR [%]	# sig	# TP	# FP	# FN	FDR [%]
1_ 1.25_3	0	0	0	0		1	0	1	30	100.00	0	0	0	0	
1_ 1.25_5	9	6	3	24	33.33	8	7	1	23	12.50	8	7	1	23	12.50
1_ 2.50_3	53	17	36	13	67.92	36	27	9	3	25.00	33	27	6	3	18.18
1_ 2.50_5	29	6	23	24	79.31	35	29	6	1	17.14	29	29	0	1	0.00
1_ 5.00_3	74	18	56	12	75.68	34	30	4	0	11.76	32	30	2	0	6.25
1_ 5.00_5	78	22	56	8	71.79	43	30	13	0	30.23	45	30	15	0	33.33
1_10.00_3	68	14	54	16	79.41	35	30	5	0	14.29	33	30	3	0	9.09
1_10.00_5	111	29	82	1	73.87	38	30	8	0	21.05	33	30	3	0	9.09
2_ 1.25_3	1	1	0	29	0.00	4	4	0	26	0.00	0	0	0	0	
2_ 1.25_5	2	2	0	28	0.00	5	5	0	25	0.00	3	3	0	27	0.00
2_ 2.50_3	46	14	32	16	69.57	32	26	6	4	18.75	30	26	4	4	13.33
2_ 2.50_5	56	19	37	11	66.07	41	29	12	1	29.27	36	26	10	4	27.78
2_ 5.00_3	55	22	33	8	60.00	36	29	7	1	19.44	29	29	0	1	0.00
2_ 5.00_5	123	26	97	4	78.86	37	30	7	0	18.92	34	30	4	0	11.76
2_10.00_3	65	16	49	14	75.38	32	30	2	0	6.25	41	30	11	0	26.83
2_10.00_5	114	24	90	6	78.95	33	30	3	0	9.09	30	30	0	0	0.00
3_ 1.25_3	0	0	0	0		2	1	1	29	50.00	1	1	0	29	0.00
3_ 1.25_5	5	2	3	28	60.00	4	4	0	26	0.00	5	4	1	26	20.00
3_ 2.50_3	45	11	34	19	75.56	35	28	7	2	20.00	31	28	3	2	9.68
3_ 2.50_5	51	16	35	14	68.63	50	29	21	1	42.00	30	29	1	1	3.33
3_ 5.00_3	56	16	40	14	71.43	31	28	3	2	9.68	24	22	2	8	8.33
3_ 5.00_5	71	19	52	11	73.24	37	30	7	0	18.92	44	30	14	0	31.82
3_10.00_3	42	17	25	13	59.52	46	30	16	0	34.78	53	30	23	0	43.40
3_10.00_5	112	28	84	2	75.00	36	30	6	0	16.67	31	29	2	1	6.45

3.5. Transcriptome analysis of *Campylobacter fetus* subspecies fetus and veneralis

To identify potential transcription start sites (TSS), as well as to gain deeper insights into the promoter structure of Campylobacterales, for both subspecies dRNA-seq, as described in Sec. 2.1.2, analysis was performed.

Initially generated sequence reads (treated and untreated) were mapped to their corresponding reference genomes (*Cff*: NC_008599.1, *Cfv*: HG004426.1) using the CLC Genomics Workbench (version 4) and the recommended default settings. Tab. 3.13 summarizes the CLC mapping results for each Campylobacter subspecies for untreated (TEX-), as well as for cDNA libraries treated (TEX+) with terminator exonuclease.

Table 3.13.: CLC Genomics Werkbench mapping report summarized for *Campylobacter fetus fetus (Cff)* and *Campylobacter fetus veneralis (Cfv)* for untreated (TEX-) and for cDNA libraries treated with terminator exonuclease (TEX+). For the reference bases mapping approach, the reference genomes NC_008599.1 for *Cff* and HG004426.1 for *Cfv* were used. Sequence yield was lower for both untreated *C. fetus* subspecies cDNA libraries, compared to the TEX+ libraries. Additionally, mapping efficiency is higher within TEX+ sequencing libraries for both subspecies.

	C. fetus fetus						*C. fetus veneralis*				
	counts	**reads**	**[%]**	**avg. len.**	**num of bases**		**counts**	**reads**	**[%]**	**avg. len.**	**num of bases**
TEX+	Reference	-	-	1,773,615	1,773,615	TEX+	Reference	-	-	994,014	1,988,028
	Mapped	5,222,345	94.12	91	475,233,395		Mapped	8,686,431	93.44	91	807,027,221
	Not mapped	326,316	5.88	91	29,694,756		Not mapped	623,046	6.56	91	56,697,186
	Total	5,548,661	100	91	504,928,151		Total	9,491,477	100	91	863,724,407
TEX-	Reference	-	-	1,773,615	1,773,615	TEX-	Reference	-	-	994,014	1,988,028
	Mapped	11,021,336	88.10	91	1,002,941,576		Mapped	11,309,554	89.61	91	1,029,169,414
	Not mapped	1,488,747	11.90	91	135,475,977		Not mapped	1,311,639	10.39	91	119,359,149
	Total	12,510,0831	100	91	1,138,417,553		Total	12,621,193	100	91	1,148,528,563

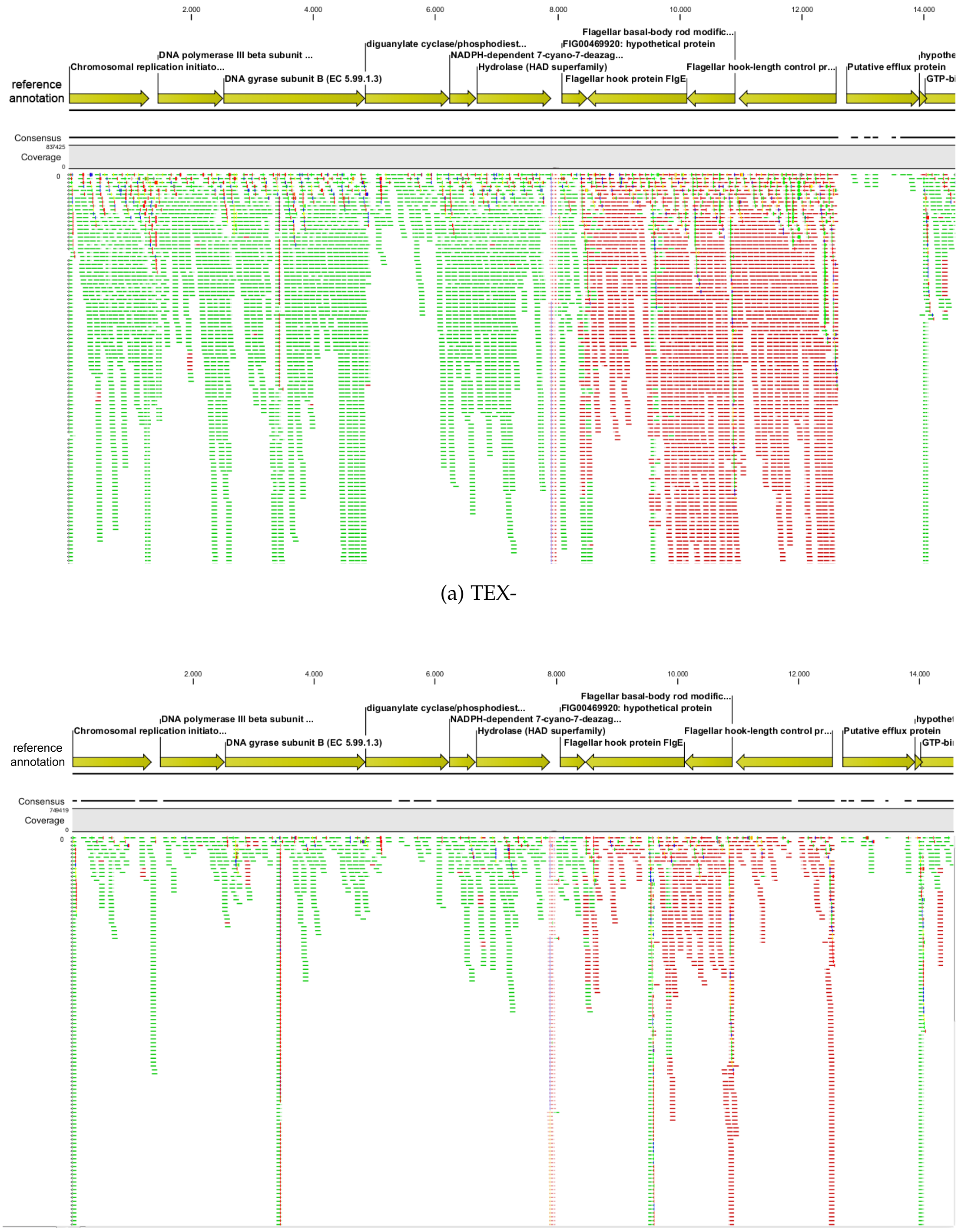
To make the CLC mapping information usable for further processing, it was exported to BAM format [161] directly via the CLC Workbench export mechanisms. SAM tools [161] were subsequently used to transform the binary representation of the mapping into a plain text format. Within the same step, the mapping information was split into separate files according to mappings on the forward or the reverse strand. To get the base counts for each position of the genome (the number of times a single position in the genome was covered by a base of one of the sequencing reads) the plain mapping information was loaded into R, for each subspecies and strand. Finally, from the tabular mapping information positions and counts were extracted and stored as an R vector object for further processing.

3.5.1. Transcription Start Site (TSS) Identification

Due to cDNA library enrichment with terminator exonuclease, primary transcripts are enriched, as seen in Fig. 3.10b. This causes a steep increase within the counts for a certain position over the length of the mapped read compared to the untreated library, as seen in Fig. 3.10a. To intensify this effect and to overcome some kind of *ambient mapping*, counts less or equal to seven had been set to zero. The algorithms for the identification of potential TSS iterates through all positions of the genome and validates the count value by calculating the mean expression for a window of 89 bps up- and downstream of the current position. Each position with an expression value bigger or equal to the mean expression of the flanking areas is treated as potential TSS. Based on the average read length of 91 bps obtained TSS are sorted in ascending order for a final distance check. All positions which are at least 91 bsp apart are considered as potential TSS for further processing and motif search.

3.5.2. Transcription Start Site Categorization

The TSS identification approach revealed 646 TSS on the leading strand and 574 TSS on the lagging strand of *C. fetus* subspecies *fetus* and 1,457 TSS on the leading strand and 1,132 TSS on the lagging strand of *C. fetus* subspecies *veneralis*, respectively. The given gene loci information, from the already annotated reference genomes, was used to assign the potential TSS into one of the five categories, specified by Sharma *et al.* [92]: (1) antisense, (2) internal, (3) orphan, (4) primary, and (5) secondary, illustrated in Fig. 3.11. A *primary* TSS is located within the next 500 bps upstream of an annotated mRNA start. A *Secondary* TSS is more than likely the same as the *primary* TSS but with smaller coverage. *Internal* TSS are found within an annotated gene on the same strand, whereas, TSS situated inside or within 100 bps of an oppositely encoded gene



(a) TEX-

(b) TEX+

Figure 3.10.: In the example of *C. fetus veneralis* the difference between the final sequence mapping based on untreated (a) or on cDNA libraries treated with terminator exonulease (b) is illustrated. Sequence reads were mapped by the CLC Genomics Workbench 5, read mapping algorithm to the corresponding reference genome, HG004426.1. Direct comparison of the two mappings visualizes once the drastic reduction of sequencing reads, as well as the enrichment of primary transcripts.

are considered as *Anitsense*. Detected TSS which do not match any of the specified criteria are summarized as *Orphans*.

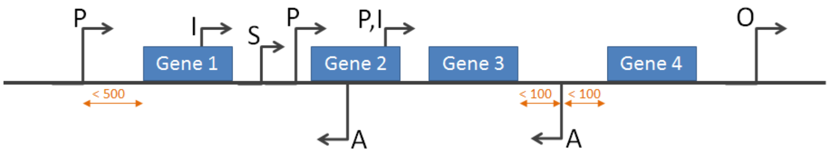


Figure 3.11.: A *primary* TSS is located in the next 500 bps upstram of an annotated mRNA start. A *Secondary* TSS is more or likely the same as the *primary* TSS but with smaller coverage. *Internal* TSS are found within an annotated gene on the same strand, whereas TSS situated inside or within 100 bps of an oppositely encoded gene are considered as *Anitsense*. *figure modified from [92].*

An R script was written for final TSS categorization. In more detail: annotation information and protein coding regions were extracted from the GenBank annotation file and loaded into R for each strand separately. All genes encoded in the same direction were sorted in ascending order. The processing starts in forward direction. For each gene g_i and the following gene g_{i+1} start and end positions, as well as flanking regions (such as 501 bps upstream of the current gene start site), were determined. First, it was determined if any TSS positions was detected within this region. For more than one TSS, it is subsequently distinguished between *primary*, *secondary*, or *internal* category. According to these specifications each TSS is likely to fit in more than one category, as seen in Fig. 3.12. For *antisense* TSS identification for each gene encoded on the complementary strand are evaluated according to the specifications illustrated in Fig. 3.11. After finishing coding region processing on the forward strand the procedure is repeated for the reverse strand. Finally, any TSS which was not assigned to any of the categories is denoted as an *orphan*. The categorization result for both *C. fetus* subspecies *veneralis* and *fetus* is given separately for each direction in Tab. 3.14.

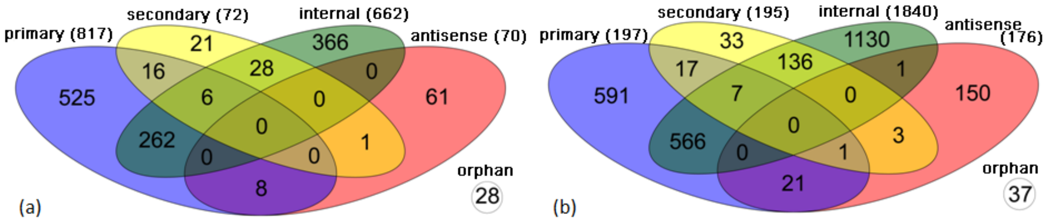


Figure 3.12.: The Venn Diagram illustrates the TSS categorization result for (a) *C. fetus veneralis* and (b) *C. fetus fetus*. According to the categories described in Fig. 3.12, TSS are likely to fit in more than one category.

Table 3.14.: TSS categorization result summary for *Cfv* and *Cff*. The table summarizes the number of found TSS for each category and separated by strand. Categorization was performed according to the illustration in Fig. 3.11. In addition to the overall sum of TSS found per category, the number of uniquely observed TSS per category is included within the presentation.

category	primary		secondary		internal		antisense		orphan	
strand	+	-	+	-	+	-	+	-	+	-
Cfv	685	518	116	81	1082	758	103	73	NA	NA
	1203		197		1840		176		37	
unique	1095		195		1837		176		37	
Cff	445	372	39	33	353	309	40	30	NA	NA
	817		72		662		70		28	
unique	716		71		662		70		28	

3.5.3. Promoter Motif Analysis

For all previously determined TSS of the two *C. fetus* subspecies *fetus* and *veneralis*, the region 60 bps upstream of the potential TSS was extracted for subsequent motif analysis. A motif within the target sequences extracts was searched using the MEME web service [157] by allowing a maximal sequence shift in both directions of 3 bps. For 797 *C. fetus veneralis* and 575 *C. fetus fetus* promoter regions, an extended Pribnow box (`tgnTAtaAT`) as the -10 motif was identified. In addition, a periodic, wave-like AT-rich signal upstream of potions -14 was found in both subspecies, as seen in Fig. 3.14 and 3.13. Sequences in which the motif was not found within a maximum sequence shift of 3 bps, may either originate from internal or secondary TSS, which are not responsible for whole operon regulation. Or these sequences originate from falsely identified TSS.

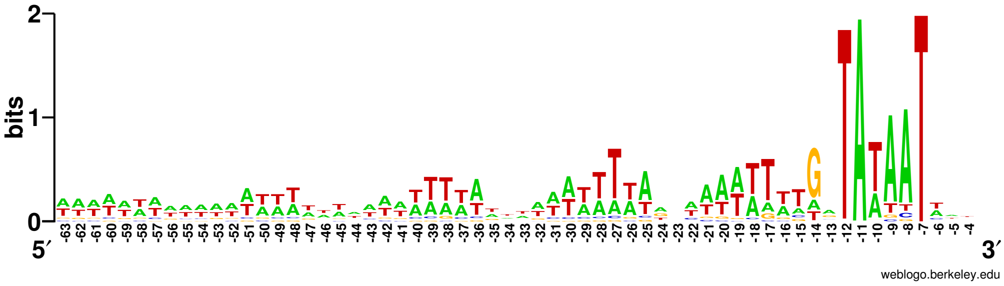


Figure 3.13.: For 797 and 575 extracted promoter regions of *C. fetus venerals* and *C. fetus fetus*, respectively, an extended Pribnow box (`tgnTAtaAT`) as the -10 motif was identified using MEME, in addition to a periodic, wave-like AT-rich signal upstream of potions -14.

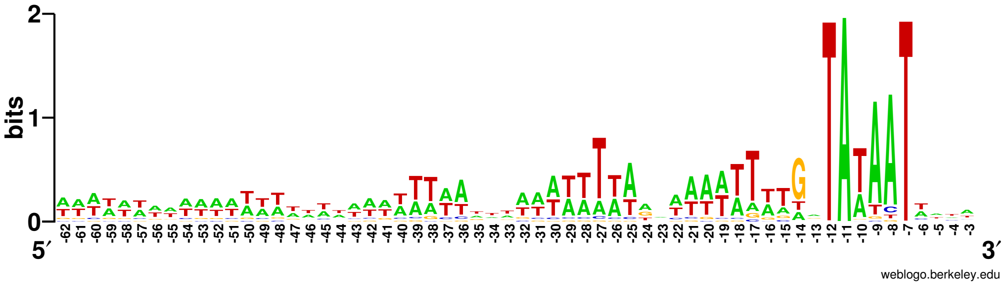


Figure 3.14.: For 575 extracted promoter regions of *C. fetus venerals*, an extended Pribnow box (`tgnTAtaAT`) as the -10 motif was identified using MEME, additionally to a periodic, wave-like AT-rich signal upstream of potions -14.

3.6. Effects of Osmotic Diarrhea on the Human Gastrointestinal Microbiome

For various diseases of the human GI tract, diarrhea is one of the most observed concomitant feature. To investigate whether alterations in the colonic microbiota is caused by the disease or by the accompanying diarrhea the team of Ass.-Prof. Gregor Gorkiewicz[68] planned and carried out the study described in Sec. 2.1.5. The raw data of this targeted amplicon sequencing base line project[69] was used to investigate the effects of different pre-processing steps, in particular focusing on the detection and removal of contaminating sequences using the novel Decontaminator application.

3.6.1. Bacterial community profile analysis

To investigate the effects of different types of contaminates on the finally determined number of OTUs, the raw data was once again analyzed, applying different pre-processing approaches prior to phylogenetic analysis. Known contaminates are chimeric, low quality, noisy, or randomly amplified sequences.

515,212 raw sequences obtained by 454 sequencing represent the basis for further processing with the microbiome analysis pipeline SnoWMAn. In particular, SnoWMAn's RDP pipeline, RDP classifier 2.5 and the Infernal [102] alignment model 2008 for bacteria were used for each analysis run using default pre-processing settings. Samples were automatically split by given MIDs, with rejection of amplicons with erroneous or no barcode match at all. Additionally, sequences with ambiguous bases (containing N's), to short sequences (less than 150 bps length), or with more than 2 mismatches within the given amplification primers (forward and reverse) were discarded within each run. According to this default pre-processing criteria 69,856 sequences had been removed from the initial amplicon set. The following paragraphs summarize the main results obtained by (1) quality filtering of the raw sequences

[68] Institute of Pathology, Medical University of Graz, Graz, Austria

[69] First targeted amplicon sequencing survey of the Microbiome Unit of the Medical University of Graz and the Bioinformatics Group of Graz University of Technology.

according to QUAL scores, (2) by removal of contaminations within the raw sequences using the Decontaminator, (3) by denoising of the raw sequences only, (4) by chimera detection and removal as the only pre-processing, and finally (5) by analysis of the fully pre-processed raw data sequence set.

Community profile analysis of raw sequences quality filtered First, the raw data was analyzed using SnoWMAn's RDP pipeline, including quality filtering according to the attached QUAL files and default RDP quality filtering settings. 70,837 were filtered according to the criteria given in Tab. 3.15, whereby only 19 sequences were removed by an average quality score of less than 20. The remaining 444,356 amplicons were assigned to 5,727 OTUs at a distance of 0.03.

Table 3.15.: Sequences filtered according to no barcode and quality, including trimming, within the community profile analysis with quality filtering only.

	no barcode	filtering and trimming	totally removed	remaining
# sequences removed	21,080	49,776	70,856	444,356

Community profile analysis of sequences decontaminated only Second, raw sequences were filtered for contaminating sequences originating not from bacterial DNA, prior to community profile analysis with SnoWMAn's RDP pipeline. Therefore taxonomic comparison using a blast like approach (BLAT) was performed by using BLAT (v.34) and the Greengenes database (release May 2009). In the next step, the BLAT output, in `blast8` format, was passed to the Decontaminator for identification and the removal of non bacterial sequences (Settings for the Decontaminator are given as supplementary information within the Appendix Tab. D.2). 27,395 sequences were removed from the initial raw sequence set, according to the criteria given in Tab. 3.16.

The remaining 487,817 decontaminated raw sequences were uploaded for community profile analysis with the RDP pipeline to SnoWMAn. Integrated quality filtering of SnoWMAn was *not performed* within this approach. 51,246 amplicons were filtered according to the basic filtering criteria. A summary of the totally filtered sequences is

Table 3.16.: Result of the decontamination procedure applied on the raw sequences. The table summarizes the number of sequences, before and after detection and removal of contaminating sequences. In addition, removed sequences are listed according to their exclusion criteria, low percentage of identity, or query coverage, and no BLAT hit. A total number of 27,395 sequences were removed from the raw data set during the decontamination step.

	before	removed	remain	low qc	low ident	low qc low ident	no BLAT hit
# sequences	515,212	27,395	487,817	0	9,364	0	18,031

presented in Tab. 3.17. The remaining 436,571 amplicons were assigned to 4,869 OTUs at a distance of 0.03.

Table 3.17.: Sequences filtered by default within the standard pre-processing step of the analysis pipeline. Prior to pre-processing contaminating sequences were removed by the Decontaminator. In total 51,246 sequences were removed because of no barcode match, ambiguous bases, or read length.

	no barcode	filtering and trimming	totally removed	remaining
# sequences removed	15,239	36,007	51,246	436,571

Community profile analysis of sequences denoised only Within the third pre-processing approach of the diarrhea study, raw sequencing data was pre-processed and quality filtered with Acacia [139], prior to community profile analysis. Acacia was applied on the raw data using the default settings, except from the quality score cutoff, which was set to twenty. 39,431 sequences were removed from the 515,212 raw sequences. A full summary of the Acacia analysis statistics is provided as supplementary information in Appendix Tab. D.3.

The remaining 475,781 sequences were used for community profile analysis with SnoWMAn's RDP pipeline. Again, integrated RDP quality filtering, using QUAL scores, is *not performed* within this analysis approach.

Table 3.18.: Sequences filtered by default within the standard pre-processing step of the analysis pipeline. Prior to pre-processing error correction (denoising) using Acacia was performed. In total 14,219 sequences were removed because of no barcode match, ambiguous bases, or read length in addition to the 39,431 sequences, removed by Acacia.

	no barcode	filtering and trimming	totally removed	remaining
# sequences removed	9,579	4,640	14,219	461,562

14,219 amplicons were filtered according to the basic filtering criteria in addition to the 39,431 sequences, removed by Acacia. A Summary of the totally filtered sequences is presented in Tab. 3.18. The remaining 461,562 amplicons were assigned to 5,350 OTUs at a distance of 0.03.

Community profile analysis of sequences chimera checked only This analysis run of the diarrhea raw sequence set comprises chimera detection and removal using UCHIME [138], in addition to the default sequence filtering process by SnoWMAn. UCHIME was applied on 444,356 sequences which remained after default sequence filtering, see Tab. 3.19. During chimera detection, 8,955 sequences were identified as potential chimeras by UCHIME used in reference based mode (reference DB: silva.gold.aligned release 104 [173]).

Table 3.19.: Sequences filtered by default within the standard pre-processing step of the analysis pipeline, as well as removed by detection and removal of chimeric sequences using UCHIME. In total 78,811 sequences, 8,955 of these are supposed to be chimeric, were removed because of no barcode match, ambiguous bases, read length, or due to identification as chimera prior to taxonomic classification.

	no barcode	filtering and trimming	chimera removal	totally removed	remaining
# sequences removed	21,080	48,776	8,955	78,811	436,401

Community profile analysis of fully preprocessed sequences Finally, the raw sequences were pre-processed in sequential order using the approaches described within the above paragraphs prior to community profile analysis with SnoWMAn. Tab. 3.20 summarizes the result of each single filtering step up to the ultimately filtered sequence set. The 466,956 remaining sequences were uploaded to SnoWMAn and once again default filtering was performed, as seen in Tab. 3.20. Thereafter, 428,811 sequences were clustered into 4,375 distinct OTUs at a distance of 0.03.

As a last point, Tab. 3.21 summarizes the main results of the previous sections in a common table.

Table 3.20.: The table summarizes the number of sequences removed during the complete pre-processing analysis chain from the raw amplicon. The number of removed sequences is given for all specified exclusion criteria, such as no barcode, low quality, read length, ambiguous bases (summarized within filtering and trimming), contaminations, noise, and chimeric sequences. A total number of 86,401 sequences have been removed by the combination of the different pre-processing approaches prior to downstream analysis.

	no barcode	filtering and trimming	contaminations	noise	chimera removal	totally removed	remaining
# sequences removed	10,654	19,804	27,395	20,861	7,687	86,401	428,811

Table 3.21.: Summary of removed and retained sequences after the different pre-processing methods and the finally obtained number of OTUs, at a cluster distance of 0.03. Raw data of the diarrhea study was analyzed using SnoWMAn's RDP pipeline (classifier version 2.5) after application of different pre-processing approach combinations.

	# raw seqs	# removed by default	# removed by pre-processing	# totally removed	# retained	# OTUs (distance 0.03)
removed by default (no prepro.)	515,212		70,856		444,356	5,727
quality filtering[70]			70,856		444,356	5,727
decontamination		51,246	27,395	78,641	436,571	4,869
denoising		14,219	39,431	53,650	461,562	5,350
chimera filtering		69,856	8,955	78,811	436,401	5,146
full pre-processing		30,458	87,101	86,401	428,811	4,375

3.6.2. Comparison of pre-processing on remaining sequences per sample

With the application of different pre-processing approaches, and their combinations on the raw sequence set, varying amounts of sequences were removed from the original sequence set. Tab. 3.22 summarizes the remaining sequences (library size) per sample after different pre-processing stages. Additionally, for each stage the remaining library size is compared to the initial amount of sequences per sample.

[70] 19 sequences were identified with an avg. qual score below 20. Additionally, these sequences do not exceed the min. sequence length. As a consequence they have been already filtered within the default trimming and pre-processing step.

Table 3.22.: Summary of library size per sample, of the diarrhea study, after application of different pre-processing stages. In addition, the table compares the number of the remaining sequences to the unfiltered library size values of the raw data set for each sample. Interestingly the number of sequences per sample was increased after denoising by Acacia only. This can be explained by the fact that pre-processing with Acacia includes error-correction within the reads as well.

	raw	only denoising			only chimera			only decontaminating			fully processed		
	# seqs	# seqs	+	[%]	# seqs	-	[%]	# seqs	-	[%]	# seqs	-	[%]
AF1	18,693	19,781	1,088	5.82	18,039	654	3.50	18,365	328	1.79	17,763	930	5.24
AF2	16,986	17,678	692	4.07	16,436	550	3.24	16,794	192	1.14	16,277	709	4.36
AF3	16,017	16,466	449	2.80	15,732	285	1.78	15,955	62	0.39	15,674	343	2.19
AF4	12,578	13,054	476	3.78	12,300	278	2.21	12,505	73	0.58	12,248	330	2.69
BF1	18,644	19,008	364	1.95	18,191	453	2.43	18,401	243	1.32	17,958	686	3.82
BF2	18,986	19,668	682	3.59	18,326	660	3.48	18,802	184	0.98	18,171	815	4.49
BF3	22,220	22,842	622	2.80	22,024	196	0.88	22,148	72	0.33	21,961	259	1.18
BF4	22,793	23,613	820	3.60	22,158	635	2.79	22,634	159	0.70	22,042	751	3.41
BM2	25,836	27,068	1,232	4.77	25,490	346	1.34	25,360	476	1.88	25,034	802	3.20
BM3	27,011	28,323	1,312	4.86	26,790	221	0.82	26,133	878	3.36	25,934	1,077	4.15
CF1	16,918	17,507	589	3.48	16,239	679	4.01	16,773	145	0.86	16,127	791	4.90
CF2	16,449	17,000	551	3.35	16,070	379	2.30	16,372	77	0.47	16,005	444	2.77
CF3	20,555	21,342	787	3.83	20,378	177	0.86	20,523	32	0.16	20,350	205	1.01
CF4	17,460	18,141	681	3.90	17,237	223	1.28	17,426	34	0.20	1,7207	253	1.47
CM2	25,875	27,147	1,272	4.92	25,712	163	0.63	25,098	777	3.10	24,952	923	3.70
CM3	24,198	25,338	1,140	4.71	24,053	145	0.60	23,554	644	2.73	23,412	786	3.36
DF1	11,394	11,664	270	2.37	11,053	341	2.99	10,693	701	6.56	10,377	1,017	9.80
DF2	18,778	19,461	683	3.64	18,203	575	3.06	18,059	719	3.98	17,508	1,270	7.25
DF3	22,483	23,170	687	3.06	21,992	491	2.18	22,240	243	1.09	21,781	702	3.22
DF4	21,626	22,403	777	3.59	20,741	885	4.09	21,363	263	1.23	20,519	1,107	5.39
DM2	25,000	26,066	1,066	4.26	24,516	484	1.94	24,483	517	2.11	24,051	949	3.95
DM3	23,856	24,822	966	4.05	23,721	135	0.57	22,890	966	4.22	22,760	1,096	4.82
classified	**444,356**	**461,562**			**435,401**			**436,571**			**428,111**		
filtered	70,856	53,650		10.41	79,811		15.49	78,641		15.26	87,101		16.91

3.6.3. Comparison of filtered sequences

Filtered sequences generated within the separate pre-processing approaches, described previously in Sec. 3.6.1, were used to investigate the overlap between different criteria (contaminations, noise, chimeras, or bad quality). Therefore sequence identifiers were extracted from FASTA headers of all discarded sequencing files and imported into R. The `VennDiagram` functionality of R was used to illustrate intersections between the different filtering groups, (1) contaminating sequences (red), (2) noise (green), (3) chimeras (blue), and (4) low quality (yellow), shown in Fig. 3.15

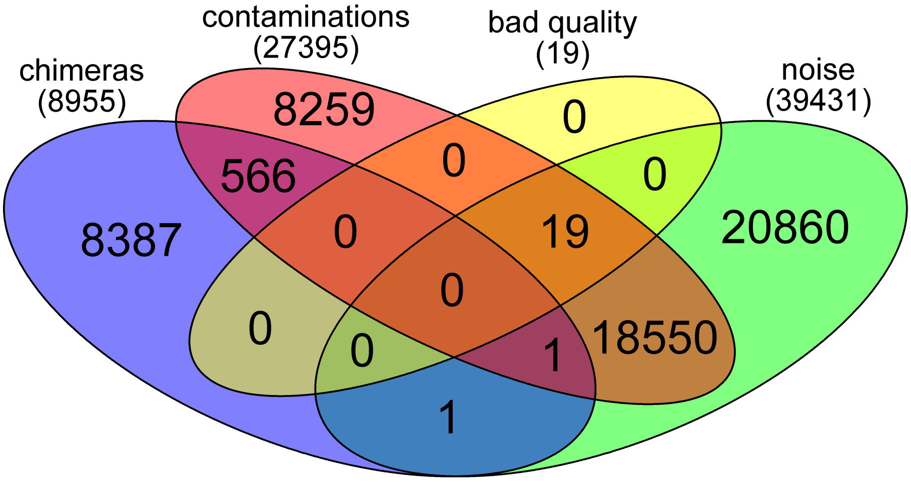


Figure 3.15.: Filtered sequences between the different pre-processing stages and approaches were visualized using a Venn diagram. It allows illustrating the overlap between filtered sequences by the different criteria: (1) contaminating sequences (red), (2) noise (green), (3) chimeras (blue), and (4) low quality (yellow).

3.7. Effects of Phospholipds on the Gastrointestinal Microbiome in Mice

The main goal of this targeted amplicon survey is to investigate the effects of phospholipids on the gastrointestinal microbiome of mice. Therefore, an experiment comprising different mice types, dietetic treatments, and material sources (details seen in Sec. 2.1.3) was planned and accomplished by the team of Prof. Peter Fickert, MD[71]. Within this thesis, the bacterial community profiles of the different sample types were determined and subsequently tested for differentially abundant features.

Tab. 3.23 summarizes abbreviations and descriptions of the different dietetic conditions, mice types, intestinal regions, and groups defined within this survey. The experimental design of the survey is given in more detail in Sec. 2.1.3.

Table 3.23.: Sample abbreviation and descriptions used within the gastrointestinal mouse survey for mice type, dietetic conditions, gastrointestinal region, source type, as well as for the more generalized intestinal region groups.

mice types		dietetic condition		intestinal region		source type		intestinal region groups	
`WT`	wild type	`N`	normal	`F`	feces	`Ile`	Ileum	`SI`	small bowel, Ile + Jej
`KO`	knock out	`E`	enriched	`M`	mucosa	`Jej`	Jejunum	`LI`	large bowel, Cae + Col
`BD`	bile-duct ligated					`Cae`	Caecum		
						`Col`	Colon		

3.7.1. Bacterial community profile analysis

1,633,199 sequences were obtained by 454 sequencing of the collected samples of the survey described in Sec. 2.1.3 whereby 818,525 originate from the first sequencing effort and 814,676 from the second. After the manual detection and removal of contaminating sequences (Decontaminator v.6) and chimeras (UCHIME, mothur v.1.31.2), noise reduction and quality filtering (Acacia, v.1.52.b0), the remaining 1,633,199 sequences were uploaded to SnoWMAn for downstream analysis. Settings and detailed results for the pre-processing are included as supplementary information in Appendix Tab. A.4 and Tab. A.5, respectively. Finally, 1,107,388 sequences were classified using

[71] Division of Gastroenterology and Hepatology, Medical University of Graz, Graz, Austria

the RDP classifier (v.2.5). Tab. 3.24 shows the sample overview of totally classified sequences, unique sequences, and detected OTUs at different cluster distances. Sample overviews for each sample grouped by sample type *Source* is included as supplementary information in Appendix Tab. A.6-A.9.

Table 3.24.: Sample overview summary for the GI mouse study. The table presents the total and unique number of finally classified sequences, as well as the obtained number of distinct OTUs, at different cluster distances.

Sample	Seqs	Unique Seqs	number of OTUs at distance						
			0.0	0.01	0.02	0.03	0.04	0.05	0.06
Total	**1,107,388**	**232,489**	**85,309**	**38,513**	**17,568**	**10,633**	**70,97**	**5,330**	**39,39**

Taxonomic distribution between mice types, dietetic treatment, sample material, and source was visualized and discussed with the team of Prof. Peter Fickert, MD. Based on the different community profiles, the main focus for the subsequent differentially abundant feature detection was set on samples originating from all different *ileum* locations.

In addition to the evaluation of samples of source type *ileum*, all samples had been generalized by assigning them to group (1) large bowel (LI), or (2) small bowel (SI). To be exact, samples of source type *caecum* and *colon* are grouped to LI and *ileum* and *jejunum* to SI. Tab. 3.25 summarizes the number of sequences for each group and the number of OTUs created at different cluster distances.

Table 3.25.: Sample overview summary, GI mouse study for subgroups SI and LI. For both subgroups the number of sequences used within the taxonomic classification step and the obtained OTUs at different cluster size values is presented. In addition, the total and unique number of classified sequences and the final number of total OTUs at different cluster distances is included.

Sample	Seqs	Unique Seqs	number of OTUs at distance						
			0.0	0.01	0.02	0.03	0.04	0.05	0.06
Total LI	660,934		61,111	30,827	15,097	9,315	6,255	4,712	3,483
Total SI	446,454		28763	12,543	5306	3,400	2,522	2058	1,676
Total	**1,107,388**	**232,489**	**85,309**	**38,513**	**17,568**	**10,633**	**7,097**	**5,330**	**3,939**

3.7.2. Analysis of differentially abundant features

The final classification result, the OTU feature table, was exported from SnoWMAn and used for differentially abundant feature detection according to the methods evaluated in Sec. 2.8. DA feature detection was accomplished for taxonomic ranks starting at the phylum down to the genus and OTU level (cluster distance 0.03, classification confidence threshold of 80 %, and "other" threshold of 2 %) for samples collected in mice's ileum (overall 62 samples), as well as for SI vs LI (SI: 80 samples, LI: 92 samples). Briefly: the feature matrix, the OTU counts per sample, were imported into R. All counts were increased by 1 to prevent taking the log from 0 and stored within a `DGEList` object. To scale the raw library sizes, the `calcNormFactors` [174] function was applied using the *relative log expression (RLE)* method [175]. Subsequently, a model matrix according to the experimental design was created. The common dispersion of all biological coefficients of variation (BCV) averaged over all OTUs, as well as the OTU-specific dispersion of the dataset, were calculated (`estimateGLMTagwiseDisp`, `estimateGLMTrendedDisp`, respectively). Prior to the likelihood ratio test (`glmLRT`) [176, 177] for the given contrasts seen in Tab. 3.26, the read counts for each feature were fit to a negative binomial generalized log-linear model (`glmFit`). To account for multiple testing and control of the Type I error (FDR), p-values were adjusted by using the method described by Benjamini and Hochberg [171]. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant.

Differentially abundant feature detection in samples of source type *ileum*

10,633 distinct OTUs (clustered at a distance of 0.03), as well as the taxonomic features determined for phylum to genus, in 12 possible contrasts were tested within the DA feature analysis using the adapted edgeR approach evaluated in Sec. 2.8, and described in more detail in Sec. 3.7.2.

The subgroup of samples of source type *ileum* comprises 62 samples. Comparisons of community profiles were made within contrasts having just **a single** varying condition. All possible contrasts according to these criteria are listed in Tab. 3.26.

Table 3.26.: Contrasts considered within DA testing of mouse study samples of source type ileum. The GI mouse study comprises 62 samples for source type ileum, which can be grouped according to type, diet, source, and region. Community profile comparisons between groups having just **a single** varying condition are considered within the DA detection.

contrasts 1-4	contrasts 5-8	contrasts 9-12
WT.N.F.Ile-BD.N.F.Ile	KO.N.F.Ile-WT.N.F.Ile	KO.N.F.Ile-KO.E.F.Ile
KO.E.M.Ile-KO.N.M.Ile	KO.N.M.Ile-BD.N.M.Ile	WT.E.M.Ile-WT.N.M.Ile
KO.E.F.Ile-WT.E.F.Ile	WT.E.M.Ile-KO.E.M.Ile	WT.N.M.Ile-KO.N.M.Ile
KO.N.F.Ile-BD.N.F.Ile	WT.N.F.Ile-WT.E.F.Ile	WT.N.M.Ile-BD.N.M.Ile

The result of DA feature analysis in samples of source type *ileum*, is summarized in Tab. 3.27. The top 30 features detected as differentially abundant, according to a FDR less than 0.05, for each taxonomic level, as well as for OTUs is given as supplementary information in Appendix Tab. A.10-A.15

Table 3.27.: Summary table of DA feature detection of the GI mouse study based on samples of source type ileum. DA feature detection was performed using the R Bioconductor package edgeR. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant. The table lists DA features detected at the phylum down to the OTU (species) level. Additionally, the number of unique features per level and contrasts is given.

	phylum	class	order	family	genus	OTU
total	29	40	37	53	42	249
unique features	9	11	13	18	17	143
unique contrasts	10	11	11	10	11	12

Differentially abundant feature detection in samples grouped by SI and LI

DA feature detection as previously described in Sec. 3.7.2 was repeated by grouping samples of source type caecum (40) and colon (52) to group *large bowel* (LI, 92 samples) and samples of source type jejunum and ileum to group *small bowel* (SI, 80 samples). Comparisons of community profiles are made within contrasts having just **a single** varying condition. 34 possible contrasts were built according to these criteria and are listed in Tab. 3.28.

The result of the DA feature analysis in samples grouped by SI and LI, is summarized in Tab. 3.29. The top 30 features detected as differentially abundant, according

Table 3.28.: The total number of samples of the GI mouse study was grouped into 2 main subgroups by merging samples of source type caecum (40) and colon (52) to group *large bowel* (LI, 92 samples) and samples of source type jejunum and ileum to group *small bowel* (SI, 80 samples). The table comprises all possible contrasts (34) between different community profiles which have just **a single** varying condition.

contrasts 1-12	contrasts 13-23	contrasts 24-34
KO.N.F.SI-WT.N.F.SI	KO.N.F.SI-KO.N.F.LI	KO.N.F.SI-KO.E.F.SI
WT.N.M.LI-KO.N.M.LI	WT.N.M.LI-WT.E.M.LI	WT.N.M.LI-WT.N.M.SI
WT.N.F.SI-BD.N.F.SI	KO.N.M.LI-BD.N.M.LI	KO.N.M.LI-KO.E.M.LI
KO.E.M.SI-KO.N.M.SI	WT.N.M.SI-BD.N.M.SI	WT.N.M.SI-KO.N.M.SI
KO.N.F.SI-BD.N.F.SI	WT.N.F.LI-WT.N.F.SI	WT.N.F.LI-KO.N.F.LI
WT.N.M.LI-BD.N.M.LI	WT.E.M.SI-WT.E.M.LI	WT.E.M.SI-KO.E.M.SI
KO.N.M.LI-KO.N.M.SI	KO.N.F.LI-KO.E.F.LI	KO.N.F.LI-BD.N.F.LI
BD.N.M.LI-BD.N.M.SI	KO.E.F.SI-WT.E.F.SI	KO.E.F.SI-KO.E.F.LI
WT.N.F.LI-WT.E.F.LI	WT.N.F.LI-BD.N.F.LI	BD.N.M.SI-KO.N.M.SI
WT.E.M.SI-WT.N.M.SI	WT.N.F.SI-WT.E.F.SI	BD.N.F.SI-BD.N.F.LI
WT.E.M.LI-KO.E.M.LI	KO.E.M.SI-KO.E.M.LI	WT.E.F.LI-KO.E.F.LI
WT.E.F.SI-WT.E.F.LI		

to a FDR less than 0.05, for each taxonomic level, as well as for OTUs is given as supplementary information in Appendix Tab. A.16-A.21

Table 3.29.: Summary table of DA feature detection of the GI mouse study based on samples of groups SI and LI. DA feature detection was performed using the R Bioconductor package edgeR. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant. The table lists DA features detected at the phylum down to the species level. Additionally, the number of unique features per level and contrasts is given.

	phylum	class	order	family	genus	OTU
total	137	135	70	207	215	2,638
unique features	17	14	11	24	27	710
unique contrasts	33	33	25	31	29	34

3.8. *Candida sp.* Colonization of the Lower Respiratory Tract in Humans

To investigate the influences, such as antibiotic treatment or medication at the intensive care unit on the bacterial, as well as on the fungal community profile of the lower respiratory tract the experiment described in Sec. 2.1.1 was accomplished. The survey focuses on *Candida sp.* in the human lower respiratory tract and how they are affected within different conditions. Within this thesis, the bacterial, as well as the fungal community profile, was determined and tested for DA features within the different treatment groups, as well as in different sample collection types. In addition, traditional BAL and tracheal secretion culture results were compared to the taxonomic classification.

Tab. 3.30 summarizes abbreviations and descriptions of the different treatment groups and sample types, the subgroups of 3B. The experimental design of the survey is given in more detail in Sec. 2.1.1.

Table 3.30.: Sample abbreviation and descriptions used within the BAL survey for main groups, as well as for subgroups of group 3B.

main groups		**subgroups of 3B**	
`1A`	control no antibiotics	`NAP`	Nosocomial-Accquired Pneumonia
`1B`	control with antibiotics	`VAP`	Ventilation-Accquired Pneumonia
`2A`	mechanically ventilated, treated at ICU, no pneumonia, no antibiotics	`CAP`	Community-Acquired Pneumonia
`2B`	mechanically ventilated, treated at ICU, no pneumonia, with antibiotics	`NTS`	No Type Specified
`3B`	mechanically ventilated, treated at ICU, pneumonia, with antibiotics		

3.8.1. Bacterial community profile analysis - BAL study

429,680 16S amplicons were obtained by 454 sequencing. Prior to high-throughput analysis with SnoWMAn's RDP pipeline, the raw sequences were filtered for contaminating sequences by the Decontaminator, as well as for chimeras using UCHIME. Additionally, the remaining data was denoised and quality filtered using Acacia, before it was uploaded to SnoWMAn's data directory. Within RDP's preprocessing step, sequences with no matching MIDs, as well as sequences with a length less than

150 bps or containing ambiguous bases (N's), were discarded. Finally, 238,990 16S amplicons remained for phylogenetic classification using the RDP Classifier (version 2.5). Complete analysis settings are included as supplementary information in Appendix Tab. F.20. The summary by main groups, as well as subgroups of group 3B, of the sample overview, presenting the number of distinct OTUs built at different cluster distances is given in Tab. 3.31. The entire sample overview for each sample is provided as supplementary information in Appendix Tab. F.3.

Table 3.31.: Summary of the sample overview for the BAL study, bacteria. Number of distinct OTUs determined at different cluster distances summarized for main groups, as well as for subgroups of group 3B.

			number of OTUs at different cluster distances						
Sample	**Seqs**	**Unique Seqs**	0.0	0.01	0.02	0.03	0.04	0.05	0.06
Total 1A	31,923		3,045	1,610	705	491	421	378	343
Total 1B	20,193		2,040	1,152	522	373	322	295	276
Total 2B	27,781		2,310	1,254	544	363	299	256	235
Total 3B	13,5497		10,381	4,702	1,899	1,211	950	786	689
Total 3B ASP	41,828		3,476	1,806	771	503	406	353	311
Total 3B NAP	19,755		2,030	1,151	536	362	305	256	234
Total 3B VAP	68,610		5,498	2,718	1,217	815	666	577	514
Total 3B CAP	5,304		424	265	104	57	51	47	42
Total	**238,990**	**31,174**	**18,238**	**7,542**	**3,071**	**1,938**	**1,471**	**1,191**	**1,028**

For a first overview about the community profile within the main groups, α-diversity scores (at a cluster distance of 0.03) according to Richness [178], Chao1 [179], Chao1 (bc) [180], Shannon [181], Evenness [178], and ACE [182] were calculated using SnoWMAn's statistics features. Tab. 3.32 summarizes the different scores for the five main subgroups. The full table, presenting different scores for each sample, is given as supplementary information in Appendix Tab. F.3.

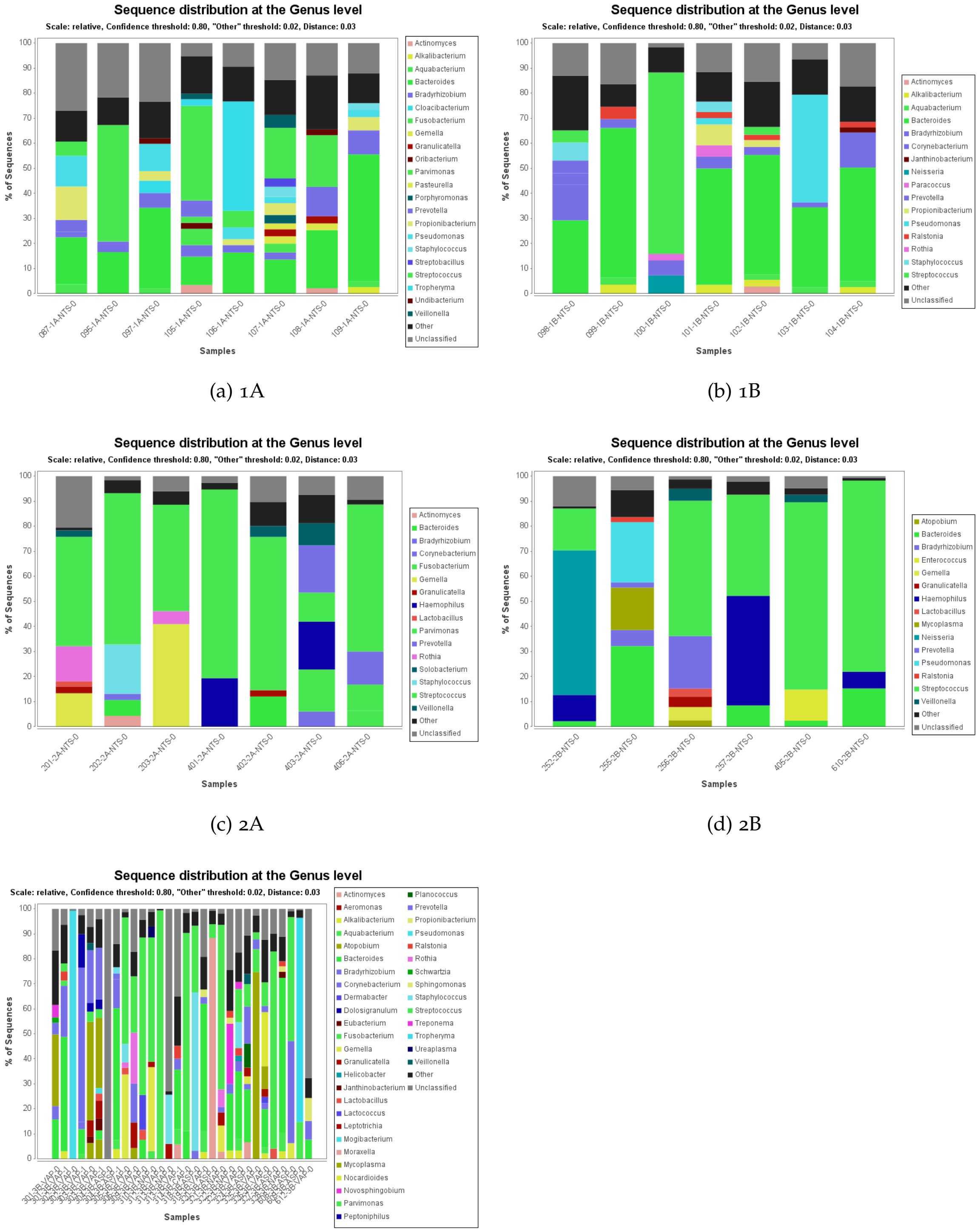
Table 3.32.: Summary of α-diversity scores according to Chao1, Chao1 (bc), Shannon, and ACE, as well as Richness and Eveness calculated for the main groups of the bacterial BAL samples.

Sample	Richness	Chao1	Chao1 (bc)	Shannon	Evenness	ACE
Total 1A	491	652.29	645.69	3.98	0.64	606.28
Total 1B	373	521.63	516.08	3.31	0.56	492.72
Total 2A	684	1,182.90	1,175.70	3.60	0.55	1,138.26
Total 2B	363	667.22	657.00	2.94	0.50	591.91
Total 3B	1,211	1,711.59	1,706.59	4.19	0.59	1,690.56

Community profiles for the main groups at the phylum level at a classification confidence of 80 %, a cluster distance of 0.03, and a threshold for *Other* of 2 % are presented as sequence distribution barcharts, seen in Fig. 3.16a-e. Charts and sequence distribution tables were directly generated and exported with SnoWMAn's visualization and statistical analysis tools. Count tables for Fig. 3.16a-e are included as supplementary information in Appendix Tab. F.5-F.9.

To get an overview about community profiles and their similarity of different samples, PCA was performed for all samples by SnoWMAn, illustrated in Fig.3.17. The major goal of PCA is to transform the given data set and reduction of dimensions, to an *alternative* data set which can be illustrated within the 2D space. This allows for community profile comparison by spatial arrangement. Samples which are located at smaller spatial distances than others share more similarities within their community profiles than samples observed further away. Generally, samples from the same group, type, or similar shared conditions are assumed to cluster together within a PCA plot; ideally they can be subsequently separated from each other. Although some kind of cluster formation can be observed within the first two components of the PCA plot (Fig. 3.17), these clusters cannot be explained by any of the defined groups of the experimental design.

In addition, community profiles of different subtypes of group 3B were illustrated using PCA, seen in Fig. 3.18. With this illustration, community profiles of subgroups are visually compared. As already observed for all samples, no clear separation or no shared community profile was found between samples of the same sub-source type (VAP, NAP, CAP, ASP).



(a) 1A

(b) 1B

(c) 2A

(d) 2B

(e) 3B

Figure 3.16.: Absolute sequence distribution at the phylum (a) down to the genus (e) level, at a classification confidence of 80 % and taxa covering more than 2 % of main BAL study groups.

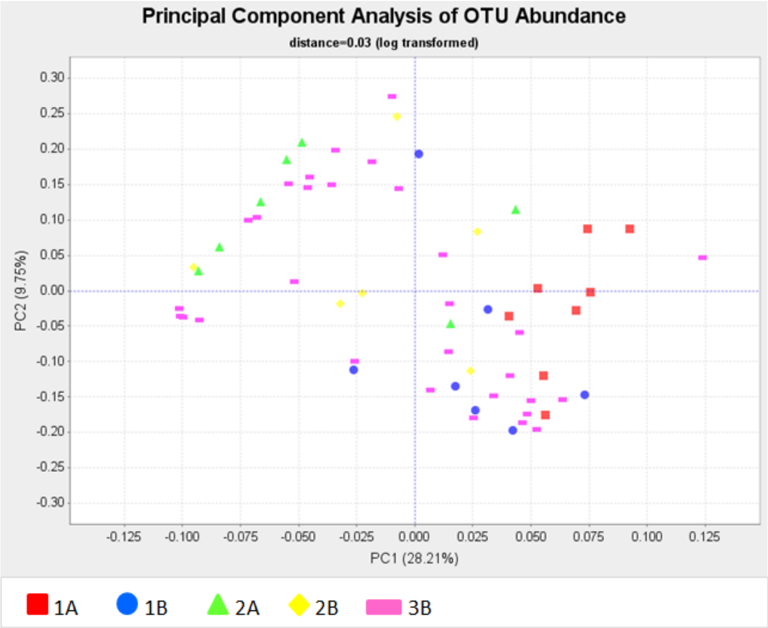


Figure 3.17.: PCA of OTU abundance for all samples of the bacterial BAL dataset. The main groups are presented by shape and color, according to the included legend. Although there seems to be a separation within the first 2 components, this is not confirmed by any of the known criteria (group 1A-3B).

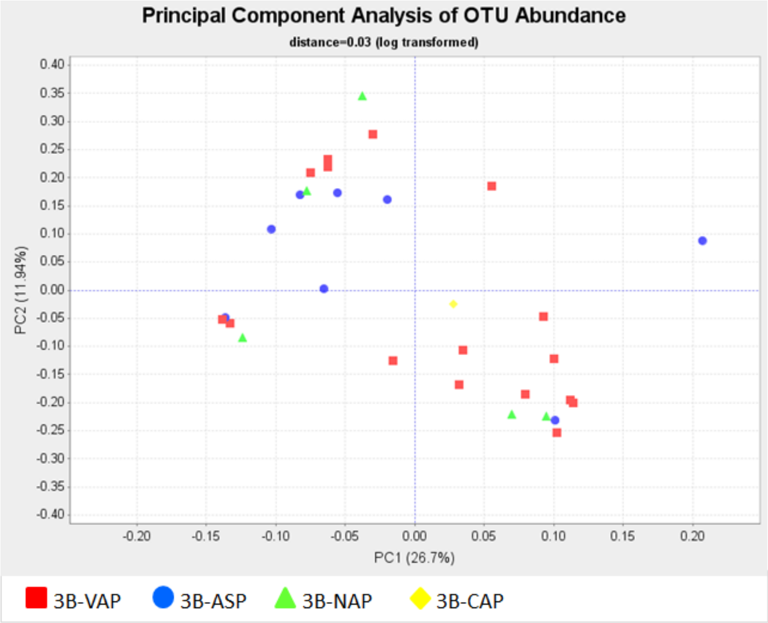


Figure 3.18.: PCA of OTU abundance for samples of group 3B. Subgroups of 3B are presented by shape and color, according to the included legend. Although there seams to be a separation within the first 2 components, this is not confirmed by any of the known subgroups (VAP, NAP, CAP, ASP).

3.8.2. Analysis of differentially abundant features

The final classification result, the OTU feature table, was exported from SnoWMAn and used for differentially abundant feature detection according to the methods evaluated in Sec. 2.8, and described in more detail in Sec. 3.7.2. DA feature detection was accomplished for taxonomic ranks starting at the phylum down to the genus and OTU level (cluster distance of 0.03, classification confidence threshold 80 %, and "other" threshold of 2 %) for samples collected in the lower respiratory tract of 58 humans. Features with an adjusted p-value (FDR) less than 0.05 were considered as statistically significant.

Differentially abundant feature detection in five main groups

578 distinct OTUs (clustered at a distance of 0.03), as well as the taxonomic features determined for the phylum to the genus in 10 possible contrasts, were tested within the DA feature analysis using an adapted edgeR approach, Sec. 3.7.2.

The experiment comprises 58 samples, which are assigned according to their treatment into five main groups: 1A (8 samples), 1B (7 samples), 2A (7 samples), 2B (6 samples), and 3B (30 samples). Changes within the community profile, between the different treatment groups in Tab. 3.33, were tested.

Table 3.33.: The experiment comprises 58 samples, which are assigned according to their treatment into five main groups: 1A (8 samples), 1B (7 samples), 2A (7 samples), 2B (6 samples), and 3B (30 samples). Changes within the community profile, between the different treatment groups presented in this table have been considered for DA feature detection.

contrasts 1-4	contrasts 5-7	contrasts 8-10
1A-1B	1A-3B	1B-3B
1A-2A	1B-2A	2A-2B
1A-2B	1B-2B	2A-3B
2B-3B		

The result of the DA analysis in bacterial samples between the main groups is summarized in Tab. 3.38. The top 30 features detected as differentially abundant, according to a FDR less than 0.05, for each taxonomic level, as well as for OTUs is given as supplementary information in Appendix Tab. F.23-F.28.

Table 3.34.: Summary table of DA feature detection of the BAL study, bacteria main groups. The DA feature detection was performed using the R Bioconductor package edgeR. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant. The table lists DA features detected at the phylum down to the species level. In addition, the number of unique features per level and contrasts is given.

	phylum	class	order	family	genus	OTU
total	17	36	87	182	178	578
unique featurs	4	11	39	39	39	258
unique contrasts	9	10	10	10	10	10

Differentially abundant feature detection in subgroups of group 3B

Community composition of 29 samples of group 3B, in respect to their sample type, was performed. Therefore, the original feature table was reduced to samples of group 3B, excluding the single sample of type CAP. Furthermore, features with no counts at all, were excluded from the feature matrix. DA feature detection as previously described in Sec. 3.7.2 was repeated by comparison of community profiles of different types of group 3B. This resulted in testing of 1,203 OTU features in 3 possible contrasts: (1) 3B.ASP-3B.VAP, (2) 3B.ASP-3B.NAP, and (3) 3B.VAP-3B.NAP. Taxonomic groups from the phylum to the genus level were also tested for DA.

The result of the DA analysis in bacterial samples between types of group 3B is summarized in Tab. 3.35.

Table 3.35.: Summary table of DA feature detection of the BAL study, bacteria and main groups. DA feature detection was performed using the R Bioconductor package edgeR. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant. The table lists DA features detected at the phylum down to the species level. In addition, the number of unique features per level and contrasts is given.

	phylum	class	order	family	genus	OTU
total	3	5	22	55	58	183
unique featurs	3	4	11	32	32	118
unique contrasts	2	2	3	3	3	3

3.8.3. Fungal community profile analysis

543,566 ITS1 amplicons were obtained by 454 sequencing. Prior to high-throughput analysis with SnoWMAn's BLAT pipeline, the raw sequences were filtered for contaminating sequences by the Decontaminator, as well as for chimeras using UCHIME. In addition, the remaining data was denoised and quality filtered using Acacia, before it was uploaded to SnoWMAn's data directory. Furthermore, 10 samples were excluded from downstream analysis due to negative PCR amplificaton. Within BLAT's preprocessing step, sequences with no matching MIDs, as well as sequences with a length less than 150 bps or containing ambiguous bases, were discarded. Finally, 466,582 ITS1 amplicons remain for phylogenetic classification using the UNITE reference database (release 15.10.2013). Complete analysis settings are included as supplementary information in Appendix Tab. F.21.

Table 3.36.: Summary and sample overview of BAL study for fungi. Number of distinct OTUs determined using SnoWMAn's BLAT pipeline, are summarized for the main groups, as well as for the subgroups of group 3B.

Sample	Sequences	Unique sequences	# OTU
Total 1A	38,653		135
Total 1B	39,412		567
Total 2A	55,536		232
Total 2B	60,243		70
Total 3B	251,022		309
Total 3B ASP	56,893		189
Total 3B NAP	31,189		96
Total 3B VAP	162,940		140
Total	444,866	51,248	855

Phylogenetic analysis of the fungal community by BLAT, for the remaining 48 samples, resulted in 855 distinct species (OTUs).

Of special interest within this analysis is the fungal community profile of the different treatment groups (1A, 1B, 2A, 2B, 3B), as well as the different sample types within group 3B (ASP, NAP, VAP). Tab. 3.36 presents the number of finally determined distinct species according to their group relationship. The number of OTUs determined for each sample is included within the supplementary information in Appendix F.22.

Richness, the number of distinct species found for each sample, is listed in Tab.

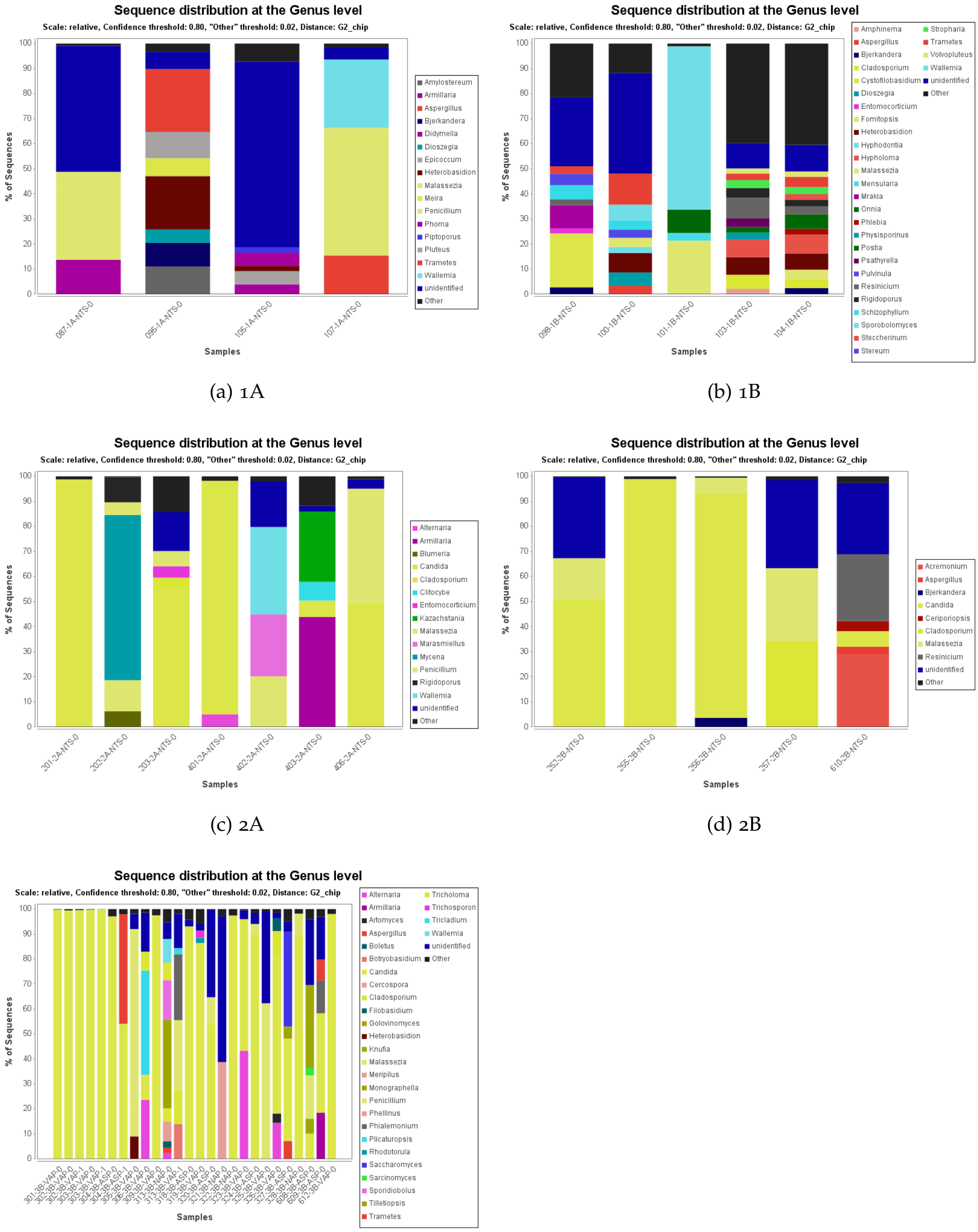
3.37, as well as α-diversity scores according to Chao(1), Chao1(bc), Shannon, Evenness, and ACE are summarized for the five main groups and subgroups of 3B in Tab. 3.37. Different α-diversity scores calculated separately for each samples are included as supplementary information in Appendix Tab. F.17.

Table 3.37.: Summary of α-diversity scores according to Chao1, Chao1 (bc), Shannon, and ACE, as well as Richness and Eveness calculated for the main groups of the fungal BAL samples.

Sample	**Richness**	**Chao1**	**Chao1** (bc)	**Shannon**	**Evenness**	**ACE**
Total 1B	567	802.76	796.52	4.70	0.74	732.92
Total 2A	232	470.10	457.00	2.47	0.45	380.96
Total 2B	70	113.56	107.80	1.68	0.40	116.55
Total 3B	309	561.02	551.45	2.61	0.45	545.88
Total 3B ASP	189	373.38	363.00	1.88	0.36	380.86
Total 3B NAP	96	120.20	117.00	2.54	0.56	118.26
Total 3B VAP	140	244.17	234.23	2.04	0.41	233.59

As the main focus of the survey is targeted to the amount of *Candida sp.*, sequence distribution for the five main groups is illustrated at the genus level, seen in Fig. 3.19a-e. Clusters which do not include more than 2 % of the total abundance are summarized to taxon *Other*. Absolute count data for this figures is given as supplementary information in Appendix Tab. F.11-F.9.

Similarities within the fungal community profile between the different samples were visualized using SnoWMAn's integrated PCA. Fig. 3.20 compares community composition amongst all samples, as well as between sample types of group 3B, seen in Fig. 3.21.



(a) 1A

(b) 1B

(c) 2A

(d) 2B

(e) 3B

Figure 3.19.: Absolute sequence distribution at the phylum (a) down to the genus (e) level, at a classification confidence of 80 % and taxa covering more than 2 % of main fungal BAL study groups.

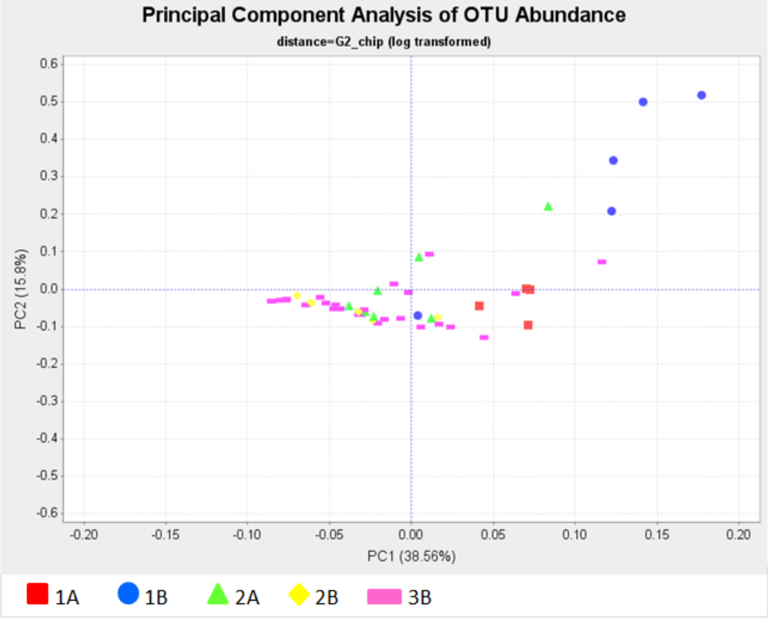


Figure 3.20.: Principle component analysis of all 48 ITS BAL samples. Subgroups of 3B are presented by shape and color, according to the included legend. Although there seems to be a separation between certain groups, not all five main groups can be spatially separated within the PCA plot.

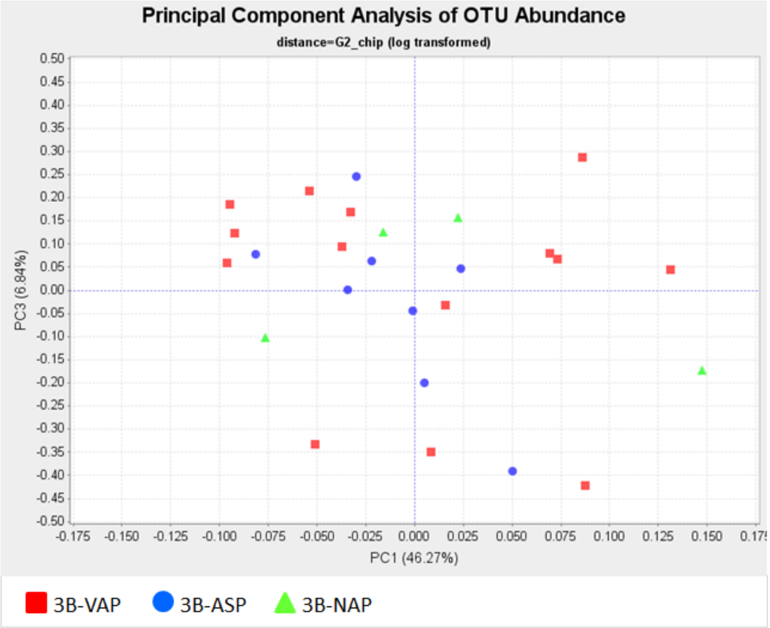


Figure 3.21.: PCA of only BAL ITS samples of group 3B. Subgroups of 3B are presented by shape and color, according to the included legend. PCA analysis of subgroups of group 3B samples illustrated that community profiles of different sample types are highly varying, and no pattern was recognized within the types.

3.8.4. Analysis of differentially abundant features

The final classification result, the OTU feature table, was exported from SnoWMAn and used for differentially abundant feature detection according to the methods evaluated in Sec. 2.8, and described in more detail in Sec. 3.7.2. The DA feature detection was accomplished for taxonomic ranks starting at the phylum down to the species ("Other" threshold of 2 %) level for samples collected in the lower respiratory tract of 58 humans. Features with an adjusted p-value (FDR) of less than 0.05 were considered as statistically significant.

Differentially abundant feature detection in five main groups

1,180 distinct OTUs (clustered at a distance of 0.03), as well as the taxonomic features determined for phylum to genus, in 10 possible contrasts, were tested within the DA feature analysis using an adapted edgeR approach, Sec. 3.7.2.

The experiment comprises 58 samples, whereby 10 samples were excluded due missing ITS amplicons during PCR. The remaining 48 are assigned according to their treatment into five main groups: 1A (4 samples), 1B (5 samples), 2A (7 samples), 2B (6 samples), and 3B (26 samples). Changes within the community profile between the different treatment groups, listed in Tab. 3.33, were tested.

The result of the DA analysis in fungal samples between the main groups is summarized in Tab. 3.38. The top 30 features detected as differentially abundant according to a FDR of less than 0.05 for each taxonomic level, as well as for OTUs, is given as supplementary information in Appendix Tab. F.29-F.34.

Differentially abundant feature detection in subgroups of group 3B

DA feature detection of 26 samples of group 3B in respect to their type was performed. Therefore, the original feature table was reduced to samples of group 3B. Furthermore, features with no counts at all, were excluded from the feature matrix. DA feature

Table 3.38.: Summary table of DA feature detection of the BAL study, fungi, main groups. DA feature detection was performed using the R Bioconductor package edgeR. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant. The table lists DA features detected at the phylum down to the species level. In addition, the number of unique features per level and contrasts is given.

	phylum	class	order	family	genus	OTU
total	7	31	82	151	191	1,180
unique featurs	3	9	17	33	47	352
unique contrasts	4	10	10	10	10	10

detection as previously described in Sec. 3.8.4 was repeated by the comparison of community profiles of different types of group 3B.

The result of the DA analysis in fungal samples between types of group 3B is summarized in Tab. 3.39.

Table 3.39.: Summary table of DA features detection in different types of group 3B of fungal samples, BAL study. DA feature detection was performed using the R Bioconductor package edgeR. Only features with an adjusted p-value less than 0.05 were considered as differentially abundant. The table lists DA features detected at the phylum down to the species level. In addition, the number of unique features per level and contrasts is given.

	phylum	class	order	family	genus	OTU
total	-	8	33	41	54	123
unique features	-	5	16	20	27	78
unique contrasts	-	3	3	3	3	3

3.8.5. Comparison of microbiological diagnostic results with results of high-throughput classification

In addition to the high-throughput classification and characterization approach, traditional bronchoalveolar lavage and tracheal secretion cultures were performed and analyzed by the team of Prof. Robert Krause, MD[72] within the BAL survey. 40 out of the initial 58 samples revealed positive cultures for known bacterial strains. Thereby 19 samples were positively tested for known pathogenic strains, such as *Klebsiella pneumoniae*, *Staphylococcus aureus*, or *Pseudomonas aeruginosa*. The complete culture result is given in detail as supplementary information in Appendix Tab. F.35.

Table 3.40.: Representative sequences used for comparison of the culture and the taxonomic classification result. For each microbial strain, detected by the culture approach, a representative sequence was selected from the public sequence repository GenBank. Species name and respective Accession number, as well as the name used within the culture result is presented within the table.

Representative species name	**RefAccession**	**culture name**
Aspergillus robustus	EF661435.1	Schimmelpilze
Candida albicans	AB437043.1	
Candida boidinii	FJ914930.1	
Candida dubliniensis	AJ865083.1	
Candida glabrata	HE993757.1	
Candida parapsilosis	FM172980.1	
Corynebacterium lactis	HE983830.1	
Enterobacter cloacae	KF535159.1	Enterobacter
Enterococcus feacalis	FJ378663.2	nicht hoemolyt. Streptococcus
Escherichia coli	J01859.1	
Haemophilus influenzae	AY613741.1	
Klebsiella pneumoniae	KC990817.1	Klebsiella sp
Klebiella oxytoca	AB626120.1	
Neisseria sp. oral strain	AY005028.1	Neisseria
Proteus mirabilis	KF535110.1	
Pseudomonas aeruginosa	KJ156527.1	
Staphylococcus aureus	DQ630753.1	
Staphylococcus lugdunensis	AY903258.1	Koagulase negative Staphylokokken
Streptococcus mitis	NR_028664.1	alpha haemolyt. Streptococcus
Streptococcus pneumoniae	GU326244.1	
Streptococcus viridans	AF076036.1	

BAL and tracheal secretion culture results revealed 21 distinct bacterial or fungal strains. For group names such as *alpha haemolyt. Streptococcus* a representative sequence was manually selected and used for further processing. Tab. 3.40 lists all 21

[72] Section of Infectious Diseases and Tropical Medicine, Medical University of Graz, Graz, Austria

reference sequences and the corresponding *culture name* in the case of group names. Reference sequences were manually selected, according to the given cultures names, and extracted from GenBank.

To compare the culture results with the high-throughput analysis result, preprocessed and split sequence information was exported from SnoWMAn to distinct `FASTA` files. Reference sequences were merged to a single database file which was indexed using `formatdb`. In the next step, a simple *bash* script was used to run BLAT analysis for each sample (16S and ITS) against the created *culture reference DB*. BLAT output was further processed using a custom R script which selects the best BLAT hit for each sequence and piles them up (counting of how often a specific hit occurs).

Tab. 3.41 compares the high-throughput classification with the culture result, for all *Candida sp.* positive samples and a corresponding abundance that covers more than 2 % of the average library size.

The comparison was evaluated in respect to the agreement on detected *Candida sp.*. The *Sync* column distinguishes between **match** ($\surd$, *Candida sp.* was detected as the most abundant feature), **partial match** ($\sim$, more than one high abundant *Candida sp.* was not detected either by culture or by molecular characterization), or **no match** (X, high abundant *Candida sp.* sequence count within the molecular approach, but negative culture result). For low abundant[73] *Candida sp.*, data not shown in Tab. 3.41, no corresponding positive culture result was available.

[73] less than 2 % of the average library size

Table 3.41.: Comparison of the taxonomic classification with the BAL culture result, for all samples which contain more than 2 % of the average library size sequences, identified as *Candida sp. Sync* column evaluates the comparison by **match** (√, *Candida sp.* was detected as the most abundant feature), **partial match** (∼, more than one high abundant *Candida sp.* was not detected either by culture or by molecular characterization), or **no match** (X, high abundant *Candida sp.* sequence count within the molecular approach, but negative culture result).

	Representative sequence and Accession number					culture result	bacterial count	sync
	A. robustus EF661435.1	*C. albicans* AB437043.1	*C. dubliniensis* AJ865083.1	*C. glabrata* HE993757.1	*C. parapsilosis* FM172980.1			
201-2A-NTS-0	0	17	8,638	0	0	*C. boidinii* *C. dubliniensis*	1.00E+03	∼
203-2A-NTS-0	1	725,	158	0	1	negative		X
252-2B-NTS-0	0	5,826	144	0	0	negative		X
255-2B-NTS-0	0	8,228	195	0	0	*C. albicans* *Schimmelpilz*	1.00E+03 1.00E+03	∼
256-2B-NTS-0	0	8,107	202	0	0	negative		X
301-3B-VAP-0	0	16,093	220	0	6	*C. albicans*	1.00E+03	√
302-3B-VAP-0	0	10,805	195	0	0	*C. albicans*	1.00E+02	√
302-3B-VAP-1	0	5,036	67	0	0	negative		X
303-3B-VAP-0	0	7,168	122	0	0	*C. albicans*	1.00E+03	√
303-3B-VAP-1	0	18,638	214	0	0	*C. albicans*	1.00E+01	√
304-3B-ASP-0	261	11,376	110	0	0	negative		X
304-3B-ASP-1	0	5,891	90	2	0	negative		X
309-3B-VAP-0	1	21,969	262	0	0	*C. albicans*		√
318-3B-ASP-0	0	1,293	13	1,233	0	*C. albicans* *C. glabrata*	1.00E+03 1.00E+06	√
319-3B-VAP-0	0	15,826	155	0	7	*C. albicans*	1.00E+01	√
322-3B-NAP-0	0	11,903	149	0	0	negative		X
323-3B-VAP-0	1	11,443	125	0	2	*C. albicans*	1.00E+01	√
324-3B-ASP-0	0	18,318	198	3	1	negative		X
326-3B-VAP-0	0	8,412	80	0	0	*C. albicans*	1.00E+01	√
327-3B-ASP-0	142	797	0	0	0	negative		X
328-3B-NAP-0	0	1,219	10	14	0	negative		X
401-2A-NTS-0	0	10,035	190	0	1	*C. albicans*	1.00E+03	√
403-2A-NTS-0	0	492	1	0	0	negative		X
405-2B-NTS-0	0	14,623	332	0	13	*C. albicans*	1.00E+05	√
406-2A-NTS-0	0	1,530	32	0	0	negative		X
609-3B-ASP-0	0	2,914	60	0	0	negative		X
612-3B-VAP-0	0	6,867	3,185	0	1	*C. albicans*	1.00E+02	∼

4. Discussion

The following sections summarize, review, and critically discuss the results obtained related to the specific objectives for this thesis. In the course of this research work, new approaches, methods, and tools for important steps in the entire high-throughput characterization and classification process of complex microbial communities have been created, evaluated, adopted, or extended. At the experimental design level, the effects of sequencing library normalization on the final community profile and its diversity was investigated. Subsequently, the *Decontaminator*, an effective tool for the removal of contaminating sequences from the target data sets is introduced as a major improvement during sequence pre-processing. For the core step, the taxonomic classification, an internal transcribed spacer (ITS) reference database for fungal sequences was created. Tests of the ITS amplicon classification with a hand curated *in-silico* amplified and fully annotated ITS mock community, showed good results for *reference based* classification and *de-novo* OTU picking approaches based on the UNITE ITS reference sequences. Statistical analysis of determined community profiles was extended by methods for differentially abundant feature detection. Therefor, Metastats, edgeR, and limma+voom, were evaluated using simulated count data, revealing that the linear modeling approaches outperform Metastats for bigger library sizes and fold change values. Based on this evaluation result, real community profiles obtained from analyses conducted within this thesis were tested for differentially abundant features. Finally, with the transcriptome analysis of two *Campylobacter fetus* subspecies, the typical ϵ-proteobacterial promotor motif was also confirmed for *C. fetus sp.* Moreover, this kind of analysis introduces a future direction for more detailed investigation of specific members of a microbial community.

4.1. Investigation of How Sequencing Library Normalization Affects the Final Community Profile and its Diversity

Within this thesis, the effects of library normalization on the final community profile and its diversity was investigated. The number of DNA copies per species of a metagenomic sample may differ by several orders of magnitude. This can be easily explained by the community composition, as more abundant species are more likely to be collected than the less abundant ones. Consequently, DNA of rare species is less often amplified during PCR. To overcome this *"ousting"* of rare species, DNA library normalization can be performed.

The effects of sequencing library normalization was studied, using the example of the community profile of a normalized and a standard shotgun sequencing library, sampled at an Alpine peat bog habitat. The main focus of the survey was the analysis of functional systems covered by bacteria. However, sampling of bacterial DNA in an eukaryotic host leads very likely to contaminations by host DNA.

The first examination of the sequence distribution at the phylum level revealed that normalization of sequencing libraries affects overall domain distribution with a shift to more eukaryotic sequences and a higher percentage of unassigned sequences. These observations had been confirmed using the χ^2-test [163]. Additionally, rarefaction analysis illustrated that, although the standard library comprises more sequences, Richness [178] is higher within the community profile based on the normalized sequencing libraries. This effect can be explained by the fact that within the normalized sequencing libraries, low abundant species are ousted by high abundant species to a lesser extent. Furthermore, rarefaction analysis illustrates that further sample collection or deeper sequencing would still increase species count, as rarefaction curves have not reached saturation yet.

Surprisingly, the contaminating, eukaryotic fraction of the metagenomic community profile was not that large as expected. Additionally, DNA of the host system *S. maggelanicum* (phylum Streptophyta) revealed only to 16 % of overall eukaryotic DNA. This can be explained by the two stage mechanical filtering procedure, which was performed during sample processing. It seems as this simple purification approach facilitates successful filtering of fragments from the moss host. The remaining amount of contaminating DNA originates mainly from fungal and animal material, which is also common in peat bogs. This is also reflected within the shift to eukaryotic sequences within the normalized community profile, as these species belong to the less abundant inhabitants of the investigated habitat.

Conclusively, with this experimental design on sequencing library normalization it was shown that, although the effects on the contaminating sequences within the community profile had not been that drastic as expected, less abundant species are covered to a higher extent after library normalization.

4.2. Development of an Application for Identification and Removal of Contaminating Sequences

The Decontamiantor was implemented as a platform independent standalone JAVA [130] command line tool for integration into the pre-processing step of the different analysis pipelines of SnoWMAn [67]. Apart from the integration into SnoWMAn, the structure and interfaces, provided by the Decontaminator implementation, allow for easy standalone or use within other pipeline systems, such as mothur [69], CloVR [68], or QIIME [70].

The major goal of the Decontaminator is to identify and remove contaminating sequences in targeted amplicon sequencing projects. The removal of sequences which do not originate from the targeted source (contaminations), noise, as well as of sequences of poor quality, or artificial sequences (chimeras), is a very crucial step in the pre-processing of targeted amplicon sequencing data [183]. All these kinds of "unwanted" sequence fragments falsify the result. Especially the final number of OTUs, and subsequent community diversity is artificially increased (OTU inflation), which can lead to false conclusions [183].

Firstly, the Decontaminator was evaluated by using a small dataset of true 16S amplicons in a two-step procedure. Within this evaluation, it was also shown, by *in silico* amplification, that the universal 16S marker gene primers lead to random amplification of DNA within the human genome. This might be also true for other eukaryotic host genomes to a certain extent [184–186]. These sequences were used in combination with the true 16S amplicons for the first evaluation step. As expected, the decontamination result revealed that these randomly amplified fragments do not have BLAT [97] hits at all. This confirms that contaminating sequences can be identified by using their marker gene structure using a BLAT similarity search. In the second evaluation step, in addition to the contaminating sequences, manually created chimeras have been added to the true 16S test fragment set in order to investigate how they are treated by the Decontaminator approach.

Chimeric sequences are a mixture of two or more 16S sequence fragments which are merged by accident during PCR. Hence, it was expected that chimeric sequences show a low query coverage within the BLAT result, as they match different reference sequences partially. This is true for chimeras formed by fragments of distinct species. Chimeras which are products of similar fragments or of related species, still show high-scoring BLAT hits and high percentage of query coverage.

The same test datasets were used to investigate the detection and removal of contaminations by the standalone version of DeconSeq [80] (v.0.4.3). Although DeconSeq is more tailored towards finding contaminations within metagenomic datasets, it can be also used for genomic data as long as read length is $>$ 150 bps.

The direct comparison of the Decontaminator and DeconSeq showed, that both approaches were able to identify all randomly amplified human contaminations correctly, whereas, chimeras were only detected by the Decontaminator. Despite DeconSeq also facilitates optimization by the percentage of identity and query coverage, even for high query coverage values, none of the chimeras was identified.

In addition, the Decontaminator was used to filter a 16S targeted amplicon sequencing raw data set, originating from the diarrhea study, (described in Sec. 2.1.5) [96]. Exemplified by this survey, different pre-processing approaches and combinations in respect to the effect on finally observed OTUs and community diversity, have been evaluated. The comparison of filtered sequences according to chimera detection, noise (incl. quality) filtering, and decontamination of the raw sequences revealed, that nearly half of noisy and low quality sequences, as well as about 7 % of chimeric sequences were already removed by the Decontaminator.

By comparing the number of finally obtained OTUs at different stages of pre-processing, it was revealed that the number of OTUs decreases by one quarter, in raw sequences compared to the fully pre-processed data set. Although each different pre-process step decreases the finally observed OTU counts, decontamination was shown to have the biggest effect. This is explained by the Decontaminator filtering criteria which allows via cutoffs for percentage identity and percentage query coverage to discard sequences of poor quality, or short reads, as well. BLAT results with low percentage identity can be explained by sequencing noise or homopolymimeric regions which are also filtered during denoising using tools such as Acacia [139].

Conclusively, the Decontaminator, a novel tool developed within this thesis, was shown to be an effective approach for the identification and removal of contaminating sequences and partially for chimeric sequences. The major advantage of the Decontaminator compared to other tools such as DeconSeq is that the origin of the contaminating sequences can remain unknown. The basic assumption of the Decontamintor is, that contaminating fragments do not follow the characteristic marker gene structure. This also means that for each marker gene only one reference database has to be stored and maintained for sequence similarity analysis.

Furthermore, the Decontaminator is implemented as a JAVA application, which allows for platform independent usage. The generic structure of the implementation enables easy adaption and extension to other input formats or evaluation mechanisms than the standard `blast8` BLAT output. Although the Decontaminator was developed for the integration into SnoWMAn's analysis pipelines, the provided command line interface can also be easily used within other pipelines such as mothur, CloVR, QIIME, or independently on the command line only, prior to phylogenetic analysis. The identification and removal of contaminating sequences is one of the first pre-processing steps. Hence, samples are mostly not de-multiplexed or trimmed. As a consequence artificial oligos which might be ligated to the amplified fragments, such as MIDs or primers, are still part of the sequence. The Decontaminator facilitates consideration of theses oligos by specific off-set parameters which are used for calculation of the true query coverage.

4.3. Integration and Evaluation of Resources for Fungal Community Analysis

Within this thesis, a fungal reference database, based on the UNITE [82] system for DNA-based fungal species circumscriptions, was created and incorporated into SnoWMAn's BLAT pipeline. Currently, fungal community analysis is supported by common approaches such as mothur and QIIME by using the provided UNITE sequence collection, as well as by CloVR [68] using a custom ITS reference sequence set. In addition, an ITS training set for the RDP classifier [110] was announced already at the end of last year, and released officially just during finishing this thesis[74]. As a consequence, the evaluation of resources for fungal community profiling was done on the beta version of the RDP ITS classifier. Nevertheless, although these newly evaluation results were not shown within this thesis the obtained results are discussed within the following paragraphs.

Although existing tools, methods, and pipelines for fungal community profile analysis are continuously improving, community analysis based on ITS amplicons is still in its infancy. Therefore, an *in silico* ITS mock community was created for the evaluation of this newly integrated resource, as well as for already existing or upcoming resources for fungal ITS characterization and classification.

Although the large subunit (LSU) has a longer tradition in mycology for phylogenetic analysis of fungal species [187–190], the ITS region was introduced by Schoch *et al.* [58] as the universal barcode for fungi. Recent comparisons by Porras-Alfaro *et al.* [83] also confirmed that LSU and ITS regions show similar classification accuracies.

[74] RDP classifier 2.8, release 8. July 2014

More general consideration of the community profile analysis in bacterial community studies addressed the *3 % gold standard* for species discrimination, which is based on the 16S ribosomal small subunit (SSU) [191]. This 97 % sequence similarity level was also confirmed as a good threshold for species discrimination in fungi of the phylum Basidomycota [192]. In contrast, for phylum Ascomycota, it was shown that a cluster distance of 0.02 (98 % sequence similarity within the clusters) reaches better species discrimination compared to 0.03 [192]. In addition, these cluster similarity cut-off values were confirmed for ITS1, as well as for ITS2 fragments [192]. Nevertheless, the 97 % similarity level for species level approximation, is also supported by the results of previous analysis [193–196].

The *in silico* mock communities were created by *in silico* amplification from a set of manually selected sequences, provided by Henrik R. Nilsson[75] and Kessy Abarenkov[76]. Although for all fungal sequences in this set it was guaranteed that the full ITS1 and ITS2 region was covered some of the sequences did not generate an amplicon during the *in silico* amplification. This can be explained by a lack of primer binding sites within the flanking 18S or LSU region or the fact that the used primers are biased towards certain species. Subsequent annotation of the successful amplified ITS mock sequences revealed that only one third of all mock sequences have an explicit annotation down to species level. Precisely, two thirds of all mock sequences contain at least one ambiguous description for one or more taxonomic level, such as *"derived from ..."*, *"unidentified"*, *"Incertae sedis"*, or *"uncultured fungus"*. These incomplete fungal annotations are a consequence of the difficulties in fungal species identification by traditional methods. A large proportion of fungi cannot be kept in culture and thus cannot be examined by traditional culture dependent methods; other fungi do not seem to produce tangible fruiting bodies, such that morphological examination becomes hard or impossible.

In addition, fungal annotations suffer from inconsistencies in name declaration that are caused by fungi which occur in several morphological forms, resulting in different names for the sexual and asexual or vegetative reproduction of the same fungus [197].

[75] Department of Biological and Environmental Sciences, University of Gothenburg, Sweden

[76] Natural History Museum, University of Tartu, Estonia

Furthermore, sequence names contained in public repositories, such as GenBank, essentially rely on accurate and correct user input. Thus, a significant number of sequences within this repository was discovered with erroneous, ambiguous, or imprecise annotations the so-called "dark taxa" [198, 199]. However, by annotation improvement approaches such as the initiative started by Nilsson *et al.* [81] or the currently introduced standards and protocols for sequence data quality improvement from Schoch *et al.* [197], will help to improve references which were used for the identification of fungi, and subsequently also for annotation and characterization.

For now, available methods, tools, and reference systems have to be continuously updated and evaluated, through approaches such as the introduced ITS mock communities. Within this thesis, two different taxonomic classification approaches and their reference sequence sets were investigated by the analysis of the created mock communities.

First, the annotated ITS1/2 *in silico* mock was used to test SnoWMAn's BLAT pipeline using the newly incorporated ITS reference DB. This evaluation revealed that all sequences were identified as kingdom fungi. Further consideration of the classification result at different taxonomic levels showed that annotation within the reference system is very poor for lower taxonomic levels such as family, genus, or species. Almost 20 % of total counts were assigned to an *unidentified* fungal family level, which increases to a quarter of ambiguous species annotations such as *"uncultured Glomos", "uncultured fungus"*, or similar descriptions.

The same analysis was repeated for the ITS1 mock community. Interestingly, the amount of ambiguous annotations at lower taxonomic levels decreases to ~13 % at the genus level. It seems as shorter fragments are more suitable for *reference based* OTU picking. This might be based on the underlying homology search for which it is more likely to find good matching results for shorter fragments.

Second, the *beta* version of the RDP ITS classifier [83, 100], trained on manually curated ITS sequences (based on UNITE), was evaluated using the ITS1/2 mock community. For the beta version of the classifier this analysis showed an increasing amount of *unclassified* sequences ranging from 15 % at the phylum level to almost

50 % at the genus level, at a classification confidence threshold of 80 %, and a cluster distance of 0.03. Hence, this approach was still far from allowing full characterization of fungal communities.

Subsequently the RDP ITS classifier (beta version) was applied to the ITS1 mock community. From this analysis similar results compared to the previously evaluated ITS1/2 mock were obtained. In particular, the amount of unclassified sequences increased nearly to 50 % at the genus level.

Also the comparison of the final taxonomic classification, at different phylogenetic levels, between the BLAT and the RDP classifier approach (beta version) resulted in overall promising results for BLAT (percentage of correct identified sequences > 80 % at all phylogenetic levels), whereas the number of correctly identified sequences decreases from 85 % at the phylum to 40 % at the genus level for the beta version of the RDP ITS classifier. This might be caused more or less by the amount of unclassified sequences, which ranges from ∼13 % at the phylum level to ∼50 % at the genus level for both mock communities. With the official release of the RDP classifier 2.8, including the updated training set, based on the UNITE reference sequences, as well as on ITS sequences obtained from the Warcup[77] collection, these observations were mainly confirmed.

For the UNITE training set, the amount of unclassified sequences ranged from 1.5 % at the phylum level to about 4 % at the genus level. In contrast, for the Warcup training set, the fraction of unclassified sequences starts at about 5 % at the phylum level and reaches almost 25 % at the genus level. Although the amount of unclassified sequences was drastically decreased the percentage of correct identified sequences resulted to less than 50 % at the family and genus level for the RDP classifier 2.8 trained on the UNITE sequence set. The same classifier version trained on the sequences from the Warcup collection showed high levels of correct identified sequences at the phylum and class level, but for lower taxonomic levels nearly half of all sequences were annotated incorrectly. The better performance of the RDP classifier 2.8 trained on the Warcup

[77] Warcup is a version from an active curatorial effort kindly provided by Paul Greenfield, Vinita Deshpande and colleagues of the Australian CSIRO (manuscript in preparation)

collection is very likely based on the more sensitive selection of training sequences, as only sequences with a full and unambiguous taxonomic annotation were chosen. The comparison of correctly identified sequences between the two mock communities showed overall better results for the ITS1 mock.

Conclusively, the novel reference DB for community profile analysis using SnoWMAn's BLAT pipeline was shown as confident classification resource for ITS amplicons, although fungal annotations suffer from classification deficiencies at lower taxonomic levels. In addition, the reference based OTU picking approach used by BLAT outperformed the *de novo* OTU picking and classification method of the RDP classifier although both approaches rest on the same reference sequence collection.

For a more elaborate classification approach using the newly introduced ITS classification resource, SnoWMAn's UCLUST pipeline needs to be adopted and extended for BLAT. This would allow for *de novo* OTU picking using the clustering approach of UCLUST and subsequent BLAT classification of each cluster representative with the created ITS reference DB. In addition, the training sets of the RDP classifier have to be extended and improved to allow for confident classifications also at lower taxonomic levels.

4.4. Evaluation and Adaption of Methods for Differentially Abundant Feature Detection

The evaluation of three different approaches, (1) Metastats [144], (2) edgeR [149], and (3) limma+voom [154, 155], for differentially abundant feature detection on simulated count data, confirmed that methods already well-established for differentially expressed gene detection in microarray or RNA-seq experiments (edgeR and limma+voom), can be applied for the analysis on community profile count data as well.

There are two crucial things which have to be considered. First, as the final feature matrix which represents the community profile, is very sparse (counts for most features are zero), this might cause problems in logarithmic calculations. To overcome this common structure of the data, a small value can be add to each count value of the feature matrix. Second, basic assumptions on the data distribution have to be satisfied in order to guarantee valid results. To account for the correct distribution of the count data, it can be transformed, for example using voom, prior to processing with limma.

Within this thesis, different methods for DA feature detection were compared, as well as in addition, their performance on different sample sizes, effect sizes, number of samples, and replicates was evaluated.

Unlike the results of McMurdie and Holmes [84] the smallest amount of max. sequences per samples (2000) resulted in overall bad results for all three methods. Especially for log fold change (logFC) values of 1.25, regardless of the sample size, or the replicate number, DA feature detection failed. For bigger effect sizes (logFC1e^{-}10.0, and logFC = 1e^{-}5.0), between 20 and 40 differentially abundant features were detected within the 2000 targets. However, the false discovery rate (FDR) exceeded 50 % for results created by limma+voom and edgeR.

The bad performance on count data, which show only nominal difference within their differentially abundant features, is continued even for big sample sizes. Whereas, edgeR and limma+voom are able to detect a small fraction of the DA features, for data with smaller group sizes and number of replicates at high confidence (low FDR), the number of not detected truly DA abundant features still remains higher than 90 %. For edgeR and limma+voom the number of false negatives (FN) decreases with increasing library size and samples per group, independently from the number of replicates. Furthermore, for both methods, it was shown that almost all truly DA features had been detected, for effect sizes bigger than 1.25, at low rates of false positives (FP). In contrast, the number of DA features by Metastats increases for bigger library sizes, but this is linear to the number of FPs. For almost all results obtained by Metastats the false discovery rate was 50 % or higher.

Interestingly, for small sample sizes over all simulated count data scenarios, Metastats showed the best results, with the lowest average FDR. With increasing sample sizes, this benefit diminishes. With the comparison of the average FDR for all three methods, and bigger library sizes, Metastas was outperformed by edgeR, as well as by limma+voom. Further comparison of edgeR and limma+voom revealed that limma+voom performs better than edgeR as library size increases.

The analysis of the real targeted amplicon sequencing datasets, originating from the GI mouse, as well as from the BAL study, showed that very sparse data cannot be properly transformed by the voom method and consequently not be processed with limma. Alternatively methods based on other count data distributions such as edgeR, which was evaluated within this thesis, or DESeq2 [200] or metagenomicSeq [145], evaluated by McMurdie and Holmes, can be used as long as the analyzed count data is correctly distributed.

By using different simulation criteria, the evaluation approach revealed, that features with higher effect sizes are more likely to be detected, independently of replicate or sample size. Further consideration within the two approaches showed, that the number of false positives cannot be further reduced by bigger sample size per group for increasing replicate numbers. This was observed for both methods and moderate library sizes.

Conclusively, the evaluation performed within this thesis showed, that both approaches, edgeR and limma+voom, which were originally developed for the analysis of RNA-seq and microarray experiments, are applicable for DA feature detection in microbiome, as well as in metagenome count data, as long as the data distribution conforms with the basic assumptions. Furthermore, correction of Type I errors, using methods such as introduced by Benjamini and Hochberg [171], are absolutely necessary for limiting the number of FPs within the final result. In addition to a more sophisticated multiple testing correction which is offered by both approaches, they allow for the comparison of more groups, as well as for more complex experimental designs than Metastats.

4.5. High-throughput Characterization of Bacterial and Fungal Community Profiles of the BAL Survey

With the analysis of the bronchoalveolar lavage (BAL) survey data the first targeted amplicon sequencing study based on fungal ribosomal DNA using SnoWMAn, was accomplished within this thesis. The characterization of fungal community profiles by molecular techniques is not as well-established as for bacteria and archaea. As a consequence, resources for automated high-throughput classification and characterization are lagging behind [201]. Nevertheless, fungal community profiling is rapidly emerging as the role of the fungal microbiome as a cofactor in health and disease has been underestimated so far [202]. Additionally, current studies showed that the majority of fungi can not be detected by traditional culture approaches [203–205]. Hence, characterization approaches of the *healthy* fungal microbiome, also called *mycobiome* [202], at different body sites [206–210], as well as fungal community profiling within different diseases such as cystic fibroses [203], chronic obstructive pulmonary disease (COPD) [209], or inflammatory bowel disease (IBD) [207], become more and more the focus of attention.

In general *Candida sp.* are carried by almost all human without causing disease [211]. They have been identified together with other fungal species in the oral cavity of healthy humans [212]. In addition, previous investigations on BAL samples, taken from lung transplantation patients, also revealed high portions of *Candida sp.* compared to healthy individuals [213]. This was also observed in patients who were admitted to or developed pneumonia at the intensive care unit (ICU) [214].

Currently, the biggest resource for ITS sequences is the UNITE [82] system for DNA-based fungal species circumscriptions. By using their reference sequence collection and annotations, a reference database for SnoWMAn's BLAT pipeline was implemented and tested within this thesis. Subsequently, this newly introduced resource was used for determining the fungal community profile of the BAL study samples. In addition to the fungal community profile, DNA from the same samples was used for investigations on the bacterial community composition.

Not just classification resources are still under development, but also sequencing preparation techniques are struggling with the protocols of universal amplification of the ITS marker gene. Not surprisingly, about 17 % of the samples showed missing ITS amplicons during PCR, and therefore had to be excluded from downstream analysis and evaluation. Furthermore, the comparison of the libraries sizes between the ITS and the 16S samples revealed that, although sequencing of the ITS samples results in almost twice as many reads as the 16S sequencing approach, variance in library size is much higher within the ITS data.

The phylogenetic analysis of the ITS samples showed that for all samples which received antibiotic treatment apart from the control group who did not show any clinical, radiological, or laboratory evidence for an infectious disease at the sampling time point, fungi of genus *Candida* as the most abundant feature, as shown in Sec. 3.8.3. This observation reaffirms the results of Bousbia *et al.* [214]. The overall comparison of sample composition resulted in a high variability within samples of the same group. This was also confirmed by the PCA, seen in Fig. 3.20, as no clear separation according to the main groups was possible.

The bacterial community profile of the control group was dominated mainly by *Bacteroides*. For almost half of the control group samples, considerable amounts of *Streptococcus* were also found. Interestingly, from the relative sequence distribution it seems that high amounts of *Streptococcus* decrease or even eliminate the *Bacteroides* population, which was previously observed within the control group. This high amount of genus *Streptococcus* was also found for almost all samples of group 2, with no or very sparse population of *Bacteroides*. Group 3 was identified as the groups with the most varying community profile. About half of the samples were dominated again by *Streptococcus*, whereas no common prevalent genus or shared pattern was identified within the other half.

By the comparison of different α-diversity scores of groups 1 and 2, it was nicely shown that antibiotic treatment reduces bacterial richness. Interestingly, the number of distinct bacterial species, the obtained number of OTUs, was drastically increased between the patients of the control and the ICU group. This might be based on the one hand on some kind of disease or on the other hand on the medical treatment at the ICU.

4.5.1. DA abundant feature analysis of BAL community profiles

The determined community profiles for bacteria, as well as for fungi were tested for differential abundant features using edgeR. The first evaluation by limma+voom confirmed that not properly distributed count data causes inconclusive results, even by transformation using the voom function. As a consequence edgeR was used for community profiles evaluation at different taxonomic levels, from phylum down to the OTU (species) level. First, the profiles of the 5 main groups were tested. At the genus level a total of 46 community features were tested as differentially abundant within all different group comparisons. *Candida* was reaffirmed as most the *"up regulated"* feature (logFC > 1e^{-}11) for all groups compared to the control group, independently of antibiotic treatment. *Penicillium* was detected as most the *"down regulated"* genus level feature (logFC > 1e^{-}10) within samples of group 1A compared to samples of both conditions of group 2.

The comparison of genus level features, within the different types of group 3B, revealed that almost 50 % of the differentially abundant features were regulated in the same direction, when NAP and VAP were compared to ASP. Hence, it is likely that these features are triggered by factors of medical treatment at the hospital. Interestingly, almost all features which were detected as differentially abundant between ASP and NAP or ASP and VAP were shown to be either *up-* or *down regulated* between NAP and VAP, which might be another indication for the importance of the medical treatment on the alterations of the community profile.

39 bacterial features were tested as differentiability abundant within all main treatment groups. *Neisseria* and *Haemophilus* revealed to be the most *"down regulated"* features at the genus level (logFC > 1e^{-}8.5), tested in group 2B compared to group 3B. As both of this groups are treated at the ICU and with antibiotics this *down regulation* might be triggered by the pneumonia. Interestingly, *Neisseria* was also

shown to be *"up regulated"* in samples without pneumonia but treated with antibiotics. Additionally, a similar effect was observed for the genus *Mycoplasma*. It was tested as *up regulated* feature in group 3 samples compared to, group 1A, 1B, as well as to 2A. This observation enforce that these alterations within the community profile are triggered by the disease, pneumonia, an not by the medical treatment. Further investigations on comparisons within group 1 and 2, towards the effect of antibiotic treatment, resulted for both groups in an *up regulation* of *Neisseria* and a *down regulation* of *Fusobacterium*.

Further considerations within subgroups of group 3B revealed, that features which are not regulated the same direction between NAP and VAP compared to ASP, are mainly contrary regulated between NAP and VAP. Hence, this behavior might point towards an association between these features.

A very interesting follow-up analysis would be the combined analysis of the bacterial and fungal community compositions. Surveys which comprise the bacterial, as well as the fungal community profile are not very common, but not to say unique. This kind of association or dissociation would allow to explain whether bacteria, or fungi are occupying the habitat of the other.

4.5.2. Comparison of traditional BAL and tracheal secretion culture results with high-throughput characterization and classification

For the targeted amplicon sequencing data originating from the BAL study which was analyzed within this thesis, also traditional BAL and tracheal secretion culture tests were available. This allowed for the cross comparison of the obtained results from both analysis approaches.

The design of the comparison procedure focused on the microbial strains identified previously by the traditional cultures. Hence, other microbial strains contained within the samples were neglected. Independent from a positive or a negative test result all samples were used within this evaluation approach.

The comparison of the results towards *Candida spp.* colonization revealed that for more than 50 % of the samples analyzed by molecular techniques, *Candida albicans* was identified as the most abundant community species. Additionally, a smaller amount of *Candida dubliniensis* was identified for almost all these samples. In contrast, only a quarter was tested positive for *Candida spp.* colonization using traditional BAL and tracheal secretion cultures. However, all culture results which were positively tested for *Candida spp.* colonization were confirmed by the molecular characterization approach.

Further considerations, especially of the samples which showed a *negative* culture result, revealed that for most of them a significant amount of at least one of the strains in question was identified.

The comparison of bacterial strains detected by traditional BAL and tracheal secretion culture with the taxonomic classification result revealed that for the vast majority of samples, BAL and tracheal secretion culture results were confirmed by molecular identification. But, for some of the samples, other strains than identified by culture dependent techniques were detected as well. This could be explained on the one hand by some kind of specificity, for some of the microbial strains in questions, of the used media or on the other hand by the dominance of one community species which inhibits grows of others. However, it is also likely that a particular strain or the mixture of strains generally hampers the cultivation approach.

The comparisons above were performed on a present/absent level, it would have been more informative also to compare species abundance between both techniques. Unfortunately, there is no base line study or ratio available to directly compare count values with the corresponding bacterial count. Therefore, an extra targeted amplicon sequencing survey and related culture experiment, with controlled bacterial strains, on the same samples, would be necessary to determine the relation between these two quantitative numbers.

With this kind of comparison of bacterial strain identification by either molecular techniques or by traditional culture dependent methods, it was shown that molecular techniques coincide with traditional approaches, but allow for a more complete community profile of the investigated habitat/sample.

Conclusively, this targeted amplicon sequencing survey helped to establish microbial community studies of fungi using SnoWMAn, as well as supported deeper insights into the *Candida spp.* colonization of the lower respiratory tract in humans, under different conditions of medical treatment. In addition, the combination of the fungal with the bacterial community profiles demonstrates a new direction for future analysis.

4.6. High-throughput Characterization of the Bacterial Community Profiles of the Phospholipid Survey in Mouse

After the conducted community profile analysis, the data was initially evaluated by the team of Prof. Peter Fickert[78]. The major goal was to identify phylogenetic groups which change between different experimental groups, summarized in Tab. 3.23, in order to get deeper insights into the effects of phospholipids on the gastrointestinal microbiome. These first evaluations revealed that alterations of the GI microbiome caused by phospholipids are more likely to occur in the upper colonic region, such as the ileum. Therefore, further downstream analysis was focused on samples of the source type ileum. In addition, the analysis on the community profiles, showed more similar community structure within samples of the upper intestinal tract, jejunum and ileum, and the lower intestinal tract, caecum and colon. Therefore, samples were additionally combined according to their intestinal region.

During the initial phylogenetic community profile analysis for some of the fecal samples extracted from the ileum, *unclassified* was identified as the most abundant community feature. Further investigation of sequences of these *unclassified* OTUs revealed eukaryotic DNA originating mainly from plant material. This is not unlikely, as for upper intestinal regions, the digestion process is not entirely completed. Thus, it

[78] Division of Gastroenterology and Hepatology, Medical University of Graz, Graz, Austria

was also shown that random amplifications with the used universal bacterial marker gene primers occur in eukaryotic DNA other than human and mouse. To remove this unwanted sequences, the targeted amplicon data set was pre-processed using the novel Decontaminator approach. Thereby, the relative amount of unclassified sequences, as well as the number of OTUs, was drastically reduced. The remaining amount of *unclassified* amplicons was manually evaluated with BLAST. Here it was confirmed that these *unclassified* OTUs originate from uncultured, *unclassified* 16S rRNA.

The direct comparison to DeconSeq reaffirmed that not all sources of contaminating sequences can be removed by the DeconSeq approach without prior knowledge. In particular for the sequencing data which originated from fecal material the amount of filtered sequences differs by server orders of magnitude between the Decontaminator and DeconSeq. This is not surprising as it was shown that the main source of contaminating material in feces originates from plant residues rather than from the mouse host.

The visualization of the different community profiles, by PCA [103], MDS [215] plots, or by β-diversity according to Bray-Curtis [216] (comparisons not shown within this thesis), revealed a high degree of inter group and sample variation. As some kind of clustering of different sample types was observed within the visualization by PCA and MDS plots, a bias according to extraction day or time, person or location were investigated (data is not presented within this thesis). However, with these variables no association of the formed clustered could be established.

4.6.1. DA abundant feature analysis of GI mouse survey community profiles

The community profiles were tested for differential abundant features using edgeR, at different taxonomic levels, from the phylum down to the OTU (species) level.

First, community profiles of samples of the source type ileum were tested. Although sequence distribution charts, as well as PCA plots, did not look very promising, for almost all conditions, DA features were detected (see Tab. 3.27). At the genus level, 17 different features were detected amongst almost all contrasts.

The genus *Akkermansia* was observed as being significantly increased (logFC > 1e^{-}4) in KO mice for both material types (feces and mucosa) under the phospholipid enriched diet, as well as compared to WT mice, which were fed with the enriched diet. The highest increase, of *Akkermansia* was detected within the comparison of WT and KO mice of source type feces and the enriched diet (logFC ∼1e^{-}8).

This might be interesting because bacteria of the genus *Akkermansia* were related to obesity in former studies within the gut microbiome of mice [217]. *Staphylococcus* and *Streptococcus* were detected as highly *up regulated* (logFC 1e^{-}6.5-1e^{-}8) under the enriched diet of KO and WT mice in samples of material feces. Interestingly, strains of genus *Pseudomonas* were drastically *down regulated* (logFC ∼1e^{-}6) in BD mice under normal conditions of material type mucosa, compared to WT, as well as to KO mice. The most *down regulated* genus, *Enterococcus*, was observed within the comparison of WT and KO mice under an enriched diet in material type feces. A similar effect for *Enterococcus* was observed for comparisons of WT mice under a normal diet of material type feces, to BD and KO mice. Hence, this shift in *Enterococcus* abundance seems to be triggered by mice phenotype rather than by diet.

Several genus level features have been detected as differentially abundant between comparisons of samples under different dietetic conditions. Therefore, *Mycoplasma* and the already mentioned *Akkermansia* were the only ones that showed significant changes within both material types. Species of type *Mycoplasma* have been linked to different intestinal cancer types by previous studies [218, 219]. Interestingly, *Mycoplasma* was *down regulated* within comparisons between KO type mice for samples with an enriched diet, whereas, they were *up regulated* in WT mice samples under the same dietetic conditions (logFC ∼1e^{-}5). In addition, a high increase of *Mycoplasma* was observed between KO and BD mice which received a normal Chow diet (logFC ∼1e^{-}7.6).

Generally, effects presumably caused by dietetic treatment seems to affect more likely the community profile of the mucosa than the feces. This might be explained by digestive processes which start much earlier than in the GI tract. Presumably the delivered phospholipids are absorbed or digested before they can be processed to feces.

Second, community profiles of the introduced subgroups small and large bowels were tested for differentially abundant features at different taxonomic levels. By merging samples according to the small and large bowel, the previous observation on the effects on the sample material was confirmed. No significant features were called in contrasts towards the comparison of dietetic composition in samples originating from fecal material. In addition, differentially abundant genus level features were only called within comparisons of the group SI, in respect to the dietetic treatment. For mucosal material, three significantly changing genus level features, *Alistipes, Lactobacillus, and Lactococcus* were shown to be *down regulated* in WT mice under different diets not only within samples of the small bowel, but also of the large bowel. Interestingly, a *down regulation* of bacterial strains of genus *Helicobacter* was observed within the small bowel, for WT, as well as for KO mice, in mucosal tissues under an enriched diet.

Conclusively, the conducted study provided deeper insights into the GI microbiome of different mice types, material sources, and colonic regions under two different dietetic conditions. Although not all alterations within the community profiles, which were observed in the sequence distribution plots, could be confirmed by statistical methods, DA analysis of the obtained community profiles between different conditions revealed interesting community features for further analysis directions.

4.7. Transcriptome Analysis of the two *Campylobacter fetus* subspecies *fetus* and *veneralis*

The analysis of the transcriptome of two *Campylobacter fetus* subspecies by using dRNA-seq analysis, previous observations on promoter structure in *Helicobyter pylori* [92, 220] or related *Campylobacter jejuni* [221] species could be confirmed. In particular, the comparison of the final mappings, based on cDNA libraries treated with terminator exonuclease (TEX+) and on untreated (TEX-) sequencing reads revealed, that TEX+ sequencing enriches primary transcripts and subsequently facilitates automated

TSS identification, based on the coverage information for each position. The TSS identification process itself can be easily adapted and applied on any kind of transcriptome data, obtained by dRNA-seq (TEX+) experiments. The developed R script for TSS detection required piled-up mapping files as input, which have to be generated prior to TSS identification. Any read mapper/aligner can be used to create the initial mapping file. Within this thesis, the CLC Genomics Workbench [159] mapper was chosen according to its accuracy, efficiency, and visualization facilities compared to other popular read mapping tools, such as Bowtie [222], or BWA [223]. Although mapping information has to be manually processed prior to TSS identification, the subsequent categorization, evaluation, and TSS visualization procedures are combined as a single R [134] routine.

The categorization of the TSS showed a considerably lower number of antisense TSS in both *C. fetus* subspecies compared to *H. pylori* [92], which may also be related to TEX+ treatment of the cDNA libraries.

For the promoter motif analysis, the R routine automatically extracts the 50 bps upstream region of each detected TSS. The motif analysis by the MEME [157] webserver showed the characteristic ϵ-proteobacterial promoter signature also for both *C. fetus* subspecies. Moreover, the obtained promoter consensus sequence of both subspecies revealed an extended Pribnow box (tgnTAtaAT) at the -10 position, as promoter motif. Furthermore, it was shown that the -35 motif was replaced by a periodic AT-rich signal upstream of position -14. The observations made within this thesis are consistent with the promoter motif survey of other ϵ-proteobacteria such as *H. pylori* [92]. As a last point, the analysis reaffirmed that the promoter region is 100 % conserved between the two subspecies.

Although the transcriptome analysis workflow for TSS identification and motif analysis was not implemented as straightforward analysis workflow, it is still easy to be applied to other transcriptome data. Parameters such as average read length, coverage cut-off value, or extraction size can be easily customized within the R routine. The main advantage of separating the analysis approach is that users are able to use their favorite read mappers, and do not have to install other software. Furthermore, no additional tools for the final motif search and visualization have to be installed, as they can be used via a web-service.

Conclusively, the bacterial transcriptome was shown as an additional resource for investigations on genome composition, as well as on regulation of virulence. Hence, it should be considered as future perspectives in community profile analysis. The most abundant or the most significant changing OTU obtained by community profiling or DA detection could be further investigated by a subsequent transcriptome analysis. The obtained information on the genome composition and regulation facilitates deeper insights on what or how community changes are triggered and regulated.

5. Conclusion

All main aspects of microbial community analysis have been evaluated, adapted, or extended within this thesis. For the experimental design and the used sequencing approach it was shown, by comparison of a standard and normalized sequencing library, that library normalization affects overall community Richness. Additionally, it allows detection of less abundant community members because they are not ousted by the prevalent groups of the sampling habitat any more.

In the context of the pre-processing step of targeted amplicon sequencing studies, a tool for identification and removal of contaminating sequences was implemented and evaluated within this thesis. It could be shown that random amplifications by the universal marker gene primers are very likely and lead to OTU inflation, similar to sequencing noise, and chimeras. Furthermore, the comparison of different types of pre-processing approaches revealed, that removal of contaminating sequences has the biggest impact on the final number of observed OTUs. Moreover, the Decontaminator allows for partial removal of noisy, as well as of chimeric sequences due to control of percentage identity and percentage query coverage.

For the most important step of a microbial community study, the classification and characterization step, a new resource for fungal community profile characterization was introduce into SnoWMAn's BLAT pipeline. For evaluation of this and other resources for ITS amplicon characterization the first *in silico* ITS mock communities were created. They were subsequently used for quality control of the created BLAT ITS reference DB, as well for the beta version of the RDP ITS classifier. During creation of the mock communities the main obstacle of fungal community analysis was discovered - missing, ambiguous, or incomplete taxonomic descriptions of fungal sequences. The classification of the mock communities revealed that *reference based* OTU picking by SnoWMAn's BLAT pipeline leads to more confident taxonomic classifications as the *de novo* OTU picking approach of the RDP ITS classifier (beta version), although both approaches rest on the same ITS reference sequences collection.

Statistical analysis of the final community profile was extended by the evaluation of different methods and approaches for the detection of differentially expressed features in microarray experiments, as well as in community profile data. Within the evaluation approach it was demonstrated that edgeR and limm+voom can also be used for DA feature detection using count data, as long as fundamental assumptions such as distribution of the data are valid. In addition, the comparison of the three methods using simulated count data emphasizes that methods which are based upon linear modeling approaches and support more complicated experimental designs, such as edgeR and limma+voom are more suitable for statistical testing of differentially abundant features.

The application of the obtained knowledge, from this evaluation, to the generated feature tables of currently analyzed microbiome data sets demonstrated, that DA feature detection provides a valuable impact to the statistical analysis framework within the community profile analysis.

Finally, the transcriptome analysis of the two *Campylobacter fetus* subspecies showed future perspectives for the community profile analysis. The DA feature detection could identify the most interesting features, possible candidates, for the subsequent transcriptome analysis. This would support to get deeper insights into what regulates a certain species and as a consequence the community profile.

Conclusively, with the direct application of the gained knowledge on data, from current targeted amplicon sequencing surveys, the need for the developed tools, created resources, and evaluated methods and approaches have been shown by practical examples.

Bibliography

[1] Sanger F, Nicklen S: **DNA sequencing with chain-terminating inhibitors**. *Proc Nati Acad Sci* 1977, **74**(12):5463–5467.

[2] Maxam A, Gilbert W: **A new method for sequencing DNA**. *Proc Natl Acad Sci USA* 1977, **74**(2):560–564.

[3] Pareek CS, Smoczynski R, Tretyn A: **Sequencing technologies and genome sequencing**. *J Appl Genet* 2011, **52**(4):413–435.

[4] Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotechnol* 2008, **26**(10):1135–1145.

[5] Moore GE: **Cramming more components onto integrated circuits**. *Electronics* 1965, **38**(8):114–117.

[6] Ansorge WJ: **Next-generation DNA sequencing techniques**. *N Biotechnol* 2009, **25**(4):195–203.

[7] Voelkerding KV, Dames SA, Durtschi JD: **Next-Generation Sequencing: From Basic Research to Diagnostics**. *Clin Chem* 2009, **55**(4):641–658.

[8] Kyrpides NC: **Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream**. *Nat Biotechnol* 2009, **27**(7):627–632.

[9] Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative Metagenomics of Microbial Communities**. *Science* 2005, **308**(5721):554–557.

[10] Xu J: **Microbial Ecology in the Age of Metagenomics**. In *Handbook of molecular microbial ecology*. Edited by de Bruijn FJ. Wiley-Blackwell Hoboken, NJ; 2011:113–128.

[11] Lau JA, Lennon JT: **Rapid responses of soil microorganisms improve plant fitness in novel environments**. *Proc Natl Acad Sci USA* 2012, **109**(35):doi: 10.1073/p-nas.1202319109.

[12] Sbordone L, Bortolaia C: **Oral microbial biofilms and plaque-related diseases: microbial communities and their role in the shift from oral health to disease**. *Clin Oral Investig* 2003, **7**(4):181–188.

[13] Narihiro T, Sekiguchi Y: **Microbial communities in anaerobic digestion processes for waste and wastewater treatment: a microbiological update**. *Curr Opin Biotech* 2007, **18**(3):273 – 278.

[14] Pankhurst CE, Ophel-Keller K, Doube BM, Gupta V: **Biodiversity of soil microbial communities in agricultural systems**. *Biodivers Conserv* 1996, **5**(2):197–209.

[15] Cho I, Blaser MJ: **The human microbiome: at the interface of health and disease**. *Nat Rev Genet* 2012, **13**(4):260–270.

[16] The NIH HMP Working Group,, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M: **The NIH Human Microbiome Project**. *Genome Res* 2009, **19**(12):2317–2323.

[17] Avila M, Ojcius DM, Yilmaz O: **The oral microbiota: living with a permanent guest**. *DNA Cell Biol* 2009, **28**(8):405–411.

[18] Le CE, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jorgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clement K, Dore J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker JD, Raes J, Hansen T, Bork P, Wang J, Ehrlich SD, Pedersen O: **Richness of human gut microbiome correlates with metabolic markers**. *Nature* 2013, **500**(7464):541–546.

[19] Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le CE, Almeida M, Quinquis B, Levenez F, Galleron N, Gougis S, Rizkalla S, Batto JM, Renault P, Dore J, Zucker JD, Clement K, Ehrlich SD: **Dietary intervention impact on gut microbial gene richness**. *Nature* 2013, **500**(7464):585–588.

[20] Wade WG: **The oral microbiome in health and disease**. *Pharmacol Res* 2013, **69**(1):137–143.

[21] Christensen GJ, Bruggemann H: **Bacterial skin commensals and their role as host guardians**. *Benef Microbes* 2014, **5**(201):201–15.

[22] Streit WR, Schmitz RA: **Metagenomics–the key to the uncultured microbes**. *Curr Opin Microbiol* 2004, **7**(5):492–498.

[23] Koch C, Hoiby N: **Pathogenesis of cystic fibrosis**. *Lancet* 1993, **341**(8852):1065–9.

[24] Jha BJ, Dey S, Tamang MD, Joshy ME, Shivananda PG, Brahmadatan KN: **Characterization of Candida species isolated from cases of lower respiratory tract infection**. *Kathmandu Univ Med J* 2006, **4**(3):290–294.

[25] Kauffman CA: **Diagnosis and management of fungal urinary tract infection**. *Infect Dis Clin North Am* 2014, **28**(1):61–74.

[26] Grice EA, Segre JA: **The skin microbiome**. *Nat Rev Micro* 2011, **9**(4):244–253.

[27] Eras P, Goldstein MJ, Sherlock P: **Candida Infection of the Gastrointestinal Tract**. *Medicine* 1972, **51**(5):367–380.

[28] Richards MJ, Edwards JR, Culver DH, Gaynes RP: **Nosocomial infections in combined medical-surgical intensive care units in the United States**. *Infect Control Hosp Epidemiol* 2000, **21**(8):510–515.

[29] Meersseman W, Lagrou K, Spriet I, Maertens J, Verbeken E, Peetermans W, Wijngaerden E: **Significance of the isolation of Candida species from airway samples in critically ill patients: a prospective, autopsy study**. *Intens Care Med* 2009, **35**(9):1526–1531.

[30] Kerwat K, Rolfes C, Wulf H: **Fungal Infections in the Intensive Care Unit**. *Anasthesiol Intensivmed Notfallmed Schmerzther* 2011, **46**(11-12):744–745.

[31] Guarner J, Brandt ME: **Histopathologic diagnosis of fungal infections in the 21st century**. *Clin Microbiol Rev* 2011, **24**(2):247–280.

[32] Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome**. *Science* 2006, **312**(5778):1355–1359.

[33] Greiner T, Backhed F: **Effects of the gut microbiota on obesity and glucose homeostasis**. *Trends Endocrinol Metab* 2011, **22**(4):117–123.

[34] Musso G, Gambino R, Cassader M: **Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes**. *Annu Rev Med* 2011, **62**:361–380.

[35] Kinross JM, Darzi AW, Nicholson JK: **Gut microbiome-host interactions in health and disease**. *Genome Med* 2011, **3**(3):14.

[36] Mashaghi S, Jadidi T, Koenderink G, Mashaghi A: **Lipid Nanotechnology**. *Int J Mol Sci* 2013, **14**(2):4242–4282.

[37] von Heijne G, Rees D: **Membranes: reading between the lines**. *Curr Opin Struc Biol* 2008, **18**(4):403 – 405.

[38] Kienesberger S, Sprenger H, Wolfgruber S, Halwachs B, Thallinger GG, Perez-Perez GI, Blaser MJ, Zechner EL, Gorkiewicz G: **Comparative Genome Analysis of *Campylobacter fetus* Subspecies Revealed Horizontally Acquired Genetic Elements Important for Virulence and Niche Specificity**. *PLoS ONE* 2014, **9**(1):e85491.

[39] Wagenaar JA, van Bergen MAP, Blaser MJ, Tauxe RV, Newell DG, van Putten JPM: ***Campylobacter fetus* Infections in Humans: Exposure and Disease**. *Clin Infect Dis* 2014, doi: 10.1093/cid/ciu085.

[40] Hoffer MA: **Bovine Campylobacteriosis: A Review**. *Can Vet J* 81, **22**(11):327–330.

[41] Whipps JM: **Microbial interactions and biocontrol in the rhizosphere**. *J Exp Bot* 2001, **52**(suppl 1):487–511.

[42] Hooper LV, Gordon JI: **Commensal Host-Bacterial Relationships in the Gut**. *Science* 2001, **292**(5519):1115–1118.

[43] Hooper LV, Littman DR, Macpherson AJ: **Interactions Between the Microbiota and the Immune System**. *Science* 2012, **336**(6086):1268–1273.

[44] Smith KP, Goodman RM: **HOST VARIATION FOR INTERACTIONS WITH BENEFICIAL PLANT-ASSOCIATED MICROBES**. *Annu Rev Phytopathol* 1999, **37**:473–491.

[45] Hooper LV, Midtvedt T, Gordon JI: **How host-microbial interactions shape the nutrient environment of the mammalian intestine**. *Annu Rev Nutr* 2002, **22**:283–307.

[46] Rubin E: **Genomics of cellulosic biofuels**. *Nature* 2008, **454**(7206):841–845.

[47] Luetz S, Giver L, Lalonde J: **Engineered enzymes for chemical production**. *Biotechnol Bioeng* 2008, **101**(4):647–653.

[48] Raghoebarsing AA, Smolders AJ, Schmid MC, Rijpstra WI, Wolters-Arts M, Derksen J, Jetten MS, Schouten S, Sinninghe Damste JS, Lamers LP, Roelofs JG, Op den Camp HJ, Strous M: **Methanotrophic symbionts provide carbon for photosynthesis in peat bogs**. *Nature* 2005, **436**(7054):1153–1156.

[49] Bragina A, Berg C, Mueller H, Moser D, Berg G: **Insights into functional bacterial diversity and its effects on Alpine bog ecosystem functioning**. *Sci Rep* 2013, **3**:1955.

[50] Opelt K, Chobot V, Hadacek F, Schonmann S, Eberl L, Berg G: **Investigations of the structure and function of bacterial communities associated with Sphagnum mosses**. *Environ Microbiol* 2007, **9**(11):2795–2809.

[51] Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, Udall JA, Wilcox ER, Crandall KA: **Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics**. *Genome Biol Evol* 2011, **3**:1312–1323.

[52] Council National Research: *The New Science of Metagenomics* Revealing the Secrets of Our Microbial Planet 1 edn. Washington, DC, USA: The National Academies Press. 2007.

[53] Gilbert J, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J, Ley R, Fierer N, Field D, Kyrpides N, Glöckner F, Klenk HP, Wommack K, Glass E, Docherty K, Gallery R, Stevens R, Knight R: **The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010.** *Stand Genomic Sci* 2010, **3**(3):249–253.

[54] Moissl C, Osman S, La Duc MT, Dekas A, Brodie E, DeSantis T, Venkateswaran K: **Molecular bacterial community analysis of clean rooms where spacecraft are assembled**. *FEMS Microbiol Ecol* 2007, **61**(3):509–521.

[55] Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R: **Experimental and analytical tools for studying the human microbiome**. *Nat Rev Genet* 2012, **13**(1):47–58.

[56] Morgan XC, Huttenhower C: **Chapter 12: Human Microbiome Analysis**. *PLoS Comput Biol* 2012, **8**(12):e1002808.

[57] Patel JB: **16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory**. *Mol Diagn* 2001, **6**(4):313–321.

[58] Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W: **Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi**. *Proc Natl Acad Sci USA* 2012, **109**(16):6241–6246.

[59] Caetano-Anolles G: **Tracing the evolution of RNA structure in ribosomes**. *Nucleic Acids Res* 2002, **30**(11):2575–2587.

[60] Van de Peer Y, Chapelle S, De Wachter R: **A Quantitative Map of Nucleotide Substitution Rates in Bacterial rRNA**. *Nucleic Acids Res* 1996, **24**(17):3381–3391.

[61] Guo F, Ju F, Cai L, Zhang T: **Taxonomic Precision of Different Hypervariable Regions of 16S rRNA Gene and Annotation Methods for Functional Bacterial Groups in Biological Wastewater Treatment**. *PLoS ONE* 2013, **8**(10):e76185.

[62] Bodilis J, Nsigue-Meilo S, Besaury L, Quillet L: **Variable Copy Number, Intra-Genomic Heterogeneities and Lateral Transfers of the 16S rRNA Gene in *Pseudomonas***. *PLoS ONE* 2012, **7**(4):e35647.

[63] Sonnenberg R, Nolte A, Tautz D: **An evaluation of LSU rDNA D1-D2 sequences for their use in species identification**. *Front Zool* 2007, **4**(1):6.

[64] Liu KL, Porras-Alfaro A, Kuske CR, Eichorst SA, Xie G: **Accurate, Rapid Taxonomic Classification of Fungal Large-Subunit rRNA Genes**. *Appl Environ Microbiol* 2012, **78**(5):1523–1533.

[65] Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, van de Peer Y, Vandamme P, Thompson FL, Swings J: **Re-evaluating prokaryotic species**. *Nat Rev Micro* 2012, **2**(9):733–739.

[66] Halwachs B, Gorkiewicz G, Thallinger GG: **High-Throughput Characterization and Comparison of Microbial Communities**. In *Computational Medicine*. Edited by Trajanoski Z. Springer Vienna; 2012:37–57.

[67] Halwachs B, Höftberger J, Stocker G, Snajder R, Gorkiewicz G, Thallinger GG: **High-Throughput Characterization and Comparison of Microbial Communities**. *Biomed Tech (Berl)* 2013, doi: 10.1515/bmt–2013–4312.

[68] Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, Arze C, White JR, White O, Fricke WF: **CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing**. *BMC Bioinformatics* 2011, **12**:356.

[69] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF: **Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities**. *Appl Environ Microbiol* 2009, **75**(23):7537–7541.

[70] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder

J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data**. *Nat Meth* 2010, **7**(5):335–336.

[71] Eisen JA: **Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes**. *PLoS Biol* 2007, **5**(3):e82.

[72] Thomas T, Gilbert J, Meyer F: **Metagenomics-a guide from sampling to data analysis**. *Microb Inform Exp* 2012, **2**(1):3.

[73] Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 2000, **28**(1):27–30.

[74] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA: **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes**. *BMC Bioinformatics* 2008, **9**:386.

[75] Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**(3):377–386.

[76] Moran MA: **Metatranscriptomics: Eavesdropping on Complex Microbial Communities**. *Microbe* 2009, **4**(7):329–335.

[77] Warnecke F, Hess M: **A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts**. *J Biotechnol* 2009, **142**(1):91–95.

[78] Borries A, Vogel J, Sharma CM: **Differential RNA sequencing (dRNA-seq): Deep-sequencing-based analysis of primary transcriptomes**. In *Tag-Based Next Generation Sequencing*. Edited by Habers M, Kahl G. Wiley-VCH Verlag GmbH & Co. KGaA; 2011:109–121.

[79] Al-Haggar MMS, Khair-Allaha BA, Islam MM, Mohamed ASA: **Bioinformatics in High Throughput Sequencing: Application in Evolving Genetic Diseases**. *J Data Mining Genomics Proteomics* 2013, **4**:131.

[80] Schmieder R, Edwards R: **Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets**. *PLoS ONE* 2011, **6**(3):e17288.

[81] Nilsson RH, Hyde K, Pawłowska J, Ryberg M, Tedersoo L, Aas A, Alias S, Alves A, Anderson C, Antonelli A, Arnold A, Bahnmann B, Bahram M, Bengtsson-Palme J, Berlin A, Branco S, Chomnunti P, Dissanayake A, Drenkhan R, Friberg H, Frøslev T, Halwachs B, Hartmann M, Henricot B, Jayawardena R, Jumpponen A, Kauserud H, Koskela S, Kulik T, Liimatainen K, Lindahl B, Lindner D, Liu JK, Maharachchikumbura S, Manamgoda D, Martinsson S, Neves M, Niskanen

T, Nylinder S, Pereira O, Pinho D, Porter T, Queloz V, Riit T, Sánchez-García M, de Sousa F, Stefańczyk E, Tadych M, Takamatsu S, Tian Q, Udayanga D, Unterseher M, Wang Z, Wikee S, Yan J, Larsson E, Larsson KH, Kõljalg U, Abarenkov K: **Improving ITS sequence data for identification of plant pathogenic fungi**. *Fungal Diversity* 2014, **Special Issue**:1–9.

[82] Koljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Duenas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lücking R, Martin MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, Peterson M, Poldmaa K, Saag L, Saar I, Schüßler A, Scott JA, Senes C, Smith ME, Suija A, Taylor DL, Telleria MT, Weiss M, Larsson KH: **Towards a unified paradigm for sequence-based identification of fungi**. *Mol Ecol* 2013, **22**(21):5271–5277.

[83] Porras-Alfaro A, Liu KL, Kuske CR, Xie G: **From genus to phylum: large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition**. *Appl Environ Microbiol* 2014, **80**(3):829–840.

[84] McMurdie PJ, Holmes S: **Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible**. *PLoS Comput Biol* 2014, **10**(4):e1003531.

[85] Liapikou A, Ferrer M, Polverino E, Balasso V, Esperatti M, Piñer R, Mensa J, Luque N, Ewig S, Menendez R, Niederman MS, Torres A, de Pneumologiais S: **Severe Community-Acquired Pneumonia: Validation of the Infectious Diseases Society of America/American Thoracic Society Guidelines to Predict an Intensive Care Unit Admission**. *Clin Infect Dis* 2009, **48**(4):377–385.

[86] Chastre J, Fagon JY: **Ventilator-associated pneumonia**. *Am J Respir Crit Care Med* 2002, **165**(7):867–903.

[87] Kienesberger S, Gorkiewicz G, Joainig MM, Scheicher SR, Leitner E, Zechner EL: **Development of Experimental Genetic Tools for Campylobacter fetus**. *Appl Environ Microbiol* 2007, **73**(14):4619–4630.

[88] Gorkiewicz G, Kienesberger S, Schober C, Scheicher SR, Gülly C, Zechner R, Zechner EL: **A Genomic Island Defines Subspecies-Specific Virulence Features of the Host-Adapted Pathogen Campylobacter fetus subsp. venerealis**. *J Bacteriol* 2010, **192**(2):502–517.

[89] Hum S, Quinn K, Brunner J, On SL: **Evaluation of a PCR assay for identification and differentiation of Campylobacter fetus subspecies**. *Aust Vet J* 1997, **75**(11):827–831.

[90] Wagenaar JA, van Bergen MAP, Newell DG, Grogono-Thomas R, Duim B: **Comparative Study Using Amplified Fragment Length Polymorphism Fingerprinting, PCR Genotyping, and Phenotyping To Differentiate *Campylobacter fetus* Strains Isolated from Animals**. *J Clin Microbiol* 2001, **39**(6):2283–2286.

[91] On S, Harrington C: **Evaluation of numerical analysis of PFGE-DNA profiles for differentiating Campylobacter fetus subspecies by comparison with phenotypic, PCR and 16S rDNA sequencing methods**. *J Appl Microbiol* 2001, **90**(2):285–293.

[92] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeisz S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J: **The primary transcriptome of the major human pathogen *Helicobacter pylori***. *Nature* 2010, **464**(7286):250–255.

[93] Smit JJ, Schinkel AH, Oude Elferink RP, Groen AK, Wagenaar E, Van DL, Mol CA, Ottenhoff R, van der Lugt NM, van Roon MA: **Homozygous disruption of the murine mdr2 P-glycoprotein gene leads to a complete absence of phospholipid from bile and to liver disease**. *Cell* 1993, **75**(3):451–462.

[94] Fickert P, Zollner G, Fuchsbichler A, Stumptner C, Weiglein AH, Lammert F, Marschall HU, Tsybrovskyy O, Zatloukal K, Denk H, Trauner M: **Ursodeoxycholic acid aggravates bile infarcts in bile duct-ligated and Mdr2 knockout mice via disruption of cholangioles**. *Gastroenterology* 2002, **123**(4):1238–1251.

[95] Lewis SJ, Heaton KW: **Stool form scale as a useful guide to intestinal transit time**. *Scand J Gastroenterol* 1997, **32**(9):920–924.

[96] Gorkiewicz G, Thallinger GG, Trajanoski S, Lackner S, Stocker G, Hinterleitner T, Gülly C, Högenauer C: **Alterations in the Colonic Microbiota in Response to Osmotic Diarrhea**. *PLoS ONE* 2013, **8**(2):e55817.

[97] Kent WJ: **BLAT–the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656–664.

[98] Hamady M, Knight R: **Microbial community profiling for human microbiome projects: Tools, techniques, and challenges**. *Genome Res* 2009, **19**(7):1141–1152.

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

[100] Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy**. *Appl Environ Microbiol* 2007, **73**(16):5261–5267.

[101] Lipman D, Pearson WR: **Rapid and sensitive protein similarity searches**. *Science* 1985, **227**(4693):1435–41.

[102] Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments**. *Bioinformatics* 2009, **25**(10):1335–1337.

[103] Jolliffe, IT: *Principal Component Analysis* 1 edn. New York, NYC, USA: Springer-Verlag New York, Inc. 2002.

[104] Venn J: **On the Diagrammatic and Mechanical Representation of Propositions and Reasonings**. *Philos Mag* 1880, **9**(59):1–18.

[105] DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB**. *Appl Environ Microbiol* 2006, **72**(7):5069–5072.

[106] Pace NR: **A Molecular View of Microbial Diversity and the Biosphere**. *Science* 1997, **276**(5313):734–740.

[107] Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar , Buchner A, Lai T, Steppi S, Jobb G, Foerster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, Koenig A, Liss T, Luessmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH: **ARB: a software environment for sequence data**. *Nucleic Acids Res* 2004, **32**(4):1363–1371.

[108] Hugenholtz P: **Exploring prokaryotic diversity in the genomic era**. *Genome Biol* 2002, **3**(2):1–8.

[109] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2011, **39**(Database issue):D32–D37.

[110] Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM: **The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis**. *Nucleic Acids Res* 2005, **33**(Database issue):D294–D296.

[111] Wilke A, Harrison T, Wilkening J, Field D, Glass EM, Kyrpides N, Mavrommatis K, Meyer F: **The M5nr: a novel non-redundant database containing protein**

sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* 2012, **13**:141.

[112] Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2005, **33**(Database issue):D501–D504.

[113] Reference Genome Group of the Gene Ontology Consortium: **The Gene Ontology's Reference Genome Project: A Unified Framework for Functional Annotation across Species**. *PLoS Comput Biol* 2009, **5**(7):e1000431.

[114] Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, Otillar R, Poliakov A, Ratnere I, Riley R, Smirnova T, Rokhsar D, Dubchak I: **The Genome Portal of the Department of Energy Joint Genome Institute**. *Nucleic Acids Res* 2012, **40**(Databse issue):D26–D32.

[115] Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V: **The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes**. *Nucleic Acids Res* 2005, **33**(17):5691–5702.

[116] Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie JJ, Hamelius J, Hance M, Huntington KI, Jukneliene D, Koziski J, Mackasmiel L, Mane SP, Nguyen V, Purkayastha A, Shallom J, Yu G, Guo Y, Gabbard J, Hix D, Azad AF, Baker SC, Boyle SM, Khudyakov Y, Meng XJ, Rupprecht C, Vinje J, Crasta OR, Czar MJ, Dickerman A, Eckart JD, Kenyon R, Will R, Setubal JC, Sobral BWS: **PATRIC: The VBI PathoSystems Resource Integration Center**. *Nucleic Acids Res* 2007, **35**(Database issue):D401–D406.

[117] Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P: **eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges**. *Nucleic Acids Research* 2012, **40**(D1):D284–D289.

[118] Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database* 2011, **2011**(9):bar009.

[119] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403–410.

[120] Karsch-Mizrachi I, Nakamura Y, Cochrane G: **The International Nucleotide Sequence Database Collaboration**. *Nucleic Acids Res* 2011, **40**(D1):D33–D37.

[121] Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ: **Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators**. *Proceedings of the National Academy of Sciences* 2006, **103**(52):19635–19640.

[122] Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31–46.

[123] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376–380.

[124] Nyren P: **The history of pyrosequencing**. *Methods Mol Biol* 2007, **373**:1–14.

[125] Ronaghi M, Uhlen M, Nyren P: **A sequencing method based on real-time pyrophosphate**. *Science* 1998, **281**(5375):363,365.

[126] Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P: **Real-Time DNA Sequencing Using Detection of Pyrophosphate Release**. *Anal Biochem* 1996, **242**(1):84–89.

[127] Voelkerding KV, Dames SA, Durtschi JD: **Next-Generation Sequencing: From Basic Research to Diagnostics**. *Clin Chem* 2009, **55**(4):641–658.

[128] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira CheethamR, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan

PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil CooleyR, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes FajardoKV, Scott FureyW, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw JonesTA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling NgB, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris PinkardD, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva RodriguezA, Roe PM, Rogers J, Rogert BacigalupoMC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest SohnaSJ, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ: **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature* 2008, **456**(7218):53–59.

[129] Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R: **Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms**. *ISME J* 2012, **6**(8):1621–1624.

[130] Arnold K, Gosling J, Holmes D: *The Java programming language* 4 edn. Upper Saddle River, NJ, USA: Addison-Wesley. 2008.

[131] Programming Language Popularity Official Homepage: Programming Language Popularity. http://www.langpop.com/. 2014.

[132] Prlic A, Yates A, Bliven SE, Rose PW, Jacobsen J, Troshin PV, Chapman M, Gao J, Koh CH, Foisy S, Holland R, Rimsa G, Heuer ML, Brandstatter-Muller H, Bourne PE, Willis S: **BioJava: an open-source framework for bioinformatics in 2012**. *Bioinformatics* 2012, **28**(20):2693–2695.

[133] JFreeChart Official Homepage: JFreeChart. http://www.jfree.org/jfreechart/. 2014.

[134] R Development Core Team: *R*: A Language and Environment for Statistical Computing Vienna, Austria: R Foundation for Statistical Computing. 2008.

[135] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.

[136] Bhagwat M, Young L, Robison RR: **Using BLAT to find sequence similarity in closely related genomes**. *Curr Protoc Bioinformatics* 2012, **Chapter 10**(Unit10.8):doi: 10.1002/0471250953.bi1008s37.

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

[138] Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R: **UCHIME improves sensitivity and speed of chimera detection**. *Bioinformatics* 2011, **27**(16):2194–2200.

[139] Lauren B, Stone G, Imelfort M, Hugenholtz P, Tyson GW: **Fast, accurate error-correction of amplicon pyrosequences using Acacia**. *Nat Meth* 2012, **9**(5):425–426.

[140] Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sanchez-Garcia M, Ebersberger I, de Sousa F, Amend A, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson K, Vik U, Veldre V, Nilsson RH: **Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data**. *Methods Ecology Evol* 2013, **4**(10):914–919.

[141] Baum LE, Petrie T: **Statistical Inference for Probabilistic Functions of Finite State Markov Chains**. *The Annals of Mathematical Statistics* 1966, **37**(6):1554–1563.

[142] Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessiere J, Taberlet P, Pompanon F: **An in silico approach for the evaluation of DNA barcodes**. *BMC Genomics* 2010, **11**:434.

[143] Wu S, Manber U: **Agrep - a fast approximate pattern-matching tool**. In *Proceedings of USENIX Technical Conference: USENIX Winter 1992 Technical Conference: 20-24 January, 1992*. Edited by Association U: USENIX; 1992:153–162 San Francisco, CA, USA.

[144] White JR, Nagarajan N, Pop M: **Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples**. *PLoS Comput Biol* 2009, **5**(4):e1000352.

[145] Paulson JN, Stine OC, Bravo HC, Pop M: **Differential abundance analysis for microbial marker-gene surveys**. *Nat Meth* 2013, **10**(12):1200–1202.

[146] Ruijter JM, Van Kampen AH, Baas F: **Statistical evaluation of SAGE libraries: consequences for experimental design**. *Physiol Genomics* 2002, **11**(2):37–44.

[147] Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data**. *Bioinformatics* 2002, **18**(11):1454–1461.

[148] Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci USA* 2003, **100**(16):9440–9445.

[149] Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139–140.

[150] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res* 2008, **18**(9):1509–1517.

[151] Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW: **Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model**. *Proc Natl Acad Sci USA* 2008, **105**(51):20179–20184.

[152] Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: **ChIP-seq accurately predicts tissue-specific activity of enhancers**. *Nature* 2009, **457**(7231):854–858.

[153] Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.

[154] Smyth GK: **Limma: linear models for microarray data**. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. Springer New York, USA; 2005:397–420.

[155] Law CW, Chen Y, Shi W, Smyth GK: **Voom: precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome Biol* 2014, **15**(2):R29.

[156] Murtagh F, Contreras P: **Methods of Hierarchical Clustering**. *J Data Mini Know Disc* 2012, **2**(1):86–97.

[157] Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology: 14-17 August 1994* . Edited by Altman R, Brutlag D, Karp P, Lathrop R, Searls D: AAAI Press; 1992:28–36.

[158] Schneider T, Stephens R: **Sequence Logos: A New Way to Display Consensus Sequences**. *Nucleic Acids Res.* 1990, **18**:6097–6100.

[159] CLC Bio AS, Aarhus, Denmark: The CLC Genomics Workbench. http://www.clcbio.com/products/clc-genomics-workbench/. 2014.

[160] CLC Bio AS, Aarhus, Denmark: The CLC Main Workbench. http://www.clcbio.com/products/clc-main-workbench/. 2014.

[161] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078–2079.

[162] Ondov B, Bergman N, Phillippy A: **Interactive metagenomic visualization in a Web browser**. *BMC Bioinformatics* 2011, **12**:385.

[163] Hope ACA: **A simplified Monte Carlo significance test procedure**. *J R Statist Soc B* 1968, **30**(3):289–300.

[164] Massey FrankJ: **The Kolmogorov-Smirnov Test for Goodness of Fit**. *J Am Statist Assoc* 1951, **46**(253):68–78.

[165] Wang Y, Qian PY: **Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies**. *PLoS ONE* 2009, **4**(10):e7401.

[166] Mrozek D, Malysiak-Mrozek B, Siaznik A: **search GenBank: interactive orchestration and ad-hoc choreography of Web services in the exploration of the biomedical resources of the National Center For Biotechnology Information**. *BMC Bioinformatics* 2013, **14**:73.

[167] Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2009, **37**(Special Issue):D5–D15.

[168] Martin K, Rygiewicz P: **Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts**. *BMC Microbiology* 2005, **5**(1):28.

[169] White T, Bruns T, Lee S, Taylor J 1990: *Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics* 315–322: Academic Press San Diego.

[170] Smyth GK, Ritchie M, Thorne N, Wettenhall J, Wei S, Yifang H: **Linear models for microarray data**. 18. Feb. 2014 edn. 2002, http://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf.

[171] Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing**. *J R Statist Soc B* 1995, **57**(1):289–300.

[172] Robinson M, McCarthy D, Chen Y, Smyth GK: **edger: differential expression analysis of digital gene expression data**. 31. Aug. 2013 edn. 2008, http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf.

[173] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools**. *Nucleic Acids Research* 2013, **41**(D1):D590–D596.

[174] Bullard J, Purdom E, Hansen K, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC Bioinformatics* 2010, **11**(1):94.

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

[176] McCarthy DJ, Chen Y, Smyth GK: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation**. *Nucleic Acids Research* 2012, **40**(10):4288–4297.

[177] Lund SP, Nettleton D, McCarthy DJ, Smyth GK: **Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates**. *Stat Appl Genet Mol Biol* 2012, **11**(5):doi: 10.1515/1544–6115.1826.

[178] Magurran, Anne, E: *Measuring Biological Diversity* 1 edn. Oxford, UK: Blackwell Science Ltd. 2004.

[179] Chao A: **Non-parametric estimation of the number of classes in a population**. *Scand J Stat* 1984, **11**:265–270.

[180] Chao A, Chazdon RL, Colwell RK, Shen TJ: **A new statistical approach for assessing similarity of species composition with incidence and abundance data**. *Ecology Letters* 2005, **8**(2):148–159.

[181] Shannon, CE and Weaver, W: *The mathematical theory of communication* 1 edn. Urbana, IL, USA: University of Illinois Press. 1949.

[182] Chazdon R, Colwell R, Denslow J, Guariguata M: **Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica**. In *Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies*. Edited by Dallmeier F, JA C. Parthenon Publishing; Paris; 1998:285–309.

[183] Kunin V, Engelbrektson A, Ochman H, Hugenholtz P: **Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates**. *Environ Microb* 2010, **12**(1):118–123.

[184] Huws S, Edwards J, Kim E, Scollan ND: **Specificity and sensitivity of eubacterial primers utilized for molecular profiling of bacteria within complex microbial ecosystems**. *J Microbiol Methods* 2007, **70**(3):565–9.

[185] Baker G, Smith J, Cowan D: **Review and re-analysis of domain-specific 16S primers**. *J Microbiol Methods* 2003, **55**(3):541 – 555.

[186] Kumar PS, Brooker MR, Dowd SE, Camerlengo T: **Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing**. *PLoS ONE* 2011, **6**(6):e20956.

[187] Arnold AE, Miadlikowska J, Higgins KL, Sarvate SD, Gugger P, Way A, Hofstetter V, Kauff F, Lutzoni F: **A Phylogenetic Estimation of Trophic Transition Networks for Ascomycetous Fungi: Are Lichens Cradles of Symbiotrophic Fungal Diversification?** *Syst Biol* 2009, **58**(3):283–297.

[188] James TY, Letcher PM, Longcore JE, Mozley-Standridge SE, Porter D, Powell MJ, Griffith GW, Vilgalys R: **A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota)**. *Mycologia* 2006, **98**(6):860–871.

[189] James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung GH, Johnson D, O/'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schuszler A, Longcore JE, O/'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lucking R, Budel B, Geiser

DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R: **Reconstructing the early evolution of Fungi using a six-gene phylogeny**. *Nature* 2006, **443**(7113):818–822.

[190] Weber CF, Vilgalys R, Kuske CR: **Changes in fungal community composition in response to elevated atmospheric CO2 and nitrogen fertilization varies with soil horizon**. *Front Microbiol* 2013, 4(78).

[191] Stackerbrandt E, Goebel BM: **Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology**. *Int J Syst Bacteriol* 1994, **44**(4):846–849.

[192] Blaalid R, Kumar S, Nilsson RH, Abarenkov K, Kirk PM, Kauserud H: **ITS1 versus ITS2 as DNA metabarcodes for fungi**. *Mol Ecol Resour* 2013, **13**(2):218–224.

[193] O'Brien HE, Parrent JL, Jackson JA, Moncalvo JM, Vilgalys R: **Fungal Community Analysis by Large-Scale Sequencing of Environmental Samples**. *Appl Environ Microbiol* 2005, **71**(9):5544–5550.

[194] Morris MH, Smith ME, Rizzo DM, Rejmánek M, Bledsoe CS: **Contrasting ectomycorrhizal fungal communities on the roots of co-occurring oaks (*Quercus spp.*) in a California woodland**. *New Phytol* 2008, **178**(1):167–176.

[195] Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH: **An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity**. *New Phytol* 2009, **181**(2):471–477.

[196] Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U: **454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases**. *New Phytol* 2010, **188**(1):291–301.

[197] Schoch CL, Robbertse B, Robert V, Vu D, Cardinali G, Irinyi L, Meyer W, Nilsson RH, Hughes K, Miller AN, Kirk PM, Abarenkov K, Aime MC, Ariyawansa HA, Bidartondo M, Boekhout T, Buyck B, Cai Q, Chen J, Crespo A, Crous PW, Damm U, De Beer ZW, Dentinger BTM, Divakar PK, Dueñas M, Feau N, Fliegerova K, García MA, Ge ZW, Griffith GW, Groenewald JZ, Groenewald M, Grube M, Gryzenhout M, Gueidan C, Guo L, Hambleton S, Hamelin R, Hansen K, Hofstetter V, Hong SB, Houbraken J, Hyde KD, Inderbitzin P, Johnston PR, Karunarathna SC, Kõljalg U, Kovács GM, Kraichak E, Krizsan K, Kurtzman CP, Larsson KH, Leavitt S, Letcher PM, Liimatainen K, Liu JK, Lodge DJ, Jennifer Luangsa-ard J, Lumbsch HT, Maharachchikumbura SS, Manamgoda D, Martín MP, Minnis AM, Moncalvo JM, Mulè G, Nakasone KK, Niskanen T, Olariaga I, Papp T, Petkovits T, Pino-Bodas

R, Powell MJ, Raja HA, Redecker D, Sarmiento-Ramirez JM, Seifert KA, Shrestha B, Stenroos S, Stielow B, Suh SO, Tanaka K, Tedersoo L, Telleria MT, Udayanga D, Untereiner WA, Diéguez Uribeondo J, Subbarao KV, Vágvölgyi C, Visagie C, Voigt K, Walker DM, Weir BS, Weiß M, Wijayawardene NN, Wingfield MJ, Xu JP, Yang ZL, Zhang N, Zhuang WY, Federhen S: **Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi**. *Database* 2014, doi: 10.1093/database/bau061.

[198] Page RD: **BioNames: linking taxonomy, texts, and trees**. *PeerJ* 2013, **1**:e190.

[199] Parr CS, Guralnick R, Cellinese N, Page RD: **Evolutionary informatics: unifying knowledge about the diversity of life**. *Trends Ecol Evol* 2012, **27**(2):94 – 103.

[200] Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2**. *bioRxiv* 2014, doi: 10.1101/002832.

[201] Cui L, Morris A, Ghedin E: **The human mycobiome in health and disease**. *Genome Med* 2013, **5**(7):63.

[202] Huffnagle GB, Noverr MC: **The emerging world of the fungal microbiome**. *Trends in Microbiol* 2014, **21**(7):334–341.

[203] Delhaes L, Monchy S, Frealle E, Hubans C, Salleron J, Leroy S, Prevotat A, Wallet F, Wallaert B, Dei-Cas E: **The airway microbiota in cystic fibrosis: a complex fungal and bacterial community-implications for therapeutic management.** *PLoS One* 2012, **7**(4):e36313.

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

[205] Chen Y, Chen Z, Guo R, Chen N, Lu H, Huang S, Wang J, Li L: **Correlation between gastrointestinal fungi and varying degrees of chronic hepatitis B virus infection.** *Diagn Micr Infec Dis* 2011, **70**:492–498.

[206] Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, Gillevet PM: **Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals**. *PLoS Pathog* 2010, **6**(1):e1000713.

[207] Ott S, Kuhbacher T, Musfeldt M, Rosenstiel P, Hellmig S, Rehman A, Drews O, Weichert W, Timmis K, Schreiber S: **Fungi and inflammatory bowel diseases: alterations of composition and diversity.** *Scand J Gastroenterol* 2008, **43**:831–841.

[208] Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, Kong HH, Segre JA: **Topographic diversity of fungal and bacterial communities in human skin**. *Nature* 2013, **498**:367-370.

[209] Bafadhel M, Mckenna S, Agbetile J, Fairs A, Desai D, Mistry V, Morley JP, Pancholi M, Pavord ID, Wardlaw AJ, Pashley CH, Brightling CE: **Aspergillus fumigatus during stable state and exacerbations of COPD**. *Eur Respir J* 2014, **43**(1):64–71.

[210] White BA, Creedon DJ, Nelson KE, Wilson BA: **The vaginal microbiome in health and disease**. *Trends Endocrinol Metab* 2014, **22**(10):389–393.

[211] Achkar JM, Fries BC: **Candida Infections of the Genitourinary Tract**. *Clin Microbiol Rev* 2010, **23**(2):253–273.

[212] Wade WG: **The oral microbiome in health and disease**. *Pharmacol Res* 2013, **69**(1):137 – 143.

[213] Charlson ES, Diamond JM, Bittinger K, Fitzgerald AS, Yadav A, Haas AR, Bushman FD, Collman RG: **Lung-enriched Organisms and Aberrant Bacterial and Fungal Respiratory Microbiota after Lung Transplant**. *Am J Respir Crit Care Med* 2014, **186**(6):536–545.

[214] Bousbia S, Papazian L, Saux P, Forel JM, Auffray JP, Martin C, Raoult D, La Scola B: **Repertoire of Intensive Care Unit Pneumonia Microbiota**. *PLoS ONE* 2012, **7**(2):e32486.

[215] Borg, I and Groenen, P: *Modern Multidimensional Scaling: theory and applications* 2 edn. New York, NYC, USA: Springer-Verlag. 2005.

[216] Bray J, JT. C: **An ordination of upland forest communities of southern Wisconsin. Ecological Monographs**. *Ecol Monogr* 1957, **27**:25–349.

[217] Owens B: **Gut microbe may fight obesity and diabetes**. *Nature* 2013, doi:10.1038/nature.2013.12975.

[218] Yang H, Qu L, Ma H, Chen L, Liu W, Liu C, Meng L, Wu J, Shou C: **Mycoplasma hyorhinis infection in gastric carcinoma and its effects on the malignant phenotypes of gastric cancer cells**. *BMC Gastroenterology* 2010, **10**(1):132.

[219] Mariotti E, Gemei M, Mirabelli P, D'Alessio F, Di Noto R, Fortunato G, Del Vecchio L: **The percentage of CD133+ cells in human colorectal cancer cell lines is influenced by Mycoplasma hyorhinis infection**. *BMC Cancer* 2010, **10**(1):120.

[220] Forsyth MH, Cover TL: **Mutational Analysis of the vacA Promoter Provides Insight into Gene Transcription in Helicobacter pylori**. *J Bacteriol* 1999, **181**(7):2261–2266.

[221] Petersen L, Larsen TS, Ussery DW, On SL, Krogh A: **RpoD Promoters in Campylobacter jejuni Exhibit a Strong Periodic Signal Instead of a -35 Box**. *J Mol Biol* 2003, **326**(5):1361 – 1372.

[222] Langmead B, Salzberg S: **Fast gapped-read alignment with Bowtie 2**. *Nat Meth* 2012, **9**:357–359.

[223] Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754–1760.

Acknowledgments

"Life is not easy for any of us. But what of that?
We must have perseverance and above all confidence in ourselves.
We must believe that we are gifted for something,
and that this thing must be attained."

- Marie Curie -

The time during a PhD and especially while writing the final manuscript was a challenging life time and an experience which would not have been possible without the support, understanding, and guidance that I received during this time from many people.

First and foremost, I would like to express my appreciation to Dr. Gerhard Thallinger, my research supervisor, for his useful and constructive guidance during the planning and development of the research projects of this thesis. In addition, I highly regard Prof. Dr. Rudolf Stollberger for undertaking the official supervision of my doctoral thesis. A special thanks also goes to Ass.-Prof. Gregor Gorkiewicz for all the valuable and fruitful discussions we had during the last years, and in particular for reviewing my thesis.

My grateful thanks are also extended to Ass.-Prof. Gregor Gorkiewicz, Prof. Christoph Högenauer, Prof. Robert Krause, and Dr. Anastasia Bragina, and their co-workers, for their assistance and advice, as well as for providing the data which was analyzed within this thesis.

A special thanks goes to Henrik R. Nilsson, PhD for providing the fungal ITS sequences which were processed within this thesis, as well as for his assistance, suggestions, and critical comments and discussions on the fungal research projects.

I would also like to thank the staff of the Center for Medical Research, Graz, Austria, for enabling me to visit their laboratories, to observe and study their daily operations in preparing and performing sequencing experiments.

I also want to mention and to appreciate the developers, research groups, and initiatives of all the open source packages, methods, tools, approaches, and resources which were used, extended, or evaluated within this thesis. All of them provided immensely valuable and free applications.

Many thanks also go out to the whole team of the Bioinformatics Group of the Institute of Knowledge Discovery at Graz University of Technology. Especially to my colleague, Christoph Fischer, for his valuable and constructive recommendations during the last years. An of course also to René Snajder for all his efforts and patience related to SnoWMAn. Thank you very much for your help.

Finally, I wish to thank my parents Erika and Josef, as well as my partner and best friend, Alexander, because they were there for me in so many ways. They were constantly supporting and encouraging me throughout my study.

Appendix

Appendix A.

Supplementary information gastrointestinal mouse study

Table A.1.: The table summarizes the total number of samples collected during the first sampling effort, of the GI mouse study, including the information about phenotype (Type: wildtype (WT), knockout (KO), bile-duct ligated (BDL)), Diet (chow (Normal)), Material (feces (F), mucosa (M)), each single mouse (individual (Ind.)), and the colonic location (Source) where the sample was collected (ileum (Ile), jejunum (Jej), caecum (Cae), colon (Col))

Sample	Type	Diet	Material	Ind.	Source
K2014-WT-N-F-Jej	WT	Normal	F	K2014	Jej
K2014-WT-N-F-Ile	WT	Normal	F	K2014	Ile
K2014-WT-N-F-Cae	WT	Normal	F	K2014	Cae
K2015-WT-N-F-Jej	WT	Normal	F	K2015	Jej
K2015-WT-N-F-Ile	WT	Normal	F	K2015	Ile
K2015-WT-N-F-Cae	WT	Normal	F	K2015	Cae
K2016-WT-N-F-Jej	WT	Normal	F	K2016	Jej
K2016-WT-N-F-Ile	WT	Normal	F	K2016	Ile
K2016-WT-N-F-Cae	WT	Normal	F	K2016	Cae
K2017-KO-N-F-Jej	KO	Normal	F	K2017	Jej
K2017-KO-N-F-Ile	KO	Normal	F	K2017	Ile
K2017-KO-N-F-Cae	KO	Normal	F	K2017	Cae
K2018-KO-N-F-Jej	KO	Normal	F	K2018	Jej
K2018-KO-N-F-Ile	KO	Normal	F	K2018	Ile
K2018-KO-N-F-Cae	KO	Normal	F	K2018	Cae
K2019-KO-N-F-Jej	KO	Normal	F	K2019	Jej
K2019-KO-N-F-Ile	KO	Normal	F	K2019	Ile
K2019-KO-N-F-Cae	KO	Normal	F	K2019	Cae
T1172-BD-N-F-Jej	BD	Normal	F	T1172	Jej
T1172-BD-N-F-Ile	BD	Normal	F	T1172	Ile
T1172-BD-N-F-Cae	BD	Normal	F	T1172	Cae
K2017-KO-N-M-Ile	KO	Normal	M	K2017	Ile
K2017-KO-N-M-Col	KO	Normal	M	K2017	Col
T1173-BD-N-F-Jej	BD	Normal	F	T1173	Jej
T1173-BD-N-F-Ile	BD	Normal	F	T1173	Ile
T1173-BD-N-F-Cae	BD	Normal	F	T1173	Cae
T1174-BD-N-F-Jej	BD	Normal	F	T1174	Jej
T1174-BD-N-F-Ile	BD	Normal	F	T1174	Ile
T1174-BD-N-F-Cae	BD	Normal	F	T1174	Cae
K2014-WT-N-M-Jej	WT	Normal	M	K2014	Jej
K2014-WT-N-M-Ile	WT	Normal	M	K2014	Ile
K2014-WT-N-M-Cae	WT	Normal	M	K2014	Cae
K2014-WT-N-M-Cae	WT	Normal	M	K2014	Cae
K2015-WT-N-M-Jej	WT	Normal	M	K2015	Jej
K2015-WT-N-M-Ile	WT	Normal	M	K2015	Ile
K2015-WT-N-M-Col	WT	Normal	M	K2015	Col
K2015-WT-N-M-Cae	WT	Normal	M	K2015	Cae
K2016-WT-N-M-Jej	WT	Normal	M	K2016	Jej
K2016-WT-N-M-Ile	WT	Normal	M	K2016	Ile
K2016-WT-N-M-Col	WT	Normal	M	K2016	Col
K2016-WT-N-M-Cae	WT	Normal	M	K2016	Cae
K2017-KO-N-M-Jej	KO	Normal	M	K2017	Jej
K2017-KO-N-M-Cae	KO	Normal	M	K2017	Cae
K2018-KO-N-M-Jej	KO	Normal	M	K2018	Jej
K2018-KO-N-M-Ile	KO	Normal	M	K2018	Ile
K2018-KO-N-M-Col	KO	Normal	M	K2018	Col
K2018-KO-N-M-Cae	KO	Normal	M	K2018	Cae
K2019-KO-N-M-Jej	KO	Normal	M	K2019	Jej
K2019-KO-N-M-Ile	KO	Normal	M	K2019	Ile
K2019-KO-N-M-Col	KO	Normal	M	K2019	Col
K2019-KO-N-M-Cae	KO	Normal	M	K2019	Cae
T2172-BD-N-M-Jej	BD	Normal	M	T2172	Jej
T2172-BD-N-M-Ile	BD	Normal	M	T2172	Ile
T2172-BD-N-M-Col	BD	Normal	M	T2172	Col
T2172-BD-N-M-Cae	BD	Normal	M	T2172	Cae
T2173-BD-N-M-Jej	BD	Normal	M	T2173	Jej
T2173-BD-N-M-Ile	BD	Normal	M	T2173	Ile
T2173-BD-N-M-Col	BD	Normal	M	T2173	Col
T2173-BD-N-M-Cae	BD	Normal	M	T2173	Cae
T2174-BD-N-M-Jej	BD	Normal	M	T2174	Jej
T1174-BD-N-M-Ile	BD	Normal	M	T1174	Ile
T1174-BD-N-M-Col	BD	Normal	M	T1174	Col
T1174-BD-N-M-Cae	BD	Normal	M	T1174	Cae

Table A.2.: The table summarizes the first part of samples collected during the second sampling effort, of the GI mouse study, including the information about phenotype (Type: wildtype (WT), knockout (KO)), Diet (chow, (Normal), 5 % Phosphatidylcholin-enriched chow (Enrich)), Material (feces (F), mucosa (M)), each single mouse (individual (Ind.)), and the colonic location (Source) where the sample was collected (ileum (Ile), caecum (Cae), colon (Col)).

Sample	Type	Diet	Material	Ind.	Source	Sample	Type	Diet	Material	Ind.	Source
K2290-WT-N-M-Ile	WT	Normal	M	K2290	Ile	MB001-WT-E-M-Ile	WT	Enrich	M	MB001	Ile
K2290-WT-N-M-Col	WT	Normal	M	K2290	Col	MB001-WT-E-M-Col	WT	Enrich	M	MB001	Col
K2290-WT-N-M-Cae	WT	Normal	M	K2290	Cae	MB001-WT-E-M-Cae	WT	Enrich	M	MB001	Cae
K2291-WT-N-M-Ile	WT	Normal	M	K2291	Ile	MB002-WT-E-M-Ile	WT	Enrich	M	MB002	Ile
K2291-WT-N-M-Col	WT	Normal	M	K2291	Col	MB002-WT-E-M-Col	WT	Enrich	M	MB002	Col
K2291-WT-N-M-Cae	WT	Normal	M	K2291	Cae	MB002-WT-E-M-Cae	WT	Enrich	M	MB002	Cae
K2292-WT-N-M-Ile	WT	Normal	M	K2292	Ile	MB003-WT-E-M-Ile	WT	Enrich	M	MB003	Ile
K2292-WT-N-M-Col	WT	Normal	M	K2292	Col	MB003-WT-E-M-Col	WT	Enrich	M	MB003	Col
K2292-WT-N-M-Cae	WT	Normal	M	K2292	Cae	MB003-WT-E-M-Cae	WT	Enrich	M	MB003	Cae
K2293-WT-N-M-Ile	WT	Normal	M	K2293	Ile	MB004-WT-E-M-Ile	WT	Enrich	M	MB004	Ile
K2293-WT-N-M-Col	WT	Normal	M	K2293	Col	MB004-WT-E-M-Col	WT	Enrich	M	MB004	Col
K2293-WT-N-M-Cae	WT	Normal	M	K2293	Cae	MB004-WT-E-M-Cae	WT	Enrich	M	MB004	Cae
K2294-WT-N-M-Ile	WT	Normal	M	K2294	Ile	MB005-WT-E-M-Ile	WT	Enrich	M	MB005	Ile
K2294-WT-N-M-Col	WT	Normal	M	K2294	Col	MB005-WT-E-M-Col	WT	Enrich	M	MB005	Col
K2294-WT-N-M-Cae	WT	Normal	M	K2294	Cae	MB005-WT-E-M-Cae	WT	Enrich	M	MB005	Cae
K2295-KO-N-M-Ile	KO	Normal	M	K2295	Ile	MB006-WT-E-M-Ile	WT	Enrich	M	MB006	Ile
K2295-KO-N-M-Col	KO	Normal	M	K2295	Col	MB006-WT-E-M-Col	WT	Enrich	M	MB006	Col
K2295-KO-N-M-Cae	KO	Normal	M	K2295	Cae	MB006-WT-E-M-Cae	WT	Enrich	M	MB006	Cae
K2296-KO-N-M-Ile	KO	Normal	M	K2296	Ile	MB007-KO-E-M-Ile	KO	Enrich	M	MB007	Ile
K2296-KO-N-M-Col	KO	Normal	M	K2296	Col	MB007-KO-E-M-Col	KO	Enrich	M	MB007	Col
K2296-KO-N-M-Cae	KO	Normal	M	K2296	Cae	MB007-KO-E-M-Cae	KO	Enrich	M	MB007	Cae
K2297-KO-N-M-Ile	KO	Normal	M	K2297	Ile	MB008-KO-E-M-Ile	KO	Enrich	M	MB008	Ile
K2297-KO-N-M-Col	KO	Normal	M	K2297	Col	MB008-KO-E-M-Col	KO	Enrich	M	MB008	Col
K2297-KO-N-M-Cae	KO	Normal	M	K2297	Cae	MB008-KO-E-M-Cae	KO	Enrich	M	MB008	Cae
K2298-KO-N-M-Ile	KO	Normal	M	K2298	Ile	MB009-KO-E-M-Ile	KO	Enrich	M	MB009	Ile
K2298-KO-N-M-Col	KO	Normal	M	K2298	Col	MB009-KO-E-M-Col	KO	Enrich	M	MB009	Col
K2298-KO-N-M-Cae	KO	Normal	M	K2298	Cae	MB009-KO-E-M-Cae	KO	Enrich	M	MB009	Cae
K2299-KO-N-M-Ile	KO	Normal	M	K2299	Ile	MB010-KO-E-M-Ile	KO	Enrich	M	MB010	Ile
K2299-KO-N-M-Col	KO	Normal	M	K2299	Col	MB010-KO-E-M-Col	KO	Enrich	M	MB010	Col
K2299-KO-N-M-Cae	KO	Normal	M	K2299	Cae	MB010-KO-E-M-Cae	KO	Enrich	M	MB010	Cae

Table A.3.: The table summarizes the second part of samples collected during the second sampling effort, of the GI mouse study, including the information about phenotype (Type) (wildtype (WT), knockout (KO)), Diet (chow, (Normal), Phosphatidylcholin-enriched chow (Enrich)), Material (feces (F), mucosa (M)), each single mouse (individual (Ind.)), each single mouse (individual (Ind.)), and the colonic location (Source) where the sample was collected (ileum (Ile), caecum (cae), colon (col)).

Sample	Type	Diet	Material	Ind.	Source	Sample	Type	Diet	Material	Ind.	Source
MB011-KO-E-M-Ile	KO	Enrich	M	MB011	Ile	K2299-KO-N-F-Col	KO	Normal	F	K2299	Col
MB011-KO-E-M-Col	KO	Enrich	M	MB011	Col	MB001-WT-E-F-Ile	WT	Enrich	F	MB001	Ile
MB011-KO-E-M-Cae	KO	Enrich	M	MB011	Cae	MB001-WT-E-F-Col	WT	Enrich	F	MB001	Col
MB012-KO-E-M-Ile	KO	Enrich	M	MB012	Ile	MB002-WT-E-F-Ile	WT	Enrich	F	MB002	Ile
MB012-KO-E-M-Col	KO	Enrich	M	MB012	Col	MB002-WT-E-F-Col	WT	Enrich	F	MB002	Col
MB012-KO-E-M-Cae	KO	Enrich	M	MB012	Cae	MB003-WT-E-F-Ile	WT	Enrich	F	MB003	Ile
K2290-WT-N-F-Ile	WT	Normal	F	K2290	Ile	MB003-WT-E-F-Col	WT	Enrich	F	MB003	Col
K2290-WT-N-F-Col	WT	Normal	F	K2290	Col	MB004-WT-E-F-Ile	WT	Enrich	F	MB004	Ile
K2291-WT-N-F-Ile	WT	Normal	F	K2291	Ile	MB004-WT-E-F-Col	WT	Enrich	F	MB004	Col
K2291-WT-N-F-Col	WT	Normal	F	K2291	Col	MB005-WT-E-F-Ile	WT	Enrich	F	MB005	Ile
K2292-WT-N-F-Ile	WT	Normal	F	K2292	Ile	MB005-WT-E-F-Col	WT	Enrich	F	MB005	Col
K2292-WT-N-F-Col	WT	Normal	F	K2292	Col	MB006-WT-E-F-Ile	WT	Enrich	F	MB006	Ile
K2293-WT-N-F-Ile	WT	Normal	F	K2293	Ile	MB006-WT-E-F-Col	WT	Enrich	F	MB006	Col
K2293-WT-N-F-Col	WT	Normal	F	K2293	Col	MB007-KO-E-F-Ile	KO	Enrich	F	MB007	Ile
K2294-WT-N-F-Ile	WT	Normal	F	K2294	Ile	MB007-KO-E-F-Col	KO	Enrich	F	MB007	Col
K2294-WT-N-F-Col	WT	Normal	F	K2294	Col	MB008-KO-E-F-Ile	KO	Enrich	F	MB008	Ile
K2295-KO-N-F-Ile	KO	Normal	F	K2295	Ile	MB008-KO-E-F-Col	KO	Enrich	F	MB008	Col
K2295-KO-N-F-Col	KO	Normal	F	K2295	Col	MB009-KO-E-F-Ile	KO	Enrich	F	MB009	Ile
K2296-KO-N-F-Ile	KO	Normal	F	K2296	Ile	MB009-KO-E-F-Col	KO	Enrich	F	MB009	Col
K2296-KO-N-F-Col	KO	Normal	F	K2296	Col	MB010-KO-E-F-Ile	KO	Enrich	F	MB010	Ile
K2297-KO-N-F-Ile	KO	Normal	F	K2297	Ile	MB010-KO-E-F-Col	KO	Enrich	F	MB010	Col
K2297-KO-N-F-Col	KO	Normal	F	K2297	Col	MB011-KO-E-F-Ile	KO	Enrich	F	MB011	Ile
K2298-KO-N-F-Ile	KO	Normal	F	K2298	Ile	MB011-KO-E-F-Col	KO	Enrich	F	MB011	Col
K2298-KO-N-F-Col	KO	Normal	F	K2298	Col	MB012-KO-E-F-Ile	KO	Enrich	F	MB012	Ile
K2299-KO-N-F-Ile	KO	Normal	F	K2299	Ile	MB012-KO-E-F-Col	KO	Enrich	F	MB012	Col

Table A.4.: The table summarizes tools and settings which were used for the analysis of the 16S GI mouse study data. Each row contains the used tool followed by the version, as well as the customized parameter. For all other parameters of the tools which are not listed within the table, the default values have been used.

	release/version	**settings**				
Decontaminator		**blat ref DB**	**perc. Identity**	**query coverage**	**MID off-set**	**primer off-set**
	v.2	GG 09May2011	95	75	0	0
UCHIME		**mode**	**reference DB**			
	mothur v.1.31.2	reference		SILVA relase 105		
Acacia		**min. avg. quality threshold**		**other settings**		
	1.52.b0	22		default		
SnoWMAn		**pipeline**	**classifier version**	**infernal model**		
	v.1.2	RDP	RDP classifier 2.5	ncbi16S_508_mod5		

Table A.5.: The table summarizes the number of filtered and remaining sequences after each pre-processing step. The 16S amplicons of the GI mouse study were noise reduced and quality filtered (Acacia), as well as filtered for contaminating sequences (Decontaminator) and chimeras (UCHIME) prior to the phylogenetic analysis.

number of sequences	G30ZHI201	G30ZHI202	G30ZHI203	HUIXUCX01	HUIXUCX02	HUIXUCX03	HUIXUCX04	**total**
raw	375833	378926	206758	267279	247950	236591	237823	**1951160**
contaminations	15279	13628	22956	26979	26529	22459	21533	**149363**
after decontamination	360554	365298	183802	240300	221421	214132	216290	**1801797**
chimeras	15279	13628	22956	26979	26529	22459	21533	**149363**
noise and low quality	41993	30465	18673	33398	26453	13008	4608	**168598**
after denoising and qual. Fitering	318561	334833	165129	206902	194968	201124	211682	**1633199**
toally removed	57272	44093	41629	60377	52982	35467	26141	**317961**
after preprocessing	318561	334833	165129	206902	194968	201124	211682	**1633199**
removed by snowman								**525811**
for classification								**1107388**

Table A.6.: Sample overview of the 16S GI mouse study samples analyzed using SnoWMAn's RDP pipeline. The table summarizes the number of the finally obtained distinct species (OTUs) for different cluster distances, for each sample of source type caecum.

Sample	Sequs	Unique Seqs	number of OTUs at different distances						
			0.0	0.01	0.02	0.03	0.04	0.05	0.06
K2015-WT-N-M-Cae	6310	1570	1069	819	568	454	383	352	320
K2298-KO-N-M-Cae	7317	1398	1012	815	548	437	374	337	310
K2018-KO-N-F-Cae	4666	2209	1624	1277	962	796	687	629	563
T1174-BD-N-M-Cae	830	344	245	214	172	158	142	130	120
T2172-BD-N-M-Cae	8109	2641	1856	1428	987	771	633	561	479
K2016-WT-N-F-Cae	6230	2269	1599	1178	826	650	554	496	442
K2019-KO-N-F-Cae	9394	3664	2556	1871	1300	1032	867	763	677
MB005-WT-E-M-Cae	5811	1252	847	692	452	373	319	292	262
MB010-KO-E-M-Cae	11298	1133	816	656	468	381	332	307	274
K2290-WT-N-M-Cae	4449	646	472	385	272	224	201	182	165
MB012-KO-E-M-Cae	7937	1460	1042	802	563	461	389	348	317
K2295-KO-N-M-Cae	10283	988	695	581	410	346	292	261	238
MB009-KO-E-M-Cae	2968	602	373	296	222	187	163	150	135
T1173-BD-N-F-Cae	10163	3203	2214	1612	1081	858	730	634	553
K2294-WT-N-M-Cae	8737	1513	990	796	476	372	326	298	275
K2299-KO-N-M-Cae	7205	1332	968	762	483	394	356	328	295
MB004-WT-E-M-Cae	8057	1866	1234	902	626	519	452	410	369
MB007-KO-E-M-Cae	10046	2460	1597	1106	720	554	464	412	368
K2292-WT-N-M-Cae	7351	1617	1156	920	688	567	471	418	372
K2297-KO-N-M-Cae	4528	941	666	539	387	320	281	254	223
K2018-KO-N-M-Cae	6161	2167	1532	1168	847	702	618	557	501
MB006-WT-E-M-Cae	4129	1225	884	738	552	464	409	372	328
MB001-WT-E-M-Cae	7262	2022	1437	1127	800	642	550	498	439
K2017-KO-N-M-Cae	2426	1335	995	831	666	571	507	460	403
MB011-KO-E-M-Cae	7205	1512	1035	754	487	395	352	330	302
MB008-KO-E-M-Cae	4846	1170	856	687	524	430	377	336	295
K2019-KO-N-M-Cae	5921	1325	939	757	546	455	384	350	308
K2015-WT-N-F-Cae	5283	2056	1343	912	598	472	402	359	331
K2014-WT-N-M-Cae	12529	4670	3155	2222	1416	1051	827	712	605
T1172-BD-N-F-Cae	11414	3847	2725	1971	1368	1075	892	780	667
K2296-KO-N-M-Cae	11959	1241	882	692	438	355	307	266	237
K2293-WT-N-M-Cae	3463	1045	802	648	510	438	388	350	313
MB003-WT-E-M-Cae	7039	1165	844	645	460	376	320	298	272
K2291-WT-N-M-Cae	6077	1104	785	604	429	366	323	298	277
K2014-WT-N-F-Cae	4885	1749	1232	871	595	501	441	402	363
T1174-BD-N-F-Cae	9513	3145	2188	1560	1075	864	712	635	533
T2173-BD-N-M-Cae	5417	1275	889	710	495	398	341	297	261
K2017-KO-N-F-Cae	2109	1166	714	616	477	398	352	320	285
MB002-WT-E-M-Cae	11317	1874	1296	970	624	512	438	394	354
K2016-WT-N-M-Cae	9376	1916	1357	1048	670	523	445	400	360
Total Cae	**280020**		**32834**	**18083**	**9209**	**6036**	**4294**	**3388**	**2630**

Table A.7.: Sample overview of the 16S GI mouse study samples analyzed using SnoWMAn's RDP pipeline. The table summarizes the number of finally obtained distinct species (OTUs) for different cluster distances, for each sample of source type colon.

Sample	Sequs	Unique Seqs	number of OTUs at different distances						
			0.0	0.01	0.02	0.03	0.04	0.05	0.06
K2290-WT-N-F-Col	2865	1258	938	805	622	541	470	429	380
MB006-WT-E-M-Col	12455	779	458	329	172	123	91	76	66
MB007-KO-E-M-Col	15348	932	551	401	218	159	124	105	92
MB003-WT-E-F-Col	14382	3290	2155	1605	1020	800	665	585	505
K2015-WT-N-M-Col	6615	1377	961	760	525	422	352	314	275
MB007-KO-E-F-Col	2258	1063	811	681	562	486	434	403	354
T2173-BD-N-M-Col	5846	1481	1029	846	617	520	452	396	346
MB011-KO-E-M-Col	8139	936	638	471	319	267	242	226	210
K2018-KO-N-M-Col	8577	4213	3174	2508	1793	1374	1103	961	800
K2294-WT-N-F-Col	3042	1083	753	659	512	433	385	355	318
K2019-KO-N-M-Col	11027	1446	918	726	508	413	349	308	269
MB008-KO-E-M-Col	9744	859	608	483	326	261	221	203	188
MB001-WT-E-F-Col	8386	2563	1826	1468	1041	843	707	631	545
MB010-KO-E-F-Col	4514	2172	1675	1355	1055	896	769	687	603
MB001-WT-E-M-Col	3842	1180	845	655	489	398	349	314	279
K2299-KO-N-M-Col	9787	973	688	563	374	305	265	232	214
K2294-WT-N-M-Col	12586	1904	1334	979	635	513	448	399	352
K2293-WT-N-M-Col	16789	929	533	368	181	125	90	69	57
T1174-BD-N-M-Col	664	240	160	138	102	85	70	61	57
T2172-BD-N-M-Col	13673	2040	1291	1000	690	549	459	404	352
MB012-KO-E-F-Col	7978	3470	2525	1993	1465	1154	948	833	717
K2295-KO-N-M-Col	7864	1139	842	687	496	417	363	335	305
MB010-KO-E-M-Col	9062	562	339	259	140	93	66	52	44
K2296-KO-N-M-Col	6402	461	317	250	152	111	95	80	70
K2291-WT-N-F-Col	6156	2804	2084	1685	1322	1095	927	814	699
K2297-KO-N-F-Col	3818	1251	904	745	556	466	397	358	313
K2291-WT-N-M-Col	16891	932	525	388	218	156	117	99	91
MB011-KO-E-F-Col	4657	2264	1715	1382	1101	926	803	719	617
K2297-KO-N-M-Col	4189	662	476	368	243	203	187	167	155
K2299-KO-N-F-Col	7125	2469	1727	1405	1008	819	688	605	524
MB005-WT-E-M-Col	8248	783	555	431	268	214	183	159	143
K2296-KO-N-F-Col	5014	1415	964	789	576	464	406	364	325
K2292-WT-N-F-Col	4299	2047	1500	1259	1001	833	721	643	560
MB002-WT-E-F-Col	2562	1354	1053	883	731	608	520	470	422
MB006-WT-E-F-Col	6230	1918	1344	1096	754	622	536	479	422
K2016-WT-N-M-Col	11735	2867	2028	1536	1027	803	667	576	488
MB005-WT-E-F-Col	4187	1229	887	758	567	480	412	375	332
K2298-KO-N-M-Col	8575	675	433	340	194	146	124	109	100
MB004-WT-E-M-Col	11133	804	515	384	243	187	154	135	124
K2295-KO-N-F-Col	3368	1629	1233	1037	844	705	621	563	494
MB009-KO-E-F-Col	4803	2326	1744	1377	1043	854	716	633	558
MB003-WT-E-M-Col	7520	582	386	317	201	159	135	120	108
MB008-KO-E-F-Col	3637	1680	1242	1005	800	667	584	523	451
K2298-KO-N-F-Col	2782	1090	810	696	559	481	428	385	352
MB012-KO-E-M-Col	7121	1130	762	514	347	298	267	248	228
K2293-WT-N-F-Col	3170	1040	737	637	471	398	343	306	278
MB002-WT-E-M-Col	4881	802	602	475	334	275	240	219	204
MB009-KO-E-M-Col	7967	1308	751	525	354	300	260	240	218
K2017-KO-N-M-Col	4971	1233	835	699	538	461	406	360	323
K2290-WT-N-M-Col	11497	835	568	423	268	207	179	162	146
K2292-WT-N-M-Col	9205	887	627	448	291	225	196	176	161
MB004-WT-E-F-Col	3328	1853	1428	1162	938	780	664	589	509
Total Col	**380914**		**34530**	**19962**	**10705**	**7024**	**4907**	**3786**	**2852**

Table A.8.: Sample overview of the 16S GI mouse study samples analyzed using SnoWMAn's RDP pipeline. The table summarizes the number of finally obtained distinct species (OTUs) for different cluster distances, for each sample of source type ileum.

Sample	Sequs	Unique Seqs	number of OTUs at different distances						
			0.0	0.01	0.02	0.03	0.04	0.05	0.06
K2019-KO-N-F-Ile	4193	699	458	349	173	91	73	65	62
MB003-WT-E-M-Ile	308	96	56	42	30	25	23	22	20
K2291-WT-N-F-Ile	7342	1449	854	645	321	183	140	120	105
MB011-KO-E-M-Ile	462	108	87	72	54	39	38	36	34
K2016-WT-N-M-Ile	5618	745	484	368	188	113	86	73	71
MB008-KO-E-F-Ile	10365	2153	1319	978	575	388	323	294	259
MB003-WT-E-F-Ile	6957	1207	668	448	205	123	98	88	79
MB012-KO-E-M-Ile	5442	1320	838	583	363	280	236	215	181
MB010-KO-E-M-Ile	289	103	70	53	41	39	37	35	34
T1172-BD-N-F-Ile	7983	1466	950	718	388	257	217	192	172
K2014-WT-N-F-Ile	5690	1097	742	543	281	182	155	138	129
K2296-KO-N-M-Ile	2408	336	233	171	92	69	57	53	49
K2298-KO-N-F-Ile	4875	1179	759	571	378	285	241	216	202
K2297-KO-N-M-Ile	1876	459	302	220	145	110	100	92	87
T2172-BD-N-M-Ile	6976	1138	734	569	333	227	184	165	150
K2014-WT-N-M-Ile	2898	546	360	278	162	121	100	93	90
MB001-WT-E-F-Ile	13986	2451	1469	877	427	282	236	209	182
K2015-WT-N-M-Ile	3927	830	495	377	211	151	120	110	100
MB009-KO-E-F-Ile	5094	1595	1090	787	555	438	369	329	295
MB010-KO-E-F-Ile	7381	1392	770	525	225	111	84	72	61
K2291-WT-N-M-Ile	2237	419	276	190	97	60	48	43	39
K2298-KO-N-M-Ile	1288	337	242	195	135	110	97	93	88
K2295-KO-N-F-Ile	6134	1475	996	766	500	362	299	267	243
K2292-WT-N-M-Ile	1790	387	254	199	110	73	61	58	56
K2295-KO-N-M-Ile	3566	806	559	425	250	185	159	143	133
K2290-WT-N-F-Ile	7777	1306	821	605	300	191	158	138	132
K2018-KO-N-F-Ile	5601	928	645	496	270	178	150	136	119
MB008-KO-E-M-Ile	1709	377	256	198	136	106	94	85	78
T2173-BD-N-M-Ile	11052	1732	1138	834	454	332	272	241	220
K2017-KO-N-F-Ile	3860	1337	714	482	283	192	164	146	135
K2290-WT-N-M-Ile	1058	212	145	117	69	46	39	38	36
MB002-WT-E-M-Ile	698	233	169	135	94	81	78	72	69
MB005-WT-E-M-Ile	420	94	65	48	32	27	26	25	24
K2292-WT-N-F-Ile	9321	1804	1017	756	373	215	172	147	131
MB009-KO-E-M-Ile	3924	966	657	486	348	288	250	223	207
MB004-WT-E-F-Ile	3879	1028	666	527	323	224	187	166	150
MB012-KO-E-F-Ile	13200	3745	2376	1673	1013	719	588	498	418
K2018-KO-N-M-Ile	5793	1527	1025	756	513	407	361	342	314
K2293-WT-N-M-Ile	1159	293	199	149	106	86	77	72	70
MB001-WT-E-M-Ile	2446	320	221	161	73	53	46	39	36
K2016-WT-N-F-Ile	18165	2044	1115	757	295	141	105	85	72
T1174-BD-N-M-Ile	514	235	166	145	116	103	96	92	92
K2019-KO-N-M-Ile	5773	1015	592	420	238	153	129	115	109
MB007-KO-E-M-Ile	4725	773	495	371	192	131	110	104	95
MB006-WT-E-M-Ile	2050	332	208	164	76	43	37	31	30
K2293-WT-N-F-Ile	5229	1220	786	597	377	270	230	205	180
K2017-KO-N-M-Ile	7328	1918	1104	800	470	334	270	226	199
K2294-WT-N-F-Ile	5474	1271	816	638	391	272	232	205	187
MB006-WT-E-F-Ile	17023	2837	1492	1004	391	198	147	121	101
K2294-WT-N-M-Ile	454	153	106	89	66	58	56	54	50
K2299-KO-N-M-Ile	3020	670	440	299	200	156	142	128	118
MB004-WT-E-M-Ile	1297	288	212	155	109	85	74	71	65
MB005-WT-E-F-Ile	5350	1159	704	526	305	209	180	157	144
K2015-WT-N-F-Ile	4412	610	417	304	143	75	50	38	35
K2297-KO-N-F-Ile	4390	1071	684	552	354	253	216	191	175
K2299-KO-N-F-Ile	6838	1784	1210	943	624	466	398	342	300
MB011-KO-E-F-Ile	10455	1764	917	665	299	158	125	108	98
T1174-BD-N-F-Ile	5260	850	569	449	201	103	77	62	56
MB002-WT-E-F-Ile	7366	1179	736	555	248	138	107	91	81
T1173-BD-N-F-Ile	9584	1556	1012	717	377	240	200	180	166
K2296-KO-N-F-Ile	7987	1439	888	657	346	207	165	146	134
MB007-KO-E-F-Ile	6123	1402	868	672	414	276	233	205	181
Total Ile	**329799**		**22199**	**10378**	**4538**	**2961**	**2221**	**1837**	**1498**

Table A.9.: Sample overview of the 16S GI mouse study samples analyzed using SnoWMAn's RDP pipeline. The table summarizes the number of finally obtained distinct species (OTUs) for different cluster distances, for each sample of source type jejunum.

Sample	Sequs	Unique Seqs	number of OTUs at different distances						
			0.0	0.01	0.02	0.03	0.04	0.05	0.06
T2173-BD-N-M-Jej	6416	1119	743	552	358	266	220	200	188
T2172-BD-N-M-Jej	6772	1221	827	647	431	333	284	253	233
K2015-WT-N-M-Jej	5097	1292	858	597	394	308	269	245	229
K2016-WT-N-M-Jej	3457	602	377	278	149	90	71	62	58
K2017-KO-N-F-Jej	6358	1481	854	574	283	175	155	131	124
K2018-KO-N-M-Jej	3303	803	535	362	243	200	175	156	148
K2014-WT-N-M-Jej	3605	519	330	250	131	68	57	50	48
K2016-WT-N-F-Jej	7239	998	580	436	190	90	68	57	47
K2015-WT-N-F-Jej	5913	831	548	410	172	90	68	53	46
T2174-BD-N-M-Jej	7103	1573	1000	650	434	337	295	261	236
T1172-BD-N-F-Jej	5021	1057	735	551	306	210	181	164	154
T1173-BD-N-F-Jej	735	179	120	100	65	48	41	39	37
K2014-WT-N-F-Jej	14839	2341	1431	947	419	237	184	152	130
T1174-BD-N-F-Jej	6553	1252	796	529	236	133	102	90	72
K2018-KO-N-F-Jej	11740	2161	1449	972	500	333	280	245	224
K2019-KO-N-F-Jej	9321	1436	827	610	297	171	134	111	99
K2017-KO-N-M-Jej	3112	578	386	309	199	155	132	115	106
K2019-KO-N-M-Jej	10071	1445	954	658	381	264	220	201	187
Total Jej	**116655**		**9306**	**4939**	**2134**	**1367**	**1069**	**900**	**780**

Table A.10.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the phylum level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA phylum level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each phylum and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Tenericutes	KO.N.M.Ile	BD.N.M.Ile	8	3	1741	0	73451.27	345.67	7.5834	0.0149
Tenericutes	KO.N.M.Ile	KO.E.M.Ile	8	6	1741	0	73451.27	358.86	7.5336	0.0001
Tenericutes	WT.N.M.Ile	BD.N.M.Ile	8	3	1502	0	62960.90	345.67	7.3613	0.0183
Verrucomicrobia	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1335	301.24	43629.64	-7.0936	0.0000
Verrucomicrobia	KO.N.F.Ile	KO.E.F.Ile	8	6	0	1335	315.66	43629.64	-6.9755	0.0000
Cyanobacteria/Chloroplast	WT.N.F.Ile	KO.N.F.Ile	8	8	880	0	32405.39	315.66	6.5384	0.0000
Cyanobacteria/Chloroplast	WT.N.F.Ile	BD.N.F.Ile	8	3	880	0	32405.39	347.02	6.3952	0.0165
Tenericutes	WT.E.F.Ile	KO.E.F.Ile	6	6	683	0	18575.56	313.01	5.7428	0.0017
Tenericutes	WT.N.F.Ile	WT.E.F.Ile	8	6	0	683	318.03	18575.56	-5.7252	0.0005
Cyanobacteria/Chloroplast	WT.E.F.Ile	KO.E.F.Ile	6	6	807	0	9182.48	313.01	4.7487	0.0020
Actinobacteria	WT.N.M.Ile	KO.N.M.Ile	8	8	223	0	9851.45	347.30	4.6838	0.0000
Actinobacteria	WT.N.M.Ile	WT.E.M.Ile	8	6	223	0	9851.45	375.08	4.5831	0.0004
Verrucomicrobia	KO.N.M.Ile	KO.E.M.Ile	8	6	0	147	347.30	8201.28	-4.4231	0.0001
TM7	WT.N.F.Ile	WT.E.F.Ile	8	6	157	0	6599.97	301.24	4.3645	0.0019
Verrucomicrobia	WT.E.M.Ile	KO.E.M.Ile	6	6	0	147	375.08	8201.28	-4.3225	0.0028
Tenericutes	WT.E.M.Ile	KO.E.M.Ile	6	6	128	0	7681.22	358.86	4.2849	0.0215
TM7	WT.N.F.Ile	KO.N.F.Ile	8	8	157	0	6599.97	315.66	4.2526	0.0004
Bacteroidetes	WT.E.M.Ile	KO.E.M.Ile	6	6	296	5698	17916.35	310751.09	-4.1130	0.0281
Proteobacteria	WT.E.F.Ile	KO.E.F.Ile	6	6	403	1192	4736.36	63905.99	-3.7173	0.0182
TM7	WT.E.F.Ile	KO.E.F.Ile	6	6	0	126	301.24	4047.77	-3.6903	0.0061
Cyanobacteria/Chloroplast	WT.N.M.Ile	KO.N.M.Ile	8	8	104	0	4783.58	347.30	3.6472	0.0076
TM7	KO.N.F.Ile	KO.E.F.Ile	8	6	0	126	315.66	4047.77	-3.5784	0.0047
Cyanobacteria/Chloroplast	WT.N.M.Ile	WT.E.M.Ile	8	6	104	0	4783.58	375.08	3.5467	0.0148
Tenericutes	WT.N.F.Ile	KO.N.F.Ile	8	8	0	179	318.03	3493.20	-3.3510	0.0234
Tenericutes	WT.N.M.Ile	WT.E.M.Ile	8	6	1502	128	62960.90	7681.22	3.0266	0.0403
TM7	WT.N.M.Ile	KO.N.M.Ile	8	8	47	0	2467.57	347.30	2.6999	0.0279
Firmicutes	WT.N.M.Ile	WT.E.M.Ile	8	6	11308	1277	483141.82	79098.82	2.6101	0.0021
TM7	WT.N.M.Ile	WT.E.M.Ile	8	6	47	0	2467.57	375.08	2.5993	0.0403
Firmicutes	WT.E.M.Ile	KO.E.M.Ile	6	6	1277	4608	79098.82	258122.23	-1.7059	0.0401

Table A.11.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the class level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA class level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each class and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Verrucomicrobiae	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1335	428.01	71897.01	-7.2392	0.0000
Bacteroidia	WT.N.F.Ile	KO.N.F.Ile	8	8	0	629	423.64	28009.61	-5.8945	0.0000
Bacteroidia	WT.E.F.Ile	KO.E.F.Ile	6	6	0	848	428.01	49070.28	-6.6883	0.0000
Actinobacteria	WT.N.M.Ile	KO.N.M.Ile	8	8	223	0	10227.14	375.29	4.6059	0.0000
Verrucomicrobiae	KO.N.F.Ile	KO.E.F.Ile	8	6	46	1335	2298.16	71897.01	-4.9252	0.0001
Deltaproteobacteria	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1135	428.01	62050.49	-7.0263	0.0001
Clostridia	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1179	428.01	65457.45	-7.1032	0.0002
Chloroplast	WT.E.F.Ile	KO.E.F.Ile	6	6	807	0	51300.09	380.72	6.9075	0.0002
Gammaproteobacteria	WT.N.M.Ile	BD.N.M.Ile	8	3	1075	0	53326.02	351.32	7.0606	0.0002
Mollicutes	WT.N.F.Ile	WT.E.F.Ile	8	6	0	683	423.64	45068.59	-6.5789	0.0002
Mollicutes	WT.E.F.Ile	KO.E.F.Ile	6	6	683	0	45068.59	380.72	6.7206	0.0002
Gammaproteobacteria	KO.N.M.Ile	BD.N.M.Ile	8	3	995	0	47958.05	351.32	6.9078	0.0003
Actinobacteria	WT.N.M.Ile	WT.E.M.Ile	8	6	223	0	10227.14	427.18	4.4422	0.0007
Deltaproteobacteria	KO.N.M.Ile	KO.E.M.Ile	8	6	0	386	375.29	20119.03	-5.5763	0.0009
Verrucomicrobiae	KO.N.M.Ile	KO.E.M.Ile	8	6	0	147	375.29	8207.77	-4.2934	0.0009
Epsilonproteobacteria	KO.N.M.Ile	KO.E.M.Ile	8	6	737	9	30454.44	1056.73	4.8084	0.0010
Bacteroidia	WT.N.M.Ile	KO.N.M.Ile	8	8	58	1081	2995.44	50302.26	-4.0433	0.0012
Mollicutes	KO.N.M.Ile	KO.E.M.Ile	8	6	1741	16	82395.54	1401.17	5.8269	0.0012
TM7_genera_incertae_sedis	KO.N.M.Ile	KO.E.M.Ile	8	6	0	122	375.29	6744.76	-4.0132	0.0013
TM7_genera_incertae_sedis	WT.N.F.Ile	KO.N.F.Ile	8	8	157	0	7846.42	383.24	4.2038	0.0014
TM7_genera_incertae_sedis	KO.N.F.Ile	KO.E.F.Ile	8	6	0	126	383.24	6962.29	-4.0353	0.0032
TM7_genera_incertae_sedis	WT.E.F.Ile	KO.E.F.Ile	6	6	0	126	428.01	6962.29	-3.8863	0.0052
Deltaproteobacteria	WT.E.M.Ile	KO.E.M.Ile	6	6	0	386	427.18	20119.03	-5.4115	0.0071
Verrucomicrobiae	WT.E.M.Ile	KO.E.M.Ile	6	6	0	147	427.18	8207.77	-4.1291	0.0071
Bacilli	WT.N.M.Ile	WT.E.M.Ile	8	6	10280	1098	482558.95	80964.30	2.5749	0.0087
TM7_genera_incertae_sedis	WT.N.F.Ile	WT.E.F.Ile	8	6	157	0	7846.42	428.01	4.0548	0.0099
Bacteroidia	WT.N.M.Ile	BD.N.M.Ile	8	3	58	347	2995.44	41305.90	-3.7592	0.0123
TM7_genera_incertae_sedis	WT.E.M.Ile	KO.E.M.Ile	6	6	0	122	427.18	6744.76	-3.8489	0.0131
Mollicutes	WT.N.F.Ile	KO.N.F.Ile	8	8	0	179	423.64	7244.07	-3.9547	0.0134
Bacteroidia	WT.E.M.Ile	KO.E.M.Ile	6	6	39	455	2546.23	25702.07	-3.2885	0.0219

Table A.12.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the order level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA the order level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each order and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Bacillales	KO.N.F.Ile	KO.E.F.Ile	8	6	0	641	413.38	46507.67	-6.6615	4.40E-07
Verrucomicrobiales	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1335	423.40	93021.55	-7.6215	3.60E-06
Bacillales	WT.E.F.Ile	KO.E.F.Ile	6	6	0	641	423.40	46507.67	-6.6219	8.75E-06
Bacteroidales	WT.N.F.Ile	KO.N.F.Ile	8	8	0	629	425.25	29850.60	-5.9785	1.04E-05
Bacteroidales	WT.E.F.Ile	KO.E.F.Ile	6	6	0	848	423.40	53829.18	-6.8334	1.06E-05
Actinomycetales	WT.N.M.Ile	KO.N.M.Ile	8	8	223	0	11084.90	363.91	4.7527	1.75E-05
Desulfovibrionales	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1131	423.40	66626.19	-7.1408	8.28E-05
Verrucomicrobiales	KO.N.F.Ile	KO.E.F.Ile	8	6	46	1335	2886.66	93021.55	-4.9905	8.93E-05
Clostridiales	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1179	423.40	73473.17	-7.2815	0.000119
Chloroplast	WT.E.F.Ile	KO.E.F.Ile	6	6	807	0	52093.91	407.17	6.8377	0.000154
Mycoplasmatales	WT.N.F.Ile	WT.E.F.Ile	8	6	0	683	425.25	45954.88	-6.5993	0.000163
Mycoplasmatales	WT.E.F.Ile	KO.E.F.Ile	6	6	683	0	45954.88	407.17	6.6568	0.000206
Pseudomonadales	WT.N.M.Ile	BD.N.M.Ile	8	3	726	0	33892.92	377.57	6.3183	0.000436
Actinomycetales	WT.N.M.Ile	WT.E.M.Ile	8	6	223	0	11084.90	412.13	4.5973	0.000454
Pseudomonadales	KO.N.M.Ile	BD.N.M.Ile	8	3	687	0	31008.88	377.57	6.1903	0.000567
Desulfovibrionales	KO.N.M.Ile	KO.E.M.Ile	8	6	0	383	363.91	19873.67	-5.5945	0.000697
Verrucomicrobiales	KO.N.M.Ile	KO.E.M.Ile	8	6	0	147	363.91	8990.00	-4.4552	0.000697
Campylobacterales	KO.N.M.Ile	KO.E.M.Ile	8	6	737	9	32726.92	1028.49	4.9456	0.000807
Mycoplasmatales	KO.N.M.Ile	KO.E.M.Ile	8	6	1741	16	79137.39	1320.07	5.8462	0.001046
Bacteroidales	WT.N.M.Ile	KO.N.M.Ile	8	8	58	1081	2995.99	50360.73	-4.0464	0.001246
Enterobacteriales	WT.N.M.Ile	BD.N.M.Ile	8	3	336	0	15845.36	377.57	5.2241	0.001332
Enterobacteriales	KO.N.M.Ile	BD.N.M.Ile	8	3	296	0	13634.96	377.57	5.0082	0.002112
Bacillales	WT.N.M.Ile	KO.N.M.Ile	8	8	79	0	4169.89	363.91	3.3518	0.005517
Desulfovibrionales	WT.E.M.Ile	KO.E.M.Ile	6	6	0	383	412.13	19873.67	-5.4377	0.005723
Verrucomicrobiales	WT.E.M.Ile	KO.E.M.Ile	6	6	0	147	412.13	8990.00	-4.2990	0.005723
Lactobacillales	WT.N.M.Ile	WT.E.M.Ile	8	6	10163	1080	482879.51	76492.36	2.6577	0.006771
Bacteroidales	WT.N.M.Ile	BD.N.M.Ile	8	3	58	347	2995.99	41220.89	-3.7584	0.008225
Mycoplasmatales	WT.N.M.Ile	BD.N.M.Ile	8	3	1516	0	70037.00	377.57	7.3619	0.008225
Mycoplasmatales	KO.N.M.Ile	BD.N.M.Ile	8	3	1741	0	79137.39	377.57	7.5379	0.009395
Chloroplast	WT.N.M.Ile	KO.N.M.Ile	8	8	104	0	5370.41	363.91	3.7133	0.013347

Table A.13.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the family level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA the family level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each family and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Staphylococcaceae	KO.N.F.Ile	KO.E.F.Ile	8	6	0	641	415.22	46760.93	-6.6511	5.63E-09
Streptococcaceae	WT.N.F.Ile	WT.E.F.Ile	8	6	0	1462	420.56	90660.32	-7.5887	3.76E-08
Enterococcaceae	WT.E.F.Ile	KO.E.F.Ile	6	6	4841	0	304396.06	415.29	9.3514	2.24E-07
Staphylococcaceae	WT.E.F.Ile	KO.E.F.Ile	6	6	0	641	409.02	46760.93	-6.6704	2.69E-07
Streptococcaceae	WT.E.F.Ile	KO.E.F.Ile	6	6	1462	0	90660.32	415.29	7.6049	7.96E-07
Verrucomicrobiaceae	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1335	409.02	90317.92	-7.6196	7.96E-07
Prevotellaceae	WT.N.M.Ile	KO.N.M.Ile	8	8	0	235	400.61	11183.01	-4.6409	4.69E-06
Enterococcaceae	WT.N.F.Ile	KO.N.F.Ile	8	8	701	0	36224.37	415.22	6.2836	1.07E-05
Propionibacteriaceae	WT.N.M.Ile	KO.N.M.Ile	8	8	147	0	7813.95	393.47	4.1457	2.30E-05
Porphyromonadaceae	WT.N.M.Ile	KO.N.M.Ile	8	8	0	129	400.61	6693.22	-3.9021	6.33E-05
Desulfovibrionaceae	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1126	409.02	76128.76	-7.3729	7.17E-05
Lachnospiraceae	WT.E.F.Ile	KO.E.F.Ile	6	6	0	918	409.02	62256.05	-7.0829	7.17E-05
Verrucomicrobiaceae	KO.N.F.Ile	KO.E.F.Ile	8	6	46	1335	2764.69	90317.92	-5.0029	8.87E-05
Prevotellaceae	KO.N.M.Ile	KO.E.M.Ile	8	6	235	0	11183.01	401.90	4.6387	0.000101
Mycoplasmataceae	WT.N.F.Ile	WT.E.F.Ile	8	6	0	683	420.56	43730.48	-6.5375	0.000116
Streptophyta	WT.E.F.Ile	KO.E.F.Ile	6	6	807	0	49617.74	415.29	6.7360	0.000203
Ruminococcaceae	WT.N.M.Ile	KO.N.M.Ile	8	8	0	165	400.61	7939.42	-4.1494	0.000223
Mycoplasmataceae	WT.E.F.Ile	KO.E.F.Ile	6	6	683	0	43730.48	415.29	6.5538	0.000283
Streptococcaceae	WT.N.M.Ile	KO.N.M.Ile	8	8	167	0	8381.87	393.47	4.2483	0.000377
Desulfovibrionaceae	KO.N.M.Ile	KO.E.M.Ile	8	6	0	383	393.47	23756.22	-5.7446	0.000412
Helicobacteraceae	KO.N.M.Ile	KO.E.M.Ile	8	6	806	9	38790.21	1014.80	5.1910	0.000412
Pseudomonadaceae	WT.N.M.Ile	BD.N.M.Ile	8	3	725	0	36670.71	405.17	6.3311	0.000449
Verrucomicrobiaceae	KO.N.M.Ile	KO.E.M.Ile	8	6	0	147	393.47	9365.57	-4.4088	0.000484
Pseudomonadaceae	KO.N.M.Ile	BD.N.M.Ile	8	3	684	0	34502.88	405.17	6.2431	0.000542
Propionibacteriaceae	WT.N.M.Ile	WT.E.M.Ile	8	6	147	0	7813.95	405.51	4.1087	0.000673
Mycoplasmataceae	KO.N.M.Ile	KO.E.M.Ile	8	6	1757	16	87393.37	1477.76	5.8385	0.001148
Enterobacteriaceae	WT.N.M.Ile	BD.N.M.Ile	8	3	336	0	17203.71	405.17	5.2414	0.001414
Enterobacteriaceae	KO.N.M.Ile	BD.N.M.Ile	8	3	296	0	15189.05	405.17	5.0617	0.002123
Lactobacillaceae	WT.N.M.Ile	WT.E.M.Ile	8	6	9874	1038	491739.69	74125.45	2.7295	0.005351
Prevotellaceae	KO.N.M.Ile	BD.N.M.Ile	8	3	235	0	11183.01	405.17	4.6254	0.005999

Table A.14.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the genus level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA genus level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each genus and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Streptococcus	WT.N.F.Ile	WT.E.F.Ile	8	6	0	1462	534.16	126520.18	-7.7195	1.11E-11
Staphylococcus	KO.N.F.Ile	KO.E.F.Ile	8	6	0	641	533.54	57090.21	-6.5738	2.57E-09
Streptococcus	WT.E.F.Ile	KO.E.F.Ile	6	6	1462	0	126520.18	533.48	7.7211	5.08E-09
Ureaplasma	WT.N.F.Ile	WT.E.F.Ile	8	6	0	448	534.16	39030.92	-6.0249	5.64E-09
Staphylococcus	WT.E.F.Ile	KO.E.F.Ile	6	6	0	641	529.51	57090.21	-6.5836	1.13E-07
Enterococcus	WT.E.F.Ile	KO.E.F.Ile	6	6	4836	0	419034.75	533.48	9.4482	1.13E-07
Ureaplasma	WT.E.F.Ile	KO.E.F.Ile	6	6	448	0	39030.92	533.48	6.0265	2.31E-07
Akkermansia	WT.E.F.Ile	KO.E.F.Ile	6	6	0	1335	529.51	119414.89	-7.6473	8.05E-07
Enterococcus	WT.N.F.Ile	KO.N.F.Ile	8	8	700	0	46855.97	533.54	6.2892	1.68E-05
Lactococcus	WT.N.M.Ile	KO.N.M.Ile	8	8	162	0	10971.67	530.99	4.2093	1.90E-05
Propionibacterium	WT.N.M.Ile	KO.N.M.Ile	8	8	147	3	10257.46	729.51	3.6967	8.54E-05
Lactococcus	WT.N.M.Ile	WT.E.M.Ile	8	6	162	0	10971.67	530.86	4.2096	0.000206
Propionibacterium	WT.N.M.Ile	WT.E.M.Ile	8	6	147	0	10257.46	530.86	4.1114	0.000206
Pseudomonas	WT.N.M.Ile	BD.N.M.Ile	8	3	689	0	46166.92	528.73	6.2793	0.000255
Pseudomonas	KO.N.M.Ile	BD.N.M.Ile	8	3	667	0	44919.28	528.73	6.2398	0.000279
Corynebacterium	WT.N.M.Ile	KO.N.M.Ile	8	8	72	0	5294.41	530.99	3.1657	0.000294
Roseburia	WT.N.M.Ile	KO.N.M.Ile	8	8	0	62	529.88	4618.50	-2.9750	0.000488
Akkermansia	KO.N.M.Ile	KO.E.M.Ile	8	6	0	147	530.99	13503.21	-4.5056	0.001726
Corynebacterium	WT.N.M.Ile	WT.E.M.Ile	8	6	72	0	5294.41	530.86	3.1660	0.001785
Mycoplasma	KO.N.M.Ile	KO.E.M.Ile	8	6	1757	16	117255.02	1944.81	5.8655	0.00199
Helicobacter	KO.N.M.Ile	KO.E.M.Ile	8	6	798	23	53320.85	2569.80	4.3395	0.00199
Mycoplasma	WT.N.F.Ile	KO.N.F.Ile	8	8	0	253	534.16	17275.57	-4.8517	0.00212
Akkermansia	WT.N.F.Ile	KO.N.F.Ile	8	8	0	130	534.16	9215.89	-3.9498	0.00212
Roseburia	KO.N.M.Ile	KO.E.M.Ile	8	6	62	0	4618.50	533.10	2.9672	0.002575
Mycoplasma	WT.N.F.Ile	WT.E.F.Ile	8	6	0	189	534.16	17205.37	-4.8458	0.002813
Lactobacillus	WT.N.M.Ile	WT.E.M.Ile	8	6	7980	912	527376.61	81646.79	2.6904	0.003009
Mycoplasma	WT.E.F.Ile	KO.E.F.Ile	6	6	189	0	17205.37	533.48	4.8474	0.006049
Akkermansia	KO.N.F.Ile	KO.E.F.Ile	8	6	130	1335	9215.89	119414.89	-3.6862	0.006244
Mycoplasma	KO.N.F.Ile	KO.E.F.Ile	8	6	253	0	17275.57	533.48	4.8534	0.009875
Akkermansia	WT.E.M.Ile	KO.E.M.Ile	6	6	0	147	530.86	13503.21	-4.5059	0.009976

Table A.15.: The table presents the top 30 differentially abundant features, determined by edgeR, based on the feature table obtained from the RDP pipeline, at the OTU level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA the OTU level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each OTU and contrast are included. Furthermore, for each OTU the RDP classification result and confidence is available from the table.

TAX	G1	G2	# G1	# G2	rcG1	rcG2	cpm G1	cpm G2	logFC	FDR	Domain	Prob	Phylum	Prob	Class	Prob	Order	Prob	Familiy	Prob	Genus	Prob
271	KO.N.F.Ile	BD.N.F.Ile	8	3	0	148	133.89	6742.10	-54.876	5.18E-17	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
271	WT.N.F.Ile	BD.N.F.Ile	8	3	0	148	134.08	6742.10	-54.858	5.35E-17	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
271	KO.N.M.Ile	BD.N.M.Ile	8	3	1	68	150.90	3173.07	-42.497	6.51E-09	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
271	WT.N.M.Ile	BD.N.M.Ile	8	3	4	68	201.74	3173.07	-38.680	1.31E-07	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
150	KO.N.F.Ile	KO.E.F.Ile	8	6	0	740	133.89	16711.24	-67.948	1.40E-07	Bacteria;	1;	Firmicutes;	1;	Bacilli;	1;	Bacillales;	1;	Staphylococcaceae;	0.99;	Staphylococcus;	0.99;
2282	KO.N.M.Ile	BD.N.M.Ile	8	3	0	38	134.16	1830.50	-36.141	1.51E-06	Bacteria;	1;	Bacteroidetes;	1;	Flavobacteria;	1;	Flavobacteriales;	1;	Flavobacteriaceae;	1;	Cloacibacterium;	0.76;
2421	KO.E.F.Ile	WT.E.F.Ile	6	6	0	554	133.59	12505.51	-63.800	2.86E-06	Bacteria;	1;	Tenericutes;	1;	Mollicutes;	1;	Mycoplasmatales;	1;	Mycoplasmataceae;	1;	Ureaplasma;	1;
1142	WT.N.M.Ile	BD.N.M.Ile	8	3	0	83	134.47	3839.37	-46.724	2.87E-06	Bacteria;	1;	Bacteroidetes;	0.96;	Bacteroidia;	0.53;	Bacteroidales;	0.53;	Porphyromonadaceae;	0.42;	Paludibacter;	0.23;
592	WT.N.M.Ile	BD.N.M.Ile	8	3	0	38	134.47	1831.75	-36.118	2.87E-06	Bacteria;	1;	Firmicutes;	0.97;	Clostridia;	0.96;	Clostridiales;	0.95;	Ruminococcaceae;	0.94;	Pseudoflavonifractor;	0.56;
100	KO.E.F.Ile	WT.E.F.Ile	6	6	0	4631	133.59	103626.90	-94.290	7.05E-06	Bacteria;	1;	Firmicutes;	1;	Bacilli;	1;	Lactobacillales;	1;	Enterococcaceae;	0.99;	Enterococcus;	0.98;
150	KO.E.F.Ile	WT.E.F.Ile	6	6	740	1	16711.24	156.35	65.943	1.32E-05	Bacteria;	1;	Firmicutes;	1;	Bacilli;	1;	Bacillales;	1;	Staphylococcaceae;	0.99;	Staphylococcus;	0.99;
298	WT.N.M.Ile	KO.N.M.Ile	8	8	0	248	134.47	4285.81	-48.304	1.46E-05	Bacteria;	1;	Firmicutes;	1;	Clostridia;	1;	Clostridiales;	1;	Lachnospiraceae;	0.97;	Lachnobacterium;	0.53;
13	WT.N.M.Ile	BD.N.M.Ile	8	3	0	79	134.47	3661.01	-46.041	1.55E-05	Bacteria;	1;	Bacteroidetes;	0.99;	Bacteroidia;	0.84;	Bacteroidales;	0.84;	Marinilabiaceae;	0.5;	Anaerophaga;	0.5;
40	WT.N.M.Ile	KO.N.M.Ile	8	8	0	1020	134.47	17209.49	-68.317	2.09E-05	Bacteria;	0.99;	Firmicutes;	0.81;	Erysipelotrichia;	0.64;	Erysipelotrichales;	0.64;	Erysipelotrichaceae;	0.64;	Allobaculum;	0.22;
637	KO.E.F.Ile	WT.E.F.Ile	6	6	0	270	133.59	6160.62	-53.606	3.24E-05	Bacteria;	1;	Firmicutes;	1;	Bacilli;	1;	Lactobacillales;	1;	Enterococcaceae;	1;	Enterococcus;	1;
393	WT.N.M.Ile	KO.N.M.Ile	8	8	172	0	3027.23	134.16	43.340	3.37E-05	Bacteria;	1;	Firmicutes;	1;	Bacilli;	1;	Lactobacillales;	1;	Streptococcaceae;	1;	Lactococcus;	1;
39	WT.N.M.Ile	KO.N.M.Ile	8	8	0	172	134.47	3013.53	-43.247	3.37E-05	Bacteria;	1;	Firmicutes;	0.96;	Clostridia;	0.94;	Clostridiales;	0.94;	Lachnospiraceae;	0.78;	Coprococcus;	0.24;
141	WT.N.M.Ile	KO.N.M.Ile	8	8	3	360	184.92	6161.53	-49.371	5.40E-05	Bacteria;	1;	Bacteroidetes;	0.94;	Sphingobacteria;	0.36;	Sphingobacteriales;	0.36;	Flammeovirgaceae;	0.2;	Limibacter;	0.12;
13	WT.N.F.Ile	BD.N.F.Ile	8	3	2	86	167.56	3973.29	-44.360	5.71E-05	Bacteria;	1;	Bacteroidetes;	0.99;	Bacteroidia;	0.84;	Bacteroidales;	0.84;	Marinilabiaceae;	0.5;	Anaerophaga;	0.5;
364	KO.N.F.Ile	WT.N.F.Ile	8	8	32	762	669.45	12931.62	-42.379	6.67E-05	Bacteria;	1;	Cyanobacteria Chloroplast;	1;	Chloroplast;	1;	Chloroplast;	1;	Streptophyta;	1;		
528	KO.N.F.Ile	WT.N.F.Ile	8	8	0	239	133.89	4135.07	-47.844	6.67E-05	Bacteria;	0.99;	Bacteroidetes;	0.96;	Bacteroidia;	0.65;	Bacteroidales;	0.65;	Marinilabiaceae;	0.33;	Anaerophaga;	0.33;
2053	WT.N.M.Ile	KO.N.M.Ile	8	8	0	167	134.47	2929.83	-42.843	7.74E-05	Bacteria;	1;	Firmicutes;	0.71;	Clostridia;	0.68;	Clostridiales;	0.65;	Ruminococcaceae;	0.21;	Ethanoligenens;	0.13;
832	WT.N.M.Ile	BD.N.M.Ile	8	3	0	88	134.47	4063.20	-47.538	8.76E-05	Bacteria;	1;	Bacteroidetes;	0.98;	Bacteroidia;	0.81;	Bacteroidales;	0.81;	Porphyromonadaceae;	0.7;	Paludibacter;	0.35;
59	WT.N.M.Ile	BD.N.M.Ile	8	3	39	2427	790.39	108481.71	-70.707	8.76E-05	Bacteria;	1;	Bacteroidetes;	0.95;	Bacteroidia;	0.47;	Bacteroidales;	0.47;	Porphyromonadaceae;	0.43;	Paludibacter;	0.25;
2282	WT.N.M.Ile	BD.N.M.Ile	8	3	6	38	235.38	1830.50	-28.739	8.76E-05	Bacteria;	1;	Bacteroidetes;	1;	Flavobacteria;	1;	Flavobacteriales;	1;	Flavobacteriaceae;	1;	Cloacibacterium;	0.76;
1665	KO.E.F.Ile	WT.E.F.Ile	6	6	133	0	3084.93	134.03	43.627	9.07E-05	Bacteria;	1;	Verrucomicrobia;	1;	Verrucomicrobiae;	1;	Verrucomicrobiales;	1;	Verrucomicrobiaceae;	1;	Akkermansia;	1;
1069	KO.E.F.Ile	WT.E.F.Ile	6	6	355	0	8018.45	134.03	57.357	9.07E-05	Bacteria;	1;	Bacteroidetes;	0.97;	Bacteroidia;	0.63;	Bacteroidales;	0.63;	Marinilabiaceae;	0.28;	Anaerophaga;	0.28;
4350	WT.N.M.Ile	KO.N.M.Ile	8	8	1	131	151.29	2327.17	-38.022	9.67E-05	Bacteria;	1;	Firmicutes;	0.96;	Clostridia;	0.94;	Clostridiales;	0.94;	Lachnospiraceae;	0.67;	Lachnobacterium;	0.08;
63	KO.N.F.Ile	WT.N.F.Ile	8	8	548	0	9310.04	134.08	59.502	0.000104	Bacteria;	1;	Firmicutes;	0.98;	Erysipelotrichia;	0.67;	Erysipelotrichales;	0.67;	Erysipelotrichaceae;	0.67;	Allobaculum;	0.33;
592	KO.N.M.Ile	BD.N.M.Ile	8	3	4	38	201.12	1831.75	-30.844	0.000104	Bacteria;	1;	Firmicutes;	0.97;	Clostridia;	0.96;	Clostridiales;	0.95;	Ruminococcaceae;	0.94;	Pseudoflavonifractor;	0.56;

Table A.16.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the phylum level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %), for samples grouped by SI (ileum and jejunum) and LI (caecum and colon). For each DA phylum level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each phylum and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Bacillales	WT.N.M.SI	BD.N.M.SI	11	6	155	0	3809.88	259.49	3.7182	1.95E-25
Mycoplasmatales	KO.N.M.LI	KO.N.M.SI	16	11	0	1757	266.29	41112.10	-7.1019	1.15E-15
Mycoplasmatales	WT.N.M.LI	WT.N.M.SI	15	11	0	2122	267.04	49852.60	-7.3761	1.25E-15
Bacillales	KO.N.F.SI	KO.E.F.SI	11	6	0	641	270.16	29216.74	-6.5912	1.37E-15
Actinomycetales	WT.N.M.LI	WT.N.M.SI	15	11	0	303	267.04	7198.46	-4.5923	6.07E-15
Enterobacteriales	WT.E.M.LI	WT.E.M.SI	12	6	0	412	268.32	18611.89	-5.9502	9.81E-14
Enterobacteriales	WT.N.M.LI	WT.N.M.SI	15	11	0	336	267.04	8386.51	-4.8084	2.80E-13
Pseudomonadales	WT.N.M.LI	WT.N.M.SI	15	11	0	726	267.04	17805.59	-5.8921	2.80E-13
Pseudomonadales	WT.E.M.LI	WT.E.M.SI	12	6	0	919	268.32	41118.45	-7.0923	3.94E-13
Pasteurellales	WT.N.F.SI	BD.N.F.SI	11	6	0	146	269.21	6867.84	-4.5132	9.35E-13
Enterobacteriales	KO.N.M.LI	KO.N.M.SI	16	11	0	296	266.29	7261.28	-4.6060	9.66E-13
Pasteurellales	KO.N.F.SI	BD.N.F.SI	11	6	0	146	270.16	6867.84	-4.5092	9.90E-13
Actinomycetales	WT.N.M.SI	KO.N.M.SI	11	11	303	0	7198.46	260.50	4.6248	1.40E-11
Campylobacterales	WT.N.M.LI	WT.N.M.SI	15	11	93026	530	1664925.14	12244.29	7.0826	1.58E-10
Verrucomicrobiales	KO.N.F.SI	KO.E.F.SI	11	6	46	1335	1403.36	59243.55	-5.3673	4.45E-10
Campylobacterales	KO.E.M.LI	KO.E.M.SI	12	6	73522	9	1655835.35	665.68	11.2166	1.21E-09
Mycoplasmatales	WT.N.F.SI	WT.E.F.SI	11	6	0	683	269.21	30502.27	-6.6577	1.36E-09
Bacillales	WT.E.F.SI	KO.E.F.SI	6	6	0	641	269.52	29216.74	-6.5937	1.65E-09
Verrucomicrobiales	WT.E.F.SI	KO.E.F.SI	6	6	0	1335	269.52	59243.55	-7.6130	1.65E-09
Bacillales	KO.E.F.LI	KO.E.F.SI	6	6	0	641	263.54	29216.74	-6.6224	2.32E-09
Desulfovibrionales	KO.E.M.LI	KO.E.M.SI	12	6	0	383	265.97	15783.82	-5.7261	2.58E-09
Campylobacterales	WT.E.M.LI	WT.E.M.SI	12	6	63144	21	1415414.82	1179.07	10.1863	4.04E-09
Bacillales	WT.N.M.LI	WT.N.M.SI	15	11	0	155	267.04	3809.88	-3.6812	6.96E-09
Xanthomonadales	WT.N.M.LI	WT.N.M.SI	15	11	0	91	267.04	2502.79	-3.0699	7.20E-09
Lactobacillales	WT.N.M.LI	WT.E.M.LI	15	12	2448	0	43253.13	268.32	7.1652	1.53E-08
Desulfovibrionales	KO.N.M.SI	KO.E.M.SI	11	6	0	383	260.50	15783.82	-5.7522	1.78E-08
Lactobacillales	WT.E.M.LI	WT.E.M.SI	12	6	0	1080	268.32	48641.60	-7.3344	2.16E-08
Chloroplast	WT.N.F.LI	WT.N.F.SI	8	11	0	2677	269.43	65625.45	-7.7609	3.29E-08
Deferribacterales	WT.N.M.LI	WT.E.M.LI	15	12	0	251	267.04	5580.13	-4.2268	1.17E-07
Desulfovibrionales	WT.E.F.SI	KO.E.F.SI	6	6	0	1110	269.52	47247.72	-7.2871	1.18E-07

Table A.17.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the class level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %), for samples grouped by SI (ileum and jejunum) and LI (caecum and colon). For each DA order level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each class and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Mollicutes	KO.N.M.LI	KO.N.M.SI	16	11	0	1757	249.07	39438.49	-7.1387	1.37E-15
Mollicutes	WT.N.M.LI	WT.N.M.SI	15	11	0	2149	248.03	47626.61	-7.4161	1.47E-15
Gammaproteobacteria	WT.N.M.LI	WT.N.M.SI	15	11	0	1167	248.03	26996.92	-6.5975	2.21E-13
Actinobacteria	WT.N.M.LI	WT.N.M.SI	15	11	0	303	248.03	7004.14	-4.6561	7.29E-13
Gammaproteobacteria	WT.E.M.LI	WT.E.M.SI	12	6	0	1355	250.64	57759.03	-7.6807	1.56E-12
Epsilonproteobacteria	WT.N.M.LI	WT.N.M.SI	15	11	93026	530	1555530.98	11173.39	7.1159	3.90E-11
Epsilonproteobacteria	KO.E.M.LI	KO.E.M.SI	12	6	73522	9	1530945.83	626.01	11.1971	3.05E-10
Verrucomicrobiae	KO.N.F.SI	KO.E.F.SI	11	6	46	1335	1326.29	53667.10	-5.3099	1.21E-09
Epsilonproteobacteria	WT.E.M.LI	WT.E.M.SI	12	6	63144	21	1327280.16	1129.42	10.1571	1.79E-09
Mollicutes	WT.N.F.SI	WT.E.F.SI	11	6	0	683	253.89	28342.30	-6.6382	1.91E-09
Verrucomicrobiae	WT.E.F.SI	KO.E.F.SI	6	6	0	1335	253.81	53667.10	-7.5588	3.80E-09
Gammaproteobacteria	WT.E.M.LI	KO.E.M.LI	12	12	0	744	250.64	14628.35	-5.7026	8.52E-09
Deltaproteobacteria	KO.E.M.LI	KO.E.M.SI	12	6	0	386	245.96	14157.43	-5.6812	9.17E-09
Chloroplast	WT.N.F.LI	WT.N.F.SI	8	11	0	2677	251.27	61284.46	-7.7629	3.24E-08
Bacilli	WT.E.M.LI	WT.E.M.SI	12	6	0	1098	250.64	47142.00	-7.3878	3.47E-08
Bacilli	WT.N.M.LI	WT.E.M.LI	15	12	2451	0	39902.02	250.64	7.1475	4.81E-08
Deltaproteobacteria	KO.N.M.SI	KO.E.M.SI	11	6	0	386	243.88	14157.43	-5.6920	6.36E-08
Deferribacteres	WT.N.M.LI	WT.E.M.LI	15	12	0	251	248.03	5224.73	-4.2376	1.35E-07
Verrucomicrobiae	KO.N.F.LI	KO.E.F.LI	8	6	0	325	248.46	13534.36	-5.6007	1.59E-07
Bacilli	WT.E.M.LI	KO.E.M.LI	12	12	0	1432	250.64	27246.59	-6.5980	2.16E-07
Verrucomicrobiae	KO.E.M.LI	KO.E.M.SI	12	6	0	147	245.96	5540.33	-4.3363	3.72E-07
Deltaproteobacteria	WT.E.F.SI	KO.E.F.SI	6	6	0	1134	253.81	45194.49	-7.3111	3.83E-07
Clostridia	WT.N.F.LI	WT.N.F.SI	8	11	27694	424	866431.47	9582.60	6.4932	3.83E-07
Bacteroidia	WT.N.F.LI	WT.N.F.SI	8	11	1223	0	38294.12	253.89	7.0716	3.83E-07
Deltaproteobacteria	KO.E.F.LI	KO.E.F.SI	6	6	0	1134	244.10	45194.49	-7.3610	6.48E-07
Verrucomicrobiae	KO.N.M.SI	KO.E.M.SI	11	6	0	147	243.88	5540.33	-4.3471	1.33E-06
Epsilonproteobacteria	KO.N.F.LI	KO.N.F.SI	8	11	367	0	11689.94	251.46	5.3748	2.28E-06
Epsilonproteobacteria	KO.N.M.LI	KO.N.M.SI	16	11	75351	2136	1182034.57	46718.09	4.6601	3.25E-06
Gammaproteobacteria	WT.N.M.SI	BD.N.M.SI	11	6	1167	0	26996.92	243.53	6.6212	3.96E-06
Mollicutes	WT.N.M.SI	BD.N.M.SI	11	6	2149	6	47626.61	500.55	6.4953	3.96E-06

Table A.18.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the order level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %), for samples grouped by SI (ileum and jejunum) and LI (caecum and colon). For each DA order level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each order and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Verrucomicrobia	KO.N.F.SI	KO.E.F.SI	11	6	0	1335	208.41	37890.04	-7.3430	7.74E-17
Tenericutes	KO.N.M.LI	KO.N.M.SI	16	11	0	1741	216.65	34026.50	-7.1299	3.77E-16
Tenericutes	WT.N.M.LI	WT.N.M.SI	15	11	0	2125	216.82	39688.00	-7.3508	6.01E-16
Actinobacteria	WT.N.M.LI	WT.N.M.SI	15	11	0	303	216.82	6136.65	-4.6613	8.02E-13
Verrucomicrobia	WT.E.F.SI	KO.E.F.SI	6	6	0	1335	193.53	37890.04	-7.4880	2.92E-10
Verrucomicrobia	KO.N.F.LI	KO.E.F.LI	8	6	0	325	215.98	11905.49	-5.6205	1.37E-08
Tenericutes	WT.N.F.SI	WT.E.F.SI	11	6	0	683	209.48	14513.76	-5.9529	2.18E-08
Cyanobacteria/Chloroplast	WT.N.F.LI	WT.N.F.SI	8	11	0	2677	216.71	50793.25	-7.7070	2.30E-08
Verrucomicrobia	KO.E.M.LI	KO.E.M.SI	12	6	0	147	218.23	5428.51	-4.4813	5.47E-08
Verrucomicrobia	KO.N.M.SI	KO.E.M.SI	11	6	0	147	213.01	5428.51	-4.5144	1.30E-07
Deferribacteres	WT.N.M.LI	WT.E.M.LI	15	12	0	251	216.82	4661.96	-4.2692	1.70E-07
Tenericutes	KO.N.M.SI	KO.E.M.SI	11	6	1741	0	34026.50	221.83	7.1001	4.41E-07
Tenericutes	WT.N.M.SI	BD.N.M.SI	11	6	2125	0	39688.00	223.69	7.3150	4.81E-07
Verrucomicrobia	WT.E.F.LI	KO.E.F.LI	6	6	0	325	216.19	11905.49	-5.6193	8.81E-07
Tenericutes	KO.N.M.SI	BD.N.M.SI	11	6	1741	0	34026.50	223.69	7.0931	9.06E-07
TM7	KO.N.F.LI	KO.E.F.LI	8	6	0	190	215.98	7019.73	-4.8606	1.60E-06
Proteobacteria	WT.N.M.LI	WT.N.M.SI	15	11	93458	1834	1368628.91	33217.44	5.3635	2.69E-06
TM7	KO.N.F.SI	KO.E.F.SI	11	6	0	126	208.41	3641.71	-3.9855	6.30E-06
TM7	WT.N.F.SI	KO.N.F.SI	11	11	157	0	3276.98	208.41	3.8183	6.40E-06
Actinobacteria	WT.N.M.SI	WT.E.M.SI	11	6	303	0	6136.65	226.49	4.6056	8.44E-06
Cyanobacteria/Chloroplast	KO.N.F.LI	KO.N.F.SI	8	11	0	640	215.98	12573.07	-5.6990	1.36E-05
Tenericutes	WT.E.F.SI	KO.E.F.SI	6	6	683	0	14513.76	206.74	5.9687	1.96E-05
Tenericutes	WT.E.M.LI	WT.E.M.SI	12	6	0	128	216.80	4884.05	-4.3365	2.52E-05
TM7	WT.E.F.LI	KO.E.F.LI	6	6	0	190	216.19	7019.73	-4.8594	2.96E-05
Tenericutes	WT.E.F.LI	WT.E.F.SI	6	6	0	683	216.19	14513.76	-5.9069	4.73E-05
Cyanobacteria/Chloroplast	KO.N.F.SI	KO.E.F.SI	11	6	640	0	12573.07	206.74	5.7591	6.59E-05
TM7	WT.N.F.LI	WT.N.F.SI	8	11	0	157	216.71	3276.98	-3.7616	8.45E-05
Verrucomicrobia	WT.E.M.SI	KO.E.M.SI	6	6	0	147	226.49	5428.51	-4.4335	0.000116
Deferribacteres	WT.N.M.LI	KO.N.M.LI	15	16	0	138	216.82	2072.77	-3.1097	0.000122
Actinobacteria	KO.N.M.LI	KO.N.M.SI	16	11	0	71	216.65	1332.54	-2.5165	0.00016

Table A.19.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the family level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %), for samples grouped by SI (ileum and jejunum) and LI (caecum and colon). For each DA family level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each family and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Staphylococcaceae	KO.N.F.SI	KO.E.F.SI	11	6	0	641	313.17	34447.82	-6.6199	2.04E-20
Mycoplasmataceae	KO.N.M.LI	KO.N.M.SI	16	11	0	1757	297.12	47461.42	-7.1502	9.52E-16
Mycoplasmataceae	WT.N.M.LI	WT.N.M.SI	15	11	0	2122	302.26	56370.19	-7.3760	1.51E-15
Prevotellaceae	WT.N.M.LI	KO.N.M.LI	15	16	0	2379	302.26	42101.82	-6.9555	3.83E-15
Propionibacteriaceae	WT.N.M.LI	WT.N.M.SI	15	11	0	215	302.26	6113.16	-4.1803	5.84E-14
Enterobacteriaceae	WT.E.M.LI	WT.E.M.SI	12	6	0	412	303.76	21246.02	-5.9639	9.44E-14
Enterobacteriaceae	WT.N.M.LI	WT.N.M.SI	15	11	0	336	302.26	9478.32	-4.8086	3.64E-13
Pseudomonadaceae	WT.N.M.LI	WT.N.M.SI	15	11	0	725	302.26	20081.56	-5.8886	4.44E-13
Pseudomonadaceae	WT.E.M.LI	WT.E.M.SI	12	6	0	918	303.76	46829.06	-7.1024	5.69E-13
Enterobacteriaceae	KO.N.M.LI	KO.N.M.SI	16	11	0	296	297.12	8287.83	-4.6383	7.64E-13
Pasteurellaceae	KO.N.F.SI	BD.N.F.SI	11	6	0	146	313.17	7872.65	-4.4997	7.64E-13
Pasteurellaceae	WT.N.F.SI	BD.N.F.SI	11	6	0	146	314.44	7872.65	-4.4937	8.29E-13
Prevotellaceae	KO.N.F.LI	KO.N.F.SI	8	11	2225	0	79880.64	313.17	7.8325	1.06E-12
Staphylococcaceae	KO.E.F.LI	KO.E.F.SI	6	6	0	641	279.46	34447.82	-6.7646	1.47E-12
Enterococcaceae	WT.E.F.SI	KO.E.F.SI	6	6	4841	0	230530.90	312.67	9.3633	1.79E-12
Staphylococcaceae	WT.E.F.SI	KO.E.F.SI	6	6	0	641	307.24	34447.82	-6.6451	1.79E-12
Enterococcaceae	WT.E.F.LI	WT.E.F.SI	6	6	0	4841	294.53	230530.90	-9.4398	2.43E-12
Enterococcaceae	WT.N.F.SI	KO.N.F.SI	11	11	701	0	20513.58	313.17	5.8732	2.02E-11
Helicobacteraceae	WT.N.M.LI	WT.N.M.SI	15	11	93026	530	1932421.62	13921.79	7.1121	5.21E-11
Propionibacteriaceae	WT.N.M.SI	KO.N.M.SI	11	11	215	0	6113.16	296.08	4.2059	7.89E-11
Helicobacteraceae	KO.E.M.LI	KO.E.M.SI	12	6	73521	9	1922255.76	778.77	11.2115	4.86E-10
Verrucomicrobiaceae	KO.N.F.SI	KO.E.F.SI	11	6	46	1335	1669.67	67775.16	-5.3180	8.16E-10
Prevotellaceae	WT.N.F.LI	KO.N.F.LI	8	8	0	2225	303.15	79880.64	-7.8740	1.38E-09
Verrucomicrobiaceae	WT.E.F.SI	KO.E.F.SI	6	6	0	1335	307.24	67775.16	-7.6209	1.43E-09
Helicobacteraceae	WT.E.M.LI	WT.E.M.SI	12	6	63144	21	1637626.27	1347.40	10.2048	1.76E-09
Mycoplasmataceae	WT.N.F.SI	WT.E.F.SI	11	6	0	683	314.44	33341.17	-6.5677	2.95E-09
Xanthomonadaceae	WT.N.M.LI	WT.N.M.SI	15	11	0	91	302.26	2943.08	-3.1202	5.59E-09
Bacteroidaceae	WT.E.M.LI	KO.E.M.LI	12	12	1490	0	35970.35	305.35	6.7155	9.48E-09
Lactobacillaceae	WT.E.M.LI	WT.E.M.SI	12	6	0	1039	303.76	53789.40	-7.3017	1.01E-08
Enterococcaceae	WT.N.F.LI	WT.N.F.SI	8	11	0	701	303.15	20513.58	-5.9146	1.08E-08

Table A.20.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the genus level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %), for samples grouped by SI (ileum and jejunum) and LI (caecum and colon). For each DA the genus level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR, for each genus and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Staphylococcus	KO.N.F.SI	KO.E.F.SI	11	6	0	641	471.24	50387.96	-6.5745	3.60E-21
Ureaplasma	WT.N.F.SI	WT.E.F.SI	11	6	0	448	471.65	34558.99	-6.0307	2.09E-19
Ureaplasma	KO.N.M.SI	BD.N.M.SI	11	6	0	0	467.03	463.55	0.0097	3.58E-19
Ureaplasma	WT.N.M.SI	BD.N.M.SI	11	6	0	0	465.19	463.55	0.0045	3.58E-19
Ureaplasma	WT.N.M.LI	WT.E.M.LI	15	12	0	0	467.71	467.34	0.0010	3.58E-19
Staphylococcus	KO.N.M.SI	BD.N.M.SI	11	6	0	0	467.03	463.55	0.0097	1.39E-18
Staphylococcus	WT.N.M.SI	BD.N.M.SI	11	6	0	0	465.19	463.55	0.0045	1.39E-18
Staphylococcus	WT.N.M.LI	WT.E.M.LI	15	12	0	0	467.71	467.34	0.0010	1.39E-18
Mycoplasma	WT.N.M.LI	WT.N.M.SI	15	11	0	2103	467.71	88665.31	-7.3986	8.68E-15
Propionibacterium	WT.N.M.LI	WT.N.M.SI	15	11	0	215	467.71	9545.34	-4.1910	3.20E-13
Lactococcus	WT.N.M.LI	KO.N.M.LI	15	16	659	0	20964.07	465.42	5.3276	4.80E-13
Staphylococcus	WT.E.F.SI	KO.E.F.SI	6	6	0	641	466.72	50387.96	-6.5870	9.12E-13
Enterococcus	WT.E.F.SI	KO.E.F.SI	6	6	4836	0	369625.90	471.02	9.4484	9.12E-13
Staphylococcus	KO.E.F.LI	KO.E.F.SI	6	6	0	641	453.72	50387.96	-6.6230	1.13E-12
Enterococcus	WT.E.F.LI	WT.E.F.SI	6	6	0	4836	453.66	369625.90	-9.4965	1.47E-12
Helicobacter	WT.N.M.LI	WT.N.M.SI	15	11	92834	528	2911106.30	22321.86	7.0228	1.60E-12
Ureaplasma	WT.E.F.SI	KO.E.F.SI	6	6	448	0	34558.99	471.02	6.0324	4.61E-12
Ureaplasma	WT.E.F.LI	WT.E.F.SI	6	6	0	448	453.66	34558.99	-6.0805	4.66E-12
Pseudomonas	WT.E.M.LI	WT.E.M.SI	12	6	0	892	467.34	69994.28	-7.0587	1.20E-11
Enterococcus	WT.N.F.SI	KO.N.F.SI	11	11	701	0	30202.54	471.24	5.8372	2.05E-11
Mycoplasma	KO.N.M.LI	KO.N.M.SI	16	11	33	1757	1394.34	74640.71	-5.6795	7.31E-11
Helicobacter	KO.E.M.LI	KO.E.M.SI	12	6	73314	23	2876899.82	2270.23	10.2711	9.24E-11
Lactococcus	WT.N.M.LI	WT.E.M.LI	15	12	659	0	20964.07	467.34	5.3223	1.40E-10
Streptococcus	KO.N.F.SI	BD.N.F.SI	11	6	0	712	471.24	55833.48	-6.7223	2.02E-10
Helicobacter	WT.E.M.LI	WT.E.M.SI	12	6	63012	43	2469005.00	3831.18	9.3107	7.27E-10
Odoribacter	KO.N.F.LI	BD.N.F.LI	8	3	0	68	453.73	10768.16	-4.4026	8.52E-10
Streptococcus	WT.N.F.SI	KO.N.F.SI	11	11	776	0	33699.95	471.24	5.9948	9.96E-10
Butyricicoccus	KO.N.F.LI	KO.N.F.SI	8	11	202	0	11843.12	471.24	4.4911	2.60E-09
Oscillibacter	KO.N.F.LI	KO.N.F.SI	8	11	219	0	12828.54	471.24	4.6056	1.09E-08
Enterococcus	WT.N.F.LI	WT.N.F.SI	8	11	0	701	453.90	30202.54	-5.8853	1.11E-08

Table A.21.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from the RDP pipeline, at the OTU level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %), for samples grouped by SI (ileum and jejunum) and LI (caecum and colon). For each DA the OTU level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, the statistical result parameters of edgeR, the logFC and the FDR for each OTU and contrast are included. For each OTU, the RDP classification result and confidence is available from the table.

TAX	G1	G2	number samples G1	number samples G2	raw counts G1	raw counts G2	cpm G1	cpm G2	logFC	FDR	Domain	Domain Prob	Phylum	Phylum Prob	Class	Class Prob	Order	Order Prob	Familiy	Family Prob	Genus	Genus Prob
459	WT.N.M.LI	BD.N.M.LI	15	6	0	147	59.93	1528.41	-4.5097	4.98E-53	Bacteria;	1;	Firmicutes;	0.99;	Clostridia;	0.99;	Clostridiales;	0.99;	Ruminococcaceae;	0.99;	Oscillibacter;	0.54;
143	KO.N.F.LI	BD.N.F.LI	8	3	0	143	59.92	2915.59	-5.4382	2.69E-51	Bacteria;	1;	Firmicutes;	0.98;	Clostridia;	0.96;	Clostridiales;	0.96;	Geosporobacter;	0.21;		
271	WT.N.M.LI	BD.N.M.LI	15	6	1	2936	63.92	29389.33	-8.6852	1.11E-48	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
271	KO.N.M.LI	BD.N.M.LI	16	6	1	2936	63.68	29389.33	-8.6902	7.35E-44	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
143	BD.N.F.SI	BD.N.F.LI	6	3	0	143	59.94	2915.59	-5.4379	9.06E-43	Bacteria;	1;	Firmicutes;	0.98;	Clostridia;	0.96;	Clostridiales;	0.96;	Geosporobacter;	0.21;		
1414	KO.N.M.LI	KO.N.M.SI	16	11	3	0	71.17	59.94	0.2223	8.70E-41	Bacteria;	0.99;	Bacteroidetes;	0.97;	Sphingobacteria;	0.68;	Sphingobacteriales;	0.68;	Sphingobacteriaceae;	0.3;	Pseudosphingobacterium;	0.3;
143	WT.N.F.LI	BD.N.F.LI	8	3	15	143	172.31	2915.59	-4.0231	2.49E-39	Bacteria;	1;	Firmicutes;	0.98;	Clostridia;	0.96;	Clostridiales;	0.96;	Geosporobacter;	0.21;		
354	WT.N.F.LI	BD.N.F.LI	8	3	2	94	74.91	1936.78	-4.5605	3.58E-37	Bacteria;	1;	Firmicutes;	0.99;	Clostridia;	0.98;	Clostridiales;	0.98;	Lachnospiraceae;	0.77;	Lachnobacterium;	0.33;
354	KO.N.F.LI	BD.N.F.LI	8	3	5	94	97.34	1936.78	-4.2126	1.03E-34	Bacteria;	1;	Firmicutes;	0.99;	Clostridia;	0.98;	Clostridiales;	0.98;	Lachnospiraceae;	0.77;	Lachnobacterium;	0.33;
785	WT.N.F.LI	BD.N.F.LI	8	3	0	124	59.92	2535.45	-5.2373	9.39E-34	Bacteria;	1;	Firmicutes;	0.92;	Clostridia;	0.92;	Clostridiales;	0.92;	Ruminococcaceae;	0.61;	Sporobacter;	0.19;
785	KO.N.F.LI	BD.N.F.LI	8	3	0	124	59.92	2535.45	-5.2373	9.39E-34	Bacteria;	1;	Firmicutes;	0.92;	Clostridia;	0.92;	Clostridiales;	0.92;	Ruminococcaceae;	0.61;	Sporobacter;	0.19;
354	BD.N.F.SI	BD.N.F.LI	6	3	0	94	59.94	1936.78	-4.8498	4.74E-33	Bacteria;	1;	Firmicutes;	0.99;	Clostridia;	0.98;	Clostridiales;	0.98;	Lachnospiraceae;	0.77;	Lachnobacterium;	0.33;
799	WT.N.M.LI	BD.N.M.LI	15	6	14	196	115.87	2017.89	-4.0372	1.01E-32	Bacteria;	1;	Bacteroidetes;	1;	Bacteroidia;	0.73;	Bacteroidales;	0.73;	Rikenellaceae;	0.73;	Alistipes;	0.73;
271	WT.N.F.LI	BD.N.F.LI	8	3	0	8696	59.92	173697.13	-11.3312	2.65E-32	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
271	KO.N.F.LI	BD.N.F.LI	8	3	0	8696	59.92	173697.13	-11.3312	2.65E-32	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
587	WT.N.F.LI	BD.N.F.LI	8	3	0	116	59.92	2375.88	-5.1438	3.63E-32	Bacteria;	1;	Firmicutes;	0.93;	Clostridia;	0.8;	Clostridiales;	0.78;	Ruminococcaceae;	0.18;	Pseudoflavonifractor;	0.1;
587	KO.N.F.LI	BD.N.F.LI	8	3	0	116	59.92	2375.88	-5.1438	3.63E-32	Bacteria;	1;	Firmicutes;	0.93;	Clostridia;	0.8;	Clostridiales;	0.78;	Ruminococcaceae;	0.18;	Pseudoflavonifractor;	0.1;
1414	WT.N.M.LI	BD.N.M.LI	15	6	17	132	127.86	1378.56	-3.3562	4.76E-32	Bacteria;	0.99;	Bacteroidetes;	0.97;	Sphingobacteria;	0.68;	Sphingobacteriales;	0.68;	Sphingobacteriaceae;	0.3;	Pseudosphingobacterium;	0.3;
586	WT.N.F.LI	BD.N.F.LI	8	3	0	199	59.92	4031.98	-5.9050	7.16E-32	Bacteria;	1;	Firmicutes;	0.98;	Clostridia;	0.97;	Clostridiales;	0.97;	Anaerosporobacter;	0.34;		
586	KO.N.F.LI	BD.N.F.LI	8	3	0	199	59.92	4031.98	-5.9050	7.16E-32	Bacteria;	1;	Firmicutes;	0.98;	Clostridia;	0.97;	Clostridiales;	0.97;	Anaerosporobacter;	0.34;		
271	BD.N.M.SI	KO.N.M.SI	6	11	2512	3	25153.75	76.28	8.2305	1.08E-28	Bacteria;	1;	Firmicutes;	0.95;	Clostridia;	0.94;	Clostridiales;	0.92;	Lachnospiraceae;	0.81;	Marvinbryantia;	0.23;
1065	WT.N.M.LI	BD.N.M.LI	15	6	0	108	59.93	1138.81	-4.0876	2.54E-28	Bacteria;	1;	Firmicutes;	0.86;	Clostridia;	0.86;	Clostridiales;	0.86;	Peptococcus;	0.36;		
2901	WT.N.F.LI	BD.N.F.LI	8	3	0	119	59.92	2436.02	-5.1797	4.17E-27	Bacteria;	1;	Firmicutes;	1;	Clostridia;	1;	Clostridiales;	1;	Anaerosporobacter;	0.1;		
2901	KO.N.F.LI	BD.N.F.LI	8	3	0	119	59.92	2436.02	-5.1797	4.17E-27	Bacteria;	1;	Firmicutes;	1;	Clostridia;	1;	Clostridiales;	1;	Anaerosporobacter;	0.1;		
1163	WT.N.M.LI	BD.N.M.LI	15	6	0	93	59.93	988.97	-3.8855	1.88E-26	Bacteria;	1;	Firmicutes;	1;	Clostridia;	0.98;	Clostridiales;	0.98;	Lachnospiraceae;	0.96;	Coprococcus;	0.42;
799	KO.N.M.LI	BD.N.M.LI	16	6	5	196	78.66	2017.89	-4.5552	1.42E-24	Bacteria;	1;	Bacteroidetes;	1;	Bacteroidia;	0.73;	Bacteroidales;	0.73;	Rikenellaceae;	0.73;	Alistipes;	0.73;
785	WT.N.M.LI	BD.N.M.LI	15	6	0	41	59.93	469.51	-2.8227	3.45E-24	Bacteria;	1;	Firmicutes;	0.92;	Clostridia;	0.92;	Clostridiales;	0.92;	Ruminococcaceae;	0.61;	Sporobacter;	0.19;
2901	WT.N.M.LI	BD.N.M.LI	15	6	0	43	59.93	489.49	-2.8819	3.69E-24	Bacteria;	1;	Firmicutes;	1;	Clostridia;	1;	Clostridiales;	1;	Anaerosporobacter;	0.1;		
2053	KO.N.M.LI	KO.N.M.SI	16	11	0	1256	59.93	6903.70	-6.6796	4.29E-24	Bacteria;	1;	Firmicutes;	0.71;	Clostridia;	0.68;	Clostridiales;	0.65;	Ruminococcaceae;	0.21;	Ethanoligenens;	0.13;
1335	WT.N.M.LI	BD.N.M.LI	15	6	0	198	59.93	2037.87	-4.9230	2.91E-23	Bacteria;	0.99;	Firmicutes;	0.83;	Clostridia;	0.83;	Clostridiales;	0.83;	Peptococcus;	0.36;		

A.1. Sample extraction protocol of GI mouse study

Protokoll Mikrobiom Tierversuch
mdr 2 -/- und mdr 2 +/+
Probenentnahme bei 4 Wochen und 8 Wochen

<u>Proben:</u>

1 Dickdarm:
Gesamten Dickdarm mit Ligaturen abbinden und entfernen

Inzision des Dickdarm und gesamten Darminhalt in steriles Eppendorfröhrchen geben und einfrieren.

Anschl. Ligaturen lösen und Dickdarm längs aufschneiden. Dickdarm in 10ml sterilem NaCl 0,9% 2xig waschen und von Stuhlresten befreien (nur in NaCl Lösung spülen, nicht aber mit Instrument ausstreichen!) . Dickdarm in einem Teil einfrieren.

2 Dünndarm:

Dünndarm mit Ligaturen in proximalen und distalen Dünndarm unterteilen und abbinden und prox. und distalen in einem Teil entfernen.

a. Inzision des prox. Dünndarm und gesamten Darminhalt in steriles Eppendorfröhrchen geben und einfrieren.
Anschl. Ligaturen lösen und prox. Dünndarm längs aufschneiden. Dünndarm in 10ml sterilem NaCl 0,9% 2xig waschen und von Stuhlresten befreien (nur in NaCl Lösung spülen, nicht aber mit Instrument ausstreichen!). Prox. Dünndarm in einem Teil einfrieren.

b. Inzision des dist. Dünndarm und Darminhalt in steriles Eppendorfröhrchen geben und einfrieren.
Anschl. Ligaturen lösen und dist. Dünndarm längs aufschneiden. Dünndarm in 10ml sterilem NaCl 0,9% 2xig waschen und von Stuhlresten befreien (nur in NaCl Lösung spülen, nicht aber mit Instrument ausstreichen!) Dist. Dünndarm in einem Teil einfrieren

3 Leber:
Einen Leberteil in sterilem Eppendorf einfrieren.

4. Stuhl

Bei Tieren die in Woche 8 getötet werden in Woche 4 eine Stuhlprobe entnehmen und in steriles Eppendorf geben und einfrieren.

<u>Zwischen allen Arbeitsschritten Instrumente mit Wasser von Stuhlresten reinigen und anschl. abflammen.</u>

<u>Alle Proben bei -20°C einfreieren.</u>

A.2. Surgical procedures and mice treatment protocol of GI mouse study

Ad Punkt 14: Art des Eingriffes bzw. der Behandlungen

Dauer: Die in der Folge beschriebenen experimentellen Manipulationen erfolgen bei 25-30 g schweren (ca. 2 Monate alten) Mäusen über den Zeitraum von 3 bis 6 Wochen. Eine genaue Auflistung der Versuchsdauer der unterschiedlichen Eingriffe und Behandlungen ist der beigefügten Tabelle (Tabelle 1) zu entnehmen.

Fütterung von Phosphatidylcholin-angereicherter Nahrung:
Die Fütterung einer Phospholipid-angereicherten Nahrung folgt nach Aufbereitung einer solchen mittels Untermischung von Phospholipiden in eine Maus Standard Diät, wobei eine Verwendung von nahrungsmittelindustriell verwendeten Phospholipiden erfolgt, die ernährungstechnisch für die Versuchstiere unbedenklich sind. Bezüglich der Aufbereitung solcher Diäten liegt im Labor des Antragstellers ausreichend Erfahrung vor.

Gallengangsligatur: Nach Isoflurannarkose wird eine mediane Oberbauch-Laparotomie (ca. 1,5 cm Länge) durchgeführt und die Leberpforte dargestellt. Nach Untertunnelung mit einer Pinzette wird der Gallengang mit sterilen Seidenfäden zwischen den beiden Ligaturen durchtrennt (Trauner et al. 1997, Fickert et al. 2002, Wagner et al. 2003). Zusätzlich wird eine Cholezystektomie durchgeführt um eine Ruptur der Gallenblase als Folge des Gallerückstaus und damit eine biliäre Peritonitis zu verhindern. Danach erfolgt der schichtweise Wundverschluss mit atraumatischem Nahtmaterial. Der operative Eingriff selbst dauert ca. 10-15 Minuten. Um ein intraoperatives Auskühlen der Tiere zu verhindern, werden diese auf einer Heizmatte zur Erhaltung der Körpertemperatur von 37°C platziert. In regelmäßigen Abständen wird der Darm mit 37°C warmen 0.9%-igem NaCl befeuchtet. Um die Austrocknung der Augen zu vermeiden, wird unmittelbar vor der Operation Augensalbe (Oleovit) eingetragen. Bei den schein-ligierten Tieren (Sham) wird wie oben beschrieben eine Oberbauch-Laparatomie mit Untertunnelung des Gallenganges ohne Ligatur und Cholezystektomie durchgeführt. Sämtliche Eingriffe werden zur Vermeidung von Infektionen unter sterilen Bedingungen durchgeführt. Zur postoperativen Schmerztherapie erhalten die Tiere Nolvagin (15 Trpf./100ml Trinkwasser) über sechs Tage. Schein-operierte Kontrollen werden mit Ausnahme der Gallengangsligatur und –durchtrennung völlig identisch behandelt. Die Tiere werden 3 Tage bzw. 4 und 8 Wochen nach der Gallengangsligatur unter Vollnarkose getötet.

Organ-Entnahme: Nach Anästhesie mit Isofluran werden die vollnarkotisierten Tiere durch Dekapitation (Enthauptung) getötet, was der Sammlung von Blut zur weiteren biochemischen

Analyse dient. Nach medianer Laparotomie werden die Leber, die Nieren, sowie entsprechende Darmabschnitte entnommen.

Narkose: Die Tiere werden mittels dem Narkosegas Isofluran über die gesamte Dauer des operativen Eingriffs in tiefer Narkose gehalten. Die Einleitung erfolgt in einer Induktionskammer (Fa. Rothacher), die auf einer Seite den Normanschuss für die Narkoseeinspeisung (Combi-vet®-Anästhesiegerät inklusive elektronischem Durchflussmesser und Verdampfer, Fa. Rothacher), sowie auf andere Seite den Anschluss für die passive Narkosegasabsagung aufweist. Die Durchflussrate beträgt 4 Vol% Isofluran (Forane® Abbot) und 1,5 Sauerstoff/min. Zur Aufrechthaltung der Narkose wird eine Maske und Isofluran 2 Vol % verwendet. Zur Schmerzbehandlung wird Tramadol 20mg/kg (Tramabene®) unmittelbar vor der OP subcutan verabreicht.

Analgesie (postoperativ nach Gallengangsligaturen, Ligatur der A. hepatica und Hepatektomie): Zur präoperativen Schmerzbehandlung wird Tramadol 20mg/kg (Tramabene®) unmittelbar vor der OP subcutan verabreicht. Nach allen chirurgischen Eingriffen ist eine standardisierte postoperative Analgesie der Versuchstiere mittels Nolvagin (15 Trpf./100 ml Trinkwasser) über sechs Tage vorgesehen.

Tötung: Die vollnarkotisierten Tiere werden durch Dekapitation (Enthauptung) getötet, wobei dabei auch Blut zur weitern biochemischen Analytik gesammelt wird. Auf Grund der sehr geringen Gesamt-Blutmenge bei der Maus stellt dies auch eine sehr effiziente Methode zur Blutkollektion dar.

Berechnung der Versuchtier-Anzahl:

Die Versuchsplanung mit genauer Festlegung der notwendigen Tierzahl und des zeitlichen Ablaufes erfolgt in enger Absprache mit den Kollegen vom Institut für Medizinische Informatik, Statistik und Dokumentation (Univ. Prof. Dr. Berghold, Medizinische Universität Graz). Die Versuche erfolgen sequentiell, d.h. primär werden Pilotstudien durchgeführt (siehe unten). Basierend auf unseren langjährigen Erfahrungen mit Tierexperimenten im Bereich der biliären Physiologie und im Bereich der Cholangiopathien unter Verwendung von Mäusen sind zu einer besseren Einschätzung tatsächlich notwendiger Tierversuchszahlen in Pilotexperimenten Versuchszahlen von 5 Mäusen pro Arm ausreichend. Nach Durchführung dieser Untersuchungen kann dann anhand dieser präliminären Daten eine genaue Abschätzung nach Beziehung des Statistikers getroffen werden. Der ausgearbeitete Versuchsaufbau kann bei Bedarf anhand der Daten der Pilotstudien unter Beratung der Statistiker modifiziert werden und ein Nachtrag bzw. weitere Planung an die

Tierversuchbehörde erfolgen. Damit ist eine Minimierung der nötigen Tierzahl zur Erreichung aussagekräftiger Experimente möglich.

Tabelle 1

Experiment	Diät	Versuchsdauer	Anzahl
Wildtyp Maus	Chow	3 Wochen 6 Wochen	10 10
Wildtyp Maus	1% PC	3 Wochen 6 Wochen	10 10
Wildtyp Maus	2,5 % PC	3 Wochen 6 Wochen	10 10
Mdr2-/- Mäuse	1% PC	3 Wochen 6 Wochen	10 10
Mdr2-/- Mäuse	2,5% PC	3 Wochen 6 Wochen	10 10
Mdr2-/- Mäuse	Antibiose	3 Wochen 6 Wochen	10 10
CBDL Wildtyp Mäuse	Chow	3 Wochen	10
CBDL Wildtyp Mäuse	0,5% PC	3 Wochen	10
CBDL Wildtyp Mäuse	1% PC	3 Wochen	10
CBDL Wildtyp Mäuse	2,5% PC	3 Wochen	10
CBDL Wildtyp Mäuse	Antibiose	3 Wochen	10

Literatur (in alphabetischer Reihenfolge):

Fickert P, Zollner G, Fuchsbichler A et al. Ursodeoxycholic acid aggravates bile infarcts in bile duct-ligated and Mdr2 knockout mice via disruption of cholangioles. Gastroenterology 2002;123:1238-51.

Greiner T, Bäckhed F. Effects of the Gut Microbiota on Obesity and Glucose Homeostasis. Trends in Endocrinology and Metabolism. 2011 Apr 22(4):117-23. 23. Epub 2011 Feb 23. Review.

Musso G, Gambino R, Cassader M. Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. Annual Review of Medicine 2011 Feb 18;62:361-80. Review.

Rak K, Rader DJ. Cardiovascular disease: the diet-microbe morbid union. Nature. 2011 Apr 7;472(7341):40-1.

Trauner M, Arrese M, Soroka CJ, Ananthanarayanan M, Koeppel TA, Schlosser SF, Suchy FJ, Keppler D, Boyer JL. The rat canalicular conjugate export pump (Mrp2) is down-regulated in intrahepatic and obstructive cholestasis. Gastroenterology. 1997;113:255-64.

Wagner M, Fickert P, Zollner G, Fuchsbichler A, Silbert D, Tsybrovskyy O, Zatloukal K, Guo GL, Schuetz JD, Gonzalez FJ, Marschall HU, Denk H, Trauner M. Role of farnesoid X receptor in determining hepatic ABC transporter expression and liver injury in bile duct-ligated mice. Gastroenterology. 2003;125:825-38.

Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, Feldstein AE, Britt EB, Fu X, Chung YM, Wu Y, Schauer P, Smith JD, Allayee H, Tang WH, DiDonato JA, Lusis AJ, Hazen SL. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. 2011 Apr 7;472(7341):57-63.

Appendix B.

Supplementary information metagenome moss study

Table B.1.: Summary of the Kolmogorov-Smirnov test results for the metagenome moss study. The raw count data of the selected subsystems for metagenomes of the *S. magellanicum*, higher plants and peat soils, were tested for normal distribution using the one-sample Kolmogorov-Smirnov test.

Metagenomes	# subsystems	p-value	statistic
Peat soils	990	< 2.2e-16	0.9368
Higher platns	990	< 2.2e-16	0.9744
S. magellanicum	198	< 2.2e-16	0.9949

Table B.2.: DA functional subsystems (adj. p-values < 0.05) between the *S. magellanicum* and higher plant metagenomes, part I. The table summarizes the statistical testing result of DA feature detection using limma+voom. The detected subsystems are sorted descending, according the logFC values. Subsystems which were tested as differentially abundant between both, the *S. magellanicum*/plant metagenomes and the *S. magellanicum*/peat soils metagenomes, are highlighted with bold text.

Subsystmes level 1	Subsystems level 2	LogFC	AveExpr	t-val	p-val	Adj. p-val
S. magellanicum/higher plants metagenomes						
Stress response	**Dessication stress**	**-8,52**	**3,88**	**-5,73**	**0,00**	**0,0007**
Dormancy and sporulation	**Spore DNA protection**	**-5,14**	**3,24**	**-4,67**	**0,00**	**0,0022**
Phages, prophages, transposable elements, plasmids	Gene Transfer Agent (GTA)	-4,53	8,11	-11,83	0,00	0,0000
Membrane transport	Protein secretion system, type IV	-3,04	6,74	-9,68	0,00	0,0000
Membrane transport	Protein secretion system, type VII (chaperone/usher pathway, CU)	-2,30	9,94	-11,28	0,00	0,0000
Cofactors, vitamins, prostetic groups, pigments	**Coenzyme B**	**-2,25**	**3,49**	**-2,60**	**0,02**	**0,0473**
Clustering-based subsystems	Putative GGDEF domain protein related to agglutinin secretion	-1,99	6,56	-6,40	0,00	0,0003
Iron acquisition and metabolism	Siderophores	-1,80	9,24	-6,03	0,00	0,0005
Clustering-based subsystems	**Hypothetical associated with RecF**	**-1,74**	**7,76**	**-12,82**	**0,00**	**0,0000**
Clustering-based subsystems	**Tricarboxylate transporter**	**-1,72**	**10,17**	**-10,23**	**0,00**	**0,0000**
Virulence, diesease and defense	Invasion and intracellular resistance	-1,72	5,76	-4,89	0,00	0,0017
Iron acquisition and metabolism	-	-1,62	12,76	-4,82	0,00	0,0019
Membrane transport	Protein secretion system, type I	-1,47	6,63	-7,02	0,00	0,0002
Clustering-based subsystems	**D-tyrosyl-tRNA(Tyr) deacylase (EC 3.1....) cluster**	**-1,33**	**8,89**	**-8,03**	**0,00**	**0,0001**
Membrane transport	-	-1,23	13,24	-4,92	0,00	0,0017
Fatty acids, lipids and isoprenoids	Triacylglycerols	-1,18	7,62	-5,15	0,00	0,0012
Virulence, diesease and defense	Bacteriocins, ribosomally synthesized antibacterial peptides	-0,99	7,20	-5,11	0,00	0,0013
Carbohydrates	Aminosugars	-0,98	10,93	-6,54	0,00	0,0003
Membrane transport	Protein secretion system, type V	-0,84	8,89	-7,12	0,00	0,0002
Cofactors, vitamins, prostetic groups, pigments	Coenzyme M	-0,73	6,35	-2,80	0,02	0,0351
Clustering-based subsystems	**Oxidative stress**	**-0,67**	**7,22**	**-4,06**	**0,00**	**0,0054**
Secondary metabolism	Bacterial cytostatics, differentiation factors and antibiotics	-0,64	7,47	-3,10	0,01	0,0229
Clustering-based subsystems	Biosynthesis of galactoglycans and related lipopolysacharides	-0,63	12,26	-4,91	0,00	0,0017
Clustering-based subsystems	Probably organic hydroperoxide resistance related hypothetical protein	-0,62	8,55	-4,33	0,00	0,0036
Clustering-based subsystems	Probably Ybbk-related hypothetical membrane proteins	-0,59	8,43	-5,38	0,00	0,0010
Stress response	Periplasmic stress	-0,55	9,62	-5,73	0,00	0,0007
Clustering-based subsystems	Three hypotheticals linked to lipoprotein biosynthesis	-0,55	9,02	-7,05	0,00	0,0002
Stress response	**-**	**-0,51**	**11,43**	**-9,31**	**0,00**	**0,0000**
Clustering-based subsystems	Hypothetical in lysine biosynthetic cluster	-0,51	9,79	-7,87	0,00	0,0001
Carbohydrates	Sugar alcohols	-0,46	12,46	-3,99	0,00	0,0059
Motility and chemotaxis	-	-0,46	11,64	-2,86	0,02	0,0332
Clustering-based subsystems	Pyruvate kinase associated cluster	-0,45	9,63	-6,74	0,00	0,0002
Clustering-based subsystems	Hypothetical lipase related to phosphatidate metabolism	-0,44	8,92	-4,26	0,00	0,0040
Nucleosides and nucleotides	Detoxification	-0,43	10,37	-7,46	0,00	0,0001
Clustering-based subsystems	DNA polymerase III epsilon cluster	-0,41	9,52	-5,20	0,00	0,0012
Motility and chemotaxis	Flagellar motility in Prokaryota	-0,40	12,72	-3,84	0,00	0,0074
Cofactors, vitamins, prostetic groups, pigments	Riboflavin, FMN, FAD	-0,39	10,69	-7,33	0,00	0,0001
Virulence, diesease and defense	**-**	**-0,39**	**11,86**	**-5,46**	**0,00**	**0,0009**
Virulence, diesease and defense	Adhesion	-0,39	8,89	-3,83	0,00	0,0074
Clustering-based subsystems	Putative asociate of RNA polymerase sigma-54 factor rpoN	-0,38	10,57	-3,98	0,00	0,0060
Clustering-based subsystems	recX and regulatory cluster	-0,37	8,68	-3,35	0,01	0,0154
Miscellaneous	-	-0,31	12,18	-6,60	0,00	0,0002
Regulation and cell signaling	-	-0,30	13,34	-5,77	0,00	0,0007
Clustering-based subsystems	tRNA sulfuration	-0,25	9,26	-3,27	0,01	0,0175
Cell wall and capsule	-	-0,23	13,52	-3,39	0,01	0,0146
Clustering-based subsystems	Probably GTP or GMP signaling related	-0,22	10,98	-3,00	0,01	0,0266
Cell wall and capsule	Gram-positive cell wall components	-0,21	10,20	-2,70	0,02	0,0415
Respiration	-	-0,20	12,99	-3,27	0,01	0,0175
Clustering-based subsystems	Ribosomal protein L28P relates to a set of uncharacterized proteins	-0,20	10,26	-3,02	0,01	0,0258

Table B.3.: DA functional subsystems (adj. p-values < 0.05) between the *S. magellanicum* and higher plant metagenomes, part I. The table summarizes the statistical testing result of DA feature detection using limma+voom. The detected subsystems are sorted descending, according the logFC values. Subsystems which were tested as differentially abundant between both, the *S. magellanicum*/plant metagenomes and the *S. magellanicum*/peat soils, metagenomes are highlighted with bold text.

Subsystmes level 1	Subsystems level 2	LogFC	AveExpr	t-val	p-val	Adj. p-val
S. magellanicum/higher plants metagenomes						
Photosynthesis	**Electron transport and photophosphorylation**	**1,90**	**7,48**	**9,00**	**0,00**	**0,0000**
Photosynthesis	Light-harvesting complexes	1,76	5,03	4,35	0,00	0,0035
Secondary metabolism	Plant alkaloids	1,54	7,16	3,73	0,00	0,0088
RNA metabolism	**-**	**1,53**	**8,74**	**5,16**	**0,00**	**0,0012**
Clustering-based subsystems	**Related to menaquinone-cytochrome C reductase**	**1,31**	**4,46**	**3,67**	**0,00**	**0,0095**
Clustering-based subsystems	**Carotenoid biosynthesis**	**1,15**	**8,72**	**11,03**	**0,00**	**0,0000**
Virulence, diesease and defense	Toxins and superantigens	1,01	5,37	3,55	0,00	0,0114
Clustering-based subsystems	Sarcosine oxidase	0,93	8,54	4,01	0,00	0,0058
Clustering-based subsystems	Molybdopterin oxidoreductase	0,75	8,66	2,95	0,01	0,0287
Amino acids and derivatives	-	0,75	10,27	3,43	0,01	0,0139
Stress response	Acid stress	0,70	9,11	7,08	0,00	0,0002
Secondary metabolism	Biologically active compounds in metazoan cell defence and differentiation	0,66	9,75	4,17	0,00	0,0046
Phages, prophages, plasmids, transposable elements	Pathogenicity islands	0,66	11,15	8,29	0,00	0,0001
Respiration	ATP synthases	0,64	11,43	7,15	0,00	0,0002
Clustering-based subsystems	Chromosome replication	0,59	8,81	4,73	0,00	0,0020
Metabolism of aromatic compounds	**Anaerobic degradation of aromatic compounds**	**0,57**	**11,22**	**8,79**	**0,00**	**0,0000**
Secondary metabolism	**Plant hormones**	**0,57**	**11,05**	**8,05**	**0,00**	**0,0001**
Clustering-based subsystems	**Methylamine utilization**	**0,57**	**11,63**	**7,37**	**0,00**	**0,0001**
Clustering-based subsystems	Putative isoquinoline 1-oxidoreductase subunit	0,55	9,64	5,21	0,00	0,0012
Protein metabolism	Protein folding	0,51	12,59	5,63	0,00	0,0007
Clustering-based subsystems	Probably pyrimidine biosynthesis-related	0,51	8,84	5,91	0,00	0,0006
Membrane transport	**Sugar Phosphotransferase Systems, PTS**	**0,50**	**11,04**	**9,51**	**0,00**	**0,0000**
Regulation and cell signaling	Quorum sensing and biofilm formation	0,47	9,58	4,72	0,00	0,0020
Membrane transport	**Protein secretion system, type II**	**0,47**	**11,01**	**5,33**	**0,00**	**0,0010**
RNA metabolism	Transcription	0,46	13,33	5,76	0,00	0,0007
Carbohydrates	Fermentation	0,45	13,30	9,17	0,00	0,0000
Nucleosides and nucleotides	-	0,45	11,73	3,09	0,01	0,0229
Secondary metabolism	Aromatic amino acids and derivatives	0,45	8,27	2,63	0,02	0,0456
Clustering-based subsystems	Choline bitartrate degradation, putative	0,41	9,37	5,12	0,00	0,0013
Carbohydrates	One-carbon metabolism	0,38	14,17	7,60	0,00	0,0001
Amino acids and derivatives	**Branched-chain amino acids**	**0,38**	**14,08**	**7,85**	**0,00**	**0,0001**
Fatty acids, lipids and isoprenoids	**Isoprenoids**	**0,37**	**12,73**	**5,44**	**0,00**	**0,0009**
Fatty acids, lipids and isoprenoids	-	0,35	12,27	4,62	0,00	0,0023
Clustering-based subsystems	Two related proteases	0,34	10,00	2,77	0,02	0,0367
Carbohydrates	**Central carbohydrate metabolism**	**0,34**	**15,23**	**8,05**	**0,00**	**0,0001**
Carbohydrates	**CO2 fixation**	**0,32**	**13,46**	**7,04**	**0,00**	**0,0002**
Metabolism of aromatic compounds	**-**	**0,31**	**11,22**	**4,72**	**0,00**	**0,0020**
Protein metabolism	**Protein degradation**	**0,30**	**13,36**	**4,17**	**0,00**	**0,0046**
Carbohydrates	**Organic acids**	**0,28**	**13,05**	**4,80**	**0,00**	**0,0019**
Fatty acids, lipids and isoprenoids	Fatty acids	0,28	13,73	6,59	0,00	0,0002
Stress response	**Heat shock**	**0,28**	**12,31**	**4,98**	**0,00**	**0,0015**
Clustering-based subsystems	TldD cluster	0,26	9,90	2,85	0,02	0,0334
Amino acids and derivatives	Lysine, threonine, methionine, and cysteine	0,25	14,51	6,47	0,00	0,0003
Nucleosides and nucleotides	**Pyrimidines**	**0,24**	**13,25**	**5,66**	**0,00**	**0,0007**
Clustering-based subsystems	Ribosome-related cluster	0,22	10,64	2,64	0,02	0,0456
Metabolism of aromatic compounds	Peripheral pathways for catabolism of aromatic compounds	0,21	13,05	2,60	0,02	0,0473
Virulence, diesease and defense	Detection	0,21	11,46	2,58	0,03	0,0478
Amino acids and derivatives	Histidine metabolism	0,20	11,81	3,40	0,01	0,0146
Protein metabolism	Protein processing and modification	0,20	13,30	2,64	0,02	0,0456
Sulfur metabolism	Inorganic sulfur assimilation	0,20	11,95	3,70	0,00	0,0092
Phages, prophages, transposable elements, plasmids	Phages, prophages	-0,18	13,00	-2,82	0,02	0,0351
Amino acids and derivatives	Alanine, serine, and glycine	0,17	13,15	3,10	0,01	0,0229
Phosphorus metabolism	-	0,16	12,99	3,68	0,00	0,0093
Amino acids and derivatives	Arginine, urea cycle, polyamines	0,16	13,48	2,58	0,03	0,0478
Cofactors, vitamins, prostetic groups, pigments	NAD and NADP	-0,15	12,02	-2,58	0,03	0,0478
Cofactors, vitamins, prostetic groups, pigments	Pyridoxine	0,14	11,86	2,81	0,02	0,0351
Cofactors, vitamins, prostetic groups,pigments	Folate and pterines	0,12	14,93	3,13	0,01	0,0220

Table B.4.: DA functional subsystems (adj. p-values < 0.05) between *S. magellanicum* and peat soil metagenomes. The table summarizes the statistical testing result of DA feature detection using limma+voom. The subsystems are sorted descending according to the logFC values. Subsystems which were as tested differentially abundant between both, the *S. magellanicum*/plant metagenomes and the *S. magellanicum*/peat soils, metagenomes are highlighted with bold text.

Subsystmes level 1	Subsystems level 2	LogFC	AveExpr	t-val	p-val	Adj. p-val
S. magellanicum/peat soils metagenomes						
Stress response	**Dessication stress**	**-10,57**	**3,88**	**-6,88**	**0,00**	**0,0013**
Dormancy and sporulation	**Spore DNA protection**	**-7,68**	**3,24**	**-6,84**	**0,00**	**0,0013**
Cofactor, vitamins, prostetic groups, pigments	**Coenzyme B**	**-5,70**	**3,49**	**-6,47**	**0,00**	**0,0018**
Respiration	Reverse electron transport	-4,32	4,51	-4,47	0,00	0,0114
Phages, prophages, plasmids, transposable elements	-	-3,54	5,71	-5,76	0,00	0,0033
Respiration	Sodium ion-coupled energetics	-3,31	6,35	-4,37	0,00	0,0114
Secondary metabolism	Plant octadecanoids	-2,88	3,52	-5,40	0,00	0,0042
Clustering-based subsystems	Proteasome related clusters	-2,84	4,25	-4,55	0,00	0,0114
Clustering-based subsystems	**Tricarboxylate transporter**	**-2,44**	**10,17**	**-7,80**	**0,00**	**0,0013**
Clustering-based subsystems	**Related to menaquinone-cytochrome C reductase**	**-2,14**	**4,46**	**-5,70**	**0,00**	**0,0033**
Motility and chemotaxis	Social motility and nonflagellar swimming in bacteria	-1,70	4,72	-4,18	0,00	0,0150
RNA metabolism	**-**	**-1,46**	**8,74**	**-4,09**	**0,00**	**0,0165**
Clustering-based subsystems	**D-tyrosyl-tRNA(Tyr) deacylase cluster**	**-1,38**	**8,89**	**-3,60**	**0,00**	**0,0298**
Clustering-based subsystems	**Oxidative stress**	**-1,33**	**7,22**	**-4,41**	**0,00**	**0,0114**
Clustering-based subsystems	**Hypothetical associated with RecF**	**-1,11**	**7,76**	**-3,43**	**0,01**	**0,0335**
DNA metabolism	-	-0,94	11,27	-4,40	0,00	0,0114
Clustering-based subsystems	Nucleotidyl-phosphate metabolic cluster	-0,68	11,11	-3,32	0,01	0,0394
Stress response	**-**	**-0,53**	**11,43**	**-4,92**	**0,00**	**0,0074**
Virulence, diesease and defense	**-**	**-0,45**	**11,86**	**-3,45**	**0,01**	**0,0335**
Phages, prophages, plasmids, transposable elements	**Phages, prophages**	**-0,32**	**13,00**	**-3,16**	**0,01**	**0,0478**
Nucleosides and nucleotides	**Pyrimidines**	**0,24**	**13,25**	**3,45**	**0,01**	**0,0335**
Carbohydrates	**Central carbohydrate metabolism**	**0,24**	**15,23**	**4,07**	**0,00**	**0,0165**
Amino acids and derivatives	**Branched-chain amino acids**	**0,30**	**14,08**	**3,95**	**0,00**	**0,0185**
Carbohydrates	**CO2 fixation**	**0,33**	**13,46**	**4,39**	**0,00**	**0,0114**
Carbohydrates	**Organic acids**	**0,35**	**13,05**	**3,55**	**0,00**	**0,0305**
Stress response	**Heat shock**	**0,35**	**12,31**	**3,59**	**0,00**	**0,0298**
Protein metabolism	**Protein degradation**	**0,41**	**13,36**	**3,43**	**0,01**	**0,0335**
Clustering-based subsystems	**Methylamine utilization**	**0,46**	**11,63**	**3,21**	**0,01**	**0,0450**
Fatty acids, lipids and isoprenoids	**Isoprenoids**	**0,47**	**12,73**	**3,98**	**0,00**	**0,0184**
Secondary metabolism	**Plant hormones**	**0,47**	**11,05**	**3,25**	**0,01**	**0,0435**
Metabolism of aromatic compounds	**Anaerobic degradation of aromatic compounds**	**0,50**	**11,22**	**3,77**	**0,00**	**0,0232**
Metabolism of aromatic compounds	**-**	**0,56**	**11,22**	**3,92**	**0,00**	**0,0187**
Cofactor, vitamins, prostetic groups, pigments	**Tetrapyrroles**	**0,79**	**12,84**	**6,13**	**0,00**	**0,0024**
Membrane transport	**Sugar Phosphotransferase Systems, PTS**	**0,84**	**11,04**	**6,96**	**0,00**	**0,0013**
Membrane transport	**Protein secretion system, type II**	**1,07**	**11,01**	**4,96**	**0,00**	**0,0074**
Clustering.based subsystems	**Carotenoid biosynthesis**	**1,67**	**8,72**	**5,51**	**0,00**	**0,0039**
Photosynthesis	**Electron transport and photophosphorylation**	**2,56**	**7,48**	**4,46**	**0,00**	**0,0114**

B.1. eurofins Library Normalization Protocol

Normalization of one shotgun library for Illumina sequencing

Starting Material

One Illumina TrueSeq Shotgun Library from moss communities.

Table 1: Description of the sample

No.	Sample	Description	Vol. (µl)	Total amount (pg)
1	M2	Illumina TrueSeq Shotgun Library	6	49

Library amplification

The library was amplified with PCR (number of cycles indicated in Table 2) using a proof reading enzyme (see Fig. 1, N0) and SBS3 and SBS8 sequencing primers.

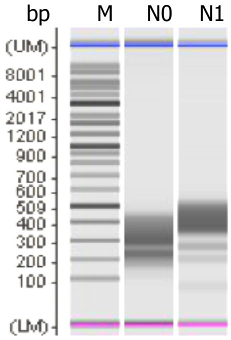


Figure 1: Analysis of the PCR-amplified N0 and N1 library on a Shimadzu MultiNA microchip electrophoresis system. M = 100 bp ladder

Normalization

Normalization was carried out by one cycle of denaturation and reassociation of the DNA, resulting in the N1-library. Reassociated ds-DNAs were separated from the remaining ss-DNAs (normalized DNA) by passing the mixture over a hydroxylapatite column. After hydroxylapatite chromatography, the ss-DNAs were PCR amplified (see Fig.1, N1 and Table 2 for number of cycles and barcode).

Size fractionation

For Illumina sequencing, the tagged N1 library was eluted from a preparative agarose gel in the size range of 300 –500 bp. An aliquot of the size fractionated library was analyzed by capillary electrophoresis (Fig. 2).

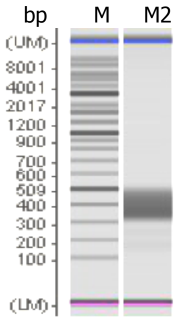


Figure 2: Analysis of the size fractionated N1 library on a Shimadzu MultiNA microchip electrophoresis system. M = 100 bp ladder

Description of the normalized library

The library has a size range of 300 – 500 bp. The primers used for PCR amplification were designed for TruSeq sequencing according to the instructions of Illumina.

The following adapter sequences flank the DNA insert:

TruSeq_Sense_primer
5´- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’

TruSeq_Antisense_ NNNNNN_primer   Barcode
5’-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3’.

The total length of the flanking sequences is 122 bp.

Table 2: Properties of the library

No.	**1**
Sample	**M2**
Barcode	GCCAAT
Cycles N0	15
Cycles N1	6
Conc. (ng/µl)	48
Volume (µl)	20

Appendix C.

Supplementary information Decontaminator

```
usage: java -jar Decontaminator.jar -q <querySequenceFile> -o <filteredOutputFile>
            -b <blatResultFile> -i <percentageIdentityThreshold> -c
            <percentageQueryCoverageThreshold>

            In case of very big files, it might be necessary to increase java memory
            to eg min 1G and max to 12G, -Xms1024m -Xmx12288m, respectively.

 -b <arg>     BLAT mapping, default blast8 format
 -bcl <arg>   length of used barcodes. Value is used to "correct" query coverage calculation
 -c <arg>     threshold for percentage identiy of the BLAT mapping result.
              BLAT results below this threshold are
              discarded
 -fa <arg>    file path to write the query coverage histogram to (default: result file,
              plus "_alignment_length.png"
 -fc <arg>    file path to write the charts data to "_charts_data.tsv"
 -fp <arg>    file path to write the percentage identity histogram to
              (default: result file, plus "_identity.png")
 -fq <arg>    file path to write the query coverage histogram to (default: result file, plus
              "_coverage.png"
 -i <arg>     threshold for percentage identiy of the BLAT mapping result. BLAT results below
              this threshold are discarded
 -nl <arg>    query sequences identified as pontential contaminations are not written to
              file separatly
 -nn          query sequences identified as pontential contaminations are not written to
              file separatly
 -o <arg>     file name and path for the filtered output file, fasta formated
 -pl <arg>    length of used primer (in case of multiple primeres, use length of longest).
              Value is used to "correct" query coverage calculation.
 -q <arg>     file path to the query sequence file in fasta format which should be deconatminated
```

Table C.1.: Decontaminator settings used for the three step evaluation process. Firstly, the raw *true* 16S amplicon set was tested for contaminating sequences using the given seconds. Subsequently, the same set was mixed with 25 sequences which originate from the human host and finally, with 57 manually created chimeric sequences.

	cutoff % identity	cutoff QC	barcode length	primer length
clean set	80	55	6	14
(a) incl. 25 human seqs	80	55	0	0
(b) incl. 57 chimeric seqs	80	55	6	0

Appendix D.

Supplementary information diarrhea study

Table D.1.: Sample summary of the diarrhea study. Fecal (F) samples were collected from four patients (A-D) on four different timepoints (TP; 1-4). Additionally, Mucosal (M) biopsy tissue was taken from three of these patients (B-D) at two timepoints (TP; 2,3) of the experiment.

Sample	Patient	TP	Type	Sample	Patient	TP	Type
AF1	A	1	F	BF4	B	4	F
AF2	A	2	F	DF1	D	1	F
AF3	A	3	F	DF2	D	2	F
AF4	A	4	F	DF3	D	3	F
CF1	C	1	F	DF4	D	4	F
CF2	C	2	F	BM2	B	2	M
CF3	C	3	F	BM3	B	3	M
CF4	C	4	F	DM2	D	2	M
BF1	B	1	F	DM3	D	3	M
BF2	B	2	F	CM2	C	2	M
BF3	B	3	F	CM3	C	3	M

Table D.2.: Summary of settings which were used for detection and removal of contaminating sequences by the Decontaminator, in the 16S raw data of the diarrhea study.

version	BLAT ref DB	% identity	cutoff QC	barcode length	primer length
v.5	GreneGenes May 2011	80	55	6	14

Table D.3.: Summary of analysis statics of Acacia, for removal and correction of low quality sequences, as well as of sequencing noise, for the 16S diarrhea study data.

Description	FS0825402	FJC1PKF02	sum
Mean length (before filtering)	229.0	239.0	
Length SD (before filtering)	51.862	34.913	
Length SD collapsed (before filtering)	76.5483	67.908	
# Seqs usable	164680	311101	475781
# Seqs thrown out	23093	16338	39431
# Low quality	33	13	64
# Outside length range	16180	13455	29635
# with early N's	14101	7158	21259
# collapsed too short	12074	5426	17500
# Unique sequences	50307	71479	121786
# Singletons	19524	25004	44528
# Reference sequences	27796	35659	63455
# Sequences corrected	30679	40726	71405

Appendix E.

Supplementary information ITS mock community

Table E.1.: True sequence distribution of the ITS1/2 mock at the phylum level.

pyhlum	# sequences	
	counts	[%]
Ascomycota	639	46.88
Basidiomycota	403	29.57
Glomeromycota	262	19.22
Zygomycota	38	2.79
Chytridiomycota	21	1.54

Table E.2.: True sequence distribution of the ITS1/2 mock community at the species level for taxa covering more than 2 % of the total sequence abundance.

species	# sequences	
	counts	[%]
uncultured Glomus	85	6.24
uncultured Russula	32	2.35
uncultured Ascomycota	24	1.76
Metarhizium anisopliae	22	1.61
Other	1200	88.04

Table E.3.: True sequence distribution of the ITS1/2 mock community at the class level for taxa covering more than 2 % of total sequence abundance.

class	# sequences	
	counts	[%]
Agaricomycetes	329,00	24.14
Glomeromycetes	253	18.56
Sordariomycetes	220	16.14
Dothideomycetes	148	10.86
Eurotiomycetes	87	6.38
Leotiomycetes	80	5.87
unidentified	56	4.11
Incertae sedis	47	3.45
Pucciniomycetes	47	3.45
Lecanoromycetes	20	1.47
Chytridiomycetes	20	1.47
Other	56	4.11

Table E.4.: True sequence distribution of the ITS1/2 mock community at the genus level for taxa covering more than 2 % of total sequence abundance.

genus	# sequences	
	counts	[%]
unidentified	398	29.20
Inocybe	69	5.06
Fusarium	54	3.96
Cortinarius	43	3.15
Puccinia	36	2.64
Glomus	28	2.05
Metarhizium	23	1.69
Ilyonectria	23	1.69
Rhizophagus	23	1.69
Acaulospora	21	1.54
Other	645	47.32

Table E.5.: True sequence distribution of the ITS1/2 mock at the order level for taxa covering more than 2 % of total sequence abundance.

order	# sequences	
	counts	[%]
Hypocreales	181	13.28
Agaricales	178	13.06
Glomerales	160	11.74
unidentified	87	6.38
Diversisporales	72	5.28
Pleosporales	72	5.28
Russulales	64	4.70
Capnodiales	59	4.33
Pucciniales	47	3.45
Incertae sedis	39	2.86
Helotiales	33	2.42
Eurotiales	31	2.27
Polyporales	25	1.83
Erysiphales	18	1.32
Other	315	23.11

Table E.6.: True sequence distribution of the ITS1/2 mock community at the family level for taxa covering more than 2 % of total sequence abundance.

family	# sequences	
	counts	[%]
unidentified	156	11.45
Glomeraceae	153	11.23
Incertae sedis	97	7.12
Nectriaceae	87	6.38
Inocybaceae	71	5.21
Russulaceae	58	4.26
Cortinariaceae	48	3.52
Gigasporaceae	38	2.79
Pucciniaceae	36	2.64
Trichocomaceae	31	2.27
Clavicipitaceae	29	2.13
Pleosporaceae	27	1.98
Mycosphaerellaceae	24	1.76
Acaulosporaceae	22	1.61
Teratosphaeriaceae	20	1.47
Other	466	34.19

Table E.7.: True sequence distribution of the ITS1 mock at the phylum level.

phylum	# sequences	
	counts	[%]
Ascomycota	922	46.92
Basidiomycota	984	50.08
Chytridiomycota	36	1.83
Glomeromycota	23	1.17

Table E.8.: True sequence distribution of the ITS1 mock community at the genus level for taxa covering more than 2 % of total sequence abundance.

genus	# sequences	
	counts	[%]
Cortinarius	96	4.89
unidentified	72	3.66
Inocybe	65	3.31
Lactarius	48	2.44
Russula	37	1.88
Other	1647	83.82

Table E.9.: True sequence distribution of the ITS1 mock at the class level for taxa covering more than 2 % of total sequence abundance.

class	# sequences	
	counts	[%]
Agaricomycetes	858	43.66
Sordariomycetes	264	13.44
Lecanoromycetes	167	8.50
Dothideomycetes	158	8.04
Eurotiomycetes	133	6.77
Leotiomycetes	93	4.73
Pezizomycetes	72	3.66
Tremellomycetes	31	1.58
Other	189	9.62

Table E.10.: True sequence distribution of the ITS1 mock at the order level for taxa covering more than 2 % of total sequence abundance.

order	# sequences	
	counts	[%]
Agaricales	432	21.98
Russulales	116	5.90
Hypocreales	106	5.39
Polyporales	98	4.99
Boletales	71	3.61
Pezizales	71	3.61
Pleosporales	68	3.46
Xylariales	63	3.21
Capnodiales	62	3.16
Lecanorales	60	3.05
Eurotiales	54	2.75
Helotiales	46	2.34
Peltigerales	42	2,14
Verrucariales	39	1.98
Diaporthales	36	1.83
Hymenochaetales	31	1.58
Teloschistales	29	1.48
Other	541	27,53

Table E.11.: True sequence distribution of the ITS1 mock community at the family level for taxa covering more than 2 % of total sequence abundance.

family	# sequences	
	counts	[%]
Cortinariaceae	107	5.45
Russulaceae	92	4.68
Inocybaceae	68	3.46
Xylariaceae	52	2.65
Trichocomaceae	51	2.60
unidentified	41	2.09
Verrucariaceae	39	1.98
Mycosphaerellaceae	35	1.,78
Polyporaceae	33	1.68
Nectriaceae	32	1.63
Boletaceae	30	1.53
Agaricaceae	30	1.53
Fomitopsidaceae	29	1.48
Other	1326	67.48

Appendix F.

Supplementary information BAL study

Table F.1.: The table summarizes the collected 16S samples during the BAL study, including information about patient (PA), experimental group (GR, control no antibiotics (1A), control with antibiotics (1B), ICU no antibiotics (2A), ICU with antibiotics (2B), ICU, pneumonia, with antibiotics (3B)), type (no type specified (NTS), community associated pneumonia (CAP), ventilation associated pneumonia (VAP), aspiration (ASP)), and timepoint (TP).

SAMPLE	PAT	GR	TYPE	TP	SAMPLE	PAT	GR	TYPE	TP
087-1A-NTS-0	87	1A	NTS	0	609-3B-ASP-0	609	3B	ASP	0
095-1A-NTS-0	95	1A	NTS	0	301-3B-VAP-0	301	3B	VAP	0
097-1A-NTS-0	97	1A	NTS	0	301-3B-VAP-1	301	3B	VAP	1
105-1A-NTS-0	105	1A	NTS	0	302-3B-VAP-0	302	3B	VAP	0
106-1A-NTS-0	106	1A	NTS	0	302-3B-VAP-1	302	3B	VAP	1
107-1A-NTS-0	107	1A	NTS	0	303-3B-VAP-0	303	3B	VAP	0
108-1A-NTS-0	108	1A	NTS	0	303-3B-VAP-1	303	3B	VAP	1
109-1A-NTS-0	109	1A	NTS	0	304-3B-ASP-0	304	3B	ASP	0
098-1B-NTS-0	98	1B	NTS	0	304-3B-ASP-1	304	3B	ASP	1
099-1B-NTS-0	99	1B	NTS	0	305-3B-VAP-0	305	3B	VAP	0
100-1B-NTS-0	100	1B	NTS	0	306-3B-VAP-0	306	3B	VAP	0
101-1B-NTS-0	101	1B	NTS	0	309-3B-VAP-0	309	3B	VAP	0
102-1B-NTS-0	102	1B	NTS	0	310-3B-NAP-0	310	3B	NAP	0
103-1B-NTS-0	103	1B	NTS	0	312-3B-VAP-0	312	3B	VAP	0
104-1B-NTS-0	104	1B	NTS	0	313-3B-NAP-0	313	3B	NAP	0
401-2A-NTS-0	401	2A	NTS	0	313-3B-VAP-1	313	3B	VAP	1
402-2A-NTS-0	402	2A	NTS	0	314-3B-CAP-0	314	3B	CAP	0
403-2A-NTS-0	403	2A	NTS	0	318-3B-ASP-0	318	3B	ASP	0
406-2A-NTS-0	406	2A	NTS	0	319-3B-VAP-0	319	3B	VAP	0
201-2A-NTS-0	201	2A	NTS	0	320-3B-ASP-0	320	3B	ASP	0
202-2A-NTS-0	202	2A	NTS	0	321-3B-NAP-0	321	3B	NAP	0
203-2A-NTS-0	203	2A	NTS	0	322-3B-NAP-0	322	3B	NAP	0
405-2B-NTS-0	405	2B	NTS	0	323-3B-VAP-0	323	3B	VAP	0
610-2B-NTS-0	610	2B	NTS	0	324-3B-ASP-0	324	3B	ASP	0
252-2B-NTS-0	252	2B	NTS	0	325-3B-VAP-0	325	3B	VAP	0
255-2B-NTS-0	255	2B	NTS	0	326-3B-VAP-0	326	3B	VAP	0
256-2B-NTS-0	256	2B	NTS	0	327-3B-ASP-0	327	3B	ASP	0
257-2B-NTS-0	257	2B	NTS	0	328-3B-NAP-0	328	3B	NAP	0
608-3B-ASP-0	608	3B	ASP	0	612-3B-VAP-0	612	3B	VAP	0

Table F.2.: The table summarizes the collected ITS samples during the BAL study including information about patient (PA), experimental group (GR, control no antibiotics (1A); control with antibiotics (1B), ICU no antibiotics (2A), ICU with antibiotics (2B), ICU, pneumonia, with antibiotics (3B)), type (no type specified (NTS), community associated pneumonia (CAP), ventilation associated pneumonia (VAP), aspiration (ASP)), and timepoint (TP).

SAMPLE	PAT	GR	TYPE	TP	SAMPLE	PAT	GR	TYPE	TP
087-1A-NTS-0	87	1A	NTS	0	612-3B-VAP-0	612	3B	VAP	0
095-1A-NTS-0	95	1A	NTS	0	301-3B-VAP-0	301	3B	VAP	0
105-1A-NTS-0	105	1A	NTS	0	302-3B-VAP-0	302	3B	VAP	0
107-1A-NTS-0	107	1A	NTS	0	302-3B-VAP-1	302	3B	VAP	1
098-1B-NTS-0	98	1B	NTS	0	303-3B-VAP-0	303	3B	VAP	0
100-1B-NTS-0	100	1B	NTS	0	303-3B-VAP-1	303	3B	VAP	1
101-1B-NTS-0	101	1B	NTS	0	304-3B-ASP-0	304	3B	ASP	0
103-1B-NTS-0	103	1B	NTS	0	304-3B-ASP-1	304	3B	ASP	1
104-1B-NTS-0	104	1B	NTS	0	305-3B-VAP-0	305	3B	VAP	0
401-2A-NTS-0	401	2A	NTS	0	306-3B-VAP-0	306	3B	VAP	0
402-2A-NTS-0	402	2A	NTS	0	309-3B-VAP-0	309	3B	VAP	0
403-2A-NTS-0	403	2A	NTS	0	313-3B-NAP-0	313	3B	NAP	0
406-2A-NTS-0	406	2A	NTS	0	313-3B-VAP-1	313	3B	VAP	1
201-2A-NTS-0	201	2A	NTS	0	318-3B-ASP-0	318	3B	ASP	0
202-2A-NTS-0	202	2A	NTS	0	319-3B-VAP-0	319	3B	VAP	0
203-2A-NTS-0	203	2A	NTS	0	320-3B-ASP-0	320	3B	ASP	0
405-2B-NTS-0	405	2B	NTS	0	321-3B-NAP-0	321	3B	NAP	0
610-2B-NTS-0	610	2B	NTS	0	322-3B-NAP-0	322	3B	NAP	0
252-2B-NTS-0	252	2B	NTS	0	323-3B-VAP-0	323	3B	VAP	0
255-2B-NTS-0	255	2B	NTS	0	324-3B-ASP-0	324	3B	ASP	0
256-2B-NTS-0	256	2B	NTS	0	325-3B-VAP-0	325	3B	VAP	0
257-2B-NTS-0	257	2B	NTS	0	326-3B-VAP-0	326	3B	VAP	0
608-3B-ASP-0	608	3B	ASP	0	327-3B-ASP-0	327	3B	ASP	0
609-3B-ASP-0	609	3B	ASP	0	328-3B-NAP-0	328	3B	NAP	0

Table F.3.: Sample overview of the bacterial BAL study samples analyzed using SnoWMAn's RDP pipeline. The table summarizes the number of obtained OTUs for different cluster distances for each sample, as well as for the main groups and of the total community profile.

Sample	Sequences	Unique Sequs	0.0	0.01	0.02	0.03	0.04	0.05	0.06
105-1A-NTS-0	3562	471	402	266	144	108	98	93	86
107-1A-NTS-0	3881	643	548	324	205	157	141	130	123
087-1A-NTS-0	3495	424	362	240	114	81	69	66	62
108-1A-NTS-0	3659	646	526	331	171	121	111	104	99
097-1A-NTS-0	4070	481	408	252	137	98	87	85	83
106-1A-NTS-0	4252	510	431	274	160	116	103	100	93
109-1A-NTS-0	4153	491	407	274	105	68	63	60	59
095-1A-NTS-0	4851	597	495	299	136	98	95	89	83
Total 1A	**31923**		**3045**	**1610**	**705**	**491**	**421**	**378**	**343**
101-1B-NTS-0	4426	536	442	281	114	70	62	61	59
098-1B-NTS-0	1464	286	254	185	127	111	107	104	103
103-1B-NTS-0	3287	442	370	232	111	80	76	74	72
102-1B-NTS-0	4388	594	483	296	119	80	74	70	66
099-1B-NTS-0	673	131	122	97	59	44	43	40	40
100-1B-NTS-0	3082	490	416	316	183	138	120	114	108
104-1B-NTS-0	2873	333	305	199	96	68	62	59	58
Total 1B	**20193**		**2040**	**1152**	**522**	**373**	**322**	**295**	**276**
401-2A-NTS-0	2843	392	312	205	123	86	75	64	57
201-2A-NTS-0	2731	679	563	384	238	165	125	107	88
202-2A-NTS-0	6382	797	589	347	121	92	84	77	73
406-2A-NTS-0	2613	415	339	200	131	107	91	78	71
403-2A-NTS-0	2949	752	642	435	296	228	187	154	143
402-2A-NTS-0	3556	456	396	271	141	97	87	83	77
203-2A-NTS-0	2522	595	487	322	199	148	120	99	90
Total 2A	**23596**	**-1**	**3162**	**1897**	**974**	**684**	**539**	**443**	**394**
405-2B-NTS-0	4555	502	417	263	109	69	61	53	48
256-2B-NTS-0	2726	634	518	341	216	162	140	121	113
252-2B-NTS-0	9331	750	583	293	123	83	67	62	58
610-2B-NTS-0	2820	213	174	93	54	43	33	27	25
257-2B-NTS-0	5010	534	427	259	116	81	70	65	58
255-2B-NTS-0	3339	404	343	206	106	78	71	71	69
Total 2B	**27781**		**2310**	**1254**	**544**	**363**	**299**	**256**	**235**
303-3B-VAP-0	3192	655	487	303	185	139	117	105	93
303-3B-VAP-1	5219	807	585	300	164	120	111	104	98
313-3B-NAP-0	3884	577	396	195	97	79	70	59	55
324-3B-ASP-0	8432	1156	851	429	200	136	118	110	102
302-3B-VAP-0	5901	266	143	59	16	12	10	9	9
327-3B-ASP-0	5393	680	559	398	201	159	134	124	113
302-3B-VAP-1	2240	251	188	91	54	45	44	41	39
328-3B-NAP-0	4033	448	344	217	83	46	44	41	39
325-3B-VAP-0	4081	556	433	291	195	158	134	128	119
320-3B-ASP-0	5426	396	284	143	76	66	61	60	58
321-3B-NAP-0	3632	734	552	373	202	138	111	96	84
314-3B-CAP-0	5304	576	424	265	104	57	51	47	42
304-3B-ASP-0	6092	405	244	96	27	11	11	11	10
310-3B-NAP-0	3847	637	445	278	155	96	84	72	66
304-3B-ASP-1	5072	524	404	248	97	56	51	48	47
319-3B-VAP-0	3211	437	347	202	81	57	53	51	49
612-3B-VAP-0	4110	340	272	148	67	50	48	46	44
313-3B-VAP-1	4293	544	420	240	118	86	75	72	68
323-3B-VAP-0	5057	614	479	281	117	79	72	66	66
318-3B-ASP-0	4903	799	610	329	187	132	110	96	83
326-3B-VAP-0	4170	554	420	234	124	90	89	85	82
301-3B-VAP-0	4170	522	412	251	133	94	86	83	79
301-3B-VAP-1	4834	534	408	245	117	84	80	76	70
306-3B-VAP-0	4034	874	682	425	269	209	181	165	156
312-3B-VAP-0	5108	291	189	63	20	12	11	11	10
608-3B-ASP-0	3571	658	495	303	151	106	88	75	66
609-3B-ASP-0	2939	314	265	162	94	74	60	55	48
305-3B-VAP-0	3868	760	521	370	176	110	85	74	63
322-3B-NAP-0	4359	496	392	216	101	75	69	65	64
309-3B-VAP-0	5122	571	427	235	110	84	78	72	68
Total 3B	**135497**		**10381**	**4702**	**1899**	**1211**	**950**	**786**	**689**
Total	**238990**	**31174**	**18238**	**7542**	**3071**	**1938**	**1471**	**1191**	**1028**

Table F.4.: α-diversity scores according to Chao1, Chao1 (bc), Shannon, ACE, as well as to Richness and Eveness calculated by SnoWMAn, based on the final bacterial community of the BAL study samples. Scores are presented for each sample, as well as for the main groups and the total community profile at a distance of 0.03.

Sample	Richness	Chao1	Chao1 (bc)	Shannon	Evenness	ACE
087-1A-NTS-0	81.00	161.67	138.75	3.14	0.72	124.77
095-1A-NTS-0	98.00	131.06	126.11	2.47	0.54	124.82
097-1A-NTS-0	98.00	143.13	132.20	3.20	0.70	119.99
105-1A-NTS-0	108.00	290.25	225.00	3.17	0.68	142.31
106-1A-NTS-0	116.00	198.57	186.13	2.59	0.54	158.55
107-1A-NTS-0	157.00	177.35	175.07	4.12	0.82	177.86
108-1A-NTS-0	121.00	160.06	154.33	3.59	0.75	146.23
109-1A-NTS-0	68.00	268.00	163.00	2.32	0.55	102.79
Total 1A	**491.00**	**652.29**	**645.69**	**3.98**	**0.64**	**606.28**
098-1B-NTS-0	111.00	121.80	120.56	3.37	0.72	124.74
099-1B-NTS-0	44.00	66.50	61.50	2.07	0.55	64.92
100-1B-NTS-0	138.00	246.90	240.14	2.24	0.45	269.59
101-1B-NTS-0	70.00	142.25	115.33	2.48	0.58	95.35
102-1B-NTS-0	80.00	98.00	93.20	2.75	0.63	93.77
103-1B-NTS-0	80.00	92.25	90.11	2.10	0.48	90.58
104-1B-NTS-0	68.00	108.50	98.60	2.46	0.58	93.99
Total 1B	**373.00**	**521.63**	**516.08**	**3.31**	**0.56**	**492.72**
201-2A-NTS-0	165.00	249.48	245.50	3.24	0.63	271.13
202-2A-NTS-0	92.00	194.08	177.00	1.57	0.35	131.86
203-2A-NTS-0	148.00	235.12	230.50	2.71	0.54	244.06
401-2A-NTS-0	86.00	174.00	164.83	1.17	0.26	152.16
402-2A-NTS-0	97.00	136.20	131.36	2.39	0.52	131.24
403-2A-NTS-0	228.00	460.26	451.13	3.45	0.64	454.56
406-2A-NTS-0	107.00	185.13	179.06	1.93	0.41	178.21
Total 2A	**684.00**	**1182.90**	**1175.70**	**3.60**	**0.55**	**1138.26**
252-2B-NTS-0	83.00	127.46	123.07	1.50	0.34	129.31
255-2B-NTS-0	78.00	101.14	97.13	2.35	0.54	94.42
256-2B-NTS-0	162.00	296.48	289.73	3.34	0.66	339.49
257-2B-NTS-0	81.00	108.56	104.33	1.82	0.41	102.36
405-2B-NTS-0	69.00	133.22	125.10	1.32	0.31	117.82
610-2B-NTS-0	43.00	68.00	64.11	0.98	0.26	73.31
Total 2B	**363.00**	**667.22**	**657.00**	**2.94**	**0.50**	**591.91**

Sample	Richness	Chao1	Chao1 (bc)	Shannon	Evenness	ACE
301-3B-VAP-0	94.00	158.00	134.00	3.11	0.68	108.62
301-3B-VAP-1	84.00	99.13	95.00	2.44	0.55	92.94
302-3B-VAP-0	12.00	30.00	19.50	0.07	0.03	27.40
302-3B-VAP-1	45.00	49.90	48.50	1.85	0.49	49.96
303-3B-VAP-0	139.00	193.15	190.48	2.76	0.56	221.91
303-3B-VAP-1	120.00	140.63	138.90	3.10	0.65	146.22
304-3B-ASP-0	11.00	23.25	18.00	0.03	0.01	36.54
304-3B-ASP-1	56.00	65.00	61.00	2.25	0.56	60.87
305-3B-VAP-0	110.00	183.63	177.56	2.26	0.48	170.83
306-3B-VAP-0	209.00	328.12	323.43	3.16	0.59	337.26
309-3B-VAP-0	84.00	106.56	103.00	1.66	0.38	99.22
310-3B-NAP-0	96.00	162.67	156.00	1.75	0.38	157.34
312-3B-VAP-0	12.00	36.50	22.50	0.05	0.02	27.71
313-3B-NAP-0	79.00	139.50	134.65	1.64	0.37	180.07
313-3B-VAP-1	86.00	100.40	97.00	3.28	0.74	100.36
314-3B-CAP-0	57.00	99.25	83.00	1.11	0.27	66.65
318-3B-ASP-0	132.00	219.03	213.48	1.67	0.34	225.41
319-3B-VAP-0	57.00	67.13	64.20	2.36	0.58	66.50
320-3B-ASP-0	66.00	109.56	103.80	0.66	0.16	95.61
321-3B-NAP-0	138.00	210.32	204.00	2.92	0.59	188.72
322-3B-NAP-0	75.00	79.00	78.11	3.01	0.70	83.79
323-3B-VAP-0	79.00	103.50	97.20	2.99	0.68	93.30
324-3B-ASP-0	136.00	212.06	202.60	3.49	0.71	202.49
325-3B-VAP-0	158.00	302.11	293.05	1.56	0.31	277.63
326-3B-VAP-0	90.00	102.00	99.43	3.23	0.72	98.45
327-3B-ASP-0	159.00	249.73	245.72	1.33	0.26	268.25
328-3B-NAP-0	46.00	64.00	53.50	2.01	0.53	50.22
608-3B-ASP-0	106.00	223.04	212.00	1.94	0.42	185.51
609-3B-ASP-0	74.00	305.13	254.60	0.98	0.23	153.65
612-3B-VAP-0	50.00	66.67	61.25	1.56	0.40	58.79
Total 3B	**1211.00**	**1711.59**	**1706.59**	**4.19**	**0.59**	**1690.56**

Table F.5.: Relative sequence distribution at the phylum level determined with SnoWMAn's RDP pipeline, at a classification confidence of 80 %, a cluster distance of 0.03, for clusters with more than 2 % overall abundance. The table presents counts for all samples of group 1A of the 16S amplicon set.

	087-1A-NTS-0	095-1A-NTS-0	097-1A-NTS-0	105-1A-NTS-0	106-1A-NTS-0	107-1A-NTS-0	108-1A-NTS-0	109-1A-NTS-0
Actinomyces	0.00	0.00	0.00	3.40	0.00	0.00	2.13	0.00
Alkalibacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.60
Aquabacterium	3.69	0.00	2.09	0.00	0.00	0.00	0.00	2.24
Bacteroides	18.80	16.57	32.16	11.37	16.39	13.58	23.18	50.73
Bradyrhizobium	2.15	4.21	5.90	4.58	0.00	2.83	0.00	9.61
Cloacibacterium	0.00	0.00	4.96	0.00	0.00	0.00	0.00	0.00
Fusobacterium	0.00	0.00	0.00	6.51	0.00	3.50	0.00	0.00
Gemella	0.00	0.00	0.00	0.00	0.00	2.96	2.71	0.00
Granulicatella	0.00	0.00	0.00	0.00	0.00	2.78	2.84	0.00
Oribacterium	0.00	0.00	0.00	2.39	0.00	0.00	0.00	0.00
Parvimonas	0.00	0.00	0.00	2.44	0.00	0.00	0.00	0.00
Pasteurella	0.00	0.00	0.00	0.00	0.00	2.34	0.00	0.00
Porphyromonas	0.00	0.00	0.00	0.00	0.00	3.38	0.00	0.00
Prevotella	4.75	0.00	0.00	6.46	2.96	0.00	11.78	0.00
Propionibacterium	13.36	0.00	3.78	0.00	2.38	4.69	0.00	5.30
Pseudomonas	12.27	0.00	10.91	0.00	4.70	2.53	0.00	2.94
Staphylococcus	0.00	0.00	0.00	0.00	0.00	4.07	0.00	2.62
Streptobacillus	0.00	0.00	0.00	0.00	0.00	3.32	0.00	0.00
Streptococcus	5.67	46.53	0.00	37.84	6.59	20.20	20.61	0.00
Tropheryma	0.00	0.00	0.00	2.64	43.70	0.00	0.00	0.00
Undibacterium	0.00	0.00	2.14	0.00	0.00	0.00	2.30	0.00
Veillonella	0.00	0.00	0.00	2.16	0.00	5.13	0.00	0.00
Other	12.36	11.05	14.74	15.02	13.83	13.99	21.59	11.85
Unclassified	26.95	21.65	23.32	5.19	9.45	14.69	12.87	12.11

Table F.6.: Relative sequence distribution at the phylum level determined with SnoWMAn's RDP pipeline, at a classification confidence of 80 %, a cluster distance of 0.03, for clusters with more than 2 % overall abundance. The table presents counts for all samples of group 1B of the 16S amplicon set.

	098-1B-NTS-0	099-1B-NTS-0	100-1B-NTS-0	101-1B-NTS-0	102-1B-NTS-0	103-1B-NTS-0	104-1B-NTS-0
Actinomyces	0.00	0.00	0.00	0.00	2.87	0.00	0.00
Alkalibacterium	0.00	3.57	0.00	3.50	2.62	0.00	2.61
Aquabacterium	0.00	2.82	0.00	0.00	2.12	2.56	2.33
Bacteroides	29.23	59.73	0.00	46.43	47.74	31.88	45.35
Bradyrhizobium	14.14	3.57	0.00	4.74	3.19	2.01	14.03
Corynebacterium	4.64	0.00	0.00	0.00	0.00	0.00	0.00
Janthinobacterium	0.00	0.00	0.00	0.00	0.00	0.00	2.09
Neisseria	0.00	0.00	7.24	0.00	0.00	0.00	0.00
Paracoccus	0.00	0.00	0.00	4.52	0.00	0.00	0.00
Prevotella	5.12	0.00	5.97	0.00	0.00	0.00	0.00
Propionibacterium	0.00	0.00	0.00	8.34	2.76	0.00	0.00
Pseudomonas	0.00	0.00	0.00	2.53	0.00	42.99	0.00
Ralstonia	0.00	4.90	0.00	2.44	2.12	0.00	2.16
Rothia	0.00	0.00	2.63	0.00	0.00	0.00	0.00
Staphylococcus	7.17	0.00	0.00	4.16	0.00	0.00	0.00
Streptococcus	4.85	0.00	72.45	0.00	3.12	0.00	0.00
Other	21.86	9.06	10.12	11.82	18.07	14.15	14.17
Unclassified	12.98	16.34	1.59	11.52	15.38	6.42	17.26

Table F.7.: Relative sequence distribution at the phylum level determined with SnoWMAn's RDP pipeline, at a classification confidence of 80 %, a cluster distance of 0.03, for clusters with more than 2 % overall abundance. The table presents counts for all samples of group 2B of the 16S amplicon set.

	252-2B-NTS-0	255-2B-NTS-0	256-2B-NTS-0	257-2B-NTS-0	405-2B-NTS-0	610-2B-NTS-0
Atopobium	0.00	0.00	2.53	0.00	0.00	0.00
Bacteroides	2.16	32.17	0.00	8.46	2.37	15.25
Bradyrhizobium	0.00	6.41	0.00	0.00	0.00	0.00
Enterococcus	0.00	0.00	0.00	0.00	12.43	0.00
Gemella	0.00	0.00	5.28	0.00	0.00	0.00
Granulicatella	0.00	0.00	4.15	0.00	0.00	0.00
Haemophilus	10.48	0.00	0.00	43.73	0.00	6.70
Lactobacillus	0.00	0.00	3.23	0.00	0.00	0.00
Mycoplasma	0.00	16.95	0.00	0.00	0.00	0.00
Neisseria	57.80	0.00	0.00	0.00	0.00	0.00
Prevotella	0.00	2.13	20.98	0.00	0.00	0.00
Pseudomonas	0.00	24.02	0.00	0.00	0.00	0.00
Ralstonia	0.00	2.01	0.00	0.00	0.00	0.00
Streptococcus	16.66	0.00	54.04	40.44	74.80	76.31
Veillonella	0.00	0.00	4.73	0.00	3.01	0.00
Other	0.84	10.78	3.78	5.23	2.59	1.06
Unclassified	12.06	5.54	1.28	2.14	4.81	0.67

Table F.8.: Relative sequence distribution at the phylum level determined with SnoWMAn's RDP pipeline, at a classification confidence of 80 %, a cluster distance of 0.03, for clusters with more than 2 % overall abundance. The table presents counts for the first part of samples of group 3B for the 16S amplicon set.

	301-3B-VAP-0	301-3B-VAP-1	302-3B-VAP-0	302-3B-VAP-1	303-3B-VAP-0	303-3B-VAP-1	304-3B-ASP-0	304-3B-ASP-1	305-3B-VAP-0	306-3B-VAP-0
Actinomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aeromonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Alkalibacterium	0.00	3.04	0.00	0.00	0.00	0.00	0.00	3.82	0.00	0.00
Aquabacterium	0.00	0.00	0.00	2.01	0.00	0.00	0.00	3.67	0.00	0.00
Atopobium	0.00	0.00	0.00	0.00	6.33	7.74	0.00	0.00	0.00	4.29
Bacteroides	15.78	45.74	0.00	9.82	0.00	3.62	0.00	52.78	0.00	0.00
Bradyrhizobium	5.30	20.44	0.00	2.90	0.00	0.00	0.00	11.47	0.00	0.00
Corynebacterium	0.00	0.00	0.00	61.79	0.00	0.00	0.00	2.33	0.00	0.00
Dermabacter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dolosigranulum	0.00	0.00	0.00	13.35	0.00	0.00	0.00	0.00	0.00	0.00
Eubacterium	0.00	0.00	0.00	0.00	2.51	4.87	0.00	0.00	0.00	0.00
Fusobacterium	0.00	2.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.76	0.00
Granulicatella	0.00	0.00	0.00	0.00	6.52	7.03	0.00	0.00	0.00	10.24
Helicobacter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Janthinobacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lactobacillus	0.00	0.00	0.00	0.00	0.00	2.87	0.00	0.00	2.66	0.00
Lactococcus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leptotrichia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mogibacterium	0.00	0.00	0.00	0.00	0.00	2.18	0.00	0.00	0.00	0.00
Moraxella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mycoplasma	28.68	0.00	0.00	0.00	39.47	28.07	0.00	0.00	0.00	0.00
Nocardioides	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Novosphingobium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parvimonas	0.00	0.00	0.00	0.00	4.07	3.49	0.00	0.00	0.00	0.00
Peptoniphilus	0.00	0.00	0.00	0.00	3.51	3.85	0.00	0.00	0.00	0.00
Planococcus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Prevotella	4.68	0.00	0.00	0.00	21.12	20.71	0.00	0.00	0.00	15.59
Propionibacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pseudomonas	0.00	0.00	99.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ralstonia	0.00	3.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Rothia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.15	20.33
Schwartzia	2.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sphingomonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Staphylococcus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.58	7.39	0.00
Streptococcus	0.00	3.19	0.00	0.00	0.00	0.00	0.00	0.00	50.62	22.56
Treponema	4.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tropheryma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ureaplasma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Veillonella	0.00	0.00	0.00	0.00	2.79	0.00	0.00	0.00	0.00	0.00
Other	21.70	15.45	0.22	7.63	6.55	11.52	0.18	9.35	2.12	9.82
Unclassified	16.71	6.37	0.44	2.50	7.14	4.04	99.82	14.00	1.29	17.18

Table F.9.: Relative sequence distribution at the phylum level determined with SnoWMAn's RDP pipeline, at a classification confidence of 80 %, a cluster distance of 0.03, for clusters with more than 2 % overall abundance. The table presents counts for the second part of samples of group 3B for the 16S amplicon set.

	309-3B-VAP-0	310-3B-NAP-0	312-3B-VAP-0	313-3B-NAP-0	313-3B-VAP-1	314-3B-CAP-0	318-3B-ASP-0	319-3B-VAP-0	320-3B-ASP-0	321-3B-NAP-0
Actinomyces	0.00	0.00	0.00	0.00	5.75	0.00	0.00	0.00	0.00	2.75
Aeromonas	0.00	0.00	0.00	5.95	0.00	0.00	0.00	0.00	0.00	0.00
Alkalibacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.71	0.00	0.00
Aquabacterium	0.00	0.00	0.00	0.00	6.13	0.00	0.00	8.28	0.00	0.00
Atopobium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bacteroides	7.61	0.00	0.00	0.00	23.81	11.12	0.00	51.04	0.00	0.00
Bradyrhizobium	0.00	0.00	0.00	0.00	4.38	0.00	0.00	2.74	0.00	0.00
Corynebacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dermabacter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dolosigranulum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eubacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fusobacterium	0.00	3.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemella	0.00	33.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.55
Granulicatella	0.00	2.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.26
Helicobacter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Janthinobacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lactobacillus	3.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lactococcus	13.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leptotrichia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mogibacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Moraxella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	88.33	0.00
Mycoplasma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Nocardioides	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Novosphingobium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parvimonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Peptoniphilus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Planococcus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Prevotella	0.00	0.00	0.00	0.00	0.00	0.00	3.20	0.00	0.00	2.20
Propionibacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pseudomonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ralstonia	0.00	0.00	0.00	0.00	5.17	0.00	0.00	0.00	0.00	0.00
Rothia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.05
Schwartzia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sphingomonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.96	0.00	0.00
Staphylococcus	0.00	0.00	0.00	19.70	0.00	0.00	63.37	0.00	0.00	0.00
Streptococcus	63.06	49.75	99.41	0.00	0.00	79.24	26.70	0.00	5.55	65.80
Treponema	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tropheryma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ureaplasma	0.00	4.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Veillonella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other	7.07	5.90	0.53	1.44	19.66	8.07	5.65	13.08	5.44	4.57
Unclassified	4.37	1.14	0.06	72.91	35.10	1.56	1.08	19.18	0.68	1.82

Table F.10.: Relative sequence distribution at the phylum level determined with SnoWMAn's RDP pipeline, at a classification confidence of 80 %, a cluster distance of 0.03, for clusters with more than 2 % overall abundance. The table presents counts for the third part of samples of group 3B for the 16S amplicon set.

	322-3B-NAP-0	323-3B-VAP-0	324-3B-ASP-0	325-3B-VAP-0	326-3B-VAP-0	327-3B-ASP-0	328-3B-NAP-0	608-3B-ASP-0	609-3B-ASP-0	612-3B-VAP-0
Actinomyces	0.00	0.00	6.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aeromonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Alkalibacterium	3.30	3.28	0.00	0.00	2.13	0.00	3.00	0.00	0.00	0.00
Aquabacterium	0.00	4.63	0.00	0.00	2.37	0.00	7.49	0.00	0.00	0.00
Atopobium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bacteroides	22.71	27.03	21.11	0.00	15.37	0.00	61.91	0.00	0.00	7.69
Bradyrhizobium	3.90	4.07	2.24	0.00	2.35	0.00	0.00	0.00	0.00	7.47
Corynebacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dermabacter	0.00	0.00	0.00	0.00	2.59	0.00	0.00	0.00	0.00	0.00
Dolosigranulum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Eubacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fusobacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gemella	0.00	0.00	3.02	0.00	0.00	0.00	0.00	6.24	0.00	0.00
Granulicatella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Helicobacter	0.00	2.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Janthinobacterium	0.00	0.00	0.00	0.00	0.00	0.00	2.45	0.00	0.00	0.00
Lactobacillus	0.00	0.00	0.00	0.00	0.00	4.06	0.00	0.00	0.00	0.00
Lactococcus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leptotrichia	0.00	0.00	3.45	0.00	3.12	0.00	0.00	0.00	0.00	0.00
Mogibacterium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Moraxella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mycoplasma	0.00	0.00	0.00	74.74	9.09	0.00	0.00	0.00	0.00	0.00
Nocardioides	0.00	0.00	0.00	0.00	21.65	0.00	0.00	0.00	0.00	0.00
Novosphingobium	24.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parvimonas	0.00	0.00	0.00	9.14	0.00	0.00	0.00	0.00	0.00	0.00
Peptoniphilus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Planococcus	0.00	0.00	9.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Prevotella	0.00	0.00	14.84	3.95	2.49	0.00	0.00	40.88	0.00	0.00
Propionibacterium	2.41	0.00	0.00	0.00	0.00	0.00	2.16	0.00	0.00	9.17
Pseudomonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ralstonia	2.68	3.03	0.00	0.00	0.00	0.00	2.03	0.00	0.00	0.00
Rothia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Schwartzia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sphingomonas	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Staphylococcus	0.00	10.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Streptococcus	0.00	13.25	8.94	2.79	9.42	78.90	0.00	49.59	14.70	0.00
Treponema	0.00	2.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tropheryma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	81.73	0.00
Ureaplasma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Veillonella	0.00	0.00	4.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other	16.38	11.75	15.39	6.79	17.10	7.19	9.87	2.41	3.10	7.96
Unclassified	24.48	17.48	10.65	2.60	12.30	9.85	11.08	0.87	0.48	67.71

Table F.11.: Relative sequence distribution at the genus level determined with SnoWMAn's BLAT pipeline, at a classification confidence of 80 % and clusters with more than 2 % overall abundance. The table presents counts for all samples of group 1A of the ITS BAL amplicon set.

	087-1A-NTS-0	095-1A-NTS-0	105-1A-NTS-0	107-1A-NTS-0
Amylostereum	0.00	10.95	0.00	0.00
Armillaria	13.72	0.00	0.00	0.00
Aspergillus	0.00	0.00	0.00	15.42
Bjerkandera	0.00	9.50	0.00	0.00
Didymella	0.00	0.00	3.91	0.00
Dioszegia	0.00	5.33	0.00	0.00
Epicoccum	0.00	0.00	5.27	0.00
Heterobasidion	0.00	21.33	2.07	0.00
Malassezia	28.94	0.00	0.00	48.94
Meira	0.00	7.15	0.00	0.00
Penicillium	6.14	0.00	0.00	2.03
Phoma	0.00	0.00	5.33	0.00
Piptoporus	0.00	0.00	2.01	0.00
Pluteus	0.00	10.52	0.00	0.00
Trametes	0.00	25.09	0.00	0.00
Wallemia	0.00	0.00	0.00	27.26
unidentified	50.32	6.91	74.21	4.84
Other	0.87	3.22	7.20	1.51

Table F.12.: Relative sequence distribution at the genus level determined with SnoWMAn's BLAT pipeline, at a classification confidence of 80 % and clusters with more than 2 % overall abundance. The table presents counts for all samples of group 1B of the ITS BAL amplicon set.

	098-1B-NTS-0	100-1B-NTS-0	101-1B-NTS-0	103-1B-NTS-0	104-1B-NTS-0
Amphinema	0.00	0.00	0.00	2.13	0.00
Aspergillus	0.00	3.31	0.00	0.00	0.00
Bjerkandera	2.74	0.00	0.00	0.00	2.42
Cladosporium	0.00	0.00	0.00	5.60	3.66
Cystofilobasidium	21.44	0.00	0.00	0.00	0.00
Dioszegia	0.00	5.34	0.00	0.00	0.00
Entomocorticium	2.06	0.00	0.00	0.00	0.00
Fomitopsis	0.00	0.00	0.00	0.00	3.68
Heterobasidion	0.00	7.78	0.00	6.98	6.34
Hyphodontia	0.00	2.40	0.00	0.00	0.00
Hypholoma	0.00	0.00	0.00	7.15	7.66
Malassezia	0.00	3.65	21.33	0.00	0.00
Mensularia	0.00	0.00	3.09	0.00	0.00
Mrakia	9.17	0.00	0.00	0.00	0.00
Onnia	0.00	0.00	9.33	0.00	0.00
Phlebia	0.00	0.00	0.00	0.00	2.32
Physisporinus	0.00	0.00	0.00	2.72	0.00
Postia	0.00	0.00	0.00	2.19	5.74
Psathyrella	0.00	0.00	0.00	3.51	0.00
Pulvinula	0.00	3.22	0.00	0.00	0.00
Resinicium	2.36	0.00	0.00	8.15	3.14
Rigidoporus	0.00	0.00	0.00	3.90	2.72
Schizophyllum	5.70	3.72	0.00	0.00	0.00
Sporobolomyces	0.00	6.29	0.00	0.00	0.00
Steccherinum	0.00	0.00	0.00	0.00	2.25
Stereum	4.58	0.00	0.00	0.00	0.00
Stropharia	0.00	0.00	0.00	3.20	2.79
Trametes	2.99	12.36	0.00	2.59	4.10
Volvopluteus	0.00	0.00	0.00	2.06	2.11
Wallemia	0.00	0.00	65.17	0.00	0.00
unidentified	27.49	40.25	0.00	10.08	10.64
Other	21.47	11.68	1.08	39.74	40.41

Table F.13.: Relative sequence distribution at the genus level determined with SnoWMAn's BLAT pipeline, at a classification confidence of 80 % and clusters with more than 2 % overall abundance. The table presents counts for all samples of group 2A of the ITS BAL amplicon set.

	201-2A-NTS-0	202-2A-NTS-0	203-2A-NTS-0	401-2A-NTS-0	402-2A-NTS-0	403-2A-NTS-0	406-2A-NTS-0
Alternaria	0.00	0.00	0.00	4.99	0.00	0.00	0.00
Armillaria	0.00	0.00	0.00	0.00	0.00	43.91	0.00
Blumeria	0.00	6.25	0.00	0.00	0.00	0.00	0.00
Candida	98.79	0.00	55.81	93.23	0.00	6.52	49.38
Cladosporium	0.00	0.00	3.76	0.00	0.00	0.00	0.00
Clitocybe	0.00	0.00	0.00	0.00	0.00	7.46	0.00
Entomocorticium	0.00	0.00	4.57	0.00	0.00	0.00	0.00
Kazachstania	0.00	0.00	0.00	0.00	0.00	28.02	0.00
Malassezia	0.00	12.38	0.00	0.00	20.20	0.00	45.72
Marasmiellus	0.00	0.00	0.00	0.00	24.74	0.00	0.00
Mycena	0.00	65.91	0.00	0.00	0.00	0.00	0.00
Penicillium	0.00	5.11	6.05	0.00	0.00	0.00	0.00
Rigidoporus	0.00	9.94	0.00	0.00	0.00	0.00	0.00
Wallemia	0.00	0.00	0.00	0.00	34.86	0.00	0.00
unidentified	0.00	0.00	15.55	0.00	18.23	2.30	3.73
Other	1.21	0.41	14.25	1.78	1.97	11.79	1.17

Table F.14.: Relative sequence distribution at the genus level determined with SnoWMAn's BLAT pipeline, at a classification confidence of 80 % and clusters with more than 2 % overall abundance. The table presents counts for all samples of group 2B of the ITS BAL amplicon set.

	252-2B-NTS-0	255-2B-NTS-0	256-2B-NTS-0	257-2B-NTS-0	610-2B-NTS-0
Acremonium	0.00	0.00	0.00	0.00	28.89
Aspergillus	0.00	0.00	0.00	0.00	3.11
Bjerkandera	0.00	0.00	3.60	0.00	0.00
Candida	51.08	98.92	89.90	0.00	6.22
Ceriporiopsis	0.00	0.00	0.00	0.00	4.00
Cladosporium	0.00	0.00	0.00	34.26	0.00
Malassezia	16.24	0.00	5.96	29.09	0.00
Resinicium	0.00	0.00	0.00	0.00	26.67
unidentified	32.35	0.00	0.00	35.16	28.44
Other	0.33	1.08	0.54	1.50	2.67

Table F.15.: Relative sequence distribution at the genus level determined with SnoWMAn's BLAT pipeline, at a classification confidence of 80 % and clusters with more than 2 % overall abundance. The table presents counts for all samples of group 3B (part I) of the ITS BAL amplicon set.

	301-3B-VAP-0	302-3B-VAP-0	302-3B-VAP-1	303-3B-VAP-0	303-3B-VAP-1	304-3B-ASP-0	304-3B-ASP-1	305-3B-VAP-0	306-3B-VAP-0	309-3B-VAP-0	313-3B-NAP-0	313-3B-VAP-1	318-3B-ASP-0
Alternaria	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	23.61	0.00	2.31	0.00	0.00
Armillaria	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Artomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aspergillus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.16	0.00	0.00
Boletus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.57	0.00	0.00
Botryobasidium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.97	0.00
Candida	99.82	99.47	99.63	99.82	99.94	97.10	54.08	0.00	10.02	97.53	0.00	0.00	93.07
Cercospora	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.79	0.00	0.00
Cladosporium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.40	13.59	0.00
Filobasidium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Golovinomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Heterobasidion	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.97	0.00	0.00	0.00	0.00	0.00
Knufia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Malassezia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.08	0.00	0.00	0.00	27.97	0.00
Meripilus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monographella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	35.44	0.00	0.00
Penicillium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	70.97	0.00	0.00	0.00	0.00	0.00
Phellinus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phialemonium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	26.43	0.00
Plicaturopsis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	41.76	0.00	0.00	0.00	0.00
Rhodotorula	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Saccharomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sarcinomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sporidiobolus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.89	0.00	0.00
Tilletiopsis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Trametes	0.00	0.00	0.00	0.00	0.00	0.00	43.89	0.00	0.00	0.00	0.00	0.00	0.00
Tricholoma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.57	0.00	7.07	0.00	0.00
Trichosporon	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tricladium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.43	0.00
Wallemia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.42	0.00	0.00
unidentified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.10	15.59	0.00	6.75	13.69	2.65
Other	0.18	0.53	0.37	0.18	0.06	2.90	2.03	1.90	1.45	2.47	5.19	1.91	4.29

Table F.16.: Relative sequence distribution at the genus level determined with SnoWMAn's BLAT pipeline, at a classification confidence of 80 % and clusters with more than 2 % overall abundance. The table presents counts for all samples of group 3B (part II) of the ITS BAL amplicon set.

	319-3B-VAP-0	320-3B-ASP-0	321-3B-NAP-0	322-3B-NAP-0	323-3B-VAP-0	324-3B-ASP-0	325-3B-VAP-0	326-3B-VAP-0	327-3B-ASP-0	328-3B-NAP-0	608-3B-ASP-0	609-3B-ASP-0	612-3B-VAP-0
Alternaria	0.00	0.00	0.00	0.00	43.30	0.00	0.00	14.43	0.00	0.00	0.00	0.00	0.00
Armillaria	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.45	0.00
Artomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.72	0.00	0.00	0.00	0.00	0.00
Aspergillus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.09	0.00	0.00	0.00	0.00
Boletus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Botryobasidium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Candida	86.43	53.99	0.00	97.45	52.66	90.31	0.00	62.12	41.05	89.86	10.10	39.84	98.07
Cercospora	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cladosporium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.96	0.00	0.00	0.00	0.00	0.00
Filobasidium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.12	0.00	0.00	0.00	0.00	0.00
Golovinomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.85	0.00	0.00	0.00	0.00
Heterobasidion	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Knufia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.94	0.00	0.00
Malassezia	0.00	10.74	0.00	0.00	0.00	3.74	62.30	0.00	0.00	8.33	17.34	0.00	0.00
Meripilus	0.00	0.00	10.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Monographella	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Penicillium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phellinus	0.00	0.00	28.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phialemonium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.16	0.00
Plicaturopsis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Rhodotorula	2.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Saccharomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.89	0.00	0.00	0.00	0.00
Sarcinomyces	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.97	0.00	0.00
Sporidiobolus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tilletiopsis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.25	0.00	0.00
Trametes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.34	0.00
Tricholoma	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Trichosporon	2.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tricladium	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wallemia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
unidentified	2.75	34.97	58.49	0.00	3.58	4.45	36.68	2.17	4.16	0.00	26.31	17.18	0.00
Other	5.83	0.31	2.75	2.55	0.46	1.50	1.01	1.48	4.94	1.81	4.10	3.03	1.93

Table F.17.: α-diversity scores according to Chao1, Chao1 (bc), Shannon, ACE, as well as to Richness and Eveness calculated by SnoWMAn, based on the final fungal community profile of the BAL study samples. Scores are presented for each sample, as well as for the main groups and the total community profile at the species level.

Sample	Richness	Chao1	Chao1 (bc)	Shannon	Evenness	ACE
087-1A-NTS-0	30.00	NaN	36.00	1.85	0.54	33.18
095-1A-NTS-0	46.00	78.67	68.75	2.37	0.62	64.37
105-1A-NTS-0	50.00	64.40	61.00	1.46	0.37	63.88
107-1A-NTS-0	34.00	74.50	52.00	1.64	0.47	47.06
Total 1A	**135.00**	**207.25**	**197.33**	**3.03**	**0.62**	**178.27**
098-1B-NTS-0	183.00	226.20	222.38	3.38	0.65	217.25
100-1B-NTS-0	90.00	220.67	184.50	3.31	0.74	135.39
101-1B-NTS-0	25.00	31.13	29.20	1.09	0.34	36.40
103-1B-NTS-0	288.00	380.48	375.62	4.48	0.79	359.08
104-1B-NTS-0	258.00	324.27	320.22	4.62	0.83	312.77
Total 1B	**567.00**	**802.76**	**796.52**	**4.70**	**0.74**	**732.92**
201-2A-NTS-0	33.00	58.00	48.00	0.46	0.13	41.71
202-2A-NTS-0	18.00	NaN	33.00	1.13	0.39	23.52
203-2A-NTS-0	104.00	168.80	161.27	2.08	0.45	163.23
401-2A-NTS-0	23.00	43.25	35.00	0.33	0.10	34.89
402-2A-NTS-0	21.00	27.25	24.33	1.59	0.52	25.62
403-2A-NTS-0	81.00	157.56	147.11	2.14	0.49	123.79
406-2A-NTS-0	28.00	53.60	48.00	0.94	0.28	72.09
Total 2A	**232.00**	**470.10**	**457.00**	**2.47**	**0.45**	**380.96**
252-2B-NTS-0	17.00	35.00	24.50	1.10	0.39	27.63
255-2B-NTS-0	18.00	30.25	25.00	0.09	0.03	24.96
256-2B-NTS-0	20.00	26.13	24.20	0.43	0.15	33.03
257-2B-NTS-0	23.00	35.25	30.00	1.56	0.50	31.80
405-2B-NTS-0	13.00	15.00	14.20	0.10	0.04	16.27
610-2B-NTS-0	14.00	14.67	14.25	1.91	0.72	15.61
Total 2B	**70.00**	**113.56**	**107.80**	**1.68**	**0.40**	**116.55**
301-3B-VAP-0	9.00	NaN	9.00	0.04	0.02	9.26
302-3B-VAP-0	23.00	151.00	83.00	0.13	0.04	96.03
302-3B-VAP-1	7.00	NaN	8.00	0.08	0.04	8.84
303-3B-VAP-0	5.00	NaN	5.00	0.07	0.05	6.99
303-3B-VAP-1	5.00	NaN	5.00	0.01	0.01	5.86
304-3B-ASP-0	19.00	28.00	24.00	0.18	0.06	26.03
304-3B-ASP-1	51.00	123.90	109.50	0.92	0.23	117.90
305-3B-VAP-0	17.00	17.50	17.00	1.33	0.47	17.62
306-3B-VAP-0	15.00	16.50	15.75	1.59	0.59	19.23
309-3B-VAP-0	22.00	72.00	44.50	0.16	0.05	51.86
313-3B-NAP-0	65.00	77.00	74.43	2.31	0.55	75.09
313-3B-VAP-1	20.00	28.00	23.00	1.87	0.62	25.64
318-3B-ASP-0	18.00	20.67	19.50	0.93	0.32	21.26
319-3B-VAP-0	54.00	68.00	65.38	0.79	0.20	67.02
320-3B-ASP-0	7.00	NaN	7.00	1.40	0.72	8.11
321-3B-NAP-0	14.00	22.00	17.00	1.64	0.62	21.37
322-3B-NAP-0	21.00	39.00	28.50	0.22	0.07	28.93
323-3B-VAP-0	22.00	46.50	32.50	1.10	0.36	33.41
324-3B-ASP-0	20.00	29.00	25.00	0.47	0.16	27.12
325-3B-VAP-0	9.00	9.50	9.00	0.93	0.42	9.50
326-3B-VAP-0	20.00	22.00	20.50	1.35	0.45	22.61
327-3B-ASP-0	35.00	47.50	44.00	1.74	0.49	51.16
328-3B-NAP-0	8.00	8.17	8.00	0.45	0.22	8.67
608-3B-ASP-0	19.00	27.00	22.00	1.78	0.61	23.25
609-3B-ASP-0	78.00	154.05	145.36	1.79	0.41	163.20
612-3B-VAP-0	16.00	25.00	21.00	0.82	0.29	28.00
Total 3B	**309.00**	**561.02**	**551.45**	**2.61**	**0.45**	**545.88**

Table F.18.: The table summarizes the number of filtered and remaining sequences after each pre-processing step. The ITS amplicons of the BAL study were noise reduced and quality filtered by Acacia, as well as filtered for contaminating sequences using the Decontaminator prior to phylogenetic analysis.

number of sequences	H70LSVG01	H70LSVG03	**total**
raw	256956	286610	**543566**
noise and low quality	26258	29137	**55395**
after denoising and qual. Fitering	230698	257473	**488171**
toally removed	26258	29137	**55395**
after preprocessing			**488171**
removed by snowman			**21589**
for classification			**466582**
not classified (contaminations)			**21716**
finally classified			**444866**

Table F.19.: The table summarizes the number of filtered and remaining sequences after each pre-processing step. 16S amplicons of the BAL study were noise reduced and quality filtered by Acacia, as well as filtered for contaminating sequences using the Decontaminator prior to phylogenetic analysis.

number of sequences	H70LSVG02	H70LSVG04	**total**
raw	168409	261271	**429680**
noise and low quality	17068	43689	**60757**
after denoising and qual. Fitering	151341	217582	**368923**
contaminations	1953	3083	**5036**
after decontamination	149388	214499	**363887**
after preprocessing	149388	214499	**363887**
removed by snowman			**124897**
for classification			**238990**

Table F.20.: The table summarizes tools and settings which were used within 16S BAL study pre-processing, as well as for the final phylogenetic classification.

	release/version	**settings**				
Decontaminator		**blat ref db**	**perc. Identity**	**query coverage**	**offset MID**	**offset primer**
	v.5	GG 09May2011	95	75	10	20
uchime		**mode**		**reference db**		
	mothur v.1.31.2	reference	SILVA relase 105			
Acacia		**min. avg. quality threshold**		**other settings**		
	1.52.b0	22		default		
SnoWMAn		**pipeline**	**classifier version**	**infernal model**		
	v.1.2	RDP	RDP classifier 2.5		ncbi16S_508_mod5	

Table F.21.: The table summarizes tools and settings which were used within ITS BAL study pre-processing, as well as for final the phylogenetic classification.

	release/version	**settings**	
Acacia		**min. avg. quality threshold**	**other settings**
	1.52.b0	22	default
SnoWMAn		**pipeline**	**reference db**
	v.1.2	BLAT	unite IST 15.Oct.2013

Table F.22.: Sample overview of the fungal BAL study samples analyzed using SnoWMAn's BLAT pipeline. The table summarizes the number of finally obtained distinct species (OTUs) for different cluster distances for each sample, as well as for the main groups and of the total community profile.

Sample	Sequences	Unique Seqs	# OTUs
105-1A-NTS-0	8411	849	50
107-1A-NTS-0	10673	1423	34
087-1A-NTS-0	10355	1473	30
095-1A-NTS-0	9214	1122	46
Total 1A	**38653**		**135**
101-1B-NTS-0	9714	1243	25
100-1B-NTS-0	8873	1253	90
098-1B-NTS-0	6599	1201	183
103-1B-NTS-0	7033	1275	288
104-1B-NTS-0	7193	1450	258
Total 1B	**39412**		**567**
406-2A-NTS-0	3163	413	28
401-2A-NTS-0	10964	1149	23
403-2A-NTS-0	1824	491	81
402-2A-NTS-0	11301	1409	21
203-2A-NTS-0	13182	1648	104
201-2A-NTS-0	8834	1200	33
202-2A-NTS-0	6268	681	18
Total 2A	**55536**		**232**
405-2B-NTS-0	15053	1856	13
257-2B-NTS-0	15550	2342	23
256-2B-NTS-0	9230	993	20
255-2B-NTS-0	8506	853	18
252-2B-NTS-0	11679	1287	17
610-2B-NTS-0	225	50	14
Total 2B	**60243**		**70**

Sample	Sequences	Unique Seqs	# OTUs
303-3B-VAP-0	7287	682	5
326-3B-VAP-0	13674	1214	20
313-3B-NAP-0	16152	1679	65
303-3B-VAP-1	18848	1038	5
324-3B-ASP-0	20498	2040	20
327-3B-ASP-0	2185	709	35
302-3B-VAP-0	11006	1128	23
301-3B-VAP-0	16346	1924	9
302-3B-VAP-1	5110	465	7
328-3B-NAP-0	1380	164	8
306-3B-VAP-0	898	124	15
325-3B-VAP-0	6417	861	9
320-3B-ASP-0	326	49	7
321-3B-NAP-0	1308	203	14
608-3B-ASP-0	1684	370	19
304-3B-ASP-0	11828	1050	19
304-3B-ASP-1	11025	1070	51
319-3B-VAP-0	18487	1769	54
609-3B-ASP-0	7457	903	78
612-3B-VAP-0	10243	1309	16
313-3B-VAP-1	4812	743	20
323-3B-VAP-0	21973	1897	22
322-3B-NAP-0	12349	1195	21
305-3B-VAP-0	5118	682	17
318-3B-ASP-0	1890	467	18
309-3B-VAP-0	22721	1852	22
Total 3B	**251022**		**309**

Table F.23.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's RDP pipeline, at the phylum level. For each DA phylum level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each phylum and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Tenericutes	1A	3B	8	30	0	7574	235.04	50753.75	-7.6267	0.0002332
Cyanobacteria/Chloroplast	1A	3B	8	30	0	3601	235.04	30019.08	-6.8720	0.0002332
Tenericutes	2A	3B	7	30	0	7574	260.93	50753.75	-7.4550	0.0008855
Cyanobacteria/Chloroplast	2A	3B	7	30	0	3601	260.93	30019.08	-6.6979	0.0008855
Tenericutes	1B	3B	7	30	0	7574	261.16	50753.75	-7.4538	0.0008872
Cyanobacteria/Chloroplast	1B	3B	7	30	0	3601	261.16	30019.08	-6.6968	0.0008872
Fusobacteria	1B	2A	7	7	0	668	261.16	25095.07	-6.4382	0.0012495
Tenericutes	1A	2B	8	6	0	566	235.04	24880.08	-6.5997	0.0028923
Fusobacteria	2A	2B	7	6	668	0	25095.07	330.96	6.1458	0.0040784
Tenericutes	2A	2B	7	6	0	566	260.93	24880.08	-6.4280	0.0040784
Fusobacteria	2A	3B	7	30	668	648	25095.07	3371.97	2.8736	0.0042141
Fusobacteria	1A	1B	8	7	781	0	12161.84	261.16	5.3993	0.0069189
Cyanobacteria/Chloroplast	2B	3B	6	30	0	3601	330.96	30019.08	-6.4019	0.0077994
Tenericutes	1B	2B	7	6	0	566	261.16	24880.08	-6.4268	0.0081709
Fusobacteria	1A	2B	8	6	781	0	12161.84	330.96	5.1069	0.0127958
Actinobacteria	1A	2B	8	6	4099	107	144357.85	5483.82	4.7078	0.0204568
Fusobacteria	1B	3B	7	30	0	648	261.16	3371.97	-3.5646	0.0306517

Table F.24.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's RDP pipeline, at the class level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA class level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each class and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Erysipelotrichia	2A	3B	7	30	267	0	9108.48	296.48	4.8980	1.10E-13
Flavobacteria	1A	3B	8	30	280	0	8293.09	296.48	4.7482	2.61E-12
Erysipelotrichia	1A	2A	8	7	0	267	193.37	9108.48	-5.4037	3.77E-06
Erysipelotrichia	1B	2A	7	7	0	267	241.09	9108.48	-5.0876	5.80E-05
Mollicutes	1A	3B	8	30	0	7574	193.37	52243.32	-7.8931	0.000132
Chloroplast	1A	3B	8	30	0	3601	193.37	30676.47	-7.1265	0.000156
Flavobacteria	1A	2A	8	7	280	0	8293.09	274.23	4.7929	0.000186
Flavobacteria	1A	1B	8	7	280	0	8293.09	241.09	4.9433	0.000214
Fusobacteria	1B	2A	7	7	0	668	241.09	22382.35	-6.3765	0.000764
Mollicutes	2A	3B	7	30	0	7574	274.23	52243.32	-7.4391	0.000876
Erysipelotrichia	2A	2B	7	6	267	0	9108.48	327.54	4.7082	0.000877
Chloroplast	2A	3B	7	30	0	3601	274.23	30676.47	-6.6720	0.001014
Mollicutes	1B	3B	7	30	0	7574	241.09	52243.32	-7.5973	0.00117
Chloroplast	1B	3B	7	30	0	3601	241.09	30676.47	-6.8297	0.00117
Fusobacteria	2A	3B	7	30	668	648	22382.35	2685.71	3.0246	0.001921
Flavobacteria	1A	2B	8	6	280	0	8293.09	327.54	4.5627	0.001979
Mollicutes	1A	2B	8	6	0	566	193.37	22158.20	-6.6578	0.001979
Fusobacteria	1A	1B	8	7	781	0	12870.10	241.09	5.5808	0.002919
Negativicutes	1B	2B	7	6	0	280	241.09	17647.76	-6.0314	0.003735
Fusobacteria	2A	2B	7	6	668	0	22382.35	327.54	5.9890	0.004373
Mollicutes	2A	2B	7	6	0	566	274.23	22158.20	-6.2039	0.005943
Negativicutes	2B	3B	6	30	280	607	17647.76	1806.14	3.2209	0.005973
Chloroplast	2B	3B	6	30	0	3601	327.54	30676.47	-6.4401	0.005973
Mollicutes	1B	2B	7	6	0	566	241.09	22158.20	-6.3620	0.00698
Fusobacteria	1A	2B	8	6	781	0	12870.10	327.54	5.1933	0.00934
Clostridia	2A	2B	7	6	734	0	24569.99	327.54	6.1194	0.011818
Betaproteobacteria	2B	3B	6	30	5552	4635	274367.28	32458.19	3.0774	0.012837
Negativicutes	1A	1B	8	7	367	0	5821.49	241.09	4.4432	0.016204
Clostridia	2B	3B	6	30	0	1741	327.54	13739.11	-5.2832	0.020638
Negativicutes	1B	2A	7	7	0	156	241.09	5441.34	-4.3446	0.021174

Table F.25.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's RDP pipeline, at the order level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA order level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each order and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Pasteurellales	2B	3B	6	30	4474	0	222478.73	271.75	9.5355	2.98E-30
Pasteurellales	2A	3B	7	30	1142	0	45082.68	271.75	7.2330	4.02E-20
Erysipelotrichales	2A	3B	7	30	267	0	8809.22	271.75	4.8911	6.98E-14
Neisseriales	2B	3B	6	30	5393	150	279862.92	1331.05	7.6693	5.77E-13
Flavobacteriales	1A	3B	8	30	280	0	7602.28	271.75	4.6742	3.19E-12
Rhodobacterales	1B	3B	7	30	200	0	6377.08	271.75	4.4276	2.12E-11
Pasteurellales	1A	2B	8	6	0	4474	215.30	222478.73	-9.8160	8.02E-10
Pasteurellales	1B	2B	7	6	0	4474	246.42	222478.73	-9.6524	1.83E-08
Pasteurellales	1A	2A	8	7	0	1142	215.30	45082.68	-7.5136	5.88E-07
Neisseriales	1A	2B	8	6	0	5393	215.30	279862.92	-10.1469	3.42E-06
Pasteurellales	1B	2A	7	7	0	1142	246.42	45082.68	-7.3500	6.11E-06
Erysipelotrichales	1A	2A	8	7	0	267	215.30	8809.22	-5.1675	7.69E-06
Erysipelotrichales	1B	2A	7	7	0	267	246.42	8809.22	-5.0058	6.02E-05
Rhodobacterales	1A	1B	8	7	0	200	215.30	6377.08	-4.7036	7.09E-05
Flavobacteriales	1A	2A	8	7	280	0	7602.28	265.94	4.6881	0.000192
Flavobacteriales	1A	1B	8	7	280	0	7602.28	246.42	4.7893	0.0002
Rhodobacterales	1B	2A	7	7	200	0	6377.08	265.94	4.4407	0.000227
Mycoplasmatales	1A	3B	8	30	0	7518	215.30	60253.38	-7.9320	0.000248
Chloroplast	1A	3B	8	30	0	3601	215.30	33441.37	-7.0830	0.000302
Actinomycetales	1A	2B	8	6	4086	0	109182.17	282.65	8.4453	0.000326
Actinomycetales	2B	3B	6	30	0	9090	282.65	80839.95	-8.0119	0.000388
Burkholderiales	1B	2A	7	7	1049	0	34842.76	265.94	6.8768	0.00047
Burkholderiales	2A	3B	7	30	0	3762	265.94	28038.19	-6.5643	0.000544
Mycoplasmatales	2A	3B	7	30	0	7518	265.94	60253.38	-7.6658	0.000593
Flavobacteriales	1A	2B	8	6	280	0	7602.28	282.65	4.6109	0.000599
Burkholderiales	1A	2A	8	7	1015	0	27171.25	265.94	6.5183	0.000765
Rhizobiales	1A	2A	8	7	1339	0	36515.05	265.94	6.9439	0.000765
Chloroplast	2A	3B	7	30	0	3601	265.94	33441.37	-6.8169	0.000777
Rhizobiales	2A	3B	7	30	0	3550	265.94	26287.57	-6.4710	0.000777
Erysipelotrichales	2A	2B	7	6	267	0	8809.22	282.65	4.8273	0.00086

Table F.26.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's RDP pipeline, at the family level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA family level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR for each family and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Pasteurellaceae	2B	3B	6	30	4474	0	197220.02	258.57	9.4097	1.02E-29
Enterococcaceae	2B	3B	6	30	566	0	25531.24	258.57	6.4621	3.44E-21
Pasteurellaceae	2A	3B	7	30	1142	0	41156.46	258.57	7.1501	1.04E-19
Erysipelotrichaceae	2A	3B	7	30	267	0	9131.30	258.57	4.9850	4.41E-14
Flavobacteriaceae	1A	3B	8	30	280	0	8382.33	258.57	4.8599	5.61E-13
Rhodobacteraceae	1B	3B	7	30	200	0	6899.16	258.57	4.5824	6.36E-12
Neisseriaceae	2B	3B	6	30	5393	150	237521.96	1564.74	7.2177	2.48E-11
Pasteurellaceae	1A	2B	8	6	0	4474	240.84	197220.02	-9.5003	3.34E-09
Porphyromonadaceae	1A	3B	8	30	131	0	3935.20	258.57	3.7783	9.35E-09
Lachnospiraceae	1A	3B	8	30	111	0	3865.46	258.57	3.7435	1.03E-08
Oxalobacteraceae	1B	3B	7	30	96	0	3842.30	258.57	3.7379	1.99E-08
Pasteurellaceae	1B	2B	7	6	0	4474	253.04	197220.02	-9.4369	5.09E-08
Enterococcaceae	1A	2B	8	6	0	566	240.84	25531.24	-6.5526	1.78E-07
Enterococcaceae	1B	2B	7	6	0	566	253.04	25531.24	-6.4891	1.55E-06
Pasteurellaceae	1A	2A	8	7	0	1142	240.84	41156.46	-7.2407	2.28E-06
Enterococcaceae	2A	2B	7	6	0	566	258.73	25531.24	-6.4597	3.46E-06
Neisseriaceae	1A	2B	8	6	0	5393	240.84	237521.96	-9.7679	1.23E-05
Fusobacteriaceae	2A	3B	7	30	664	117	23022.31	1279.26	4.1376	1.27E-05
Pasteurellaceae	1B	2A	7	7	0	1142	253.04	41156.46	-7.1772	1.59E-05
Erysipelotrichaceae	1A	2A	8	7	0	267	240.84	9131.30	-5.0756	2.47E-05
Erysipelotrichaceae	1B	2A	7	7	0	267	253.04	9131.30	-5.0121	0.000121
Rhodobacteraceae	1A	1B	8	7	0	200	240.84	6899.16	-4.6730	0.000173
Flavobacteriaceae	1A	2A	8	7	280	0	8382.33	258.73	4.8576	0.0002
Flavobacteriaceae	1A	1B	8	7	280	0	8382.33	253.04	4.8870	0.00027
Rhodobacteraceae	1B	2A	7	7	200	0	6899.16	258.73	4.5800	0.000272
Mycoplasmataceae	1A	3B	8	30	0	7518	240.84	62966.42	-7.8529	0.000328
Streptophyta	1A	3B	8	30	0	3601	240.84	31185.43	-6.8400	0.000502
Moraxellaceae	1A	3B	8	30	0	4793	240.84	43177.15	-7.3088	0.000529
Erysipelotrichaceae	2A	2B	7	6	267	0	9131.30	264.43	4.9537	0.000571
Flavobacteriaceae	1A	2B	8	6	280	0	8382.33	264.43	4.8287	0.000581

Table F.27.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's RDP pipeline, at the genus level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA genus level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each genus and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Neisseria	2B	3B	6	30	5393	0	226976.2	257.4067	9.617026	3.25E-28
Haemophilus	2B	3B	6	30	3358	0	147558.7	257.4067	8.995943	1.93E-28
Tropheryma	1A	1B	8	7	1858	0	57783.19	254.4233	7.65788	0.002311
Tropheryma	1A	2A	8	7	1858	0	57783.19	257.9026	7.640587	0.00138
Tropheryma	1A	2B	8	6	1858	0	57783.19	260.4866	7.627699	0.004058
Pseudomonas	1B	2A	7	7	1413	0	52918.95	257.9026	7.513699	0.006437
Staphylococcus	2A	2B	7	6	1260	0	45659.89	260.4866	7.288209	0.01152
Haemophilus	2A	3B	7	30	1114	0	41285.96	257.4067	7.159177	3.07E-20
Bradyrhizobium	1B	2A	7	7	984	0	35585.91	257.9026	6.941835	0.001618
Pseudomonas	1A	2A	8	7	1073	0	32942.67	257.9026	6.830449	0.010782
Bradyrhizobium	1A	2A	8	7	1006	0	31605.92	257.9026	6.770816	0.00138
Propionibacterium	1A	2A	8	7	1023	0	31245.5	257.9026	6.754429	0.000834
Propionibacterium	1A	2B	8	6	1023	0	31245.5	260.4866	6.74152	0.002524
Enterococcus	2B	3B	6	30	566	0	25889.39	257.4067	6.486371	6.88E-22
Fusobacterium	2A	2B	7	6	662	0	23335.66	260.4866	6.321238	0.002377
Parvimonas	2A	2B	7	6	614	0	21846.01	260.4866	6.226103	0.004772
Rothia	2A	2B	7	6	513	0	19578.43	260.4866	6.067825	0.007973
Propionibacterium	1B	2A	7	7	369	0	12972.02	257.9026	5.488648	0.004835
Propionibacterium	1B	2B	7	6	369	0	12972.02	260.4866	5.475739	0.014786
Fusobacterium	1A	1B	8	7	368	0	11285.66	254.4233	5.305421	0.002766
Fusobacterium	1A	2B	8	6	368	0	11285.66	260.4866	5.275264	0.004815
Actinomyces	2A	2B	7	6	274	0	10131.03	260.4866	5.119863	0.021571
Solobacterium	2A	3B	7	30	259	0	9184.349	257.4067	4.996856	1.99E-14
Solobacterium	2A	2B	7	6	259	0	9184.349	260.4866	4.980835	0.000283
Neisseria	1B	3B	7	30	223	0	8649.386	257.4067	4.907619	2.76E-09
Neisseria	1B	2A	7	7	223	0	8649.386	257.9026	4.90474	0.001584
Paracoccus	1B	3B	7	30	200	0	7147.432	257.4067	4.636616	1.91E-12
Paracoccus	1B	2A	7	7	200	0	7147.432	257.9026	4.633582	0.000184
Paracoccus	1B	2B	7	6	200	0	7147.432	260.4866	4.620582	0.000665
Cloacibacterium	1A	1B	8	7	202	0	6501.51	254.4233	4.51247	0.000544

Table F.28.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's RDP pipeline, at the OTU level (classification confidence 80 %, cluster distance 0.03, "other" threshold 2 %). For each DA OTU level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts (rcG1, rcG2) and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each OTU and contrast are included. In addition, for each OTU the RDP classification result and confidence are available from the table.

TAX	G1	G2	# G1	# G2	rc G1	rc G2	cpm G1	cpm G2	logFC	FDR	Phylum	Class	Order	Familiy	Genus	
86	2B	3B	6	30	5353	1	151605.95	175.13	9.5932	3.94E-33	Proteobacteria;	Betaproteobacteria;	Neisseriales;	Neisseriaceae;	Neisseria;	0.93;
76	2B	3B	6	30	3313	23	93908.90	299.12	8.1957	3.06E-25	Proteobacteria;	Gammaproteobacteria;	Pasteurellales;	Pasteurellaceae;	Haemophilus;	0.91;
171	2B	3B	6	30	418	0	12032.16	169.49	5.9821	4.00E-25	Bacteroidetes;	Bacteroidia;	Bacteroidales;	Bacteroidaceae;	Bacteroides;	1;
301	2B	3B	6	30	408	1	11712.03	175.15	5.9012	1.23E-24	Proteobacteria;	Gammaproteobacteria;	Pasteurellales;	Pasteurellaceae;	Pasteurella;	0.43;
640	2B	3B	6	30	356	0	10240.95	169.49	5.7501	5.34E-24	Proteobacteria;	Gammaproteobacteria;	Pasteurellales;	Pasteurellaceae;	Haemophilus;	0.69;
359	2B	3B	6	30	294	1	8486.97	175.13	5.4376	4.51E-23	Proteobacteria;	Gammaproteobacteria;	Pasteurellales;	Pasteurellaceae;	Pasteurella;	0.59;
53	2B	3B	6	30	482	1	13805.49	175.13	6.1381	4.53E-22	Firmicutes;	Bacilli;	Lactobacillales;	Enterococcaceae;	Enterococcus;	0.99;
613	2B	3B	6	30	237	0	6848.70	169.49	5.1713	5.71E-20	Firmicutes;	Bacilli;	Lactobacillales;	Streptococcaceae;	Streptococcus;	1;
290	2A	3B	7	30	175	0	4396.57	169.49	4.5343	1.60E-16	Actinobacteria;	Actinobacteria;	Actinomycetales;	Corynebacteriaceae;	Corynebacterium;	0.99;
76	2A	3B	7	30	1040	23	25339.63	299.12	6.3067	2.54E-16	Proteobacteria;	Gammaproteobacteria;	Pasteurellales;	Pasteurellaceae;	Haemophilus;	0.91;
27	2A	3B	7	30	138	3	3503.19	186.40	4.0856	2.50E-15	Bacteroidetes;	Bacteroidia;	Bacteroidales;	Prevotellaceae;	Prevotella;	1;
2730	1B	3B	7	30	200	0	5019.26	169.49	4.7243	5.29E-15	Proteobacteria;	Alphaproteobacteria;	Rhodobacterales;	Rhodobacteraceae;	Paracoccus;	0.97;
329	2A	3B	7	30	206	8	5145.39	214.67	4.4533	1.57E-14	Firmicutes;	Bacilli;	Lactobacillales;	Streptococcaceae;	Streptococcus;	1;
4776	1A	3B	8	30	102	0	2331.32	169.49	3.6248	1.29E-11	Proteobacteria;	Gammaproteobacteria;	Pseudomonadales;	Pseudomonadaceae;	Pseudomonas;	0.93;
226	1A	3B	8	30	120	0	2705.57	169.49	3.8381	2.20E-11	Fusobacteria;	Fusobacteria;	Fusobacteriales;	Leptotrichiaceae;	Streptobacillus;	0.98;
1621	2A	3B	7	30	155	12	3913.55	237.25	3.9284	2.85E-11	Proteobacteria;	Betaproteobacteria;	Neisseriales;	Neisseriaceae;	Neisseria;	0.75;
1752	1A	3B	8	30	92	0	2118.41	169.49	3.4879	3.51E-11	Actinobacteria;	Actinobacteria;	Actinomycetales;	Propionibacteriaceae;	Propionibacterium;	1;
123	1B	3B	7	30	177	5	4461.55	197.78	4.3557	4.23E-11	Firmicutes;	Bacilli;	Bacillales;	Staphylococcaceae;	Staphylococcus;	1;
61	2A	3B	7	30	522	59	12778.88	502.44	4.6113	5.14E-11	Firmicutes;	Bacilli;	Lactobacillales;	Streptococcaceae;	Streptococcus;	1;
86	1B	3B	7	30	256	1	6356.79	175.13	5.0219	1.72E-10	Proteobacteria;	Betaproteobacteria;	Neisseriales;	Neisseriaceae;	Neisseria;	0.93;
4955	1B	3B	7	30	85	0	2230.68	169.49	3.5618	1.45E-09	Proteobacteria;	Alphaproteobacteria;	Rhizobiales;	Bradyrhizobiaceae;	Bradyrhizobium;	0.96;
147	1A	3B	8	30	83	0	1923.54	169.49	3.3505	3.34E-09	Bacteroidetes;	Sphingobacteria;	Sphingobacteriales;	Sphingobacteriaceae;	Pseudosphingobacterium;	0.43;
2927	1A	3B	8	30	68	0	1609.35	169.49	3.0960	3.34E-09	Firmicutes;	Bacilli;	Lactobacillales;	Carnobacteriaceae;	Carnobacterium;	0.98;
1587	1A	3B	8	30	73	0	1712.18	169.49	3.1846	3.34E-09	Proteobacteria;	Betaproteobacteria;	Burkholderiales;	Burkholderiales_incertae_sedis;	Tepidimonas;	0.9;
2947	1A	3B	8	30	82	6	1902.40	203.37	3.0988	4.92E-09	Fusobacteria;	Fusobacteria;	Fusobacteriales;	Leptotrichiaceae;	Leptotrichia;	1;
2574	1A	3B	8	30	82	1	1902.40	175.15	3.2926	4.92E-09	Proteobacteria;	Gammaproteobacteria;	Pseudomonadales;	Pseudomonadaceae;	Pseudomonas;	0.71;
1641	1A	3B	8	30	221	32	4840.31	350.52	3.7095	5.31E-09	Bacteroidetes;	Flavobacteria;	Flavobacteriales;	Flavobacteriaceae;	Cloacibacterium;	0.93;
####	1A	3B	8	30	61	0	1460.99	169.49	2.9583	1.05E-08	Proteobacteria;	Gammaproteobacteria;	Pseudomonadales;	Moraxellaceae;	Psychrobacter;	1;
1104	1A	3B	8	30	129	16	2895.80	259.67	3.3762	1.43E-08	Bacteroidetes;	Bacteroidia;	Bacteroidales;	Porphyromonadaceae;	Porphyromonas;	0.98;
998	1A	3B	8	30	128	21	2881.26	288.24	3.2297	2.41E-08	Proteobacteria;	Gammaproteobacteria;	Enterobacteriales;	Enterobacteriaceae;	Escherichia/Shigella;	0.44;

Table F.29.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's BLAT pipeline, at the phylum level (classification confidence 80 %, "other" threshold 2 %). For each DA phylum level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each phylum and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Glomeromycota	1B	3B	5	26	782	0	1777.01	247.56	3.4023	2.60E-05
Basidiomycota	1B	3B	5	26	19746	26386	831212.57	99653.99	3.0601	0.008294
Glomeromycota	1B	2B	5	6	782	0	1777.01	222.06	3.4013	0.01194
Glomeromycota	1B	2A	5	7	782	0	1777.01	210.52	3.3578	0.014295
Ascomycota	1B	3B	5	26	2302	200693	95992.92	1227167.85	-3.6762	0.015398
Glomeromycota	1A	1B	4	5	0	782	136.05	1777.01	-3.9945	0.024425
Ascomycota	1B	2B	5	6	2302	37782	95992.92	1212548.01	-3.6589	0.03303

Table F.30.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's BLAT pipeline, at the class level (classification confidence 80 %, "other" threshold 2 %). For each DA class level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each class and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Glomeromycetes	1B	3B	5	26	782	0	23722.06	154.07	7.1472	7.98E-19
Glomeromycetes	1B	2A	5	7	782	0	23722.06	159.87	7.0799	1.97E-06
Saccharomycetes	1B	3B	5	26	0	182994	124.45	1103194.42	-13.0071	2.45E-06
Saccharomycetes	1B	2B	5	6	0	37545	124.45	1017497.37	-12.8905	6.28E-06
Glomeromycetes	1B	2B	5	6	782	0	23722.06	160.30	7.0761	6.28E-06
Saccharomycetes	1B	2A	5	7	0	28501	124.45	661281.01	-12.2688	8.26E-06
Saccharomycetes	1A	3B	4	26	0	182994	147.55	1103194.42	-12.7225	9.18E-05
Saccharomycetes	1A	2B	4	6	0	37545	147.55	1017497.37	-12.6058	0.000136
Saccharomycetes	1A	2A	4	7	0	28501	147.55	661281.01	-11.9841	0.000234
Glomeromycetes	1A	1B	4	5	0	782	147.55	23722.06	-7.1832	0.000496
Sordariomycetes	1A	2A	4	7	1665	0	60177.90	159.87	8.4223	0.001576
Tremellomycetes	1A	2A	4	7	505	0	18354.95	159.87	6.7104	0.001576
Leotiomycetes	2A	3B	7	26	393	137	8630.70	849.73	3.3114	0.001903
Microbotryomycetes	2A	3B	7	26	0	4066	159.87	10461.70	-5.9019	0.001903
Sordariomycetes	2A	3B	7	26	0	8810	159.87	23820.52	-7.0863	0.001903
Dothideomycetes	2A	3B	7	26	0	3269	159.87	8408.19	-5.5875	0.001903
Tremellomycetes	1A	2B	4	6	505	0	18354.95	160.30	6.7067	0.003742
Leotiomycetes	1B	2A	5	7	0	393	124.45	8630.70	-6.0224	0.004353
Leotiomycetes	2A	2B	7	6	393	0	8630.70	160.30	5.6202	0.004451
Microbotryomycetes	2B	3B	6	26	0	4066	160.30	10461.70	-5.8983	0.006149
Dothideomycetes	2B	3B	6	26	0	3269	160.30	8408.19	-5.5839	0.006149
Microbotryomycetes	1B	3B	5	26	0	4066	124.45	10461.70	-6.2854	0.00778
Exobasidiomycetes	2A	3B	7	26	0	581	159.87	3376.76	-4.2747	0.008041
Eurotiomycetes	1A	2B	4	6	4855	35	183277.98	1040.43	7.4371	0.009306
Leotiomycetes	1A	2A	4	7	0	393	147.55	8630.70	-5.7277	0.013542
Tremellomycetes	2A	3B	7	26	0	549	159.87	2924.61	-4.0700	0.01544
Exobasidiomycetes	2B	3B	6	26	0	581	160.30	3376.76	-4.2710	0.019501
Exobasidiomycetes	1B	3B	5	26	0	581	124.45	3376.76	-4.6651	0.023308
Tremellomycetes	2B	3B	6	26	0	549	160.30	2924.61	-4.0663	0.032321
Microbotryomycetes	1A	3B	4	26	0	4066	147.55	10461.70	-6.0059	0.034042

Table F.31.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's BLAT pipeline, at the order level (classification confidence 80 %, "other" threshold 2 %). For each DA order level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each order and contrast is included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Malasseziales	2B	3B	6	26	4751	0	116765.03	146.69	9.4705	2.38E-26
Glomerales	1B	3B	5	26	780	0	23111.81	146.69	7.1343	8.24E-20
Hypocreales	1A	3B	4	26	1665	0	59827.98	146.69	8.5059	3.56E-18
Atheliales	1A	3B	4	26	230	0	8385.74	146.69	5.6739	2.14E-14
Hypocreales	1B	3B	5	26	735	0	18191.68	146.69	6.7898	2.33E-13
Thelephorales	1B	3B	5	26	154	0	4374.76	146.69	4.7381	8.53E-11
Malasseziales	2A	2B	7	6	0	4751	145.84	116765.03	-9.4768	3.15E-08
Glomerales	1B	2A	5	7	780	0	23111.81	145.84	7.1404	1.01E-06
Saccharomycetales	1B	3B	5	26	0	182989	136.58	1023075.25	-12.6926	1.90E-06
Glomerales	1B	2B	5	6	780	0	23111.81	146.13	7.1378	3.71E-06
Malasseziales	1B	2B	5	6	0	4751	136.58	116765.03	-9.5619	3.71E-06
Saccharomycetales	1B	2B	5	6	0	37544	136.58	916261.76	-12.5336	4.70E-06
Hypocreales	1A	2A	4	7	1665	0	59827.98	145.84	8.5123	5.53E-06
Atheliales	1A	2A	4	7	230	0	8385.74	145.84	5.6798	1.13E-05
Saccharomycetales	1B	2A	5	7	0	28499	136.58	595292.29	-11.9114	1.40E-05
Saccharomycetales	1A	3B	4	26	0	182989	140.64	1023075.25	-12.6547	3.81E-05
Atheliales	1A	2B	4	6	230	0	8385.74	146.13	5.6772	5.09E-05
Malasseziales	1A	2B	4	6	0	4751	140.64	116765.03	-9.5239	5.09E-05
Saccharomycetales	1A	2B	4	6	0	37544	140.64	916261.76	-12.4957	5.93E-05
Hypocreales	1B	2A	5	7	735	0	18191.68	145.84	6.7962	6.27E-05
Saccharomycetales	1A	2A	4	7	0	28499	140.64	595292.29	-11.8735	0.000104
Russulales	1A	2A	4	7	2966	0	97136.55	145.84	9.2113	0.00011
Hymenochaetales	1B	2A	5	7	1258	0	37191.49	145.84	7.8266	0.000116
Thelephorales	1B	2A	5	7	154	0	4374.76	145.84	4.7440	0.000116
Hymenochaetales	1B	3B	5	26	1258	309	37191.49	2013.13	4.1973	0.000141
Atheliales	1A	1B	4	5	230	0	8385.74	136.58	5.7653	0.00022
Glomerales	1A	1B	4	5	0	780	140.64	23111.81	-7.1876	0.00022
Agaricales	1B	2B	5	6	3209	0	86676.79	146.13	9.0444	0.000326
Hymenochaetales	1B	2B	5	6	1258	0	37191.49	146.13	7.8240	0.000326
Thelephorales	1B	2B	5	6	154	0	4374.76	146.13	4.7415	0.000326

Table F.32.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's BLAT pipeline, at the family level (classification confidence 80 %, "other" threshold 2 %). For each DA family level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each family and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Ophiocordycipitaceae	1A	3B	4	26	1665	0	60257.48	144.77	8.5328	2.72E-29
Pluteaceae	1A	3B	4	26	941	0	33101.05	144.77	7.6690	9.20E-26
Schizophyllaceae	1B	3B	5	26	972	0	26942.62	144.77	7.3721	9.99E-26
Hapalopilaceae	1A	3B	4	26	911	0	32050.32	144.77	7.6225	1.00E-25
Meripilaceae	2A	3B	7	26	4639	0	94764.12	144.77	9.1858	2.56E-22
Glomeraceae	1B	3B	5	26	780	0	22667.06	144.77	7.1227	5.25E-22
Stereaceae	1A	3B	4	26	1017	0	35762.92	144.77	7.7805	9.60E-22
Strophariaceae	1B	3B	5	26	452	0	12249.72	144.77	6.2365	1.00E-20
Meruliaceae	1B	3B	5	26	1216	0	32872.17	144.77	7.6590	9.29E-20
Fomitopsidaceae	1B	3B	5	26	628	0	17018.50	144.77	6.7101	5.78E-19
Cortinariaceae	1B	3B	5	26	318	0	8664.84	144.77	5.7380	1.73E-18
Atheliaceae	1A	3B	4	26	230	0	8352.52	144.77	5.6847	1.49E-16
Tremellaceae	1A	3B	4	26	498	0	17585.17	144.77	6.7572	2.26E-16
Stereaceae	1B	3B	5	26	440	0	12112.00	144.77	6.2199	9.77E-16
Tremellaceae	1B	3B	5	26	481	0	13402.24	144.77	6.3656	1.02E-15
Psathyrellaceae	1B	3B	5	26	323	0	8740.05	144.77	5.7504	4.32E-14
Meripilaceae	1B	3B	5	26	525	0	14246.84	144.77	6.4538	8.07E-13
Peniophoraceae	1B	3B	5	26	156	0	4439.67	144.77	4.7756	1.28E-12
Fomitopsidaceae	1A	3B	4	26	208	0	7567.25	144.77	5.5425	1.36E-12
Thelephoraceae	1B	3B	5	26	154	0	4239.10	144.77	4.7105	2.14E-12
Psathyrellaceae	1A	3B	4	26	211	0	7674.33	144.77	5.5628	2.67E-12
Mycenaceae	1B	3B	5	26	148	0	4128.88	144.77	4.6722	2.97E-12
Ophiocordycipitaceae	1A	2A	4	7	1665	0	60257.48	144.79	8.5326	1.60E-09
Hapalopilaceae	1A	2A	4	7	911	0	32050.32	144.79	7.6223	6.00E-09
Pluteaceae	1A	2A	4	7	941	0	33101.05	144.79	7.6688	6.00E-09
Schizophyllaceae	1B	2A	5	7	972	0	26942.62	144.79	7.3719	1.67E-08
Ophiocordycipitaceae	1A	2B	4	6	1665	0	60257.48	146.20	8.5200	3.02E-08
Meruliaceae	2A	3B	7	26	166	0	3467.12	144.77	4.4215	7.41E-08
Hapalopilaceae	1A	2B	4	6	911	0	32050.32	146.20	7.6097	8.00E-08
Pluteaceae	1A	2B	4	6	941	0	33101.05	146.20	7.6562	8.00E-08

Table F.33.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's BLAT pipeline, at the genus level (classification confidence 80 %, "other" threshold 2 %). For each DA genus level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each genus and contrast are included.

TAX	G1	G2	num samples G1	num samples G2	rawCounts G1	rawCounts G2	cpm G1	cpm G2	logFC	FDR
Ophiocordyceps	1A	3B	4	26	1665	0	59830.94	143.52	8.5341	4.07E-30
Amylostereum	1A	3B	4	26	1017	0	36251.41	143.52	7.8116	3.70E-27
Pluteus	1A	3B	4	26	941	0	33553.04	143.52	7.7001	8.19E-27
Bjerkandera	1A	3B	4	26	911	0	32487.90	143.52	7.6535	1.01E-26
Rigidoporus	2A	3B	7	26	4639	0	95173.01	143.52	9.2036	4.75E-26
Mensularia	1B	3B	5	26	1257	0	36310.56	143.52	7.8139	4.03E-25
Schizophyllum	1B	3B	5	26	705	0	20255.60	143.52	6.9724	5.62E-25
Hypholoma	1B	3B	5	26	343	0	9928.04	143.52	5.9450	6.93E-20
Clitocybe	1A	3B	4	26	335	0	12152.33	143.52	6.2361	7.47E-20
Resinicium	1B	3B	5	26	915	0	26240.28	143.52	7.3456	1.05E-18
Kazachstania	2A	3B	7	26	687	0	14129.59	143.52	6.4534	3.60E-18
Postia	1B	3B	5	26	341	0	9879.34	143.52	5.9379	6.64E-18
Physisporinus	1B	3B	5	26	294	0	8537.36	143.52	5.7277	4.91E-17
Dioszegia	1A	3B	4	26	498	0	17824.43	143.52	6.7882	9.92E-17
Stereum	1B	3B	5	26	270	0	7852.10	143.52	5.6073	1.48E-16
Piptoporus	1A	3B	4	26	208	0	7599.56	143.52	5.5601	1.50E-16
Porotheleum	1B	3B	5	26	267	0	7752.76	143.52	5.5890	1.56E-16
Dioszegia	1B	3B	5	26	479	0	13795.07	143.52	6.4189	3.33E-16
Grandinia	1B	3B	5	26	249	0	7239.74	143.52	5.4904	3.51E-16
Blumeria	2A	3B	7	26	393	0	8194.10	143.52	5.6686	1.47E-15
Psathyrella	1B	3B	5	26	274	0	7952.27	143.52	5.6255	6.43E-14
Mycena	1B	3B	5	26	148	0	4368.67	143.52	4.7640	6.10E-13
Psathyrella	1A	3B	4	26	210	0	7671.25	143.52	5.5737	1.31E-12
Peniophora	1B	3B	5	26	138	0	4083.15	143.52	4.6669	1.55E-12
Ophiocordyceps	1A	2A	4	7	1665	0	59830.94	143.54	8.5339	1.04E-09
Amylostereum	1A	2A	4	7	1017	0	36251.41	143.54	7.8114	2.80E-09
Bjerkandera	1A	2A	4	7	911	0	32487.90	143.54	7.6534	2.80E-09
Pluteus	1A	2A	4	7	941	0	33553.04	143.54	7.6999	2.80E-09
Ophiocordyceps	1A	2B	4	6	1665	0	59830.94	143.71	8.5324	2.07E-08
Mensularia	1B	2A	5	7	1257	0	36310.56	143.54	7.8137	2.30E-08

Table F.34.: The table presents the top 30 differentially abundant features determined by edgeR, based on the feature table obtained from SnoWMAn's BLAT pipeline, at the OTU level (classification confidence 80 %, "other" threshold 2 %). For each DA OTU level feature (TAX), the contrast (G1, G2), the number of samples per contrast group, as well as the raw counts (rcG1, rcG2) and counts per million (cpm) normalized by library size, per contrast group, are given. Additionally, statistical result parameters of edgeR, the logFC and the FDR, for each OTU and contrast are included. In addition for each OTU, the BLAT classification result according to the available UNITE (release October 2013) annotation, is added to the table.

TAX	G1	G2	# G1	#G2	rc G1	rc G2	cpm G1	cpm G2	logFC	FDR	Phylum	Class	Order	Familiy	Genus	Species
269	1A	3B	4	26	1947	2	60646.11	134.03	8.6638	4.09E-33	Basidiomycota;	Agaricomycetes;	Polyporales;	Polyporaceae;	Trametes;	Trametes_gibbosa;
443	1A	3B	4	26	946	1	29530.40	129.28	7.6727	2.55E-32	Basidiomycota;	Agaricomycetes;	Russulales;	Amylostereaceae;	Amylostereum;	Amylostereum_areolatum;
150	1A	3B	4	26	749	0	23406.72	124.50	7.3860	1.40E-31	Basidiomycota;	Agaricomycetes;	Agaricales;	Pluteaceae;	Pluteus;	Pluteus_brunneidiscus;
477	1A	3B	4	26	658	0	20578.02	124.50	7.2003	9.74E-31	Basidiomycota;	Exobasidiomycetes;	Incertae sedis;	Incertae sedis;	Meira;	Meira_sp_07F1061;
417	1A	3B	4	26	1646	2	51289.63	134.07	8.4219	9.74E-31	Ascomycota;	Eurotiomycetes;	Eurotiales;	Trichocomaceae;	Aspergillus;	Aspergillus_sp_CCF_4264;
454	1A	3B	4	26	835	1	26080.00	129.26	7.4936	9.43E-30	Basidiomycota;	Agaricomycetes;	Polyporales;	Meruliaceae;	Bjerkandera;	Bjerkandera_fumosa;
226	1B	3B	5	26	381	0	9539.63	124.50	6.0922	1.57E-29	Basidiomycota;	Agaricomycetes;	Hymenochaetales;	Hymenochaetaceae;	Mensularia;	Mensularia_radiata;
435	1B	3B	5	26	907	0	22676.68	124.50	7.3402	1.80E-28	Basidiomycota;	Agaricomycetes;	Hymenochaetales;	Hymenochaetaceae;	Onnia;	Onnia_tomentosa;
725	1B	3B	5	26	236	0	5856.52	124.50	5.3905	1.80E-28	Basidiomycota;	Agaricomycetes;	Russulales;	Stereaceae;	Stereum;	Stereum_gausapatum;
842	2A	3B	7	26	4148	7	73926.59	158.06	8.7347	1.19E-27	Basidiomycota;	Agaricomycetes;	Agaricales;	Mycenaceae;	Mycena;	Mycena_olida;
609	1B	3B	5	26	570	0	14238.26	124.50	6.6693	2.03E-27	Basidiomycota;	Microbotryomycetes;	Sporidiobolales;	Incertae sedis;	Sporobolomyces;	Sporobolomyces_ruberrimus;
562	1B	3B	5	26	471	1	11501.89	129.26	6.3138	2.03E-26	Basidiomycota;	Agaricomycetes;	Polyporales;	Meripilaceae;	Rigidoporus;	Rigidoporus_crocatus;
131	1B	3B	5	26	660	0	16472.84	124.50	6.8795	2.65E-26	Ascomycota;	Sordariomycetes;	Hypocreales;	unidentified;	unidentified;	Hypocreales_sp_H4429;
850	1B	3B	5	26	499	0	12481.12	124.50	6.4795	2.65E-26	Basidiomycota;	Tremellomycetes;	Tremellales;	unidentified;	unidentified;	Tremellales_sp;
22	2A	3B	7	26	2796	0	49871.41	124.50	8.4768	4.39E-26	Basidiomycota;	Agaricomycetes;	Agaricales;	Marasmiaceae;	Marasmiellus;	Marasmiellus_palmivorus;
758	1B	3B	5	26	372	1	9109.81	129.26	5.9781	1.55E-25	Basidiomycota;	Agaricomycetes;	Polyporales;	Fomitopsidaceae;	Fomitopsis;	Fomitopsis_pinicola;
489	1B	3B	5	26	342	0	8385.02	124.50	5.9070	3.09E-25	Basidiomycota;	Agaricomycetes;	Agaricales;	Strophariaceae;	Hypholoma;	Hypholoma_sublateritium;
712	1A	3B	4	26	351	0	11035.04	124.50	6.3023	6.73E-25	Basidiomycota;	Agaricomycetes;	Agaricales;	unidentified;	unidentified;	Agaricales_sp;
511	1B	3B	5	26	345	1	8619.22	129.26	5.8979	9.90E-25	unidentified;	unidentified;	unidentified;	unidentified;	unidentified;	Fungi_sp;
831	1B	3B	5	26	1259	20	30950.42	219.89	7.0386	1.28E-24	Basidiomycota;	Agaricomycetes;	Russulales;	Bondarzewiaceae;	Heterobasidion;	Heterobasidion_abietinum;
136	1B	3B	5	26	631	12	15368.48	181.55	6.2855	4.79E-24	Basidiomycota;	Agaricomycetes;	Agaricales;	Strophariaceae;	Hypholoma;	Hypholoma_fasciculare;
475	1B	3B	5	26	222	7	5503.17	157.86	4.9920	4.79E-24	Basidiomycota;	Agaricomycetes;	Polyporales;	Meruliaceae;	Phlebia;	Phlebia_rufa;
280	1A	3B	4	26	597	0	18681.86	124.50	7.0610	6.17E-23	Ascomycota;	Eurotiomycetes;	Eurotiales;	Trichocomaceae;	unidentified;	uncultured_Penicillium;
116	1A	3B	4	26	443	2	13894.83	134.05	6.5390	6.17E-23	Ascomycota;	Dothideomycetes;	Pleosporales;	Pleosporaceae;	Epicoccum;	Epicoccum_nigrum;
726	1A	3B	4	26	272	1	8579.36	129.29	5.8912	1.35E-22	Ascomycota;	Dothideomycetes;	Pleosporales;	Incertae sedis;	Phoma;	Phoma_versabilis;
721	1B	3B	5	26	682	23	16882.76	234.29	6.0791	3.24E-22	Basidiomycota;	Agaricomycetes;	Agaricales;	Schizophyllaceae;	Schizophyllum;	Schizophyllum_commune;
611	1B	3B	5	26	145	1	3638.73	129.28	4.6582	5.53E-22	Basidiomycota;	Agaricomycetes;	Polyporales;	Meruliaceae;	Steccherinum;	Steccherinum_ochraceum;
252	1A	3B	4	26	208	0	6589.94	124.50	5.5599	7.73E-22	Basidiomycota;	Agaricomycetes;	Agaricales;	Pluteaceae;	Pluteus;	Pluteus_cervinus;
494	1B	3B	5	26	318	0	8000.33	124.50	5.8388	1.39E-21	unidentified;	unidentified;	unidentified;	unidentified;	unidentified;	fungal_sp_12S_1;
824	1B	3B	5	26	172	0	4301.86	124.50	4.9468	1.90E-21	Basidiomycota;	Agaricomycetes;	Russulales;	Bondarzewiaceae;	Heterobasidion;	Heterobasidion_sp_Cui2;

Table F.35.: Summary of the BAL study BAL culture result. For each sample the culture result is described by bacterial strains and respective germination number. Positive cultures, containing pathogenic strains are highlighted by bold text.

Samplename	Culture	Patho	Spezies	BacCount
087-1A-NTS-0	NEG	N	Non	0
095-1A-NTS-0	**POS**	**N**	**alpha haemolyt. Streptococcus**	**1.00E+01**
097-1A-NTS-0	NEG	N	Non	0
098-1B-NTS-0	NEG	N	Non	0
099-1B-NTS-0	NEG	N	Non	0
100-1B-NTS-0	**POS**	**N**	**alpha haemolyt. Streptococcus; nicht hoemolyt. Streptococcus**	**1.00E+05; 1.00E+05**
101-1B-NTS-0	NEG	N	Non	0
102-1B-NTS-0	NEG	N	Non	0
103-1B-NTS-0	**POS**	**Y**	**Pseudomonas aeruginosa**	**1.00E+05**
104-1B-NTS-0	NEG	N	Non	0
105-1A-NTS-0	**POS**	**N**	**alpha haemolyt. Streptococcus; nicht hoemolyt. Streptococcus**	**1.00E+03**
106-1A-NTS-0	**POS**	**N**	**Koagulase negative Staphylokokken**	**1.00E+02**
107-1A-NTS-0	NEG	N	Non	0
108-1A-NTS-0	**POS**	**N**	**alpha haemolyt. Streptococcus; Neisseria**	**1.00E+03; 1.0E+03**
109-1A-NTS-0	**POS**	**N**	**Koagulase negative Staphylokokken; Neisseria**	**100;10**
201-2A-NTS-0	**POS**	**Y**	**Candida boidinii; Candida dubliniensis; Klebsiella pneumoniae**	**1.00E+03; 1.00E+04**
202-2A-NTS-0	**POS**	**N**	**Koagulase negative Staphylokokken**	**1.00E+05**
203-2A-NTS-0	**POS**	**N**	**Koagulase negative Staphylokokken; Neisseria**	**1.00E+03; 1.00E+03**
252-2B-NTS-0	NEG	N	Non	0
255-2B-NTS-0	**POS**	**Y**	**Candida albicans; Schimmelpilz; Pseudomonas aeruginosa**	**1.00E+03; 1.00E+03; 1.00E+03**
256-2B-NTS-0	NEG	N	Non	0
257-2B-NTS-0	NEG	N	Non	0
301-3B-VAP-0	**POS**	**N**	**Candida albicans**	**1.00E+03**
301-3B-VAP-1	**POS**	**N**	**Candida albicans**	**1.00E+02**
302-3B-VAP-0	**POS**	**N**	**Candida albicans**	
302-3B-VAP-1	**POS**	**Y**	**Pseudomonas aeruginosa**	**1.00E+06**
303-3B-VAP-0	**POS**	**N**	**Candida albicans**	**1.00E+03**
303-3B-VAP-1	**POS**	**N**	**Candida albicans**	**1.00E+01**
304-3B-ASP-0	NEG	N	Non	0
304-3B-ASP-1	**POS**	**Y**	**Escherichia coli**	**1.00E+06**
305-3B-VAP-0	**POS**	**Y**	**Staphylococcus aureus; Escherichia coli**	**1.00E+02; 1.00E+02**
306-3B-VAP-0	**POS**	**N**	**alpha haemolyt. Streptococcus; Neisseria**	
309-3B-VAP-0	**POS**	**N**	**Candida albicans**	
310-3B-NAP-0	**POS**	**Y**	**Escherichia coli**	**1.00E+02**
312-3B-VAP-0	**POS**	**Y**	**Staphylococcus aureus; Candida parapsilosis**	**1.00E+01**
313-3B-NAP-0	**POS**	**Y**	**Klebsiella sp; Enterobacter cloacae**	**1.00E+06; 1.00E+06**
313-3B-VAP-1	**POS**	**Y**	**Klebsiella pneumoniae**	**1.00E+01**
314-3B-CAP-0	NEG	N	Non	0
318-3B-ASP-0	**POS**	**Y**	**Enterobacter cloacae; Staphylococcus aureus; Candida albicans; Candida glabrata**	**1.00E+01; 1.00E+03; 1.00E+03; 1.00E+06**
319-3B-VAP-0	**POS**	**N**	**Candida albicans**	**1.00E+01**
320-3B-ASP-0	**POS**	**Y**	**Escherichia coli**	**1.00E+02**
321-3B-NAP-0	NEG	N	Non	0
322-3B-NAP-0	NEG	N	Non	0
323-3B-VAP-0	**POS**	**N**	**Candida albicans**	**1.00E+01**
324-3B-ASP-0	NEG	N	Non	0
325-3B-VAP-0	**POS**	**Y**	**Proteus mirabilis; Streptococcus viridans**	**0.00E+00**
326-3B-VAP-0	**POS**	**N**	**Candida albicans**	**1.00E+01**
327-3B-ASP-0	**POS**	**Y**	**Klebsiella oxytoca; Enterobacter cloacae**	**1.00E+05; 1.00E+05**
328-3B-NAP-0	NEG	N	Non	0
401-2A-NTS-0	**POS**	**Y**	**Escherichia coli; Candida albicans**	**1.00E+05; 1.00E+03**
402-2A-NTS-0	**POS**	**N**	**Koagulase negative Staphylokokken; Neisseria; Enterobacter**	**10000000; 10000000; 1.00E+5**
403-2A-NTS-0	**POS**	**Y**	**Klebsiella pneumoniae; Haemophilus influenzae; Corynebacterium**	**1.00E+06; 1.00E+05; 1.00E+06**
405-2B-NTS-0	**POS**	**N**	**Candida albicans; Staphylococcus aureus**	**1.00E+05; 1.00E+04**
406-2A-NTS-0	**POS**	**Y**	**Pseudomonas aeruginosa**	**1.00E+05**
608-3B-ASP-0	NEG	N	Non	0
609-3B-ASP-0	**POS**	**Y**	**Staphylococcus aureus**	**1.00E+04**
610-2B-NTS-0	**POS**	**Y**	**Streptococcus pneumoniae; Haemophilus influenzae**	**1.00E+06; 1.00E+06**
612-3B-VAP-0	**POS**	**N**	**Candida albicans**	**1.00E+02**

Table F.36.: Comparison result of the high-throughput classification vs traditional BAL cultures, for fungal strains. All samples of the BAL study having either a positive or negative culturing result are listed within this table. For each sample the number of observed amplicons for the fungal reference species, *Aspergillus robustus, Candida albicans, Candida dubliniensis, Candida glabrata,* and *Candida parapsilosis*, is presented.

	Representative sequence and Accession number				
	Aspergillus robustus	Candida albicans	Candida dubliniensis	Candida glabrata	Candida parapsilosis
	EF661435.1	AB437043.1	AJ865083.1	HE993757.1	FM172980.1
087-1A-NTS-0	9	0	0	0	0
095-1A-NTS-0	0	1	0	0	0
098-1B-NTS-0	9	23	1	0	0
100-1B-NTS-0	276	19	0	0	0
101-1B-NTS-0	0	0	10	0	0
103-1B-NTS-0	8	36	0	0	0
104-1B-NTS-0	17	0	0	0	0
105-1A-NTS-0	0	39	2	0	0
107-1A-NTS-0	2	6	0	0	0
201-2A-NTS-0	0	17	8638	0	0
202-2A-NTS-0	0	0	0	7	0
203-2A-NTS-0	1	7251	158	0	1
252-2B-NTS-0	0	5826	144	0	0
255-2B-NTS-0	0	8228	195	0	0
256-2B-NTS-0	0	8107	202	0	0
257-2B-NTS-0	0	1	0	0	0
301-3B-VAP-0	0	16093	220	0	6
302-3B-VAP-0	0	10805	195	0	0
302-3B-VAP-1	0	5036	67	0	0
303-3B-VAP-0	0	7168	122	0	0
303-3B-VAP-1	0	18638	214	0	0
304-3B-ASP-0	261	11376	110	0	0
304-3B-ASP-1	0	5891	90	2	0
305-3B-VAP-0	0	7	0	0	0
306-3B-VAP-0	6	89	2	0	0
309-3B-VAP-0	1	21969	262	0	0
313-3B-NAP-0	323	5	0	0	0
313-3B-VAP-1	0	3	0	0	0
318-3B-ASP-0	0	1293	13	1233	0
319-3B-VAP-0	0	15826	155	0	7
320-3B-ASP-0	0	155	21	0	0
322-3B-NAP-0	0	11903	149	0	0
323-3B-VAP-0	1	11443	125	0	2
324-3B-ASP-0	0	18318	198	3	1
325-3B-VAP-0	0	18	0	0	0
326-3B-VAP-0	0	8412	80	0	0
327-3B-ASP-0	142	797	0	0	0
328-3B-NAP-0	0	1219	10	14	0
401-2A-NTS-0	0	10035	190	0	1
402-2A-NTS-0	0	0	1	0	0
403-2A-NTS-0	0	492	1	0	0
405-2B-NTS-0	0	14623	332	0	13
406-2A-NTS-0	0	1530	32	0	0
608-3B-ASP-0	0	0	0	475	0
609-3B-ASP-0	0	2914	60	0	0
610-2B-NTS-0	0	14	0	0	0
612-3B-VAP-0	0	6867	3185	0	1

Table F.37.: Comparison result of the high-throughput classification vs traditional BAL cultures, for bacterial strains. All samples of the BAL study having either a positive or negative culturing result are listed within this table. For each sample the number of observed amplicons for the bacterial reference species, *Corynebacterium, E.coli, Enterobacter cloacae,Enterococcus faecalis,* and *Haemophilus influenzae*, is presented.

	Representative sequence and Accession number				
	Corynebacterium	**E.coli**	**Enterobacter cloacae**	**Enterococcus faecalis**	**Haemophilus influenzae**
	HE983830.1	**J01859.1**	**KF535159.1**	**FJ378663.2**	**AY613741.1**
087-1A-NTS-0	49	447	64	95	0
095-1A-NTS-0	17	0	3	8	0
097-1A-NTS-0	1	108	22	111	67
098-1B-NTS-0	67	3	1	18	3
099-1B-NTS-0	0	1	0	11	0
100-1B-NTS-0	3	0	0	62	1
101-1B-NTS-0	62	1	0	2	0
102-1B-NTS-0	55	0	0	82	15
103-1B-NTS-0	46	2	0	56	0
104-1B-NTS-0	33	0	0	58	4
105-1A-NTS-0	1	11	0	16	0
106-1A-NTS-0	21	75	3	101	2
107-1A-NTS-0	25	20	0	177	84
108-1A-NTS-0	0	6	1	223	23
109-1A-NTS-0	0	5	4	12	91
201-2A-NTS-0	0	28	1	418	0
202-2A-NTS-0	6	0	0	54	4
203-2A-NTS-0	2	0	0	990	26
252-2B-NTS-0	1	1	0	3	2066
255-2B-NTS-0	1	0	0	19	4
256-2B-NTS-0	3	0	0	250	4
257-2B-NTS-0	8	6	0	105	2199
301-3B-VAP-0	28	0	0	1	0
301-3B-VAP-1	34	3	0	0	16
302-3B-VAP-0	0	0	0	0	0
302-3B-VAP-1	6	0	0	0	0
303-3B-VAP-0	0	0	0	205	0
303-3B-VAP-1	0	0	0	358	0
304-3B-ASP-0	2	6075	5	2	0
304-3B-ASP-1	118	13	1	65	0
305-3B-VAP-0	0	1	0	1248	0
306-3B-VAP-0	7	8	0	403	0
309-3B-VAP-0	17	8	3	39	0
310-3B-NAP-0	0	1	0	1044	1
312-3B-VAP-0	0	0	0	2	0
313-3B-NAP-0	0	2126	3	28	0
313-3B-VAP-1	0	22	0	324	0
314-3B-CAP-0	12	1	0	26	0
318-3B-ASP-0	0	0	0	9	0
319-3B-VAP-0	0	2	6	46	0
320-3B-ASP-0	2	4	0	11	1
321-3B-NAP-0	1	0	0	505	1
322-3B-NAP-0	0	34	0	15	0
323-3B-VAP-0	71	6	0	20	0
324-3B-ASP-0	39	0	0	1024	0
325-3B-VAP-0	0	0	0	39	8
326-3B-VAP-0	35	45	0	110	0
327-3B-ASP-0	0	0	0	0	1
328-3B-NAP-0	47	0	0	45	0
401-2A-NTS-0	1	9	0	5	573
402-2A-NTS-0	0	0	0	146	12
403-2A-NTS-0	1	1	0	38	573
405-2B-NTS-0	1	0	0	590	0
406-2A-NTS-0	0	0	0	5	0
608-3B-ASP-0	0	0	0	193	4
609-3B-ASP-0	0	0	0	59	0
610-2B-NTS-0	0	0	0	0	199
612-3B-VAP-0	16	0	0	0	0

Table F.38.: Comparison result of the high-throughput classification vs traditional BAL cultures, for bacterial strains. All samples of the BAL study having either a positive or negative culturing result are listed within this table. For each sample the number of observed amplicons for the bacterial reference species, *Klebsiella oxytoca, Klebsiella pneumoniae, Neisseria sp. oral, Streptococcus pneumoniae,* and *Proteus mirabilis*, is presented.

	Representative sequence and Accession number				
	Klebsiella oxytoca	**Klebsiella pneumoniae**	**Neisseria sp. oral**	**Streptococcus pneumoniae**	**Proteus mirabilis**
	AB626120.1	**KC990817.1**	**AY005028.1**	**GU326244.1**	**KF535110.1**
087-1A-NTS-0	0	14	14	34	66
095-1A-NTS-0	1	19	27	24	41
097-1A-NTS-0	0	5	34	0	140
098-1B-NTS-0	0	7	22	19	16
099-1B-NTS-0	0	2	5	0	25
100-1B-NTS-0	0	0	6	153	2
101-1B-NTS-0	0	47	45	0	104
102-1B-NTS-0	0	64	40	0	81
103-1B-NTS-0	0	26	0	0	44
104-1B-NTS-0	0	5	13	0	79
105-1A-NTS-0	0	19	0	91	73
106-1A-NTS-0	0	6	0	78	71
107-1A-NTS-0	0	15	47	38	59
108-1A-NTS-0	0	6	37	64	6
109-1A-NTS-0	1	18	1	0	5
201-2A-NTS-0	7	488	0	33	0
202-2A-NTS-0	0	2	17	1	56
203-2A-NTS-0	0	0	1	70	0
252-2B-NTS-0	0	0	103	1509	1
255-2B-NTS-0	0	17	36	0	21
256-2B-NTS-0	0	1	2	99	0
257-2B-NTS-0	0	4	4	39	16
301-3B-VAP-0	0	19	60	48	93
301-3B-VAP-1	17	38	1	37	64
302-3B-VAP-0	0	0	0	0	0
302-3B-VAP-1	0	7	7	0	12
303-3B-VAP-0	0	0	0	0	0
303-3B-VAP-1	0	4	0	0	1
304-3B-ASP-0	1	0	1	0	0
304-3B-ASP-1	0	17	68	1	33
305-3B-VAP-0	0	0	0	194	0
306-3B-VAP-0	0	0	0	46	1
309-3B-VAP-0	2	2	33	1	23
310-3B-NAP-0	0	0	0	7	3
312-3B-VAP-0	0	1	0	0	0
313-3B-NAP-0	3	671	1	0	3
313-3B-VAP-1	0	241	114	0	126
314-3B-CAP-0	0	33	1	3902	17
318-3B-ASP-0	0	0	0	3	0
319-3B-VAP-0	0	15	150	0	95
320-3B-ASP-0	0	3	1	1	4700
321-3B-NAP-0	0	0	4	556	0
322-3B-NAP-0	0	7	46	0	1425
323-3B-VAP-0	0	14	58	0	125
324-3B-ASP-0	0	35	105	225	36
325-3B-VAP-0	0	0	1	10	3
326-3B-VAP-0	0	4	5	0	119
327-3B-ASP-0	0	0	4	10	3
328-3B-NAP-0	0	21	69	0	85
401-2A-NTS-0	0	0	1	2029	22
402-2A-NTS-0	0	9	43	31	0
403-2A-NTS-0	0	3	1	1	0
405-2B-NTS-0	0	1	4	1	3
406-2A-NTS-0	0	0	39	0	0
608-3B-ASP-0	0	0	1	101	9
609-3B-ASP-0	0	0	1	89	0
610-2B-NTS-0	0	0	2	2156	0
612-3B-VAP-0	0	16	42	22	37

Table F.39.: Comparison result of the high-throughput classification vs traditional BAL cultures, for bacterial strains. All samples of the BAL study having either a positive or negative culturing result are listed within this table. For each sample the number of observed amplicons for the bacterial reference species, *Pseudomonas aeruginosa, Staphylococcus lugdunensi, Staphylococcus aureus, Streptococcus mitis,* and *Streptococcus viridans*, is presented.

	Representative sequence and Accession number				
	Pseudomonas aeruginosa	**Staphylococcus lugdunensi**	**Staphylococcus aureus**	**Streptococcus mitis**	**Streptococcus viridans**
	KJ156527.1	**AY903258.1**	**DQ630753.1**	**NR_028664.1**	**AF076036.1**
087-1A-NTS-0	543	6	0	89	74
095-1A-NTS-0	27	45	0	2171	42
097-1A-NTS-0	469	60	10	21	0
098-1B-NTS-0	14	115	0	32	19
099-1B-NTS-0	6	16	0	6	0
100-1B-NTS-0	1	0	0	1951	110
101-1B-NTS-0	113	127	183	9	0
102-1B-NTS-0	36	177	0	113	24
103-1B-NTS-0	1424	67	3	36	1
104-1B-NTS-0	32	85	0	0	0
105-1A-NTS-0	29	32	0	1168	80
106-1A-NTS-0	200	44	0	168	35
107-1A-NTS-0	96	138	25	790	28
108-1A-NTS-0	76	9	15	477	86
109-1A-NTS-0	132	65	109	58	21
201-2A-NTS-0	0	0	10	450	661
202-2A-NTS-0	5	32	1205	21	3775
203-2A-NTS-0	0	0	0	893	99
252-2B-NTS-0	7	6	1	48	0
255-2B-NTS-0	807	24	0	11	0
256-2B-NTS-0	2	0	0	1155	152
257-2B-NTS-0	8	18	0	1724	193
301-3B-VAP-0	28	74	0	28	0
301-3B-VAP-1	97	123	0	31	3
302-3B-VAP-0	5887	0	0	0	0
302-3B-VAP-1	4	9	19	3	27
303-3B-VAP-0	0	0	0	3	48
303-3B-VAP-1	30	16	39	8	58
304-3B-ASP-0	0	0	0	1	0
304-3B-ASP-1	45	158	131	26	0
305-3B-VAP-0	0	6	295	809	948
306-3B-VAP-0	1	0	1	802	54
309-3B-VAP-0	15	9	5	24	3119
310-3B-NAP-0	0	0	0	1947	11
312-3B-VAP-0	0	0	10	18	4992
313-3B-NAP-0	0	763	0	0	0
313-3B-VAP-1	59	127	0	28	0
314-3B-CAP-0	13	16	0	280	17
318-3B-ASP-0	1	1	3172	333	894
319-3B-VAP-0	32	65	4	35	13
320-3B-ASP-0	5	3	0	45	252
321-3B-NAP-0	1	1	0	1734	75
322-3B-NAP-0	27	112	1	0	81
323-3B-VAP-0	95	158	520	22	498
324-3B-ASP-0	68	48	1	179	173
325-3B-VAP-0	8	0	2	61	34
326-3B-VAP-0	76	86	0	143	52
327-3B-ASP-0	0	0	16	46	14
328-3B-NAP-0	42	116	35	0	0
401-2A-NTS-0	0	0	0	128	0
402-2A-NTS-0	8	24	12	363	1718
403-2A-NTS-0	2	0	0	54	2
405-2B-NTS-0	8	22	29	35	3137
406-2A-NTS-0	13	0	0	13	40
608-3B-ASP-0	0	4	0	1216	426
609-3B-ASP-0	0	1	4	138	191
610-2B-NTS-0	1	0	0	8	1
612-3B-VAP-0	6	33	40	0	0

Publications

DATABASE The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bas028, doi:10.1093/database/bas028

Original article

MuteinDB: the mutein database linking substrates, products and enzymatic reactions directly with genetic variants of enzymes

Andreas Braun[1,†], Bettina Halwachs[2,3,†], Martina Geier[1], Katrin Weinhandl[1], Michael Guggemos[1], Jan Marienhagen[4], Anna J. Ruff[4], Ulrich Schwaneberg[4], Vincent Rabin[1], Daniel E. Torres Pazmiño[5], Gerhard G. Thallinger[2,3,*] and Anton Glieder[1,3]

[1]Institute of Molecular Biotechnology, [2]Institute for Genomics and Bioinformatics, Graz University of Technology, [3]Austrian Centre of Industrial Biotechnology (ACIB GmbH), 8010 Graz, Austria, [4]Department of Biotechnology, RWTH Aachen University, 52074 Aachen, Germany and [5]Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, 9747 AG Groningen, the Netherlands

***Corresponding author:** Tel: +43 316 873 5343; Fax: +43 316 873 105343; Email: Gerhard.Thallinger@tugraz.at

[†]These authors contributed equally to this work.

Submitted 13 March 2012; Revised 11 May 2012; Accepted 31 May 2012

Mutational events as well as the selection of the optimal variant are essential steps in the evolution of living organisms. The same principle is used in laboratory to extend the natural biodiversity to obtain better catalysts for applications in biomanufacturing or for improved biopharmaceuticals. Furthermore, single mutation in genes of drug-metabolizing enzymes can also result in dramatic changes in pharmacokinetics. These changes are a major cause of patient-specific drug responses and are, therefore, the molecular basis for personalized medicine. MuteinDB systematically links laboratory-generated enzyme variants (muteins) and natural isoforms with their biochemical properties including kinetic data of catalyzed reactions. Detailed information about kinetic characteristics of muteins is available in a systematic way and searchable for known mutations and catalyzed reactions as well as their substrates and known products. MuteinDB is broadly applicable to any known protein and their variants and makes mutagenesis and biochemical data searchable and comparable in a simple and easy-to-use manner. For the import of new mutein data, a simple, standardized, spreadsheet-based data format has been defined. To demonstrate the broad applicability of the MuteinDB, first data sets have been incorporated for selected cytochrome P450 enzymes as well as for nitrilases and peroxidases.

Database URL: http://www.MuteinDB.org

Introduction

One of nature's fundamental mechanisms to create genetic diversity in living organisms is the creation of mutants, which, in turn, leads to evolution. Mutational events and selection of the optimal variant are essential to obtain a better catalyst. In human medicine, enzyme polymorphisms arising from evolutionary events have been identified since the 1960s (1). Physicians recognized that patients with the same disease responded differently to drugs, according to which allelic variant their genomes were carrying. This opened the road to what is nowadays called 'personalized medicine' (2). Additionally, industry desires to artificially improve enzymes through mutation and selection. To this end, efficient protein engineering tools to create tailor-made enzyme variants, named 'muteins', have been developed over the past decades (3). Muteins generated either by rational design or by directed or designed evolution were adapted to the needs of industrial processes or for completely new applications.

Increasing interest in personalized medicine and in tailor-made enzymes in the fast-growing biocatalysis industry has led to an exponential increase of literature about muteins and their influence on enzymes' kinetic properties.

In a plethora of examples, the artificial substitution of one or more amino acids in a polypeptide resulted in a significant increase or decrease of stability, turnover rate or substrate specificity for the enzyme (4). Even new reactions or activities on molecules that were not substrates for the natural parental enzyme can be caused by just a few or every single mutation. For example, esterases could be changed to hydroxynitrile lyases and epoxide hydrolases (5,6). Papain, a protease, was modified to an enzyme with efficient nitrile hydratase activity (7). Furthermore, the fatty acid hydroxylase CYP102A1 and the camphor hydroxylase CYP101 were redesigned to efficient alkane hydroxylases (8–10). By a single mutation, the broadly applied lipase CalB was modified to perform aldol additions and epoxidations (11). More recently, a transaminase showing almost no activity for a commercially interesting substrate was mutated to a highly active and selective catalyst enabling a new efficient industrial process for sitagliptin production (12).

Information about specific proteins and their muteins are widely spread in the literature. Many studies only describe single mutation and its effects without comparison to already known muteins. Possible additive effects of single amino acid changes are scarcely described or used. Even after a thorough and time-consuming literature search, researchers face the problem of assembling and presenting the data in an easy understandable and comprehensive way. Essential information may be lost such as details about potentially cooperative mutations or reactions one would not expect in certain protein families. Therefore, a web-accessible database combining available knowledge about a specific enzyme and its muteins in a single place are highly desirable. Such a database would allow researchers to access relevant information about their protein of interest in a fast and easy way and accelerate the engineering of new and improved variants.

Existing, comprehensive enzyme engineering databases such as CYPED are mainly focused on enzyme sequences and their structures (13). Only a few databases go beyond that and contain, to some extent, information about muteins and their properties. The most recently published database introducing mutein information is SuperCYP (14), which exclusively addresses human cytochrome P450s. Another example is SPROUTS (15), which provides details on the influence of point mutations on protein stability. The Protherm database (16) contains experimental thermodynamic data, and BRENDA (17), a well-known enzyme databases, includes only a small section about muteins. Finally, the Protein Mutant Database (18) includes references to mutant proteins from the literature. However, none of these databases provides kinetic characteristics of muteins and allows a fast, systematic and user-friendly way to search for known mutations and catalyzed reactions of interest. All these databases focus on enzymes and provide information about their variants from the view of the protein. Additionally, none of the existing databases is searchable by substrate or product molecule structures allowing comparison of muteins with respect to their catalytic properties.

In this article, we present the novel database MuteinDB (http://www.MuteinDB.org). It is a user-friendly graphically appealing database devoted to provide easy access to detailed information on naturally occurring and laboratory-evolved muteins as well as on the influence of mutations on kinetics of catalyzed reactions, including inhibition. It allows to search for the best biocatalyst for a given substrate, reaction or product simply by substrate name, Chemical Abstracts Service (CAS) number or molecule structure. In addition, a structure search tool offers the possibility to predict muteins which most likely accept a new substrate, if no enzyme/substrate properties were described so far.

MuteinDB overview

The MuteinDB is a platform to collect, catalog, and store experimentally derived data about muteins from publicly available sources as well as data directly submitted by the scientists. Additionally, it allows flexible searches by reaction type, molecular (sub) structures, substrate, product or mutein name. MuteinDB provides details on catalyzed reactions, kinetic data (activity, kinetic resolution) and experimental conditions used for data generation as well as for possible substrates or products consumed or produced by a reaction of choice including relevant scientific publication or patent information. Furthermore, it is possible to screen all enzyme variants for known interactions with specific inhibitors. Substrate, product and inhibitor data are linked to CAS and/or CID number (PubChem) as a unique identifier for unambiguous reference. The use of these distinct identification numbers allows even to extract information about (comparative) stereo- or enantioselectivity of individual muteins. Furthermore, once the user has identified a mutein of interest, information about its sequence, the employed expression hosts, cofactors, cosubstrates, and coproteins can be directly shown. In the sequence view, all mutations of a specific mutein are highlighted and liked to other muteins with known mutations of the same position. Additionally, the wild type sequence including all known amino acid exchanges which are again linked to muteins containing the modified position is illustrated for all muteins.

For first-time users, a comprehensive frequently asked question (FAQ) section and in-depth tutorial movies are provided. The key features are described in more detail below.

Search options

In contrast to other databases, MuteinDB allows users not only to query for muteins and mutations but also for substrates, products, or inhibitors, using the corresponding name or CAS number as well as catalyzed reaction types. The database also provides means to easily and efficiently search the data (e.g. by allowing to enter wildcards in the values) and display it in a clear, tabular form.

Structural search

Another important difference to other existing databases is the fully integrated (sub) structure search tool. It allows searching for substances with a similar structure by drawing an arbitrary chemical structure in the JME Molecule Editor (19). The database will provide all possible hits related to the drawn structure, and the user can navigate amongst them to refine the search. Based on knowledge about mutein/substrate combinations and their specific products, this for the first time also allows predictions of other possible substrates and products for known muteins which were not experimentally evaluated so far.

Individual features of MuteinDB

MuteinDB uses a mutein-based classification. A unique ID is assigned to each mutein and is linked to the reference source, the catalyzed reaction and the corresponding wild type protein. This mutein-centric approach allows more flexible and specific searches compared to the publication-based classification of the Protein Mutant Database (18) or the reaction based classification of BRENDA (17). Each reaction and publication reference can be independently surveyed, which is especially important when amino acid changes result in new functionalities. An example for such a case is the lipase CALB that was modified to a C–C bond forming enzyme for aldol additions (11).

Basic information about underlying wild type protein sequences, structures and source organism as well as compound structures, their respective references and reactions are retrieved from the public databases GenBank, PDB, UniProt, PubChem, PubMed, CrossRef, and KEGG (20–24). Wherever third party data is presented, it is linked to the corresponding database entry.

Standardized format for data collection and import

The recent introduction of experimental high-throughput techniques required the development of standardized formats for data from biological experiments. They facilitate exchange of data, their storage in publicly accessible repositories, increase experimental transparency and allow reproduction of bioinformatic analyses from publications. For example, such formats are available for DNA-microarray experiments (25), proteomics studies (26), and data deriving from qPCR experiments (27). Most of them are XML based, which can be difficult to create and manipulate. Therefore, simpler, spreadsheet-based formats have been introduced which are more accessible for the individual researcher. A prominent representative is the MAGE-TAB format for 45 DNA-microarray experiments (28).

Here, we propose a standardized spreadsheet-based data exchange format for muteins and related experimental kinetic data. The MuteinDB import spreadsheet comprises seven sections for each entry: (i) basic data; (ii) signal sequences; (iii) pH conditions; (iv) temperature conditions; (v) storage stability; (vi) reaction data and (vii) activity data. The basic data section includes the enzyme's name, the GenBank protein ID and the PDB ID (if available). Additionally, the corresponding wild-type name and the sequence mutations are illustrated for muteins. The reaction section contains the substrate and the product of the reaction (both with CAS number and name), the enzyme classification (EC) number of the reaction and the reaction type. The activity section can cover one of following types: conversion activity, enatiomeric excess or inhibition. All three types are followed by the corresponding kinetic values and the experimental conditions. The provided standards for kinetic data necessitate a minimum quality of biochemical protein data (e.g enzyme activity provided in μmol product made by μmol enzyme per minute).

A detailed description of the fields along with guidelines for data collection and a template spreadsheet are available on the MuteinDB homepage. Standardized entry of data into the spreadsheet is ensured by drop-down lists for fields with a defined value set. Drop-down lists can be extended if new values for a field are required.

For data import, the files are checked for data consistency according to the guidelines and compared with the already existing mutein data to prevent duplicate entries (Figure 1). A detailed report on the import is provided, allowing focused modification of the data to adjust it conforming to the guidelines. Upon successful import, the data is reviewed by an expert team at Graz University of Technology and feedback is provided to the submitter. After all inconsistencies are resolved, the new content is publicly released. New data can be submitted any time and is made available immediately after the review.

MuteinDB structure and implementation

The MuteinDB is implemented using Java, an object-oriented and platform-independent programming language. The application is based on a 3-tier architecture with an Oracle database as the persistence tier, an

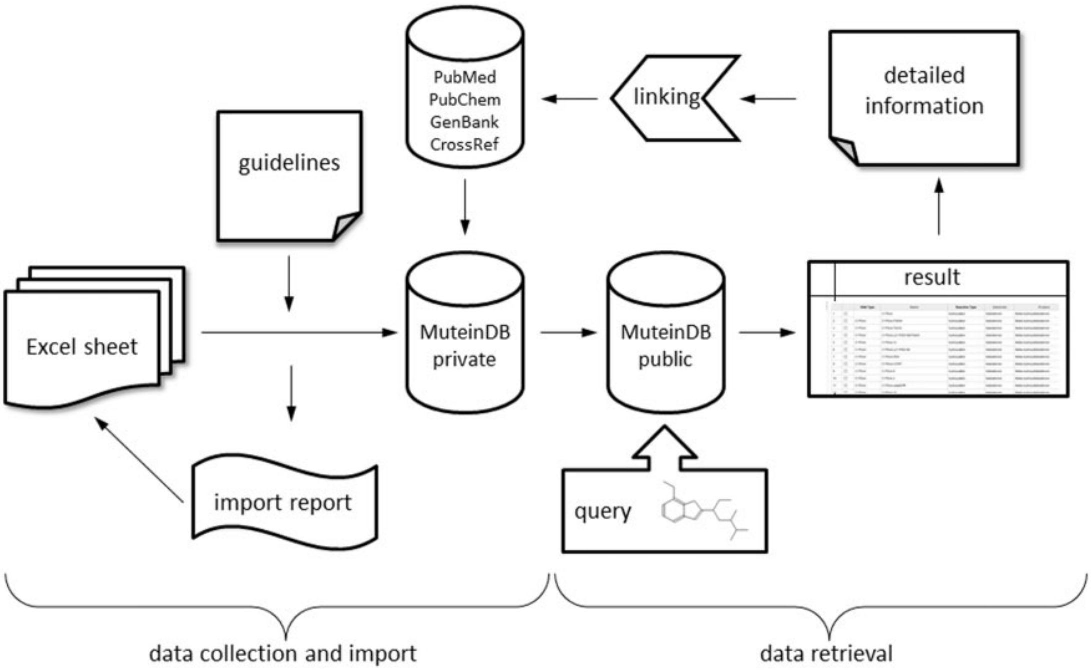


Figure 1. Schematic diagram of database structure. MuteinDB structure can be divided into two major parts. Firstly, the data collection and import structure within MuteinDB, illustrated on the left. Detailed guidelines structure and specify the correct and unified data collection as well as the data import. The standardized excel data import template guarantees data quality and consistency. During the automated data import from the data import excel sheet, metadata from third party databases such as PubMed, PubChem, GenBank and CrossRef are retrieved and added. The data import procedure ends either with a summary including imported muteins, molecules, reactions, activities or with a detailed error report. Secondly, stored public mutein data can be easily retrieved via various search mechanisms. For example, chemical structures can be used for identifying molecules of interest and their catalyzed reactions. Results are presented in tabular listings with links to third party databases or to detailed information contained in MuteinDB.

application server (JBoss) as the middle tier and a WEB interface as the client tier. Business logic is implemented using Enterprise JavaBeans 3. The web interface depends on JavaServer Faces 2, Asynchronous JavaScript and XML and JBoss Seam. The relational database schema has been designed to accommodate controlled vocabularies in form of a data dictionary. Attributes with a defined value set are linked to data dictionary entries to facilitate standardized content in the database.

For substructure search (5), the JME Molecule Editor (19) and the Chemistry Development Kid (CDK)—an open-source Java library—are used.

Use of MuteinDB

The MuteinDB was developed as a user-friendly and intuitive resource of mutein-related properties for scientists in the fields of biology, biotechnology, organic chemistry and pharmaceutical sciences. The top information bar offers 'FAQs' where users will find helpful information. Furthermore, first-time users will find tutorial movies explaining the database usage and the different MuteinDB sections.

The simplest search option 'Search by Substrate' is directly accessible via the home screen. The left side navigation bar gives access to further querying options.

Search options

(i) Substrate: enables the user to search for muteins that convert a certain substrate of interest.

(ii) Reaction: enables the user to search for specific reactions by entering a molecule name or a CAS number for the substrate and/or the product (including single enantiomers)

(iii) Structure: enables the user to draw chemical structures to search for similar or exact (sub) structure matches in either one or all of the molecule categories (substrate, product and/or inhibitor).

(iv) Inhibitor: enables the user to search for inhibitors of muteins and wild-type enzymes by entering a molecule name or a CAS number.

(v) Mutation: enables the user to search for muteins containing mutations at a certain position.
(vi) Wild type: enables the user to browse all muteins and their reactions for a defined wild-type enzyme.
(vii) Mutein: enables the user to search for all relevant reactions for a defined mutein name.

To keep the additional querying options simple and flexible, further refinements of the query can, but do not have to, be specified or selected. For example, the search can be restricted amongst others to the reaction type, the underlying wild type protein, or to a specific organism.

All text fields are equipped with 'suggest input'. While typing a box will appear and provide suggestions one can choose from. Furthermore, selected fields allow 'wildcard search' with '*' as a placeholder.

Example workflow

The ability to search for exact or similar structures is one of the unique main features of the MuteinDB. Therefore, we will describe this search type in more detail and use it as example to demonstrate the ability of MuteinDB for valuable data retrieval (Figure 2).

Selecting 'Search by Structure' will open the JME Molecule Editor Applet (19) and allows the user to draw an arbitrary chemical structure (Figure 2A). After submitting the search, results will be presented as a table listing all molecules containing the drawn structure (Figure 2B). The 'structure result' page proposes several related substrates, products and inhibitors with similar structures to the drawn molecule structure. In all result views, moving the cursor over a molecule name will show its chemical structure. Additionally, each molecule name is linked to PubChem (22). This also facilitates the search if the CAS number or the exact molecule name is unknown or if different trivial names of the molecule are commonly used.

One or several molecules can be selected via checkboxes and can be used for a subsequent search by substrate/product or inhibitor. The results are shown again in tabular form listing all muteins that convert the selected substrates or produce the selected products or are inhibited by the chosen inhibitors.

Selecting 'testosterone' from the list for a subsequent search reveals several muteins that are able to convert this steroid (Figure 2C). This supports predictions about possible transformations of testosterone derivatives where no experimental data is available so far. Hits from such searches are preferred muteins for experimental evaluation.

Information about the catalyzed reactions such as substrate, product and reaction type are presented in the 'substrate view' (Figure 2D). A link to KEGG reaction (29) is provided when a corresponding entry exists. As the kinetic data are one of the most important pieces of information stored in the database, kinetic parameters such as K_m and k_{cat} are given. Furthermore, enantiomeric access and E-values are provided if available. The view can be customized using 'edit display settings'.

As multiple publications may have reported the same reaction for a given mutein, the one stating the highest activity is shown in the main result screen. By using the expand button the data from the other reports are also shown. Clicking on the mutein name will bring up the 'mutein view' where detailed information about the mutein and the reaction are provided.

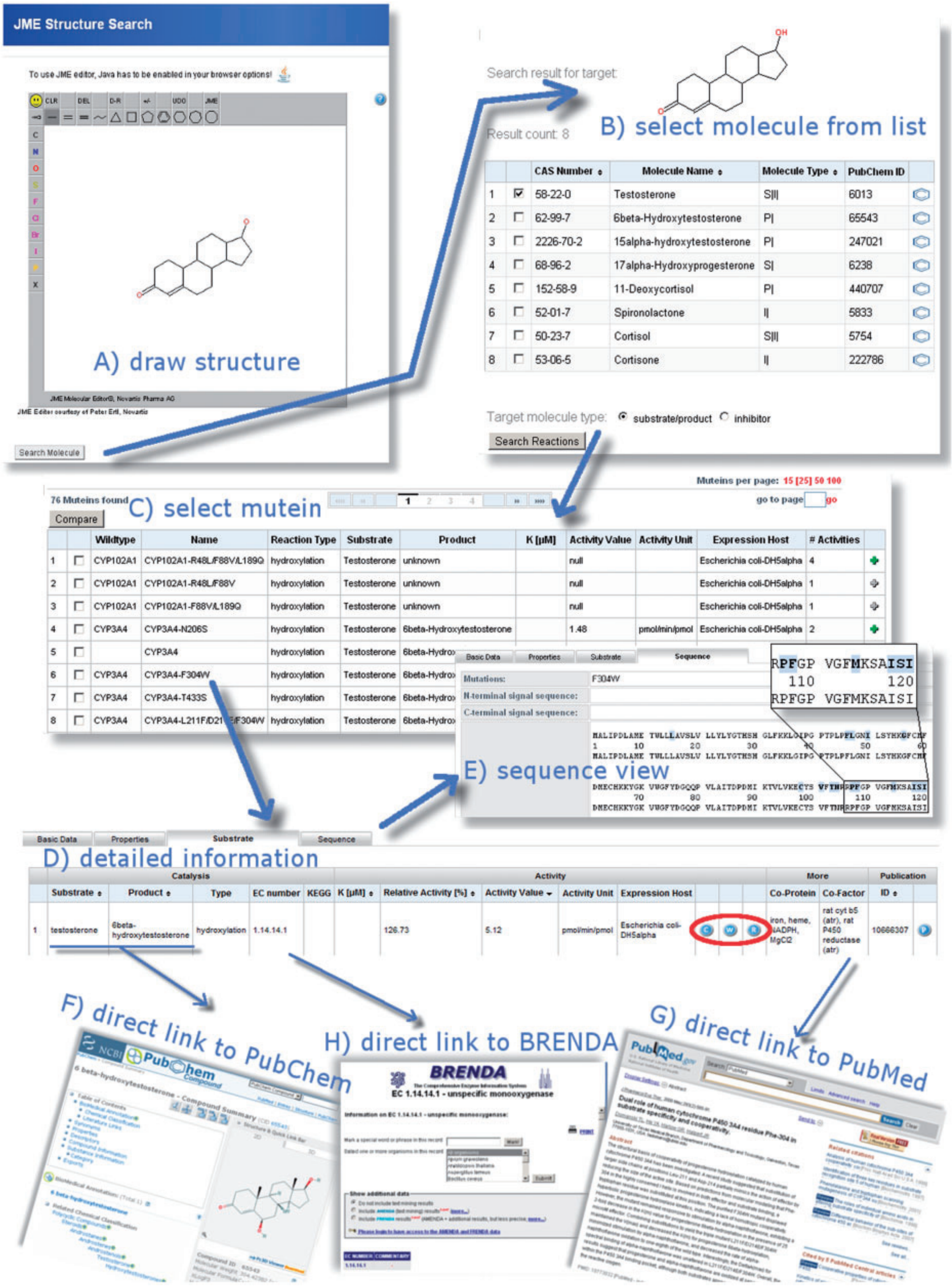
In the 'substrate section', several mouse-over buttons (Figure 2D) give further information about the catalyzed reaction. 'C' shows comments on the reaction, 'W' gives activity data of the underlying wild-type reference, 'R' shows information about reaction conditions and analysis and 'L' provides detailed information about the corresponding literature. The PubMed ID or the digital object identifier (DOI, http://crossref.org) of the publications are given and directly linked to PubMed or to the webpage associated with the digital object identifier, respectively. Additionally, the EC number is provided and linked to the comprehensive enzyme database BRENDA (Figure 2F–H).

The 'sequence section' shows the mutein sequence aligned with its corresponding wild-type sequence (Figure 2E). In the mutein sequence, the mutations are highlighted in violet. The sequence can be downloaded as FASTA format. The amino acids of the wild-type sequence highlighted in blue mark the positions of known mutations. These positions are linked to the 'enzyme mutation view'. In this view, all muteins that contain a mutation at this position are listed. Via the mutein name it is possible to navigate to the mutation view of the corresponding mutein.

Another highlight of the MuteinDB is the ability to select two or more muteins, which convert the substrate of interest or form the product of interest, for comparison in side by side view. In the 'compare view' the kinetic data of the catalyzed reaction as well as information about the mutations, expression system and involved cofactors and coproteins are displayed.

Inhibitors have a special status and may have been reported in the 'structure result' page for the inhibitor search. The results are shown in tabular form (Figure 3) listing muteins that are inhibited by the chemical compound. Instead of kinetic data, the inhibitor constant K_i or the IC_{50} value are provided. Additionally, the underlying reaction used to determine the inhibitor constant is shown.

As the same inhibitor measurements can be found in different publications, only the one with the highest inhibition constant is shown as the main result. Via the expand button, the data of the other literature sources is shown. The mutein name is again linked to the 'mutein view', where detailed information on the mutein and the inhibition reaction are provided.


JME Structure Search
To use JME editor, Java has to be enabled in your browser options!
CLR
DEL
D-R
UDO
JME
C
N
O
S
F
Cl
Br
I
P
X
O
O
A) draw structure
JME Molecular Editor©, Novartis Pharma AG
JME Editor courtesy of Peter Ertl, Novartis
Search Molecule
Search result for target:
OH
O
Result count: 8
B) select molecule from list

		CAS Number	Molecule Name	Molecule Type	PubChem ID	
1	☑	58-22-0	Testosterone	S|I	6013	
2	□	62-99-7	6beta-Hydroxytestosterone	P|	65543	
3	□	2226-70-2	15alpha-hydroxytestosterone	P|	247021	
4	□	68-96-2	17alpha-Hydroxyprogesterone	S|	6238	
5	□	152-58-9	11-Deoxycortisol	P|	440707	
6	□	52-01-7	Spironolactone	|I	5833	
7	□	50-23-7	Cortisol	S|I	5754	
8	□	53-06-5	Cortisone	|I	222786	

Target molecule type: ● substrate/product ○ inhibitor

		Wildtype	Name	Reaction Type	Substrate	Product	K [µM]	Activity Value	Activity Unit	Expression Host	# Activities	
1	□	CYP102A1	CYP102A1-R48L/F88V/L189Q	hydroxylation	Testosterone	unknown		null		Escherichia coli-DH5alpha	4	
2	□	CYP102A1	CYP102A1-R48L/F88V	hydroxylation	Testosterone	unknown		null		Escherichia coli-DH5alpha	1	
3	□	CYP102A1	CYP102A1-F88V/L189Q	hydroxylation	Testosterone	unknown		null		Escherichia coli-DH5alpha	1	
4	□	CYP3A4	CYP3A4-N206S	hydroxylation	Testosterone	6beta-Hydroxytestosterone		1.48	pmol/min/pmol	Escherichia coli-DH5alpha	2	
5	□		CYP3A4	hydroxylation	Testosterone	6beta-Hydrox						
6	□	CYP3A4	CYP3A4-F304W	hydroxylation	Testosterone	6beta-Hydrox						
7	□	CYP3A4	CYP3A4-T433S	hydroxylation	Testosterone	6beta-Hydrox						
8	□	CYP3A4	CYP3A4-L211F/D21 E/F304W	hydroxylation	Testosterone	6beta-Hydrox						

Mutations: F304W
N-terminal signal sequence:
C-terminal signal sequence:

```
MALIPDLAME TVLLLAVSLV LLYLYGTHSH GLFKKLGIPG PTPLPFLGNI LSYHKGFCMF
1        10         20         30         40         50         60
MALIPDLAME TVLLLAVSLV LLYLYGTHSH GLFKKLGIPG PTPLPFLGNI LSYHKGFCMF

DMECHKKYGK VWGFYDGQQP VLAITDPDMI KTVLVKECYS VFTNRRPFGP VGFMKSAISI
        70         80         90        100        110        120
DMECHKKYGK VWGFYDGQQP VLAITDPDMI KTVLVKECYS VFTNRRPFGP VGFMKSAISI
```

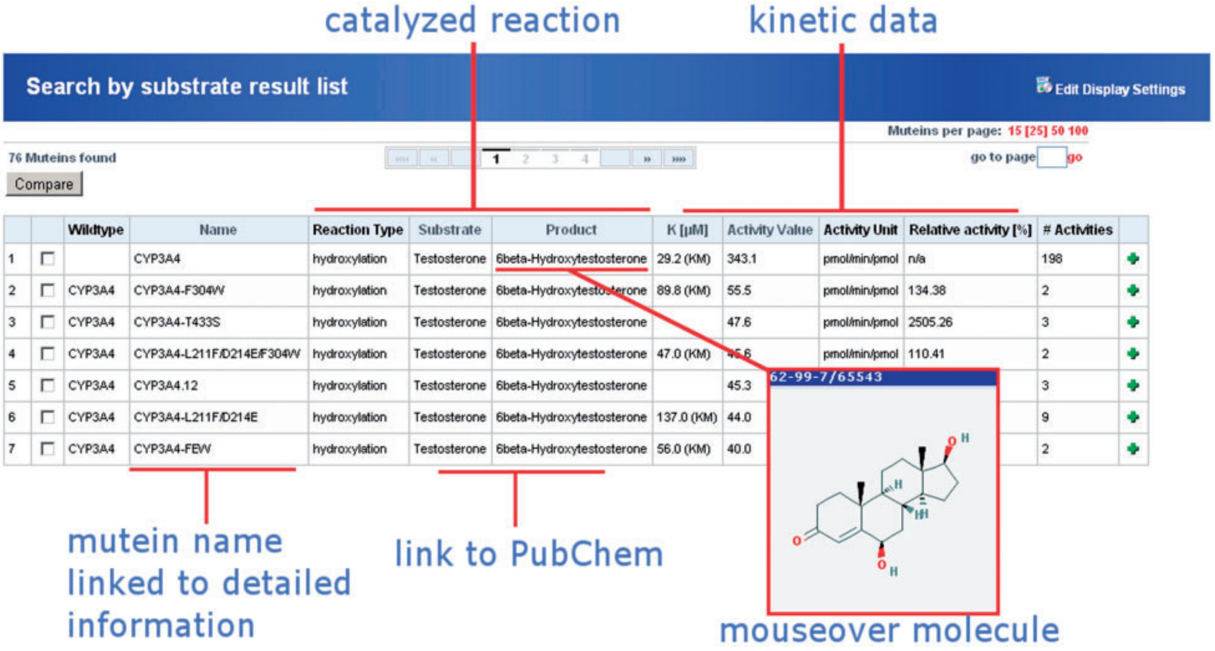
```
RPFGP VGFMKSAISI
 110          120
RPFGP VGFMKSAISI
```

	Catalysis					Activity								More		Publication	
	Substrate	Product	Type	EC number	KEGG	K [µM]	Relative Activity [%]	Activity Value	Activity Unit	Expression Host				Co-Protein	Co-Factor	ID	
1	testosterone	6beta-hydroxytestosterone	hydroxylation	1.14.14.1			126.73	5.12	pmol/min/pmol	Escherichia coli-DH5alpha	C	W	R	iron, heme, NADPH, MgCl2	rat cyt b5 (atr), rat P450 reductase (atr)	10666307	P

F) direct link to PubChem
NCBI
PubChem Compound
6 beta-hydroxytestosterone - Compound Summary
H) direct link to BRENDA
BRENDA
EC 1.14.14.1 - unspecific monooxygenase
G) direct link to PubMed
PubMed.gov
Dual role of human cytochrome P450 3A4 residue Phe-304 in substrate specificity and cooperativity.

Figure 2. MuteinDB structure search, its results and the capabilities of the MuteinDB webinterface. (**A**) The MuteinDB (sub) structure search uses the JME editor, which allows users to draw arbitrary molecular structures. (**B**) The user-drawn structure is used as seed for the following database search and shown on top of the structure search result table. In this table all molecules,

(continued)



catalyzed reaction

kinetic data

Search by substrate result list

		Wildtype	Name	Reaction Type	Substrate	Product	K [µM]	Activity Value	Activity Unit	Relative activity [%]	# Activities
1	□		CYP3A4	hydroxylation	Testosterone	6beta-Hydroxytestosterone	29.2 (KM)	343.1	pmol/min/pmol	n/a	198
2	□	CYP3A4	CYP3A4-F304W	hydroxylation	Testosterone	6beta-Hydroxytestosterone	89.8 (KM)	55.5	pmol/min/pmol	134.38	2
3	□	CYP3A4	CYP3A4-T433S	hydroxylation	Testosterone	6beta-Hydroxytestosterone		47.6	pmol/min/pmol	2505.26	3
4	□	CYP3A4	CYP3A4-L211F/D214E/F304W	hydroxylation	Testosterone	6beta-Hydroxytestosterone	47.0 (KM)	45.6	pmol/min/pmol	110.41	2
5	□	CYP3A4	CYP3A4.12	hydroxylation	Testosterone	6beta-Hydroxytestosterone		45.3			3
6	□	CYP3A4	CYP3A4-L211F/D214E	hydroxylation	Testosterone	6beta-Hydroxytestosterone	137.0 (KM)	44.0			9
7	□	CYP3A4	CYP3A4-FEW	hydroxylation	Testosterone	6beta-Hydroxytestosterone	56.0 (KM)	40.0			2

mutein name linked to detailed information

link to PubChem

mouseover molecule

Figure 3. Result display of the MuteinDB web-interface for testosterone as a substrate. Information within the result listing for each mutein is by default grouped into catalyzed reaction and kinetic data. Reaction information comprises the reaction type as well as the catalyzed substrate and product. Molecules are directly linked to their corresponding PubChem entry. Additionally, the molecule structure can be displayed by moving over the compound's name. Important kinetic parameters such as K value, activity value including its unit as well as the relative activity in (%) are directly available in the result view. All presented information and further links for each mutein or wild type is directly linked by its name.

Results and conclusions

MuteinDB is a comprehensive and carefully curated database for specific muteins and their kinetic data of catalyzed reactions including inhibition. It provides in-depth information on mutein properties combined with flexible search capabilities. The MuteinDB has been designed to be broadly applicable to proteins and their muteins from any enzyme class including those with no known catalytic function. We demonstrated this by entering data sets of several enzymes and their variants of different enzyme classes.

Presently, the understanding of the structure–function relationship of proteins is still limited. Scientists are trying to tackle the problem from different perspectives (from medicine and pharmacokinetics, to structural biology or applied biocatalysis) and are, therefore, interested in how mutations can influence catalytic properties.

By means of MuteinDB a user can find enzymes that catalyze a particular reaction not only in expected enzyme classes but also in others [e.g. a C–C bond forming mutein derived from a hydrolase (30)]. This feature helps to identify potential starting points for further enzyme engineering. Moreover, medical scientists can get information about the influence of mutations on the drug metabolism and the *in vivo* activation. This helps to predict a patient's personal response to certain administered drugs. In addition, the implemented structure search for substrates, products and inhibitors allows the prediction of structure scaffolds that could be accepted by muteins. This might provide helpful information for the development of new biocatalysts and, most probably, will facilitate drug metabolite prediction in pharmaceutical research and development.

At present MuteinDB contains several thousand reactions (Table 1) for muteins of different enzyme

Figure 2. Continued
substrates, products or inhibitors which contain the query structure are presented. A selection of these molecules can be used for a subsequent 'Search by Reaction'. (**C**) All wild type enzymes and muteins which catalyze the selected molecules are shown. (**D**) For each row of the tabular result, further information can be obtained via the mutein or wild type name. The detailed information is organized in four main categories: (i) basic data; (ii) properties; (iii) substrate and (iv) sequence. (**E**) The 'Sequence' tab of the selected mutein allows to explore the sequence of the mutein as well as the wild type sequence. Known mutations are highlighted and linked to the corresponding entries of MuteinDB. (**F**) Information in the 'Substrate' tab is linked to third party databases. For example, (**F**) molecules are linked to PubChem, (**H**) EC-Numbers to Brenda and (**G**) literature to PubMed or to its DOI location. For muteins, experimental settings and wild type activity values are available from the 'Substrate' tab.

Table 1. MuteinDB data overview

Wild-type Name	Muteins	Reactions	Activities	Publications
CYP102A1	168	909	995	42
CYP102A2	0	4	4	1
CYP2D6	98	648	1259	213
CYP3A4	124	825	1908	220
HAPMO	6	106	114	5
HRP C1	17	32	45	8
Nitrilases	8	26	26	3
NITAf	11	42	42	2
P3H	0	1	12	1
P3H type1	0	21	31	3
P3H type2	0	16	23	2
P4H	0	21	53	4
PAMO	31	309	385	10
Total: 11	444	2892	4829	422

classes. It is the largest collection of kinetic data of muteins compiled in a single database. To demonstrate the general applicability of the database, different types of enzymes from different origins have been searched in literature and imported into MuteinDB. Data were collected by searching SciFinder (www.cas.org) and PubMed (21) abstracts for specific keywords. Detailed data from texts, tables and figures were manually extracted from the matching full-text publications and were curated by a team of scientists, who enriched the published information with first-hand kinetic data wherever possible.

CYP2D6 and CYP3A4 are human liver enzymes and known to be involved in drug metabolism. Both enzymes have been chosen as primary data sets due to their pronounced polymorphism and high importance for human drug and xenobiotic metabolism. We selected CYP102A1 (BM-3) from *Bacillus megaterium* as a prokaryotic representative. This protein is one of the most mutated and investigated proteins known.

To import the data, we used the standardized spreadsheet-based import file format described previously. It contains all attributes necessary to describe a mutein and its properties.

In order to augment the database content, data collection is on-going. To make the database as comprehensive and up-to-date as possible, we are addressing the research community with a request to aid us in the collection of kinetic data sets for enzymes of different type and origin. We appreciate any contribution to the database both updates to existing data and new kinetic data sets.

Future directions

In the course of integrating new data sets, the MuteinDB will be adapted, and the guidelines for data collection will be adjusted. Feedback from end users and data collectors will ensure a continued focus on a user-friendly development.

The collection of data sets was carried out as part of the OXYGREEN (www.oxygreen.org) project, a research collaboration funded by the European Commission Seventh Framework Programme (EU FP7), and will be continued to do so. MuteinDB will be used and extended in the context of BIONEXGEN, a recently funded EU project. To ensure continuation of data collection and curation of the database, MuteinDB will be integrated into future projects.

A downloadable version of the MuteinDB is in preparation. It will be provided for companies or universities that would like to store their own data in-house. The data can be integrated into the public online database on request. The download will be available in exchange for new mutein data sets or for a fee for database curation and data collection.

Acknowledgements

We would like to thank Andrea Camattari, Peter Remler, Norbert Klempier, Kurt Faber, Peter Macheroux, Helmut Schwab and the whole OXYGREEN team for assistance and fruitful discussions. Furthermore, we want to thank Peter Ertl for providing the JME molecular editor applet.

Funding

EU-FP7 project OXYGREEN (EC grant 212281); the Austrian Ministry of Science and Research GEN-AU project BIN (FFG grant 820962). The Austrian Centre of Industrial Biotechnology (ACIB) contribution was supported by FFG, bmvit, mvwfi, ZIT, Zukunftsstiftung Tirol and Land Steiermark within the Austrian COMET programme (FFG grant 824186). Funding for open access charge: EU-FP7 project OXYGREEN (EC grant 212281).

Conflict of interest. None declared.

References

1. Ford,E.B. (1966) Genetic polymorphism. *Proc. R. Soc. Lond. B Biol. Sci.*, **164**, 350–361.
2. Shastry,B.S. (2006) Pharmacogenetics and the concept of individualized medicine. *Pharmacogenomics J.*, **6**, 16–21.
3. Lutz,S. and Bornscheuer,T.U. (2008) Protein Engineering Handbook. Wiley-VCH GmbH & Co.KGaA, Weinheim, Germany.
4. Brannigan,J.A. and Wilkinson,A.J. (2002) Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.*, **3**, 964–970.
5. Pan,K., Zhang,R., Sun,H. *et al.* (2008) An implementation of substructure search in chemical database management system. In:

Proceedings of the Third International Multi-symposiums on Computer and Computational Sciences (IMSCCS'08). IEEE Computer Society Press, pp. 203–206.

6. Padhi,S.K., Fujii,R., Legatt,G.A. *et al*. (2010) Switching from an esterase to a hydroxynitrile lyase mechanism requires only two amino acid substitutions. *Chem. Biol.*, **17**, 863–871.
7. Reddy,S.Y., Kahn,K., Zheng,Y.J. *et al*. (2002) Protein engineering of nitrile hydratase activity of papain: molecular dynamics study of a mutant and wild-type enzyme. *J. Am. Chem. Soc.*, **124**, 12979–12990.
8. Urlacher,V. and Schmid,R.D. (2002) Biotransformations using prokaryotic P450 monooxygenases. *Curr. Opin. Biotechnol.*, **13**, 557–564.
9. Peters,M.W., Meinhold,P., Glieder,A. *et al*. (2003) Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *J. Am. Chem. Soc.*, **125**, 13442–13450.
10. Glieder,A., Farinas,E.T. and Arnold,F.H. (2002) Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase. *Nat. Biotechnol.*, **20**, 1135–1139.
11. Branneby,C., Carlqvist,P., Magnusson,A. *et al*. (2003) Carbon-carbon bonds by hydrolytic enzymes. *J. Am. Chem. Soc.*, **125**, 874–875.
12. Savile,C.K., Janey,J.M., Mundorff,E.C. *et al*. (2010) Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, **329**, 305–309.
13. Fischer,M., Knoll,M., Sirim,D. *et al*. (2007) The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics*, **23**, 2015–2017.
14. Preissner,S., Kroll,K., Dunkel,M. *et al*. (2010) SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. *Nucleic Acids Res.*, **38**, D237–D243.
15. Lonquety,M., Lacroix,Z., Papandreou,N. *et al*. (2009) SPROUTS: a database for the evaluation of protein stability upon point mutation. *Nucleic Acids Res.*, **37**, D374–D379.
16. Kumar,M.D., Bava,K.A., Gromiha,M.M. *et al*. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.
17. Schomburg,I., Chang,A., Ebeling,C. *et al*. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.
18. Kawabata,T., Ota,M. and Nishikawa,K. (1999) The Protein Mutant Database. *Nucleic Acids Res.*, **27**, 355–357.
19. Ertl,P. (2010) Molecular structure input on the web. *J. Cheminform.*, **2**, 1.
20. Benson,D.A., Karsch-Mizrachi,I., Clark,K. *et al*. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
21. Wheeler,D.L., Church,D.M., Edgar,R. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35–D40.
22. Wang,Y., Xiao,J., Suzek,T.O. *et al*. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
23. The Uniprot-Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
24. Berman,H., Henrick,K., Nakamura,H. *et al*. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
25. Brazma,A., Hingamp,P., Quackenbush,J. *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
26. Taylor,C.F., Paton,N.W., Lilley,K.S. *et al*. (2007) The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.*, **25**, 887–893.
27. Bustin,S.A., Benes,V., Garson,J.A. *et al*. (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.*, **55**, 611–622.
28. Rayner,T.F., Rocca-Serra,P., Spellman,P.T. *et al*. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.
29. Kotera,M., Hirakawa,M., Tokimatsu,T. *et al*. (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.*, **802**, 19–39.
30. Li,C., Hassler,M. and Bugg,T.D. (2008) Catalytic promiscuity in the alpha/beta-hydrolase superfamily: hydroxamic acid formation, C–C bond formation, ester and thioester hydrolysis in the C–C hydrolase family. *Chembiochem.*, **9**, 71–76.

HIGH-THROUGHPUT CHARACTERIZATION AND COMPARISON OF MICROBIAL COMMUNITIES

Bettina Halwachs[1,2], Johann Höftberger[1], Gernot Stocker[1], Rene Snajder[1]
Gregor Gorkiewicz[3] and Gerhard G. Thallinger[1,2]

[1]Institute for Genomics and Bioinformatics, Graz University of Technology, Austria
[2]Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, Austria
[3]Institute of Pathology, Medical University of Graz, Austria

bettina.halwachs@acib.at

Abstract: *The analysis of the huge amount of generated sequence data as well as pyrosequencing noise and chimeric sequences originating from PCR amplification pose a considerable challenge to the individual researcher in doing microbiome studies. The unbiased knowledge about microbial community composition and -structure as well as the interactions with the human host microbiome can give important insights into its role in human health and disease. Here we introduce SnoWMAn, the high-throughput microbiome analysis pipeline and additionally investigate the effects of sequencing noise on non denoised and data denoised using two different approaches.*

Keywords: *Next Generation Sequencing, Community Composition Analysis, Denoising, OTU inflation*

Introduction

The overall goal of human microbiome studies is to represent complex community composition within a certain habitat of interest and compare it under different conditions, between time points or patients. To characterize and classify complex microbial communities gained directly from environmental samples, a certain variable region of the commonly shared 16S rRNA marker gene is directly amplified and sequenced. Before generated sequences can be classified into operational taxonomic units (OTUs) some preprocessing and filtering steps should be applied to guarantee unbiased community composition representation. Especially when working with pyrosequencing data, noise originating from longer homopolymer stretches (> 4 bps) can lead to an increase in OTUs called *OTU inflation*. Besides sequencing noise perceived diversity can be increased by chimeric 16S amplification products which were formed out of two or more sequence templates during polymerase chain reaction (PCR). During further analysis these hybrid products can be falsely interpreted as novel organisms, thus inflating apparent diversity and finally lead to false conclusions. The general microbiome data analysis workflow is illustrated in Fig. 1, where for each step a variety of tools and approaches are available. To simplify microbiome analysis from preprocessing over OTU picking to the final statistical analysis and visualization of the result, we developed the web-based analysis pipeline *SnoWMAn*. It addresses shortcomings of existing tools, such as number of sequences which can be analysed, reproducibility and usability. Additionally, SnoWMAn is unique in covering the complete analysis workflow, offering different analysis pipelines and reference databases as well as capabilities for statistical analysis and visualization.

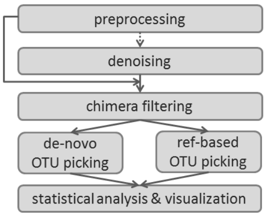


Figure 1: General microbiome analysis workflow (steps donated with a dashed arrow can be omitted).

Methods

To demonstrate the capabilities of SnoWMAn and to show the effects of pyrosequencing noise we reviewed a previous study on changes in the gut microbiome during diarrhea [1] by denoising the data with *Acacia* [2] and the *mothur* [3] implementation of *AmpliconNoise* respectively. Furthermore, contaminating sequences originating from the host genome as well as potential chimeric sequences had been removed from the amplified sequences by a BLAST approach and uchime respectively. OTUs were built using the Ribosomoal Database Project (RDP)-Pyrosequencing approach using the *Infernal alignment* v1.1 [4] and a maximal cluster similarity of 6 and similarity steps of 1 %. Additionally, quality filtering based on given quality values per base, number of Ns (discard sequences containing Ns) and length (discard sequences < 150 bp) was applied. Final taxonomic classification was done by the *RDP classifier 2.4* [5].

Results

In addition to the RDP pipeline used here, SnoWMAn users can chose according to the field of application and their study design between two *reference based OTU picking* pipelines (*BLAT, JGast*) and three *de novo OTU picking* pipelines (*mothur, RDP, UCLUST*). Depending on the selected pipeline various preprocessing- and pipeline parameters such as the applied reference database, the classification model and the clustering settings can be specified. To minimize analysis time as well as to improve result qual-

Table 1: Number of sequences and OTUs at a cluster similarity of 0.03, without denoising as well as denoised by Acacia and the mothur re-implementation of AmpliconNoise. Contaminating sequences as well as potential chimeric sequences have been also removed prior to analysis.

sample	**contam.**	**no denoising**		**AmpliconNoise**				**Acacia**			
		raw	*OTUs*	*noise*	*chimera*	*ok*	*OTUs*	*noise*	*chimera*	*ok*	*OTUs*
Feces	4	327439	3755	23810	443	301099	1939	23282	7469	296684	2295
Mucosa	3558	187773	3077	25570	2526	158291	1628	25712	1609	156913	1980

ity it is possible to filter sequences according their length, maximal mount of unidentified bases, or mean sequence quality thresholds. The newly introduced Acacia denoising tool was integrated into SnoWMAn and can be used on demand for identification and removal of noisy sequences. Not only that it was shown to be about 2000x faster than existing tools, additionally, our comparison results in modest difference between the mothur re-implementation of AmpliconNoise and Acacia (see Tab. 1). In respect to chimeric sequences SnoWMAn integrates mothur's *uchime* for optional chimera detection and removal. After all necessary and optional parameters are specified the numerical intensive analysis task is automatically started. Once the calculation is finished SnoWMAn offers various capabilities for statistical analysis and result visualization such as rarefaction curves for microbial diversity estimation and illustration of species richness (alpha diversity). Species turnover or beta diversity can be calculated and visualized using heatmaps. Comparison of individual microbiomes can be done by the integrated principal component analysis (PCA). Barcharts or piecharts can be used to represent the number of sequences for each sample and give an overview of sequence yields. Additionally, cumulative and endpoint depth of the taxonomic classification can be graphically illustrated. Line plots can be used to reveal sample composition at a specific taxonomic rank to point out compositional mircobiome changes over time. Data can be presented in relative or absolute scale for all chart types. All the generated data can be easily exported either as Excel file or as figures in PNG or SVG format. The comparison of the effects of sequencing noise on community diversity results in enormous OTU-inflation when comparing the number of OTUs resulting from denoised vs. non denoised pyrosequencing data, see Tab. 1. Surprisingly, the number of potential chimerias varies depending on the denoising approach, especially for fecal samples. Moreover, the number of OTUs varies more than expected between AmpliconNoise and Acacia.

Discussion

Here we introduced SnoWMAn as a comprehensive system for high-throughput analysis of microbial community sequencing data as well as the effects of two different denoising approaches. SnoWMAn covers the whole microbiome analysis workflow and offers the two most common analysis approaches in one single pipeline. The user-friendly and intuitive web-interface makes it a convenient resource not only for classification and characterization but also for statistical analysis, visualization and reusing or sharing of the analysis result. Furthermore, the newly integrated denoising and chimara filtering tools satisfy latest findings towards sequencing noise. Although different denoising approaches showed modest variation of noisy sequences the effect on the number of chimeric sequences needs further investigations. The modular design of SnoWMAn allows simplified extension of the classification tools to other genes than the 16S rRNA by providing appropriate reference databases and alignment models.

Acknowledgement

This work was funded in part by the Austrian Centre of Industrial Biotechnology (ACIB) supported by the Austrian BMWFJ, BMVIT, SFG, Standortagentur Tirol and ZIT through the Austrian FFG-COMET- Funding Program [FFG Grant 824186].

Bibliography

[1] G. Gorkiewicz, G. G. Thallinger, S. Trajanoski, S. Lackner, G. Stocker, T. Hinterleitner, C. Gülly, and C. Högenauer, "Alterations in the colonic microbiota in response to osmotic diarrhea," *PLoS ONE*, vol. 8, p. e55817, 02 2013.

[2] B. Lauren, G. Stone, M. Imelfort, P. Hugenholtz, and G. W. Tyson, "Fast, accurate error-correction of amplicon pyrosequences using acacia," *Nat Meth*, vol. 9, no. 5, pp. 425–426, 2012.

[3] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber, "Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Appl Environ Microbiol*, vol. 75, no. 23, pp. 7537–7541, 2009.

[4] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy, "Infernal 1.0: inference of rna alignments," *Bioinformatics*, vol. 25, no. 10, pp. 1335–1337, 2009.

[5] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Appl Environ Microbiol*, vol. 73, no. 17, pp. 5261–7, 2007.

Comparative Genome Analysis of *Campylobacter fetus* Subspecies Revealed Horizontally Acquired Genetic Elements Important for Virulence and Niche Specificity

Sabine Kienesberger[1,2,7]*, Hanna Sprenger[1,7], Stella Wolfgruber[1], Bettina Halwachs[5,6], Gerhard G. Thallinger[5,6], Guillermo I. Perez-Perez[2,3], Martin J. Blaser[2,3,4], Ellen L. Zechner[1], Gregor Gorkiewicz[7]*

1 Institute of Molecular Biosciences, University of Graz, Graz, Austria, **2** Department of Medicine, NYU Langone Medical Center, New York, New York, United States of America, **3** Department of Microbiology, NYU Langone Medical Center, New York, New York, United States of America, **4** Medical Service, VA New York Harbor Healthcare System, New York, New York, United States of America, **5** Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria, **6** Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, Graz, Austria, **7** Institute of Pathology, Medical University of Graz, Graz, Austria

Abstract

Campylobacter fetus are important animal and human pathogens and the two major subspecies differ strikingly in pathogenicity. *C. fetus* subsp. *venerealis* is highly niche-adapted, mainly infecting the genital tract of cattle. *C. fetus* subsp. *fetus* has a wider host-range, colonizing the genital- and intestinal-tract of animals and humans. We report the complete genomic sequence of *C. fetus* subsp. *venerealis* 84-112 and comparisons to the genome of *C. fetus* subsp. *fetus* 82-40. Functional analysis of genes predicted to be involved in *C. fetus* virulence was performed. The two subspecies are highly syntenic with 92% sequence identity but *C. fetus* subsp. *venerealis* has a larger genome and an extra-chromosomal element. Aside from apparent gene transfer agents and hypothetical proteins, the unique genes in both subspecies comprise two known functional groups: lipopolysaccharide production, and type IV secretion machineries. Analyses of lipopolysaccharide-biosynthesis genes in *C. fetus* isolates showed linkage to particular pathotypes, and mutational inactivation demonstrated their roles in regulating virulence and host range. The comparative analysis presented here broadens knowledge of the genomic basis of *C. fetus* pathogenesis and host specificity. It further highlights the importance of surface-exposed structures to *C. fetus* pathogenicity and demonstrates how evolutionary forces optimize the fitness and host-adaptation of these pathogens.

Citation: Kienesberger S, Sprenger H, Wolfgruber S, Halwachs B, Thallinger GG, et al. (2014) Comparative Genome Analysis of *Campylobacter fetus* Subspecies Revealed Horizontally Acquired Genetic Elements Important for Virulence and Niche Specificity. PLoS ONE 9(1): e85491. doi:10.1371/journal.pone.0085491

Editor: Thomas Alter, Free University of Berlin, Germany

Received July 9, 2013; **Accepted** November 28, 2013; **Published** January 9, 2014

Funding: This study was supported by the Austrian Science Fund FWF grants P20479 (GG & ELZ) and P24016 (ELZ), the Hygiene Fund Young Scientist grant from the Medical University of Graz (SK), the Michael Saperstein Medical Scholars Program, New York School of Medicine (SK) and GM63270 from the National Institutes of Health (MJB), and the EU-FP7 COST Action SeqAhead, EC Grant BM1006 (GGT). The Austrian Centre of Industrial Biotechnology (ACIB) contribution was supported by FFG, The Federal Ministry of Economy, Family and Youth (BMWFJ), Federal Ministry for Transport, Innovation and TechnologyBMVIT, ZIT GmbH, Zukunftsstiftung Tirol and Land Steiermark within the Austrian COMET program FFG Grant 824186 (BH, GGT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sabine.kienesberger@uni-graz.at (SK); gregor.gorkiewicz@uni-graz.at (GG)

Introduction

The ε-proteobacterial genus *Campylobacter* comprises bacteria with a high degree of niche adaptation and host tropism [1]. The species colonize mucosal surfaces and are animal and human pathogens [2]. The genomes of *Campylobacter* spp. are not large (≈1.5 Mbp) and show characteristics of genome decay typical for niche-adapted bacteria [3]. These features make *Campylobacter* species ideal model systems to study genetic contributions to niche specificity and virulence by comparative genome analysis [3]. Multi locus sequence typing (MLST) has shown that the two *C. fetus* subspecies, *C. fetus* subsp. *fetus* and *C. fetus* subsp. *venerealis*, have a clonal population structure [4] and differentiation of the taxa is only partially successful [5]. Both subspecies are important veterinary pathogens causing abortions and infertility in ruminants [6]. *C. fetus* subsp. *venerealis* is a bovine-adapted "clone" [7] causing venereal infections and epidemic abortion in cattle. Statutory preclusion of *C. fetus* subsp. *venerealis* infection underscores the importance of this veterinary pathogen [8], but human infections are rare [6]. In contrast the generalist subspecies, *C. fetus* subsp. *fetus*, colonizes the intestinal and the genital-tract of multiple hosts including sheep, cattle, birds and humans. It is an emerging human pathogen, leading to invasive infections and even death [9,10]. Most bacteremic illnesses caused by *Campylobacter* are due to *C. fetus* [9,11].

C. fetus displays two major (O-antigen based) sero-types, A and B, and a rare variant AB [12]. The sero-types correlate with the type of surface array protein (Sap) expressed by the bacterium [13] and differ in their lipopolysaccharide (LPS) composition [12,14]. The Sap-layer (S-layer) creates a paracrystalline proteinaceous cover enabling *C. fetus* to resist serum bactericidal activity, and by phase variation to overcome immune recognition [11,15,16]. Sero-type A strains expressing SapA are more frequently isolated from human blood than sero-type B strains expressing SapB. The cattle-adapted *C. fetus* subsp. *venerealis* is exclusively sero−/sap-type A (type A). Four different *Campylobacter* clades were identified using MLST [4] and represent the genotypes (I) *C. fetus* subsp. *venerealis* type A, (II) *C. fetus* subsp. *fetus* type A or (III) type B and (IV) reptile *C. fetus* type A. The reptilian clade diverges most substantially from the other three closely related genotypes.

The evolutionary interplay between microbial pathogens and their hosts is a continual process of adaptation, manifested by genomic variation of host adaptation factors, and by the gain and loss of genes via horizontal gene transfer (HGT). The underlying hypothesis for this study was that genome reduction and acquisition of relatively few novel genes has enabled *C. fetus* to adopt distinct subspecies-specific lifestyles. To evaluate this, we performed comparative genetic analyses of *C. fetus* subsp. *venerealis* (type A) and *C. fetus* subsp. *fetus* (type A), and we compared regions of the two type A strains to type B and reptile *C. fetus* strains. To gain initial insights into the transcriptional organization of *C. fetus*, differential RNA-sequencing (dRNA-seq) was performed with the sequenced strains of both subspecies. The analyses revealed many of the molecular details involved in (sub)speciation and virulence of *C. fetus* and explain the strikingly different host tropism and clinical manifestations of these pathogens.

Results

Comparative Genomics of *C. fetus* Subspecies

The genome of the bovine strain *C. fetus* subsp. *venerealis* 84-112 (type A) was sequenced generating 216.8 Mbp sequence data (≈112-fold coverage). This strain harbors a single circular chromosome 1.93 Mbp in size with GC-content of 33.3% and a circular extra-chromosomal element of 61,141 bp with GC-content of 31.5%. Until now, the only other closed *C. fetus* genome publicly available was the human isolate *C. fetus* subsp. *fetus* 82-40 (type A). That 1.77 Mbp genome also has GC-content of 33.3%. Analysis of the two genomes revealed that they are highly syntenic with 92.9% overall sequence identity. The homologous regions exhibit 99.8% DNA identity. 180 kbp were unique for strain 84-112 and 35 kbp of unique sequences were identified in strain 82-40. Including the 73 extra-chromosomal element open reading frames (*orfs*), strain 84-112 harbors 204 unique *orfs*. Nearly all represent putative type IV secretion system (T4SS) components, transposons, or hypothetical proteins **(File S4)**. The 25 *orfs* unique for strain 82-40 encode putative CRISPR associated (Cas)-proteins, LPS-biosynthetic enzymes, or hypothetical proteins **(File S4)**. General genomic characteristics are summarized in **Table 1**.

Comparative genome plots clearly illustrate that the unique DNA stretches are located in distinct genomic regions (termed variation regions, VR) scattered across the syntenic genomic core **(Figure 1)**. *C. fetus* subsp. *venerealis* 84-112 harbors 5 VRs and *C. fetus* subsp. *fetus* 82-40 harbors 3 VRs **(Figure 1, Table S1 in File S5)**. All of these regions have features indicative of horizontal acquisition including a shift in %GC-content compared to the core genome, the presence of mobility-related genes (e.g. prophages, transposases) or proximity to tRNA genes, presumably marking their insertion sites into the chromosomal backbone. Two of the VRs of strain 84-112, Venerealis Genomic Island (VGI) I and VGI II, have no counterpart in strain 82-40. Two other VRs are shared between the two subspecies. VGI III of strain 84-112 corresponds to the position of Fetus Genomic Island (FGI) I of strain 82-40. The position of VGI IV corresponds to FGI II. The respective regions of variation are not identical between the two subspecies, but are highly similar, suggesting a common origin. Notably, the VRs of strain 84-112 carry additional blocks of genes, which are predominantly prophage-related. The extra insertions appear to interrupt functional gene modules, thus VGI III can be divided into three subsections designated VGI IIIA-1, VGI IIIA-2, and VGI IIIB (see below, **Figure 2C, Table S1 in File S5**). In strain 84-112, phage-related genes or transposases flank the VGIs. Similar genes are absent in strain 82-40 except for one area on FGI II containing prophage-like features (see below, **Figure 2E**).

One VR region **(Figure 1, Figure S1)** that co-localizes in the genomes of both subspecies was designated as the Venerealis- or Fetus Subspecies Definition Region (VSDR/FSDR). These regions are marked by comparatively low GC-content (30.7% and 29.4%, respectively). They contain genes putatively involved in surface carbohydrate metabolism as analyzed below and differentiate the two subspecies.

Metabolic reconstruction based on the genome data and comparative analyses of metabolic pathways using RAST and SEED revealed only two differences between the genomes. *C. fetus* subsp. *fetus* 82-40 harbors two *orfs* putatively involved in thiamin (vitamin B1) biosynthesis, namely a phosphomethylpyrimidine kinase (EC 2.7.4.7) and the thiamin biosynthesis protein ThiC (peg.404), which are absent from the genome of *C. fetus* subsp. *venerealis* 84-112. ThiC does not appear to be specific for *C. fetus* subsp. *fetus*, however, since a *thiC* homolog is also present in the unfinished genome of *C. fetus* subsp. *venerealis* NCTC 10354. Also no other obvious differences in respiration systems, nutrient transporters and catabolic or anabolic pathways were identified. Whether more subtle genetic differences, like insertions, point mutations or variation in transcriptional control, which might influence metabolism, contribute to the different biology of *C. fetus* subspecies remains to be elucidated.

In summary, comparative genomics revealed that the two *C. fetus* subspecies are highly syntenic, but the chromosome of *C. fetus* subsp. *venerealis* 84-112 is about 9% larger. The genomic VRs distinguishing the two subspecies are located within a small number of hot-spots, displaying features typical for horizontally acquired DNA.

VGI I and II Contain T4SS-related Genes, Prophage- and Plasmid-like Features

We previously identified and characterized a pathogenicity island (PAI) in *C. fetus* subsp. *venerealis* that was absent in all 45 *C. fetus* subsp. *fetus* isolates tested [17]. The PAI contained a full set of *virB/virD4* genes prototypical for a T4SS (for review see [18]). The T4SS of strains ATCC 19438 and 84-112 mediate conjugative DNA transfer as well as host interaction [17,19]. This PAI is located in VGI I of strain 84-112 **(Figure 2A)**. VGI I also harbors the putative prophage I encompassing a region of 33.7 kb (position: 1,266,041 to 1,299,761) with 47 *orfs* and a GC-content of 35.4%.

The gene organization of VGI II is less consistent, but with conserved functional modules **(Figure 2B)**. Although T4SS-related genes are present, the system lacks *virB5* and *virB6* and may be non-functional (see below, and **Figure S3**). Under laboratory conditions, we did not detect transcription of these genes (data not shown). The gene for transposase ISHa1152 suggests a putative integration site for VGI.

FGI I and VGI III Contain the *Sap*-locus

The *sap*-locus of *C. fetus* is present in both subspecies and represents the best-characterized *C. fetus* virulence attributes [11,15,16]. In *C. fetus* subsp. *fetus* 82-40 the *sap* genes are located on FGI I close to a tRNA and putative ABC-transporter genes **(Figure 3)**. In *C. fetus* subsp. *venerealis* 84-112, the comparable region of VGI III is highly similar to FGI I. However, a block of

Table 1. *C. fetus* genome attributes, including the extra-chromosomal element.

Attributes	***Cff***	***Cfv***	***Cfv***
	strain 82-40	**strain 84-112**	**ICE_84-112**
*Genome size (bp)	1,773,615	1,926,886	61,141
*GC-content %	33.31	33.34	31.54
*coding DNA sequence (# of *orfs*)	1,769	1,992	73
*rRNA genes	6	6	–
*tRNA genes	43	43	–
Genomic Islands	2 (FGI I–II)	4 (VGI I–IV)	–
T4SS loci			
tra-like gene cluster	0	0	1
vir-like gene cluster	0	2	1
Flexible gene pool			
Integrase XERCD family	1	1	0
Integrases/recombinases	1	2	0
Insertion Elements (# of copies)	0	ISHa1152 (2)	ISHa1152 (3)
	0	ISC1904 (3)	0
Prophage-like gene clusters	1	3	0
CRISPR			
Spacers (# of copies)	21 (1)	24 (1)	0
	26 (1)		
cas-genes	*cas1-6*	0	0

*according to RAST annotation.
doi:10.1371/journal.pone.0085491.t001

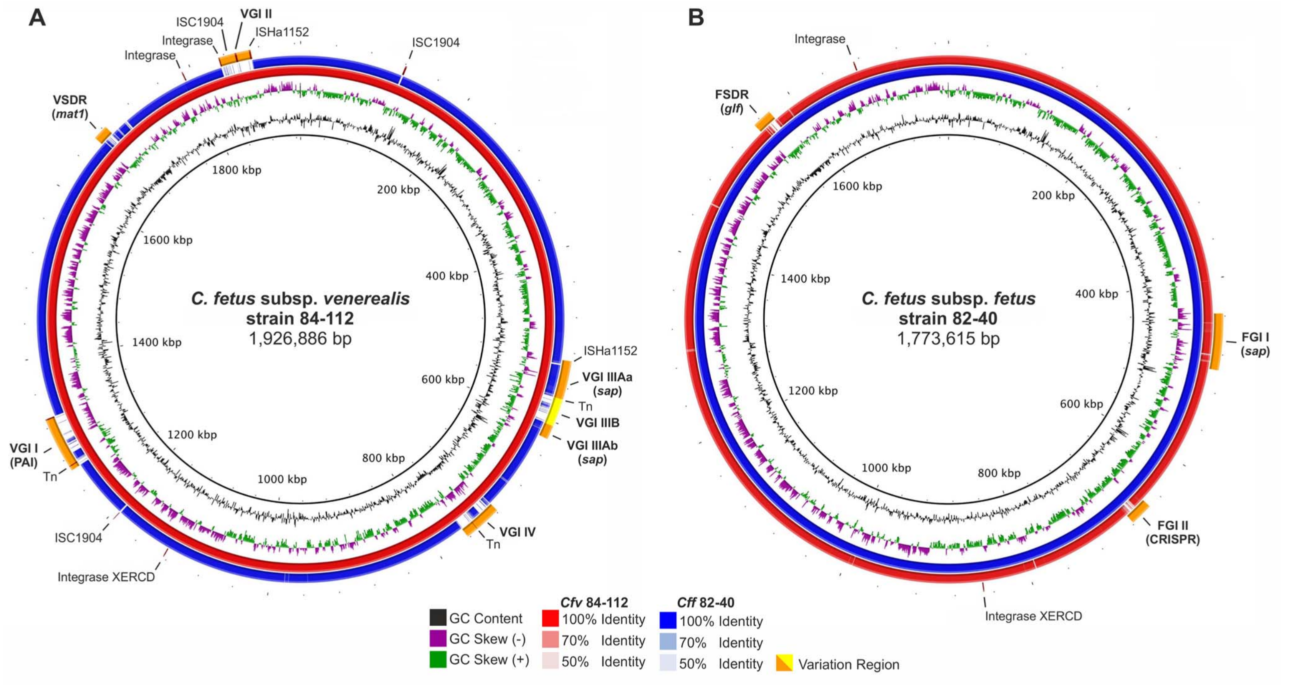


Figure 1. Genome comparisons of *C. fetus* subspecies. Plots were generated using *C. fetus* subsp. *venerealis* 84-112 (*Cfv*) as a reference (**A**) or *C. fetus* subsp. *fetus* 82-40 (*Cff*) (**B**). Inside tracks represent GC-content (ring 1) and GC-skew (ring 2). *Cff* is shown in blue and *Cfv* in red. Variation regions (VR) relative to the reference genome are indicated in orange/yellow and named according to the corresponding Genomic Island (GI) or the subspecies definition region (SDR). (V) and (F) in the feature names designate the subspecies *venerealis* and *fetus*, respectively. Important genes or features are indicated in parenthesis. Positions of selected mobility genes are indicated.
doi:10.1371/journal.pone.0085491.g001

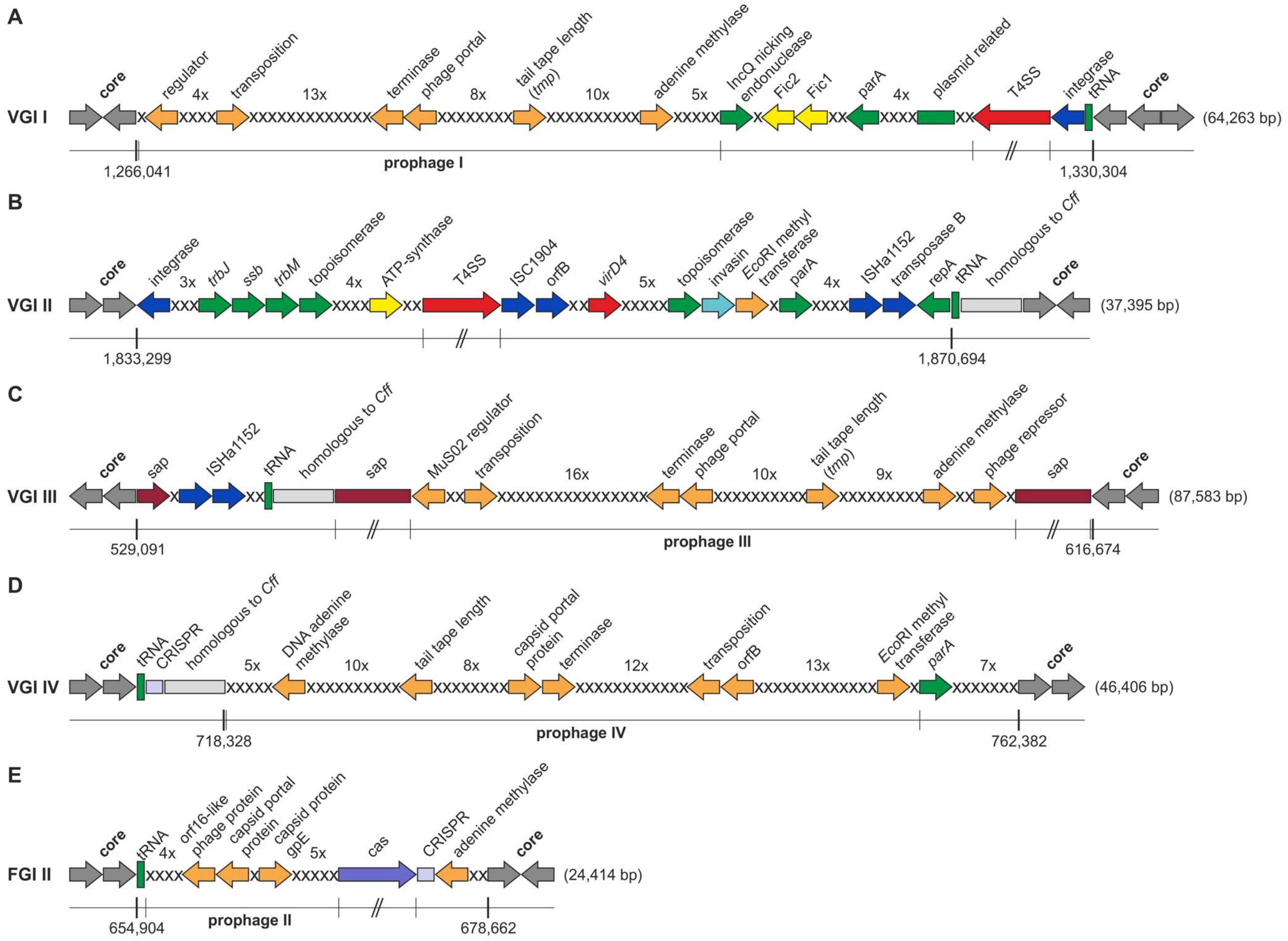


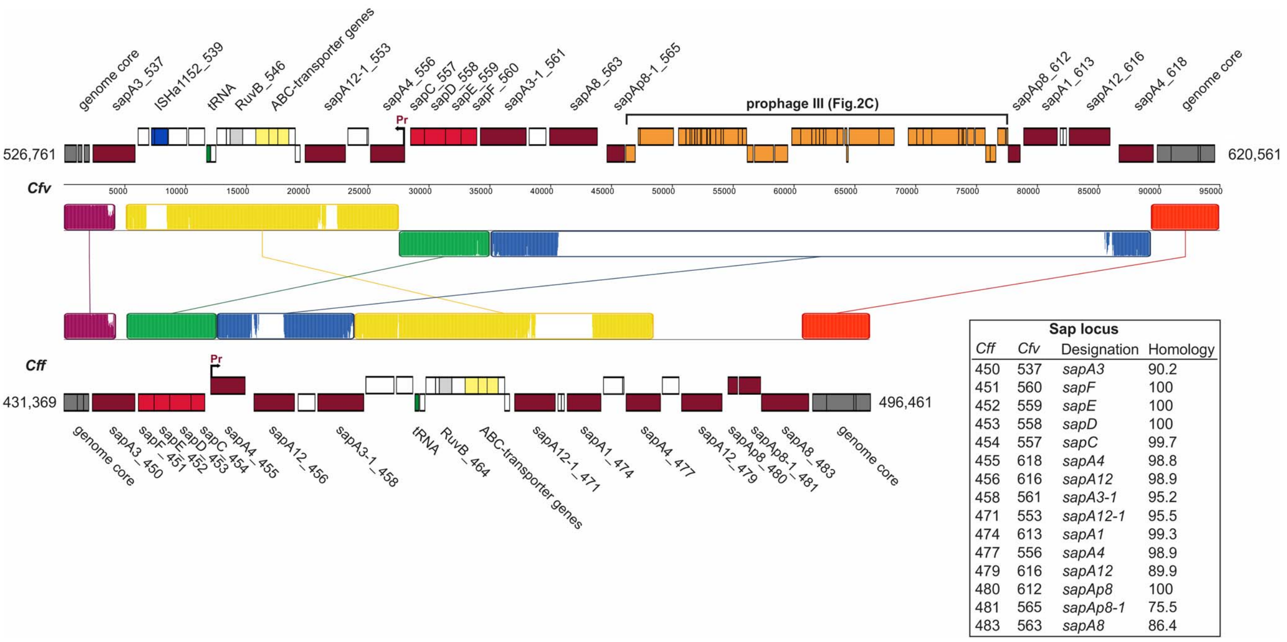
Figure 2. Comparative overview of Genomic Islands (GIs). (**A**) VGI I (PAI) with the T4SS and putative prohage I, (**B**) VGI II with a *vir*-gene cluster and plasmid-related genes, (**C**) VGI III containing the surface array protein cluster and prohage III, (**D**) VGI IV containing the CRISPR-array and prophage IV and (**E**) FGI II with prophage-related genes (prophage II) and the CRISPR-cluster (array and *cas*-genes). The GI borders to genes shared between the subspecies (grey) are indicated with nucleotide position. Gene clusters are colored as follows: phage-related genes (orange), plasmid related genes (green), integrases and transposases (blue), T4SS (red), effector proteins (yellow), surface array proteins (purple), *cas*-genes (lavender), tRNAs (green boxes); Each x represents a hypothetical protein and their numbers in tandem are indicated above.
doi:10.1371/journal.pone.0085491.g002

phage-related genes and a series of genes for hypothetical proteins indicate the presence of another prophage (**Figure 2C, Figure 3**) apparently leading to rearrangement and separation of the *sap* genes that may affect S-layer variation of *C. fetus* subsp. *venerealis* 84-112. The transcriptome analysis indicates that the insertion of prophage III did not lead to inactivation or truncation of sapAb8_612 (**Figure 4**). As in VGI I, the ISHa1152 transposase gene was detected, putatively marking a site for extra-chromosomal DNA insertion.

FGI II and VGI IV Contain CRISPR Loci

We identified CRISPR-repeats on the genomes of both *C. fetus* subspecies (**Figure S1**). In *C. fetus* subsp. *venerealis* 84-112, a single locus (nt 684,618 to 686,228) (Cfv_CRISPR) displays the typical features of a CRISPR-array with 30-bp direct repeats (DR), separated by 21 different spacers. No *cas*-homologues were identified. Two CRISPR-arrays (nt 655,350 to 656,762 and nt 674,442 to 676,187) were identified in *C. fetus* subsp. *fetus* 82-40 (Cff_CRISPR_1 and Cff_CRISPR_2), but only Cff_CRISPR_2 is in close proximity to *cas*-gene homologues. The DRs and the leader sequence are identical in both subspecies. Some spacers are shared between Cfv_CRISPR and Cff_CRISPR_1, but Cff_CRISPR_2 has no homology to Cfv_CRISPR and Cff_CRISPR_1. Sequences homologous to the spacers of the CRISPR loci were not detected in public DNA databases, thus their putative DNA targets remain unknown.

Since *Cas1* is a hallmark of dynamic CRISPR arrays, we screened 102 *C. fetus* strains for its presence. *Cas1* was detected in 19 (47.5%) of 40 subsp. *fetus* subsp. *fetus* isolates but was absent in all 62 subsp. *venerealis* isolates (Odd ratio = 110, 95% CI: 6.3 to 1,897, p = 0.0012). In strain 84-112 another prophage-like gene cluster (prophage IV) is present instead of the *cas*-genes and the second CRISPR array (**Figure 2D, Figure S1**). Interestingly, type B strains are more likely to carry the *cas1* gene (14 of 15) compared to type A strains (5 of 24) (Odd ratio = 53.2, 95% CI: 5.6 to 507.4; p = 0.0006) (**Table S6 in File S5**).



Sap locus

Cff	*Cfv*	Designation	Homology
450	537	*sapA3*	90.2
451	560	*sapF*	100
452	559	*sapE*	100
453	558	*sapD*	100
454	557	*sapC*	99.7
455	618	*sapA4*	98.8
456	616	*sapA12*	98.9
458	561	*sapA3-1*	95.2
471	553	*sapA12-1*	95.5
474	613	*sapA1*	99.3
477	556	*sapA4*	98.9
479	616	*sapA12*	89.9
480	612	*sapAp8*	100
481	565	*sapAp8-1*	75.5
483	563	*sapA8*	86.4

Figure 3. Schematic representation and structural comparison of VGI III and FGI I (sap region). MAUVE was used to compare the VRs of both subspecies for visualization of rearrangements and insertions. Regions free of rearrangements are indicated by colored colinear blocks. White color within these blocks indicates insertions or non-homologous regions. Important *orfs* are colored and labeled. S-layer genes (purple) were identified in both *C. fetus* strains. The *sap*-promoter is indicated. In *C. fetus* subsp. *venerealis* 84-112, the *sap* genes were disrupted by an inserted prophage (orange). White boxes are mainly hypothetical proteins. Detailed annotation information can be found in File S1. Genes are labeled with RAST-peg numbers and the inset table lists homologous *sap* genes of the subspecies.
doi:10.1371/journal.pone.0085491.g003

The Extra-chromosomal Element of *C. fetus* subsp. *Venerealis* 84-112 Displays Features Typical for Integrative Conjugative Elements (ICE)

The extra-chromosomal element was designated as ICE_84-112 and is the first ICE described in *C. fetus* (physical map **Figure S2**; annotation details in **File S3**). Conjugative transfer (*tra*) and other genes of apparent plasmid origin were identified but autonomous replication features were lacking. The T4SS locus, termed ICE_*trb/tra*, most likely is involved in horizontal self-transfer, based on its close relation to the broadly disseminated RP4-like systems. Several phage-related genes and transposases, including the ISHa1152 transposase, could aid chromosomal integration and excision of the ICE (**Figure 1A, Figure 2BC**). A region with structural homology to the PAI of VGI I was identified on ICE_84-112 (termed ICE_*vir*). ICE_84-112 also encodes proteins with a domain called filamentation-induced by cyclic AMP (Fic). This domain is similarly present in Fic1 and Fic2 expressed by the PAI of VGI I [17,19]. We screened our *C. fetus* collection for the presence of ICE_84-112 using the ICE specific genes *fic3* and *fic4* as PCR targets. Of 62 *C. fetus* subsp. *venerealis* strains, 7 harbored the ICE-related genes (**Table S6 in File S5**). The target genes *fic3* and *fic4* were not detected in any of the 40 *C. fetus* subsp. *fetus* strains tested. Transcriptome analysis showed expression of the majority of genes on ICE_84-112.

ICE_84-112 may replicate extra-chromosomally via a conjugative transfer replication mode, as proposed for other ICEs [20,21], since the obligatory features including a putative IncP$_{nic}$-site, an *origin of transfer*-binding protein, a relaxase, a helicase and a nicking-endonuclease were identified (**Figure S2**). According to the classification of Barcillán-Barica *et al.* [22], the putative ICE_84-112 (CDS peg.24) relaxase belongs to the MPB$_{P1}$ group (clade MOB$_{P11}$) of relaxases, displaying the typical conserved sequence motifs. Most of the MPB$_{P1}$ group of relaxases are linked to conjugative plasmids. Lee *et al.* [20] demonstrated that the chromosomally encoded *Bacillus subtilis* helicase PcrA associates with ICE*Bs1* during replication. ICE*Bs1* is defective for replication in *pcrA*-mutant strains and *pcrA* is necessary for ICE*Bs1* conjugation. PcrA orthologs, which could be recruited for replication and conjugation, are present in both *C. fetus* subspecies (84-112 CDS peg.56 & peg.1280 and 82-40 CDS peg.690 & peg.934).

dRNA-seq Identified Transcriptional Start Sites and the Typical Promoter Structure for Campylobacterales in Both Subspecies

Transcriptional start sites (TSS) annotation, performed computationally, allowed classification of TSS according to their location relative to the surrounding *orfs*. The analysis revealed a variety of transcripts with TSS located upstream and internal to their respective *orf* but also included antisense transcripts. Many TSS were simultaneously assigned to more than one category (**Figure S5**).

Sequences upstream of the annotated TSS were used to define *C. fetus* promoter motifs. *C. fetus* subsp. *venerealis* has more *orfs* than *C. fetus* subsp. *fetus* and we identified 797 promoter sequences in strain 84-112 and 575 promoter sequences in strain 82-40, with an extended Pribnow box (tgnTAtaAT) as the −10 motif in both subspecies. Consistent with other Campylobacterales [23,24] the typical bacterial −35 motif is replaced by a periodic AT-rich signal upstream of position −14 (**Figure 4AB**). This also is evident in the *sap*-locus located on genomic islands VGI III and FGI I. The intragenic promoter region between *sapC* (component of the Sap-transporter) and a respective *sap*-homologue is 100% conserved between the subspecies and only the *sap*-homolog directly downstream of the promoter is transcribed (**Figure 4C**).

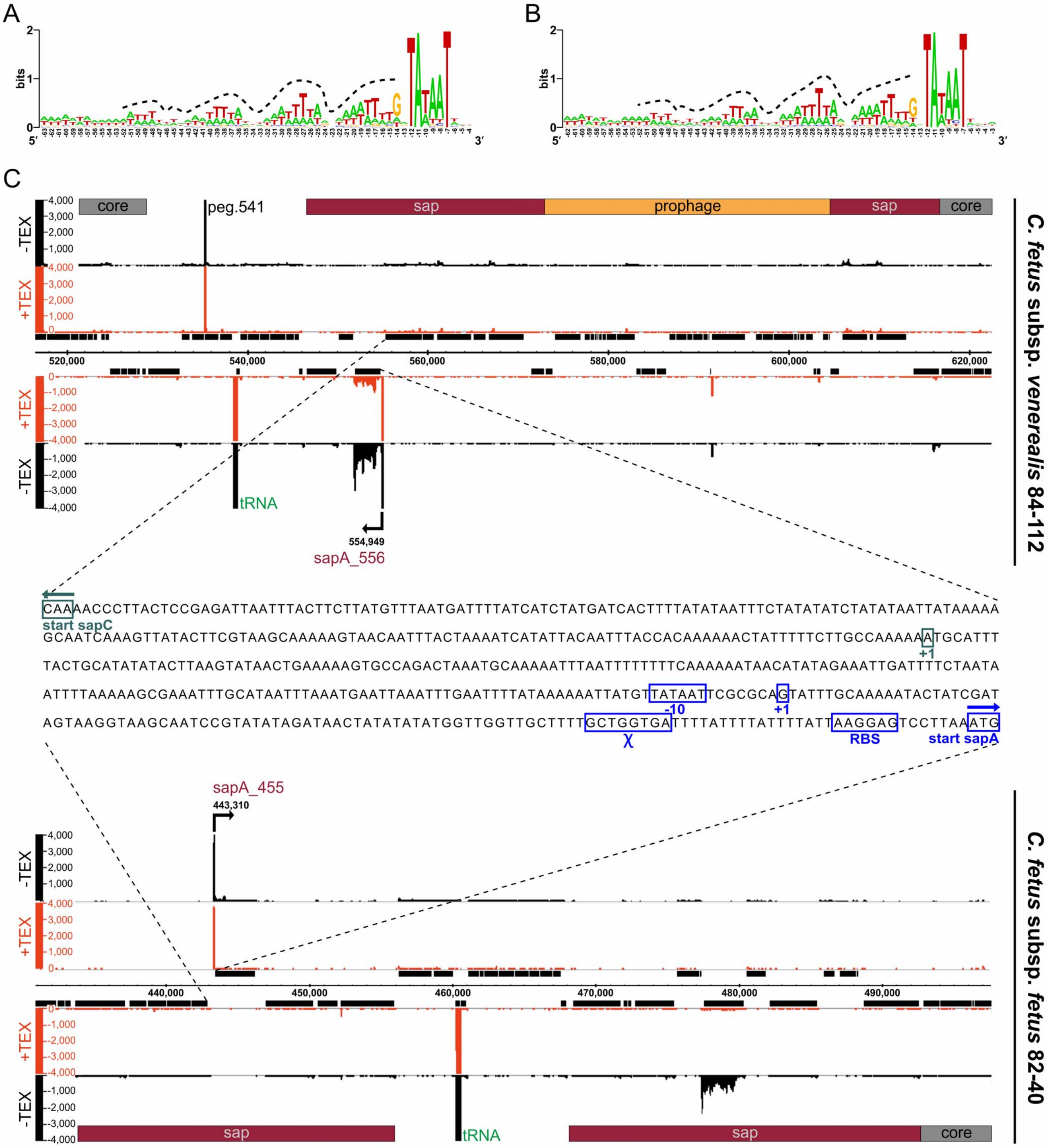


Figure 4. *C. fetus* promoter sequence and transcriptional organization of the *sap*-locus. Promoter consensus sequence for (**A**) *C. fetus* subsp. *venerealis* 84-112 (*Cfv*) and (**B**) *C. fetus* subsp. *fetus* 82-40 (*Cff*). The promoter motif is defined by an extended Pribnow box (tgnTAtaAT) at the −10 position. The −35 motif is replaced by a periodic AT-rich signal upstream of position −14 (dotted line). (**C**) Transcriptional organization of *Cfv* VG III **(top)** and *Cff* FGI I **(bottom),** identical *sap*-promoter sequence of *Cfv* and *Cff* **(middle)**.
doi:10.1371/journal.pone.0085491.g004

C. fetus subsp. *Venerealis* 84-112 Harbors T4SS-related Loci

C. fetus subsp. *venerealis* 84-112 harbors four regions showing homology to T4SS genes (**Figure S3**). Two are on the chromosome within VGI I (PAI) and II (**Figure 2AB**) and two are located on ICE_84-112 (**Figure S2**) annotated as ICE_*trb*/*tra* and ICE_*vir*. The ICE_*trb*/*tra* region differs from the other T4SS and shares homology to IncP plasmid RP4. For the ICE_*vir* region, blast searches and phylogenetic analyses using VirB4 and VirB11 [25] identified the PAI T4SS (**Table S2 in File S5**) and an as yet uncharacterized T4SS of *Campylobacter hominis* as their closest neighbor. The *vir*-genes located on VGI II did not share high homology with the *vir*-genes present on either VGI I or ICE_84-112. Instead the closest relative is a putative T4SS present in *C. rectus* RM3267, indicating a different origin. Finally, transcriptome analysis indicated that the VGI III T4SS components are not transcribed under laboratory conditions, whereas expression of the PAI T4SS (VGI I), ICE_*vir* and ICE_*trb*/*tra* was detected (data not shown, and [17]).

Genes Involved in LPS-biosynthesis Distinguish *C. fetus* Sero−/Sap-types

The subspecies definition regions contain unique genes putatively involved in LPS-biosynthesis. Although inserted at the same chromosomal position in both subspecies **(Figure 1AB)** the islands display only limited similarity **(Figure S4)**. One obvious difference was that VSDR encodes a putative maltose O-acetyltransferase (*mat1*) (cd04647) and FSDR a putative UDP-galactopyranose mutase (*glf*) (EC 5.4.99.9) **(Figure S4)**. Remarkable is the low GC-content of the VSDR and FSDR of 30.7% and 29.4%, respectively **(Table S1 in File S5)** and the absence of tRNA or apparent mobility genes.

Acetyltransferases generally catalyze the CoA-dependent acetylation of the 6-hydroxyl group of sugar substrates. Maltose O-acetyltransferases exclusively acetylate maltose and glucose. *C. fetus* type A LPS contains 74.5% mannose as well as 6.5% D-glucose [26] and thus may serve as a substrate for Mat1. UDP-galactopyranose mutase (*glf*) drives the conversion of the ring form of galactose from pyranose to furanose. The latter isomer is specifically found in glycoconjugates (including LPS) of various prokaryotic and eukaryotic pathogens, and is essential for their physiology and virulence [27,28]. To assess conservation of the subspecies-specific regions, a panel of 102 geographically and phenotypically diverse strains of *C. fetus* subspecies was screened for the presence or absence of *mat1* and *glf*. Of 62 subsp. *venerealis* isolates (all type A), 58 (93.5%) were positive for *mat1* and all were negative for *glf*. In contrast, only 16 (40%) of 40 subsp. *fetus* strains harbor *mat1* but 25 (62.5%) were positive for *glf*. The 16 subsp. *fetus* strains positive for *mat1* were all type B, whereas 24 of the 25 *glf* positive strains were type A **(Table S3, Table S6 in File S5)**. The single exception, *C. fetus* subsp. *fetus* isolate F9, which was positive for both *mat1* and *glf*, belongs to the rare group of type AB strains.

Our previous application of RDA (representational difference analysis) to *C. fetus* revealed that another LPS-biosynthesis gene (*wcbK*) encoding a putative GDP-mannose 4,6-dehydratase was exclusively present in *C. fetus* subsp. *fetus* strains [17]. In strain ATCC 27374 (type B), *wcbK* is flanked 3′ by *wbbC*, encoding a putative glycosyltransferase, and 5′ by a *sap* gene (data not shown). This region corresponds to FGI I in strain 82-40, which lacks *wcbK*. WcbK catalyzes the first step in the biosynthesis of GDP-D-rhamnose and GDP-L-fucose, and is involved in capsular polysaccharide or LPS-biosynthesis in bacteria such as *Helicobacter pylori* [29] and *C. jejuni* [30]. A PCR screen of the *C. fetus* panel confirmed that *wcbK* was not present in any of the *C. fetus* subsp. *venerealis* isolates but was exclusively detected in the 16 *C. fetus* subsp. *fetus* isolates, which were also positive for *mat1*. All of the *wcbk*+ *mat1*+ strains were type B. Thus, *C. fetus* subsp. *fetus* either carried *glf* alone in type A strains or *mat1* in combination with *wcbK* in type B strains. *C. fetus* subsp. *venerealis* (type A) only carries *mat1*. *C. fetus* subsp. *fetus* strain F9 scored positively for *mat1*, *wcbK* and *glf*.

Another phylotype of *C. fetus* is represented by reptile *C. fetus* strains, which are type A, and may represent the ancestral *C. fetus* type [4,31]. We screened four reptile isolates, which were all positive for *mat1* but lacked *glf*, *wcbK*, *virD4* and *fic1-4* **(Table S4 in File S5)**.

Finally, another enzyme of the LPS-biosynthetic pathway UDP-glucose 4-epimerase (GalE, EC 5.1.3.2) catalyzes the reversible conversion of UDP-glucose to UDP-galactose and is known to contribute to *C. jejuni* virulence [32]. Southern-blot and PCR screens of our collection showed that all 102 *C. fetus* isolates studied carried *galE*.

wcbK is Involved in LPS-biosynthesis and Accordingly should have an Impact on Acid Resistance and Serum Sensitivity in *C. fetus* subsp. *Fetus* Type B Strains

Type A strains are resistant to complement-mediated killing since C3b binding to the bacterial cell surface is inhibited by the presence of the S-layer [33,34]. It is not known why type B strains are sensitive to non-immune serum [12], despite the presence of the surface array protein. We hypothesized that *wcbK* might be linked to the susceptibility of type B strains by generating O-specific side chains where the C3b binding site is not covered by the S-layer. To test this, we first screened *C. fetus* subsp. *fetus* type A and B strains with known serum resistance phenotypes for *wcbK* and *glf* **(Table S5 in File S5)**. As hypothesized, *wcbK* was exclusively found in type B strains and correlated with serum susceptibility, whereas *glf* only was present in type A strains and correlated with serum resistance. Next we generated a non-polar *wcbK* mutant (K19) of *C. fetus* subsp. *fetus* ATCC 27374 (type B) that was deficient in LPS-production (Figure 5A). In *Vibrio cholerae* mutant strains it has been shown that providing genes in trans only partially restored LPS-production compared to wild type levels [35]. In our experiments, providing *wcbK* in trans also partially complemented LPS-production. Due to antibiotic selection throughout the experiment we can exclude the loss of the complementation vector. We next compared serum-susceptibility of mutant and wild type strains **(Figure 5BC)**. As expected, *C. fetus* subsp. *fetus* ATCC 27374 did not survive serum treatment ($\log_{10}$ kill 2.23±0.06) whereas the isogenic *wcbK* mutant strain K19 had markedly increased serum-resistance ($\log_{10}$ kill 0.86±0.05). The phenotype was partially complemented ($\log_{10}$ kill 1.23±0.10) by providing *wcbK in trans*. The serum resistant strain 82-40 (type A) was used as a control ($\log_{10}$ kill 0.27±0.01).

Type A and type B *C. fetus* strains differ in the carbohydrate composition of their LPS [12,26]. The O-antigen of type A strain has a higher molecular weight (Figure 5A) than that of type B strains. *C. fetus* strains 84-112, 82-40 and ATCC 27374 are similar in their resistance to acid (Figure 5 and results not shown). In *H. pylori* GDP-mannose 4,6-dehydratase (encoded by *wbcJ*) is important for the expression of O-antigen and for the bacterium to survive the acidic milieu of the stomach [29]. We hypothesized that the loss of LPS in the *wcbK* deficient *C. fetus* strain might result in increased acid sensitivity. Indeed, when incubated at low pH the wild type strain (ATCC 27374) survived significantly better than the *wcbK* mutant; this acid-sensitive phenotype was partially complemented by providing *wcbK in trans* (Figure 5CD).

In summary, *wcbK* is important for LPS-biosynthesis and SapB binding. Activity of this enzyme attenuates survival of the pathogen in blood, and also can provide effective protection from stomach acid en route to colonization of the intestinal niche.

Discussion

ε-Proteobacteria including *Campylobacter* and its close relative *Helicobacter* show evidence of genome reduction indicated by small genome size (≈1.5 to 2.5 Mbp) and the nearly complete absence of non-coding DNA. These features are typical for adaptation to a

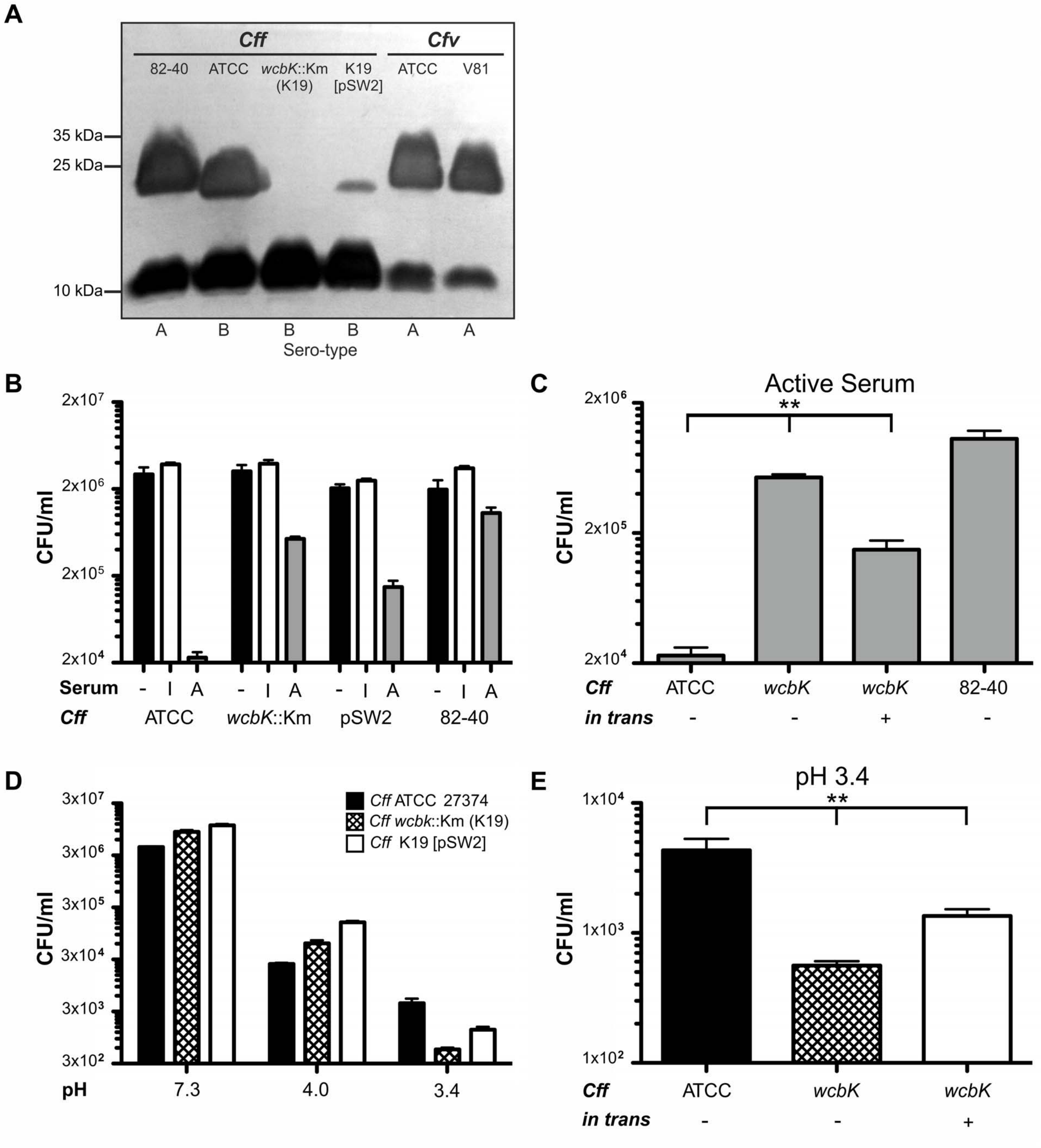


Figure 5. WcbK is important for LPS-biosynthesis, attenuates survival in blood, and promotes acid resistance. (**A**) SDS-PAGE pattern of purified LPS after silver staining. Samples were isolated from *C. fetus.* subsp. *fetus* (*Cff*) 82-40 (lane 1), *Cff* ATCC 27374 (type B) (lane 2), *wcbK* mutant K19 (*wcbK*::Km) (lane 3) and K19 [pSW2] (*wcbK in trans*) (lane 4); *C. fetus* subsp. *venerealis* (*Cfv*) ATCC 19438 (lane 5) and *Cfv* 84-112 (lane 6). (**B**) *Cff* serum resistance assays. Strains were incubated either with EMEM (-), heat-inactivated (I) or active (A) human serum and colony forming units (CFU) were counted. Results shown are for *Cff* ATCC 27374, K19 and K19 [pWS2]. *Cff* 82-40 served as a type A comparator. (**C**) Same as in (B) but for better visualization, CFU/ml obtained after treatment with active serum are displayed separately. $^{**}p<0.002$ (**D**) Acid resistance assays. *Cff* were incubated in PBS pH range 7.3 to 3.4, plated and CFU determined. Survival after exposure to different pH of the wild type, K19 and K19 [pSW2] was compared. (**E**) For better visualization, CFU/ml for the three strains after treatment with pH 3.4 were plotted separately. $^{**}p<0.003$.
doi:10.1371/journal.pone.0085491.g005

specific colonization niche and both species display strong host preference ("tropism") [36,37]. Among *Campylobacters*, *C. fetus* subspecies are an exceptional model system to study the molecular basis of pathogen-host adaptation since, despite a highly clonal structure, they display strikingly dissimilar host preferences and tissue tropism. To investigate the genetic basis underlying the distinct pathogenicity of *C. fetus* subspecies, we performed whole genome comparisons and transcriptome analyses of *C. fetus* subspecies, focusing on identifying differences that contribute to host and tissue tropism. We propose that the additional genome content of *C. fetus* subsp. *venerealis* was horizontally acquired **(Table S1 in File S5)**. The observation that genes shared between the subspecies are nearly 100% identical on the nucleotide level supports the hypothesis that HGT and not mutation or genetic drift is the predominant factor in the evolution of *C. fetus*.

To gain insights to the genetic plasticity of *C. fetus* genomes, and particularly whether the identified variation regions are conserved we compared the VGI – IV of *C. fetus* subsp. *venerealis* 84-112 to the draft genome sequences of *C. fetus* subsp. *venerealis* NCTC 10354 (ATCC 19438) [38], *C. fetus* subsp. *venerealis* Azul-94 [39] and *C. fetus* subsp. *venerealis* biovar Intermedius INTA 99/541 [40]. We identified homologous sequences in all three strains with over 90% homology on the nucleotide level. These results indicate that the GIs are at least partially present in other *venerealis* strains. However, given that many of the remaining contig boundaries are located in the variable regions, to be able to perform more detailed analysis the draft genomes will need to be closed and the sequences verified.

We focus in the current study on the description of genomic regions and genes unique to each subspecies. Genome comparisons of *C. fetus* subspecies reported previously using the draft sequences of *C. fetus* subsp. *venerealis* strains [39,41] focused mainly on the description of shared putative virulence factors or the identification of putative targets for diagnostics. Many of the genes putatively involved in adherence, invasion, motility, secretion and toxin production identified by Ali *et al.* [41] and Moolhuijzen *et al.* [39] were also present in strain 84-112 (File S1). Homologs to the antibiotic resistance gene cluster identified within a homologous genomic island in *C. fetus* subsp. *fetus* IMD 523-06 [42] were not present in *C. fetus* subsp. *venerealis* ATCC and 84-112.

Metabolic differences between *C. fetus* subspecies such as glycine tolerance, H2S production and selenite reduction have traditionally been used to discriminate the subspecies and are therefore intriguing features linked to niche adaptation. Nonetheless, metabolic modeling of the two genomes revealed no apparent subspecies differences, except a possible difference in thiamin (vitamin B1) biosynthesis. The overall metabolic capacity seems to be similar in both subspecies, consistent with our model that the described horizontally acquired genetic elements account for the different biology of *C. fetus* subspecies. However, it is important to note that subtle genetic differences, like point mutations, can inactivate genes or disrupt metabolic pathways. Therefore, nutrient utilization by the *C. fetus* subspecies remains an important priority for detailed study.

The extra-chromosomal element ICE_84-112 was identified. ICEs are plasmid-like self-transmissible mobile genetic elements, dependent on phages or transposons for inserting and excising from chromosomes, but carry their own transfer genes (*tra*-genes) for lateral transmission to other host cells. Notably the full repertoire of plasmid replication genes is typically absent. Some ICE replicate autonomously if they adopt a rolling-circle-like mechanism mediated by replication- or single-strand DNA transfer initiation factors [20,21]. In *Bacillus subtilis* helicase PcrA associates with ICE*Bs1* during replication [20]. Candidate PcrA orthologs are present in both *C. fetus* subspecies (84-112 CDS peg.56 & peg.1280 (**File S1**) and 82-40 CDS peg.690 & peg.934 (**File S2**)). The surveyed *fic3* and *fic4* genes suggest that the distribution of ICE_84-112 is quite narrow. In that case important virulence-associated characteristics are unlikely to be carried by the element, but it may be a vehicle of interspecies gene exchange.

C. fetus subsp. *fetus* 82-40 mostly lacks phage- and plasmid-related genes and this might be due to the presence of an active CRISPR cluster, protecting from invasion of foreign DNA. Although there are six core *cas*-genes, *cas1* may be of central importance in the acquisition of new spacers (for review see [43]). In contrast to *C. fetus* subsp. *venerealis* 84-112, we identified two CRISPR-arrays in strain 82-40. Since Cff_CRISPR_2 showed prototypical architecture, i.e., *cas*-genes and an AT-rich leader sequence followed by the DRs and the spacers, this CRISPR array may be functional. The presence of *cas*-genes in *C. fetus* subsp. *fetus* highlights another important subspecies difference. The occurrence of CRISPRs is linked to natural competence of bacteria [44]. That *C. fetus* subsp. *fetus* type B strains more frequently harbor putative functional CRISPRs than type A strains might have stabilized the type B phylotype and may explain why the type A clade later diverged [4] (Figure 6). All of the *C. fetus* strains that we and others have thus far tested are not naturally competent (unpublished data, [45,46]) thus a possible connection between the presence of CRISPRs and natural competence of *C. fetus* subspecies remains unresolved.

The most important genetic differences between the subspecies are cell surface structures including the S-layer and LPS. The distribution of these genes across a panel of diverse *C. fetus* isolates indicates linkage to particular pathotypes. The distinct distribution patterns detected for *wcbK*, *mat1*, and *glf* among type A and B strains support the following model **(Figure 6)**. *wcbK and glf* are subsp. *fetus*-specific genes that have been acquired more recently than *mat1* and *galE, which* represent "ancient" constituents of the *C. fetus* genome. These loci are similar in reptile *C. fetus* and *C. fetus* subsp. *venerealis* but MLST reveals that variation has emerged and that type B strains separated from type A prior to the division of *C. fetus* subsp. *fetus* and *C. fetus* subsp. *venerealis* [4]. We showed that type B strains maintained *mat1* and *galE* but diversification of phylotypes led to acquisition of *wcbK* by *C. fetus* subsp. *fetus* type B. *C. fetus* subsp. *venerealis* also maintained *mat1* and *galE*, but type A *C. fetus* subsp. *fetus,* the invasive pathotype often found in human infections, have lost *mat1* and acquired *glf*. Extended analysis of *C. fetus* evolution will require analysis of more geographically and phenotypically diverse isolates. Moreover, analysis of the newly proposed subspecies/biovar intermedius [7] may provide a missing link in the subspecies divergence.

Little is known how *C. fetus* interacts with the host immunity, but LPS and the S-layer are important for TLR4-mediated recognition [47,48]. The S-layer producing *C. rectus* induces TLR4 expression in the mouse placenta [49]. To avoid dysregulated inflammatory responses to LPS, the intestinal epithelium as well as placental tissue normally express no or low levels of TLR4 [48,50,51]. Low density of TLR4 may allow *C. fetus* to overcome the hosts' immune response and subsequently invade the host cells. Type A and type B *C. fetus* strains are different in their LPS composition and S-layer proteins [12,26]. The activity of WcbK and the putative functions of *mat1* and *glf* are linked to the S-layer. *C. fetus* subsp, *venerealis* strains (*wcbk−/glf−/mat1*+) and *C. fetus* subsp. *fetus* type A strains (*wcbK−/glf*+/*mat1*-) are serum resistant, whereas *C. fetus* subsp. *fetus* type B strains (*wcbK*+/*glf−/mat1*+) are serum sensitive. We showed that *wcbK* is essential for LPS-biosynthesis in *C. fetus* subsp. *fetus* type B strains and that loss of *wcbK* leads to increased serum resistance. This data indicates that WcbK generated side chains are important for serum sensitivity. We propose that similar to *wcbK*, the products of *mat1* and *glf* of *C. fetus* might be involved in LPS-biosynthesis by generating different O-antigen side chains, potentially influencing complement and antibody binding, acid resistance and TLR-4 recognition.

The bacterial transcriptome provides an additional reference to study genome composition as well as regulation of virulence. In the initial profile of *C. fetus* gene transcription, the characteristic ε-proteobacterial promoter signature was identified. We confirmed

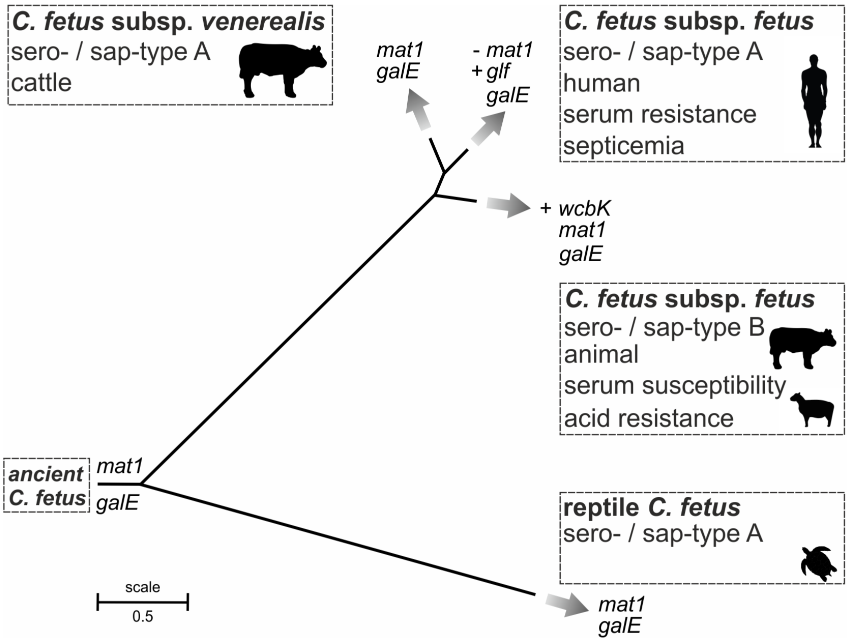


Figure 6. Phylogeny, niche specificity and virulence of *C. fetus* subspecies. MLST tree showing the phylogeny of *C. fetus*, with original scale as reported [4]. Reptile *C. fetus* represent a distinct clade harboring *mat1* and *galE*. Diversification of *C. fetus* subsp. *fetus* (*Cff*) type B happened prior to the diversification of *Cff* type A and *C. fetus* subsp. *venerealis* (*Cfv*) type A strains. *Cff* type B strains harbor *galE*, *mat1* and *wcbK*. The latter gene provides protection from acid, and this genotype is associated with animal hosts. *Cfv* type A represents the bovine clone harboring *mat1* and *galE* which is also prone to HGT. *Cff* type A have lost *mat1* but acquired *glf* correlating with serum resistance in *Cff*.
doi:10.1371/journal.pone.0085491.g006

that the promoter region is 100% conserved between the subspecies, and that one *sap* gene is predominantly transcribed under laboratory conditions. This finding is intriguing since recombination and therefore exchange of *sap*-homologs occurs frequently in this region to enable phase variation of the pathogen [11]. It has been proposed that the *sap*-region belongs to the ancestral part of the *C. fetus* core genome and not a PAI [52]. That the region is shared between both subspecies confirms ancient presence of a horizontally acquired element. Based on the significance of the S-layer for immune evasion [11,15,16], the genome insertion can be considered as a classical PAI. To date, animal models of *C. fetus* infection are not readily available. Future analyses at the transcriptiome level should investigate *C. fetus* under *in vitro* conditions resembling their colonization niche or route of infection.

Whole-genome comparisons of related pathogens of distinct characteristics, such as those described in the presented work, lay the foundation for additional mutational, functional, and animal studies that will ultimately help elucidate the mechanisms underlying the emergence of new pathogens. This study broadens knowledge of the genomic basis of *C. fetus* pathogenesis and host specificity. The most interesting differences in the genetic repertoire of the subspecies relate to cell surface structures including the S-layer and LPS and distribution of these genes is associated with certain pathotypes. This emphasizes the importance of surface-exposed structures to *C. fetus* pathogenicity and demonstrates how evolutionary forces optimize the fitness and host adaptation of these pathogens. The presence of genes like *glf* is particularly interesting as the gene product is a promising drug target, as proposed for *Leishmania* [53], and relevant since *glf* is connected to type A strains, which are more often isolated from human blood. In any event, *wcbK* and *glf* are excellent candidates applicable for reliable subspecies differentiation.

Experimental Procedures

Bacterial Strains

Campylobacter and *E. coli* strains were grown as described [45]. Antibiotic selection applied concentrations of 100 µg ml^{-1} ampicillin, 75 µg ml^{-1} nalidixic acid, or kanamycin and chloramphenicol at 25 µg ml^{-1}. Bacterial strains are listed in **Table S6** and **Table S7 in File S5**. Only *C. fetus* strains typed definitively to the subspecies level were tested in PCR screens (n = 102). Subspecies were identified biochemically as described [17].

Gene Detection

Oligonucleotides are listed in **Table S8 in File S5**. PCR amplification for surveying gene prevalence used chromosomal DNA and the following primer pairs 1/2 for *wcbK*, 3/4 for *glf* and 5/6 for *mat1*. The *sap*-type was determined with primers 7/8 and 9/10, as described [54]. Southern blots were hybridized with radiolabeled DNA probes as described [17]. Probes for *galE* and *cas1* were generated with primer pair 11/12 and 13/14 from chromosomal DNA of *C. fetus* subsp. *fetus* ATCC 27374, respectively. The same primers were used for PCR-screening for *galE* and *cas1*. *fic3* and *fic4* were amplified with primer pairs 15/16 and 17/18, respectively.

Genome Sequencing, Assembly and Annotation

A standard whole genome shot-gun and a 3-kb paired-end library were generated according to the manufacturer's recommendations (Roche Diagnostics, Vienna, Austria) using 5 µg chromosomal DNA. For each library, high-throughput pyrosequencing was performed on a Genome Sequencer FLX system (Roche) producing 145 Mb and 62.2 Mb sequence data, respectively. Read assembly applied the Newbler assembly software, version 2.6 (Roche) and resulted in 89 contigs and 11 scaffolds. One scaffold represented the circular extra-chromosomal element and the remaining 10 were grouped into 3 super-scaffolds (SSc) using the information from the 3 kbp mate-pair library and the contig-graph generated by the Newbler assembler. Additionally, PCR and Sanger sequencing was used to determine the orientation and order of contigs and the SSc. Gaps in the extra-chromosomal element and the chromosome were closed *in silico* with a custom R script [55] and with PCR. Homopolymer uncertainties from the 454-reads were corrected through mapping of the Illumina reads derived from the *C. fetus* subsp. *venerealis* 84-112 RNA to the draft sequence using CLC Genomics Workbench 5.5 (CLC Bio; Arhus, Denmark). The resulting consensus sequences and *C. fetus* subsp. *fetus* strain 82-40 were annotated and compared with Rapid Annotations using Subsystem Technology version 4.0 (RAST) [56]. Annotation tables for each strain and the extra-chromosomal element are presented in **Files S1–S3**.

Differential RNA-sequencing

Library preparation for dRNA-seq was performed as reported [23]. In brief, RNA was isolated from bacterial cells grown on CBA plates for 24 h. To construct differential cDNA library pairs, aliquots of extracted RNA from each strain was treated with Terminator-5′-phosphate-dependent exonuclease (TEX; Epicentre) to deplete processed RNAs (denoted TEX+) in addition to untreated RNA (denoted TEX-). Construction of cDNA libraries was performed by *vertis* Biotechnology AG (Munich, Germany). Libraries were sequenced using cluster amplification with the TruSeq PE Cluster Kit v.5 on a cluster station. Each library was sequenced on a single HiSeq 2000 lane using TruSeq SBS 36 Cycle Kits v.5 (Illumina, San Diego, CA) and a 91 bp single-end protocol. Sequencing image files were processed with the Sequencing Control Software (SCS) Real Time Analysis (RTA) v2.6 and CASAVA v.1.7 (Illumina). Reads were mapped to the reference genomes using the CLC Genomics workbench (CLC Bio) with default settings. Information on transcriptional start site (TSS) and promoter annotation can be found in the supplement.

Lipopolysaccharide Analysis

C. fetus strains were grown for 24 h and resuspended in buffer (10% glycerine, 20% SDS, 5% β-mercaptoethanol, 62.5 mM Tris-HCl pH 6.8, bromophenol blue) for lysis at 100°C for 10 min. Proteinase K solution was added to 6 μg/μl and samples were incubated overnight at 55°C. LPS-preparations were electrophoretically on 15% polyacrylamide gels (running buffer: 86 mM glycine, 3,5 mM SDS and 25 mM Tris pH 8). Gels were fixed overnight (25% isopropanol, 7% acetic acid) under gentle shaking. LPS was oxidized with 100 ml fixative containing 4 mmol $NaIO_4$ for 10 min. After three washing steps with H_2O for 30 min each, the gels were stained (19 mM NaOH, 1.35% NH_3, 20 mM $AgNO_3$) for 10 min, then washed three times with H_2O and immersed in developer (240 mM Na_2CO_3, preheated to 60°C, before addition of 30 μl 40% formaldehyde). The reaction was stopped with 50 mM EDTA (pH 8) for 1 h.

Serum and Acid Resistance Testing

Susceptibility of *C. fetus* strains to human serum was assessed as described [57]. All tests were performed in triplicate. Briefly, *C. fetus* was streaked on CBA plates 24 h prior to the assay and cell count was adjusted to 1×10^7 bacteria/ml, based on optical density in EMEM medium. The actual cell count was determined by plating serial dilutions. Heat-inactivated- (56°C for 30 min), or active- (thawed on ice) pooled human serum was added to the bacteria to a 10% final concentration and incubated for 1 h at 37°C. Surviving cells were counted on CBA plates after 48 h growth. For the acid resistance assays, *C. fetus* cells were harvested as described above, centrifuged, resuspended in PBS with different pH values and incubated at 37°C for 30 min. Cells were washed in PBS (pH 7.3) before the number of surviving bacteria was determined by plating serial dilutions.

Nucleotide Sequence Accession Numbers

The genome sequence of *C. fetus* subsp. *venerealis* 84-112 including the ICE element (ICE_84-112) has been deposited in EMBL Nucleotide Archive under accession numbers (HG004426 and HG004427). The genome of *C. fetus* subsp. *fetus* 82-40 used for comparative analyses has the GenBank accession number CP000487.1. dRNAseq data can be accessed via the EMBL-EBI short read archive under the accession number ERP002581.

Supporting Information

Figure S1 Comparative maps of CRISPR-related genomic islands. **(A)** *C. fetus* subsp. *venerealis* 84-112 VGI IV harbors the direct repeats with spacers (CRISPR) but lacks CRISPR-associated (*cas)*-genes. Prophage-related genes (putative prophage IV) were identified (orange) adjacent to a region identical to *C. fetus* subsp. *fetus* 82-40 Downstream of these regions the core-genome continues with a chromosomal rearrangement between the two subspecies on the 3-prime end (striped boxes). A sequence region shared between the subspecies was identified (blue box). **(B)** *C. fetus* subsp. *fetus* 82-40 FGI I carries two regions of direct repeats and spacers. *cas*-genes precede the second CRISPR-array resulting in a putatively functional CRISPR-system. One region with a prophage-like structure (orange) was identified.
(TIF)

Figure S2 Physical map of the extra-chromosomal element ICE_84-112. Shown is the GC-content (circle 1), GC-skew (circle 2) and open reading frames (circle 3). The *tra*-region (red) comprises genes putatively involved in conjugative transfer of the ICE. The *vir*-region (orange) shows putative T4SS genes with homology to the chromosomal PAI on VGI I. Genes possibly involved in autonomous replication of the ICE are named individually and labeled (green and red). Genes of predicted plasmid origin (green); phage genes and transposons (blue); putative effector proteins or toxin-antitoxin system (yellow); hypothetical proteins (grey).
(TIF)

Figure S3 Schematic representation of the apparent T4SS identified in *C. fetus* subsp. *venerealis* 84-112. (A, B, C) Represent loci with homology to *virB/virD4*-genes. **(A)** The PAI T4SS is functional in virulence and conjugative DNA transfer [1,2]. **(B)** ICE_*vir* displays a similar gene organization to VGI I but protein homologies are not strikingly high. *virD4* is truncated compared to the functional PAI homologue. **(C)** A partial set of *vir*-genes. **(D)** ICE_*trb*/*tra* genes share homology to plasmid RP4 and are putatively involved in the conjugative transfer of ICE_84-112. Homologous genes (*vir, tra*) are indicated by color.
(TIF)

Figure S4 Comparative map of *C. fetus* subspecies variation regions VSDR and FSDR. **(A)** *C. fetus* subsp. *venerealis* 84-112 VSDR and **(B)** *C. fetus* subsp. *fetus* 82-40 FSDR. MAUVE was used to compare the regions to visualize rearrangements and insertions. Regions free of rearrangements are indicated by colored colinear blocks. White regions within these blocks symbolize insertions or non-homologous regions. Important open

reading frames are colored and/or labeled accordingly. Genes unique to the subspecies, *mat1* and *glf*, are highlighted in pink.
(TIF)

Figure S5 Venn diagram of annotated TSS. (A) *C. fetus* subsp. *venerealis* 84-112 and **(B)** *C. fetus* subsp. *fetus* 82-40. TSS were categorized according to the genomic context into five classes: primary (TSS having the most cDNAs within ≈500 bp upstream of annotated mRNA start codons), secondary (TSS associated with the same gene but with fewer cDNAs), internal (TSS within an annotated gene on the same strand), antisense (TSS situated inside or within ≈100 bp of the coding region of a gene encoded on the opposite strand), or orphan (TSS without annotated genes in proximity) [3]. Numbers in parentheses indicate the TSS, which associate with only one *orf*.
(TIF)

File S1.
(XLSX)

File S2.
(XLSX)

File S3.
(XLSX)

File S4.
(XLSX)

File S5.
(DOC)

Acknowledgments

We thank B. Scherz and K. Gehmacher (Graz, Austria) for their contribution to this work and W. G. Miller (USDA, Atlanta, GA, USA) for providing sequence data of the *wcbK/gmd* region. We are grateful to C. M. Sharma and K. U. Förstner (Würzburg, Germany) for dRNA-seq and help with transcriptome data analysis.

Author Contributions

Conceived and designed the experiments: SK ELZ GG MJB GIP. Performed the experiments: SK HS SW. Analyzed the data: SK HS BH GGT GG. Contributed reagents/materials/analysis tools: SK GIP MJB. Wrote the paper: SK ELZ GG.

References

1. On SL (1996) Identification methods for *Campylobacters*, *Helicobacters*, and related organisms. Clin Microbiol Rev 9: 405–422.
2. Man SM (2011) The clinical importance of emerging *Campylobacter* species. Nat Rev Gastroenterol Hepatol 8: 669–685.
3. Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, et al. (2005) Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. PLoS Biol 3 72–85.
4. Dingle KE, Blaser MJ, Tu ZC, Pruckler J, Fitzgerald C, et al. (2010) Genetic relationships among reptile and mammalian *Campylobacter fetus* by Multilocus Sequence Typing. J Clin Microbiol 48: 977–980.
5. van Bergen MA, Simons G, van der Graaf-van Bloois L, van Putten JP, Rombout J, et al. (2005) Amplified fragment length polymorphism based identification of genetic markers and novel PCR assay for differentiation of *Campylobacter fetus* subspecies. J Med Microbiol 54: 1217–1224.
6. Thompson SA, Blaser MJ (2000) Pathogenesis of *Campylobacter fetus* infections. In: Nachamkin I, Blaser MJ, editors. *Campylobacter*, 2nd Ed. Washington, D.C.: ASM Press. 321–347.
7. van Bergen MA, Dingle KE, Maiden MC, Newell DG, van der Graaf-Van Bloois L, et al. (2005) Clonal nature of *Campylobacter fetus* as defined by multilocus sequence typing. J Clin Microbiol 43: 5888–5898.
8. van Bergen MA, Linnane S, van Putten JP, Wagenaar JA (2005) Global detection and identification of *Campylobacter fetus* subsp. *venerealis*. Rev Sci Tech 24: 1017–1026.
9. Blaser MJ (1998) *Campylobacter fetus*–emerging infection and model system for bacterial pathogenesis at mucosal surfaces. Clin Infect Dis 27: 256–258.
10. Skirrow MB, Blaser MJ (2000) Clinical aspects of *Campylobacter* infections. In: Nachamkin I, Blaser, M J., editor. *Campylobacter*, 2nd Ed. Washington, D. C.: American Society for Microbiology. 69–88.
11. Tu ZC, Gaudreau C, Blaser MJ (2005) Mechanisms underlying *Campylobacter fetus* pathogenesis in humans: surface-layer protein variation in relapsing infections. J Infect Dis 191: 2082–2089.
12. Perez-Perez GI, Blaser MJ, Bryner JH (1986) Lipopolysaccharide structures of *Campylobacter fetus* are related to heat-stable serogroups. Infect Immun 51: 209–212.
13. Thompson SA (2002) *Campylobacter* surface-layers (S-layers) and immune evasion. Ann Periodontol 7: 43–53.
14. Moran AP, O'Malley DT, Kosunen TU, Helander IM (1994) Biochemical characterization of *Campylobacter fetus* lipopolysaccharides. Infect Immun 62: 3922–3929.
15. Grogono-Thomas R, Blaser MJ, Ahmadi M, Newell DG (2003) Role of S-layer protein antigenic diversity in the immune responses of sheep experimentally challenged with *Campylobacter fetus* subsp. *fetus*. Infect Immun 71: 147–154.
16. Garcia MM, Lutze-Wallace CL, Denes AS, Eaglesome MD, Holst E, et al. (1995) Protein shift and antigenic variation in the S-layer of *Campylobacter fetus* subsp. *venerealis* during bovine infection accompanied by genomic rearrangement of sapA homologs. J Bacteriol 177: 1976–1980.
17. Gorkiewicz G, Kienesberger S, Schober C, Scheicher SR, Gully C, et al. (2010) A genomic island defines subspecies-specific virulence features of the host-adapted pathogen *Campylobacter fetus* subsp. *venerealis*. J Bacteriol 192: 502–517.
18. Bhatty M, Laverde Gomez JA, Christie PJ (2013) The expanding bacterial type IV secretion lexicon. Res Microbiol.
19. Kienesberger S, Schober Trummler C, Fauster A, Lang S, Sprenger H, et al. (2011) Interbacterial macromolecular transfer by the *Campylobacter fetus* subsp. *venerealis* type IV secretion system. J Bacteriol 193: 744–758.
20. Lee CA, Babic A, Grossman AD (2010) Autonomous plasmid-like replication of a conjugative transposon. Mol Microbiol 75: 268–279.
21. te Poele EM, Bolhuis H, Dijkhuizen L (2008) Actinomycete integrative and conjugative elements. Antonie Van Leeuwenhoek 94: 127–143.
22. Garcillan-Barcia MP, Francia MV, de la Cruz F (2009) The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol Rev 33: 657–687.
23. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, et al. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464: 250–255.
24. Petersen L, Larsen TS, Ussery DW, On SL, Krogh A (2003) RpoD promoters in *Campylobacter jejuni* exhibit a strong periodic signal instead of a −35 box. J Mol Biol 326: 1361–1372.
25. Fernandez-Lopez R, Garcillan-Barcia MP, Revilla C, Lazaro M, Vielva L, et al. (2006) Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. FEMS Microbiol Rev 30: 942–966.
26. Senchenkova SN, Shashkov AS, Knirel YA, McGovern JJ, Moran AP (1997) The O-specific polysaccharide chain of *Campylobacter fetus* serotype A lipopolysaccharide is a partially O-acetylated 1,3-linked alpha-D-mannan. Eur J Biochem 245: 637–641.
27. Oppenheimer M, Valenciano AL, Kizjakina K, Qi J, Sobrado P (2012) Chemical mechanism of UDP-galactopyranose mutase from *Trypanosoma cruzi*: a potential drug target against Chagas' disease. PLoS One 7: e32918.
28. Poulin MB, Nothaft H, Hug I, Feldman MF, Szymanski CM, et al. (2010) Characterization of a bifunctional pyranose-furanose mutase from *Campylobacter jejuni* 11168. J Biol Chem 285: 493–501.
29. McGowan CC, Necheva A, Thompson SA, Cover TL, Blaser MJ (1998) Acid-induced expression of an LPS-associated gene in *Helicobacter pylori*. Mol Microbiol 30: 19–31.
30. McCallum M, Shaw GS, Creuzenet C (2011) Characterization of the dehydratase WcbK and the reductase WcaG involved in GDP-6-deoxy-manno-heptose biosynthesis in *Campylobacter jejuni*. Biochem J 439: 235–248.
31. Tu ZC, Eisner W, Kreiswirth BN, Blaser MJ (2005) Genetic divergence of *Campylobacter fetus* strains of mammal and reptile origins. J Clin Microbiol 43: 3334–3340.
32. Fry BN, Feng S, Chen YY, Newell DG, Coloe PJ, et al. (2000) The *galE* gene of *Campylobacter jejuni* is involved in lipopolysaccharide synthesis and virulence. Infect Immun 68: 2594–2601.
33. Blaser MJ, Smith PF, Repine JE, Joiner KA (1988) Pathogenesis of *Campylobacter fetus* infections. Failure of encapsulated *Campylobacter fetus* to bind C3b explains serum and phagocytosis resistance. J Clin Invest 81: 1434–1444.
34. Pei Z, Blaser MJ (1990) Pathogenesis of *Campylobacter fetus* infections. Role of surface array proteins in virulence in a mouse model. J Clin Invest 85: 1036–1043.
35. Nesper J, Kraiss A, Schild S, Blass J, Klose KE, et al. (2002) Comparative and genetic analyses of the putative *Vibrio cholerae* lipopolysaccharide core oligosaccharide biosynthesis (*wav*) gene cluster. Infect Immun 70: 2419–2433.
36. Hofreuter D, Novik V, Galan JE (2008) Metabolic diversity in *Campylobacter jejuni* enhances specific tissue colonization. Cell Host Microbe 4: 425–433.
37. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388: 539–547.

38. Stynen AP, Lage AP, Moore RJ, Rezende AM, de Resende VD, et al. (2011) Complete genome sequence of type strain *Campylobacter fetus* subsp. *venerealis* NCTC 10354T. J Bacteriol 193: 5871–5872.
39. Moolhuijzen PM, Lew-Tabor AE, Wlodek BM, Aguero FG, Comerci DJ, et al. (2009) Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets. BMC Microbiol 9: 86.
40. Iraola G, Perez R, Naya H, Paolicchi F, Harris D, et al. (2013) Complete Genome Sequence of *Campylobacter fetus* subsp. *venerealis* Biovar Intermedius, Isolated from the Prepuce of a Bull. Genome Announc 1.
41. Ali A, Soares SC, Santos AR, Guimaraes LC, Barbosa E, et al. (2012) *Campylobacter fetus* subspecies: Comparative genomics and prediction of potential virulence targets. Gene.
42. Abril C, Brodard I, Perreten V (2010) Two novel antibiotic resistance genes, *tet*(44) and *ant(6)-Ib*, are located within a transferable pathogenicity island in *Campylobacter fetus* subsp. *fetus*. Antimicrob Agents Chemother 54: 3052–3055.
43. Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet 11: 181–190.
44. Jorth P, Whiteley M (2012) An evolutionary link between natural transformation and CRISPR adaptive immunity. MBio 3.
45. Kienesberger S, Gorkiewicz G, Joainig MM, Scheicher SR, Leitner E, et al. (2007) Development of experimental genetic tools for *Campylobacter fetus*. Appl Environ Microbiol 73: 4619–4630.
46. Tu ZC, Wassenaar TM, Thompson SA, Blaser MJ (2003) Structure and genotypic plasticity of the *Campylobacter fetus* sap locus. Mol Microbiol 48: 685–698.
47. Ryan A, Lynch M, Smith SM, Amu S, Nel HJ, et al. (2011) A role for TLR4 in *Clostridium difficile* infection and the recognition of surface layer proteins. PLoS Pathog 7: e1002076.
48. Abreu MT, Vora P, Faure E, Thomas LS, Arnold ET, et al. (2001) Decreased expression of Toll-like receptor-4 and MD-2 correlates with intestinal epithelial cell protection against dysregulated proinflammatory gene expression in response to bacterial lipopolysaccharide. J Immunol 167: 1609–1616.
49. Arce RM, Caron KM, Barros SP, Offenbacher S (2012) Toll-like receptor 4 mediates intrauterine growth restriction after systemic *Campylobacter rectus* infection in mice. Mol Oral Microbiol 27: 373–381.
50. Trinchieri G, Sher A (2007) Cooperation of Toll-like receptor signals in innate immune defence. Nat Rev Immunol 7: 179–190.
51. Gonzalez JM, Xu H, Ofori E, Elovitz MA (2007) Toll-like receptors in the uterus, cervix, and placenta: is pregnancy an immunosuppressed state? Am J Obstet Gynecol 197: 296 e291–296.
52. Tu ZC, Dewhirst FE, Blaser MJ (2001) Evidence that the *Campylobacter fetus sap* locus is an ancient genomic constituent with origins before mammals and reptiles diverged. Infect Immun 69: 2237–2244.
53. Kleczka B, Lamerz AC, van Zandbergen G, Wenzel A, Gerardy-Schahn R, et al. (2007) Targeted gene deletion of *Leishmania major* UDP-galactopyranose mutase leads to attenuated virulence. J Biol Chem 282: 10498–10505.
54. Dworkin J, Tummuru MK, Blaser MJ (1995) Segmental conservation of *sapA* sequences in type B *Campylobacter fetus* cells. J Biol Chem 270: 15093–15101.
55. R_Core_Team (2012) R: A language and environment for statistical computing. In: Computing RFfS, editor. Vienna, Austria.
56. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.
57. Blaser MJ, Smith PF, Kohler PF (1985) Susceptibility of *Campylobacter* isolates to the bactericidal activity of human serum. J Infect Dis 151: 227–235.

Research in Translation

A Physicians' Wish List for the Clinical Application of Intestinal Metagenomics

Ingeborg Klymiuk[1], Christoph Högenauer[2], Bettina Halwachs[3,4], Gerhard G. Thallinger[3,4], W. Florian Fricke[5], Christoph Steininger[6]*

1 Core Facility Molecular Biology, Center for Medical Research, Medical University of Graz, Graz, Austria, 2 Department of Gastroenterology and Hepatology, Medical University of Graz, Graz, Austria, 3 Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria, 4 Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, Graz, Austria, 5 Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, United States of America, 6 Division of Infectious Diseases, Department of Medicine I, Medical University of Vienna, Vienna, Austria

Summary Points

- Multiple infectious, autoimmune, metabolic, and neoplastic diseases have been associated with changes in the intestinal microbiome, although a cause–effect relationship is often difficult to establish.
- Here we discuss the problems, applications, and visionary requirements for the integration of microbiome analysis into clinical routine diagnostics.
- Metagenomics is increasingly used for the culture-independent and largely unbiased characterization of complex bacterial habitats at high resolution. The versatility and decreasing costs of metagenomics make this technology an interesting tool for clinical diagnostics.
- Methodological shortcomings still impede the application of metagenomics in clinical diagnostics.
- Integration of metagenomics into clinical medicine requires accepted and validated strategies for (1) translation into clinical action items; (2) sample collection, preparation, and testing; and (3) data analysis and interpretation. We highlight tasks that are of high priority from a clinical perspective for the useful medical application of metagenomics.

Introduction

The intestine is one of the most diverse and complex bacterial habitats of the human body, harboring ~1,000 bacterial phylotypes [1]. Recent studies have associated the human intestinal microbiome (i.e., the collective genomes of all intestinal microbial habitants [2]) with health and disease states, suggesting that metagenomic analysis of the intestinal microbiome could be exploited as a novel diagnostic, prophylactic, or therapeutic strategy in multiple medical specialties. For example, the identification and quantification of opportunistic pathogens in the intestinal microbiome may facilitate risk stratification in immunocompromised patients, such as in critically ill, HIV-infected or immunosuppressed (e.g., organ transplant recipients or individuals with autoimmune disease) patients. Also, the correction of intestinal dysbiosis, the pathologic imbalance of the gut microbiota, may inhibit the development and/or delay the progression of autoimmune diseases [3,4], metabolic disorders [5], and cancer [6]. The propagation of a healthy intestinal microbiota has even been shown to reduce toxicity and increase effectiveness of cancer therapies in rats [7]. In addition, standard analysis of the human intestinal microbiome in patients may enable the rapid identification of novel emerging infectious pathogens in fecal specimens, for example, in the case of an outbreak of Shiga-toxigenic *Escherichia coli* [8].

Our understanding of the human intestinal microbiome in health and disease has been revolutionized by the development of next generation sequencing and its application to metagenomics, which is the term generally used to summarize culture-independent technologies that allow the characterization of a microbiome [2]. These methods allow for the largely unbiased characterization of complex microbial communities at high resolution, including the detection of novel and uncultivable bacteria, viruses, archaea, and small eukaryotic organisms, even in compartments previously considered to be sterile, such as the urinary bladder [9]. The European MetaHIT project (http://www.metahit.eu) and the US National Institutes of Health Human Microbiome Project (http://www.hmpdacc.org) have

Research in Translation discusses health interventions in the context of translation from basic to clinical research, or from clinical evidence to practice.

Citation: Klymiuk I, Högenauer C, Halwachs B, Thallinger GG, Fricke WF, et al. (2014) A Physicians' Wish List for the Clinical Application of Intestinal Metagenomics. PLoS Med 11(4): e1001627. doi:10.1371/journal.pmed.1001627

Published April 15, 2014

Funding: The Austrian Centre of Industrial Biotechnology (ACIB) contribution was supported by FFG, BMWFJ, BMVIT, ZIT GmbH, Zukunftsstiftung Tirol and Land Steiermark within the Austrian COMET program FFG Grant 824186 (BH, GGT) and the EU COST Action SeqAhead, EC Grant BM1006 (GGT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: FMT, fecal microbiota transplantation; IBD, inflammatory bowel disease.

* E-mail: christoph.steininger@meduniwien.ac.at

Provenance: Not commissioned; externally peer reviewed

set new standards for the in-depth metagenomic characterization of the healthy human microbiota (microorganisms living inside and on humans) from different body habitats [2].

Optimizing patient outcome according to metagenomic information depends on the quality of the available information, options for translation of this information into clinical action, and the effectiveness of communication. Translation of metagenomic knowledge into clinical practice is impeded by several limitations. For example, vast amounts of information are generated by metagenomics, which has to be assorted, interpreted, and communicated to clinicians in a comprehensible format. Most clinical studies have focused on characterizing the human microbiota by its taxonomic composition using 16S rRNA–based bacterial surveys, although similar biological functions may be exerted by unrelated taxa [10]. Establishing a cause–effect relationship or using microbiome profiles as surrogate markers for diseases is accordingly difficult.

Priorities for the Application of Metagenomics in Clinical Medicine

Strategies still remain to be defined for (1) translation into clinical action items with impact on patient outcome; (2) sample collection, preparation, and testing; and (3) data analysis, interpretation, and communication. Here, we highlight the tasks that are of high priority from a clinical perspective for the useful application of metagenomics in clinical medicine.

Priority 1: Integration of Metagenomic Information with Other Clinical and Laboratory Sources of Evidence for Translation into Targeted Therapy

Metagenomic information has been associated with specific disorders in several studies. For example, clinical observations have long suggested that the intestinal microbiome plays a critical role in the pathogenesis of inflammatory bowel disease (IBD) (Crohn disease and ulcerative colitis): (1) inflammation in Crohn disease disappears if the involved bowel segment is excluded from the fecal stream and recurs after re-anastomosis with reexposure to intestinal contents [11]; (2) IBD responds at least partially to antimicrobials [12] and some probiotics (live bacteria or yeast preparations) [13]; (3) some studies have shown for IBD a decreased bacterial diversity and a shift from anti-inflammatory commensals to pro-inflammatory pathogens (dysbiosis)—particularly to an overrepresentation of proteobacteria and to a reduction in *Faecalibacterium prausnitzii* and other beneficial butyrate-producing bacteria [14–16].

While current evidence strongly suggests that the pathogenesis of IBD could be linked to the intestinal microbiota, important clinical questions remain unanswered. So far, study results analyzing microbiome changes in IBD patients were not controlled for potential confounders such as mucosal inflammation per se [17,18], accelerated intestinal transit due to diarrhea [19], or medications used for IBD treatment, for example, antibiotics and immunosuppressants [20,21]. In addition, evidence from animal models still has to be confirmed in human clinical medicine, such as the anti-inflammatory properties of *F. prausnitzii* in chronic intestinal inflammation [22]. Results from clinical studies are sometimes incongruous—initial studies of patients with ulcerative colitis showed a marked benefit from fecal microbiota transplantation (FMT) [23], but other small studies could not confirm this observation [24]. Another study showed that FMT could correct the proposed features of the dysbiotic intestinal microbiota in IBD, such as the increased abundance of proteobacteria, but did not result in significant clinical improvement [24].

Hence, metagenomics approaches have to fulfill several clinical prerequisites to have a significant impact on diagnostic, prophylactic, and therapeutic strategies. A cause–effect relationship between a defined disorder and intestinal microbiome profile has to be established beyond doubt. A clear distinction between intestinal microbiome profiles of disorders (e.g., IBD versus other causes of intestinal inflammation) on the basis of metagenomic information would greatly facilitate diagnostic strategies. Identification of significant confounders of metagenomic information (inflammation, concomitant therapy, diet, etc.) may also help in devising novel prophylactic strategies. Well-directed strategies for the targeted therapy of disorders of the intestinal microbiome have to be developed, and existing ones optimized (e.g., selection of FMT donors according to a target microbiome). For this purpose, longitudinal studies with well-defined intervention and control groups as well as adequate follow-up periods are warranted. Metagenomic information on longitudinal changes in the intestinal microbiome needs to be combined with other clinical and laboratory sources of evidence for translation into targeted therapies.

Priority 2: Standardization of Diagnostic Procedures in Sample Collection, Preparation, and Testing

Accurate sample collection, preparation, and analysis are of paramount importance for the characterization of the intestinal microbiome in health and disease. Collection of stool samples; collection of gastric, intestinal, or biliary fluid; and endoscopic mucosal biopsies are routine clinical procedures. Next generation sequencing already allows characterization of the microbial composition of a sample (e.g., by 16S rRNA gene region analysis) and of its genetic and functional potential (reviewed in [25,26]).

Nevertheless, the choice of sample, sampling procedure, and analytical workflow greatly influences the results and thus the clinical utility of metagenomic characterization. Microbiota compositions fluctuate in response to dietary and sanitary habits, age, genotype, sex, ethnicity, and use of antibiotic and other medications [27–29]. Sample contamination from other anatomic regions (e.g., from oropharynx to stomach) is difficult to avoid with currently available endoscopic tools [30]. The clinically most significant anatomic locations in relation to a specific intestinal disorder still have to be defined (e.g., fecal sample versus endoscopic biopsy, or sampling of lesions versus surrounding, unaffected mucosa in IBD). Finally, differences in sample preparation, DNA isolation, metagenomic approaches, number of reads analyzed, and sequencing instrument used have a large impact on the final results [27].

Standardization of workflows in metagenomic studies is therefore urgently needed. Sampling methods have to be developed to avoid carryover contaminations. Standards must to be adapted and optimized to specific human cohorts and diseases for a meaningful interpretation of metagenomic information.

Priority 3: Automation of Data Analysis, Interpretation, and Communication

Analysis and statistical interpretation of the data in a reproducible form are also vital for the translation of metagenomics information into clinical action items [31]. Basically, sequence reads from the sampled DNA are clustered into operational taxonomic units, which are taxonomically classified and compiled into a list of relative operational taxonomic unit abundances for each sample (reviewed in [32]). Next, the whole-community composition can be statistically evaluated and categorized for clinical purposes according to function, prevalence, absence, or alternation of particular bacterial groups. These groups of interest can range from broad taxonomic classes to specific bacterial families or species, such as the two phyla Firmicutes and Bacteroidetes, whose ratio has relevance to obesity [33]; members of the phylum Proteobacteria, whose abundance has been associated with intestinal disease states such as IBD [18]; Clostridia species that induce anti-inflammatory regulatory T-cells [34]; or tumor-inducing *Fusobacterium nucleatum* [35].

Currently, the introduction of metagenomic tools into clinical practice is facing major technical as well as biological obstacles: (1) long analysis times, (2) evolving definitions of reference microbiota, (3) missing standards of analysis methods, algorithms, and databases, (4) lack of well-defined physiological ranges, and (5) missing evidence for cause–effect relationships.

From a technical perspective, a maximum level of automation would facilitate the digest of metagenomic data into clinically meaningful information. Analysis speed is highly dependent on the number of collectively analyzed samples, and the methods and tools used. Filtering and quality improvement steps may require several days, even on medium-sized computing clusters. Hence, rapid data analysis needs a reference microbiome as a reliable standard with which to compare individual samples, reduction of analysis complexity, and, ultimately, integration of analysis algorithms and desktop sequencers into a single package. Furthermore, for meaningful interpretation and communication, results of statistical evaluations should be generated and digested into clinically relevant bits automatically in the same sequencing unit, and communicated as an analysis report to the physician within a few hours. A crucial biological point is the definition of physiological ranges of gut microbiota parameters, which are highly variable between ethnic groups, geographic locations, and different diets [36]. For the definition of reference values, representative samples from the local healthy population have to be analyzed for the relative abundance of taxonomic groups or ratios between groups, combined with relevant clinical data (see the Human Microbiome Project and the American Food Project [http://humanfoodproject.com]). This information would also provide the basis for establishing cause–effect relationships. Finally, reference values have to be updated continuously and integrated into analysis algorithms for effective translation of evolving insight into intestinal microbiota into clinical practice.

Outlook

The establishment of characteristic and thoroughly validated signatures of the intestinal microbiome allows the development of new prophylactic, therapeutic, and prognostic strategies for beneficial and targeted modification of the patient's intestinal microbiome. Most metagenomic tools required for addressing these important questions are already available, standard operating tools are under development (see the Human Microbiome Project), and insight into the human microbiome is evolving rapidly (Box 1). Modern, high-resolution, and high-throughput analysis of complex bacterial communities in clinical samples has the potential to revolutionize clinical practice. As a prerequisite, target conditions must be specified, conclusively linked with characteristic signatures of the intestinal microbiome, and thoroughly validated. In addition, sample collection, preparation, testing, analysis, and result interpretation must be standardized and widely automated, and costs per sample and turnaround times significantly reduced. The integration of metagenomic analysis into clinical diagnostics will very likely open whole new avenues to the treatment of intestinal as well as extra-intestinal diseases.

Author Contributions

Conceived and designed the experiments: IK CH BH GGT WFF CS. Analyzed the data: IK CH BH GGT WFF CS. Wrote the first draft of the manuscript: IK CH BH GGT WFF CS. Contributed to the writing of the manuscript: IK CH BH GGT WFF CS. ICMJE criteria for authorship read and met: IK CH BH GGT WFF CS. Agree with manuscript results and conclusions: IK CH BH GGT WFF CS.

Box 1. Five Key Papers on the Translation of Metagenomics into Clinical Practice

1. **Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65.** This study reports a large-scale approach to characterizing the functionality of the intestinal microbiota by cataloging human gut microbial genes, which is a prerequisite for defining health and disease states in terms of the microbiome.
2. **Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214.** This project is a trendsetting approach to establishing comprehensive metagenomic datasets of (healthy) body habitats as reference datasets and to lay the foundation for the translation of metagenomic research into diagnostic applications.
3. **Kump PK, Gröchenig HP, Lackner S, Trajanoski S, Reicht G, et al. (2013) Alteration of intestinal dysbiosis by fecal microbiota transplantation does not induce remission in patients with chronic active ulcerative colitis. Inflamm Bowel Dis 19: 2155–2165.** This was one of the first attempts not only to use FMT but also to characterize the procedure and the outcome by metagenomics.
4. **Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022–1023.** This study links the metagenomics pattern of the human intestinal microbiome to a clinical disorder and is therefore of importance for therapeutic approaches.
5. **Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, et al. (2013) Advancing our understanding of the human microbiome using QIIME. Methods Enzymol 531: 371–444.** This study describes one of the common interactive analysis tools for microbiome analysis currently used by many researchers, which might be used in the future for standardizing data analysis.

References

1. Human Microbiome Project Consortium (2012) A framework for human microbiome research. Nature 486: 215–221.
2. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804–810.
3. Sellitto M, Bai G, Serena G, Fricke WF, Sturgeon C, et al. (2012) Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. PLoS ONE 7: e33387.
4. Sjoberg V, Sandstrom O, Hedberg M, Hammarstrom S, Hernell O, et al. (2013) Intestinal T-cell responses in celiac disease—impact of celiac disease associated bacteria. PLoS ONE 8: e53414.
5. Henao-Mejia J, Elinav E, Jin C, Hao L, Mehal WZ, et al. (2012) Inflammasome-mediated dysbiosis regulates progression of NAFLD and obesity. Nature 482: 179–185.
6. Sobhani I, Tap J, Roudot-Thoraval F, Roperch JP, Letulle S, et al. (2011) Microbial dysbiosis in colorectal cancer (CRC) patients. PLoS ONE 6: e16393.
7. Lin XB, Dieleman LA, Ketabi A, Bibova I, Sawyer MB, et al. (2012) Irinotecan (CPT-11) chemotherapy alters intestinal microbiota in tumour bearing rats. PLoS ONE 7: e39764.
8. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, et al. (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA 309: 1502–1510.
9. Wolfe AJ, Toh E, Shibata N, Rong R, Kenton K, et al. (2012) Evidence of uncultivated bacteria in the adult female bladder. J Clin Microbiol 50: 1376–1383.
10. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635–1638.
11. Rutgeerts P, Goboes K, Peeters M, Hiele M, Penninckx F, et al. (1991) Effect of faecal stream diversion on recurrence of Crohn's disease in the neoterminal ileum. Lancet 338: 771–774.
12. Khan KJ, Dubinsky MC, Ford AC, Ullman TA, Talley NJ, et al. (2011) Efficacy of immunosuppressive therapy for inflammatory bowel disease: a systematic review and meta-analysis. Am J Gastroenterol 106: 630–642.
13. Kruis W, Fric P, Pokrotnieks J, Lukas M, Fixa B, et al. (2004) Maintaining remission of ulcerative colitis with the probiotic Escherichia coli Nissle 1917 is as effective as with standard mesalazine. Gut 53: 1617–1623.
14. Lepage P, Hasler R, Spehlmann ME, Rehman A, Zvirbliene A, et al. (2011) Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. Gastroenterology 141: 227–236.
15. Manichanh C, Borruel N, Casellas F, Guarner F (2012) The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol 9: 599–608.
16. Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, et al. (2008) Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A 105: 16731–16736.
17. Gill N, Ferreira RB, Antunes LC, Willing BP, Sekirov I, et al. (2012) Neutrophil elastase alters the murine gut microbiota resulting in enhanced Salmonella colonization. PLoS ONE 7: e49646.
18. Lupp C, Robertson ML, Wickham ME, Sekirov I, Champion OL, et al. (2007) Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. Cell Host Microbe 2: 119–129.
19. Gorkiewicz G, Thallinger GG, Trajanoski S, Lackner S, Stocker G, et al. (2013) Alterations in the colonic microbiota in response to osmotic diarrhea. PLoS ONE 8: e55817.
20. Dethlefsen L, Relman DA (2011) Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. Proc Natl Acad Sci U S A 108 (Suppl 1): 4554–4561.
21. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, et al. (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 13: R79.
22. Carlsson AH, Yakymenko O, Olivier I, Hakansson F, Postma E, et al. (2013) Faecalibacterium prausnitzii supernatant improves intestinal barrier function in mice DSS colitis. Scand J Gastroenterol 48: 1136–1144.
23. de Vrieze J (2013) Medical research. The promise of poop. Science 341: 954–957.
24. Kump PK, Grochenig HP, Lackner S, Trajanoski S, Reicht G, et al. (2013) Alteration of intestinal dysbiosis by fecal microbiota transplantation does not induce remission in patients with chronic active ulcerative colitis. Inflamm Bowel Dis 19: 2155–2165.
25. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, et al. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10: 599–606.
26. Soergel DA, Dey N, Knight R, Brenner SE (2012) Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J 6: 1440–1444.
27. Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, et al. (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332: 970–974.
28. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, et al. (2013) Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science 341: 1241214.
29. Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, et al. (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proc Natl Acad Sci U S A 107: 7503–7508.
30. Yang I, Nell S, Suerbaum S (2013) Survival in hostile territory: the microbiota of the stomach. FEMS Microbiol Rev 37: 736–761.
31. Fricke WF, Rasko DA (2014) Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. Nat Rev Genet 15: 49–55.
32. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, et al. (2011) Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 13: 47–58.
33. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022–1023.
34. Atarashi K, Tanoue T, Oshima K, Suda W, Nagano Y, et al. (2013) Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. Nature 500: 232–236.
35. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, et al. (2013) Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe 14: 207–215.
36. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R (2012) Diversity, stability and resilience of the human gut microbiota. Nature 489: 220–230.

Fungal Diversity
DOI 10.1007/s13225-014-0291-8

Improving ITS sequence data for identification of plant pathogenic fungi

R. Henrik Nilsson · Kevin D. Hyde · Julia Pawłowska · Martin Ryberg · Leho Tedersoo · Anders Bjørnsgard Aas · Siti A. Alias · Artur Alves · Cajsa Lisa Anderson · Alexandre Antonelli · A. Elizabeth Arnold · Barbara Bahnmann · Mohammad Bahram · Johan Bengtsson-Palme · Anna Berlin · Sara Branco · Putarak Chomnunti · Asha Dissanayake · Rein Drenkhan · Hanna Friberg · Tobias Guldberg Frøslev · Bettina Halwachs · Martin Hartmann · Beatrice Henricot · Ruvishika Jayawardena · Ari Jumpponen · Håvard Kauserud · Sonja Koskela · Tomasz Kulik · Kare Liimatainen · Björn D. Lindahl · Daniel Lindner · Jian-Kui Liu · Sajeewa Maharachchikumbura · Dimuthu Manamgoda · Svante Martinsson · Maria Alice Neves · Tuula Niskanen · Stephan Nylinder · Olinto Liparini Pereira · Danilo Batista Pinho · Teresita M. Porter · Valentin Queloz · Taavi Riit · Marisol Sánchez-García · Filipe de Sousa · Emil Stefańczyk · Mariusz Tadych · Susumu Takamatsu · Qing Tian · Dhanushka Udayanga · Martin Unterseher · Zheng Wang · Saowanee Wikee · Jiye Yan · Ellen Larsson · Karl-Henrik Larsson · Urmas Kõljalg · Kessy Abarenkov

Received: 28 March 2014 / Accepted: 18 April 2014

Summary Plant pathogenic fungi are a large and diverse assemblage of eukaryotes with substantial impacts on natural ecosystems and human endeavours. These taxa often have complex and poorly understood life cycles, lack observable, discriminatory morphological characters, and may not be amenable to in vitro culturing. As a result, species identification is frequently difficult. Molecular (DNA sequence) data have emerged as crucial information for the taxonomic identification of plant pathogenic fungi, with the nuclear ribosomal internal transcribed spacer (ITS) region being the most popular marker. However, international nucleotide sequence databases are accumulating numerous sequences of compromised or low-resolution taxonomic annotations and substandard technical quality, making their use in the molecular identification of plant pathogenic fungi problematic. Here we report on a concerted effort to identify high-quality reference sequences for various plant pathogenic fungi and to re-annotate incorrectly or insufficiently annotated public ITS sequences from these fungal lineages. A third objective was to enrich the sequences with geographical and ecological metadata. The results – a total of 31,954 changes – are incorporated in and made available through the UNITE database for molecular identification of fungi (http://unite.ut.ee), including standalone FASTA files of sequence data for local BLAST searches, use in the next-generation sequencing analysis platforms QIIME and mothur, and related applications. The present initiative is just a beginning to cover the wide spectrum of plant pathogenic fungi, and we invite all researchers with pertinent expertise to join the annotation effort.

Keywords Phytopathogenic fungi · Molecular identification · ITS · Taxonomy · Annotation

Anders Bjørnsgard Aas, Siti A. Alias, Artur Alves, Cajsa Lisa Anderson, Alexandre Antonelli, A. Elizabeth Arnold, Barbara Bahnmann, Mohammad Bahram, Johan Bengtsson-Palme, Anna Berlin, Sara Branco, Putarak Chomnunti, Asha Dissanayake, Rein Drenkhan, Hanna Friberg, Tobias Guldberg Frøslev, Bettina Halwachs, Martin Hartmann, Beatrice Henricot, Ruvishika Jayawardena, Ari Jumpponen, Håvard Kauserud, Sonja Koskela, Tomasz Kulik, Kare Liimatainen, Björn D. Lindahl, Daniel Lindner, Jian-Kui Liu, Sajeewa Maharachchikumbura, Dimuthu Manamgoda, Svante Martinsson, Maria Alice Neves, Tuula Niskanen, Stephan Nylinder, Olinto Liparini Pereira, Danilo Batista Pinho, Teresita M. Porter, Valentin Queloz, Taavi Riit, Marisol Sánchez-García, Filipe de Sousa, Emil Stefańczyk, Mariusz Tadych, Susumu Takamatsu, Qing Tian, Dhanushka Udayanga, Martin Unterseher, Zheng Wang, Saowanee Wikee and Jiye Yan contributed equally to the project and are listed in alphabetical order.

Electronic supplementary material The online version of this article (doi:10.1007/s13225-014-0291-8) contains supplementary material, which is available to authorized users.

Published online: 15 May 2014

R. H. Nilsson · C. L. Anderson · A. Antonelli · S. Martinsson · F. de Sousa · E. Larsson
Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Gothenburg, Sweden

K. D. Hyde · A. Dissanayake · R. Jayawardena · J.-K. Liu · S. Maharachchikumbura · D. Manamgoda · Q. Tian · D. Udayanga
Institute of Excellence in Fungal Research, Mae Fah Luang University, Chiang Rai 57100, Thailand

K. D. Hyde · P. Chomnunti · A. Dissanayake · R. Jayawardena · J.-K. Liu · S. Maharachchikumbura · D. Manamgoda · Q. Tian · D. Udayanga · S. Wikee
School of Science, Mae Fah Luang University, Chiang Rai 57100, Thailand

J. Pawłowska
Department of Plant Systematics and Geography, Faculty of Biology, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warsaw, Poland

M. Ryberg
Department of Organismal Biology, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

L. Tedersoo · M. Bahram · T. Riit · U. Kõljalg
Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, Tartu 51005, Estonia

A. B. Aas · H. Kauserud
Microbial Evolution Research Group, University of Oslo, Blindernveien 31, 0371 Oslo, Norway

S. A. Alias
Institute of Biological Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

A. Alves
Department of Biology, CESAM, University of Aveiro, 3810-193 Aveiro, Portugal

A. E. Arnold
School of Plant Sciences, The University of Arizona, 1140 E South Campus Drive, Forbes 303, Tucson, AZ 85721, USA

B. Bahnmann
Laboratory of Environmental Microbiology, Institute of Microbiology ASCR, Vídeňská 1083, 14220 Prague 4, Czech Republic

J. Bengtsson-Palme
Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, 413 46 Göteborg, Sweden

A. Berlin · H. Friberg · B. D. Lindahl
Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Box 7026, 750 07 Uppsala, Sweden

S. Branco
University of California at Berkeley, 321 Koshland Hall University of California, Berkeley, CA 94720-3102, USA

A. Dissanayake · R. Jayawardena · J. Yan
Institute of Plant and Environment Protection, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

R. Drenkhan
Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Kreutzwaldi, 5, 51014 Tartu, Estonia

T. G. Frøslev
Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 København K, Denmark

B. Halwachs
Institute for Genomics and Bioinformatics, Graz University of Technology, 8010 Graz, Austria

B. Halwachs
Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, 8010 Graz, Austria

M. Hartmann
Forest Soils and Biogeochemistry, Swiss Federal Research Institute WSL, Zuercherstrasse 111, 8903 Birmensdorf, Switzerland

M. Hartmann
Molecular Ecology, Institute for Sustainability Sciences, Agroscope, Reckenholzstrasse 191, 8046 Zurich, Switzerland

B. Henricot
Plant Pathology, The Royal Horticultural Society, Wisley, Woking, Surrey GU23 6QB, UK

A. Jumpponen
Division of Biology, Kansas State University, Manhattan, KS 66506, USA

S. Koskela
Metapopulation Research Group, Department of Biosciences, University of Helsinki, PO Box 65, 00014 Helsinki, Finland

T. Kulik
Department of Diagnostics and Plant Pathophysiology, University of Warmia and Mazury in Olsztyn, Plac Lodzki 5, Olsztyn 10-957, Poland

K. Liimatainen · T. Niskanen
Plant Biology, Department of Biosciences, University of Helsinki, P.O. Box 65, 00014 Helsinki, Finland

D. Lindner
US Forest Service, Northern Research Station, Center for Forest Mycology Research, One Gifford Pinchot Drive, Madison, WI, USA

M. A. Neves
Departamento Botânica, PPG Biologia de Fungos, Algas e Plantas, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil

S. Nylinder
Department of Botany, Swedish Natural History Museum, Svante Arrhenius väg 7, 10405 Stockholm, Sweden

O. L. Pereira · D. B. Pinho
Departamento de Fitopatologia, Universidade Federal de Viçosa,
Viçosa, Minas Gerais 36570-900, Brazil

T. M. Porter
Department of Biology, McMaster University, Hamilton, ON L8S
4K1, Canada

V. Queloz
ETH Zürich, Institute for Integrative Biology, CHN G 68.3,
Universitätsstrasse 16, 8092 Zürich, Switzerland

M. Sánchez-García
Department of Ecology and Evolutionary Biology, University of
Tennessee, Knoxville, TN 37996-1610, USA

E. Stefańczyk
Plant Breeding and Acclimatization Institute-National Research
Institute, Młochów Research Centre, Platanowa 19, 05831 Młochów,
Poland

M. Tadych
Department of Plant Biology and Pathology, School of
Environmental and Biological Sciences, Rutgers, The State
University of New Jersey, 59 Dudley Rd., New Brunswick,
NJ 08901, USA

S. Takamatsu
Laboratory of Plant Pathology, Faculty of Bioresources, Mie
University, 1577 Kurima-Machiya, Tsu-city 514-8507,
Japan

M. Unterseher
Institute of Botany and Landscape Ecology, Ernst-Moritz-Arndt
University, Soldmannstr. 15, 17487 Greifswald,
Germany

Z. Wang
Biostatistics Department, Yale School of Public Health, New Haven,
CT 06520, USA

K.-H. Larsson
Natural History Museum, P.O. Box 1172, Blindern 0318, Oslo,
Norway

U. Kõljalg · K. Abarenkov (✉)
Natural History Museum, University of Tartu, Vanemuise 46,
Tartu 51014, Estonia
e-mail: kessy.abarenkov@ut.ee

Introduction

Plant pathogenic fungi are a large assemblage distributed across the fungal tree of life (Stajich et al. 2009). They share a nutritional strategy that adversely affects their plant hosts, sometimes in ways that have negative repercussions for human activities. Precise knowledge of the identity of the causal agent(s) of any given plant disease is the first step toward meaningful countermeasures and disease surveillance (Rossman and Palm-Hernández 2008; Kowalski and Holdenrieder 2009; Fisher et al. 2012). In addition, recent reports of emerging plant pathogens and their cross-kingdom infections to animals and immunocompromised humans accentuate the need for accurate and quick identification in potential outbreaks (Cunha et al. 2013; Gauthier and Keller 2013; Samerpitak et al. 2014). However, it is not always easy to identify plant pathogenic fungi to the species level, as they often lack discriminatory morphological characters or cultivable life stages (Kang et al. 2010; Udayanga et al. 2012). Molecular (DNA sequence) data have emerged as a key resource in the identification of plant pathogenic fungi and carry the benefit that all fungi, regardless of life stage, morphological plasticity, and degree of cultivability, can be analyzed (Shenoy et al. 2007; Sharma et al. 2013). As a result, recent years have seen substantial progress towards a comprehensive understanding of phytopathogenic fungi in terms of taxonomy, systematics, and ecology (Dean et al. 2012; Maharachchikumbura et al. 2012; Manamgoda et al. 2012; Woudenberg et al. 2013).

DNA data, however, are not a panacea for species identification. On the contrary, taxonomically and technically compromised DNA sequences are common in the international nucleotide sequence databases (Bidartondo et al. 2008; Kang et al. 2010). This makes their use as reference data for molecular species identification difficult, particularly because many users of newly generated sequence data may not be in a position to assess whether a proposed taxonomic affiliation is reliable. As a consequence, errors and mistakes propagate over time as users adopt incorrect species names and ecological properties retrieved from sequence similarity searches (Ko Ko et al. 2011; Nilsson et al. 2012). This is especially problematic for phytopathogens, where even closely related species may differ dramatically in terms of pathogenicity, host preference, and effective countermeasures (e.g., Barnes et al. 2004; Queloz et al. 2011). Although end users do have options to propose changes in the data and metadata in the public sequence databases, few users take action when they encounter compromised sequences (Pennisi 2008; Nilsson et al. 2012).

Molecular identification of fungi usually relies, at least in the first attempts, on sequencing the nuclear ribosomal internal transcribed spacer (ITS) region, the formal fungal barcode (Schoch et al. 2012). The largest database tailored for fungal ITS sequences is UNITE (http://unite.ut.ee; Abarenkov et al. 2010a). UNITE mirrors and curates the International Nucleotide Sequence Database Collaboration (INSDC:

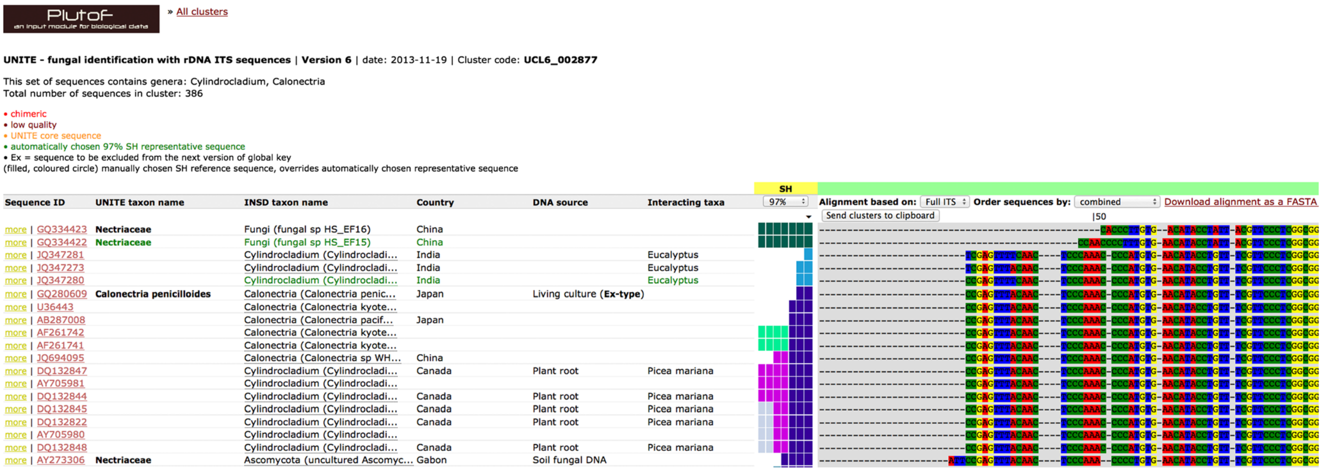
GenBank, ENA, and DDBJ; Nakamura et al. 2013) for fungal ITS sequences and offers extensive capacities for analysis and third-party annotation of sequences to its users. It has been the subject of several annotation efforts (Tedersoo et al. 2011; Bengtsson-Palme et al. 2013; Kõljalg et al. 2013), but these have in part been biased towards basidiomycetes and mycorrhizal fungi. A similar effort for plant pathogenic fungi was initiated at the symposium "Classical and molecular approaches in plant pathogen taxonomy" (10–11 September 2013, Warsaw). In addition to several of the symposium participants, other experts on various fungal lineages known to harbour plant pathogens were invited as contributors through personal networking, email, and ResearchGate (http://www.researchgate.net/). Several experts on epiphytic and endophytic fungi also participated in the effort; while these fungi may not be plant pathogenic, they are often isolated alongside, or mistaken for, plant pathogenic fungi (Unterseher et al. 2013). Moreover, many fungi showing pathogenicity in certain plants represent common endophytes in other host plants (Delaye et al. 2013). This paper reports on the outcome of the annotation effort.

Materials and methods

Using third-party sequence annotation facilities provided by the PlutoF workbench (http://plutof.ut.ee, Abarenkov et al. 2010b), the participants examined fungal lineages and ecological groups of their respective expertise in UNITE for four parameters: (i) selection of representative sequences for species, (ii) improvement of taxonomic annotations, (iii) addition of ecological metadata (chiefly host and country of collection), and (iv) compromised sequence data.

(i) Selection of representative sequences for species

UNITE clusters all public fungal ITS sequences to approximately the genus/subgenus level. A second round of clustering inside each such cluster seeks to produce molecular operational taxonomic units at approximately the species level; these are called *species hypotheses* (SHs; Fig. 1; Kõljalg et al. 2013). The species hypotheses are open for viewing and querying (http://unite.ut.ee/SearchPages.php) through uniform resource identifiers (URIs) such as "http://unite.ut.ee/sh/SH158651.06FU". As a proxy for the species hypothesis, a representative sequence is chosen automatically from the most common sequence type in the species hypothesis. Through these representative sequences, UNITE assigns a unique, stable name of the accession number type – SH158651.06FU in its shortest form for the example above – to all species hypotheses to provide a means for unambiguous reference to species-level lineages even in the absence of formal Latin names. The representative sequences are also used for non-redundant BLAST databases for molecular identification in several next-generation sequencing analysis pipelines. Depending on the algorithm, including all available fungal ITS sequences in the reference database slows down sequence similarity searches significantly, and the use of downsized, non-redundant databases with only one sequence per taxon of interest is a common solution. The representative sequences of UNITE fulfill these criteria, since they comprise a single sequence from all fungal species hypotheses recovered to date



UNITE - fungal identification with rDNA ITS sequences | **Version 6** | date: 2013-11-19 | Cluster code: **UCL6_002877**

This set of sequences contains genera: Cylindrocladium, Calonectria
Total number of sequences in cluster: 386

- chimeric
- low quality
- UNITE core sequence
- automatically chosen 97% SH representative sequence
- Ex = sequence to be excluded from the next version of global key

(filled, coloured circle) manually chosen SH reference sequence, overrides automatically chosen representative sequence

SH
97%
Alignment based on: Full ITS
Order sequences by: combined
Download alignment as a FASTA
Send clusters to clipboard
|50

Sequence ID	UNITE taxon name	INSD taxon name	Country	DNA source	Interacting taxa
more | GQ334423	**Nectriaceae**	Fungi (fungal sp HS_EF16)	China		
more | GQ334422	**Nectriaceae**	Fungi (fungal sp HS_EF15)	China		
more | JQ347281		Cylindrocladium (Cylindrocladi…	India		Eucalyptus
more | JQ347273		Cylindrocladium (Cylindrocladi…	India		Eucalyptus
more | JQ347280		Cylindrocladium (Cylindrocladi…	India		Eucalyptus
more | GQ280609	**Calonectria penicilloides**	Calonectria (Calonectria penic…	Japan	Living culture (**Ex-type**)	
more | U36443		Calonectria (Calonectria kyote…			
more | AB287008		Calonectria (Calonectria pacif…	Japan		
more | AF261742		Calonectria (Calonectria kyote…			
more | AF261741		Calonectria (Calonectria kyote…			
more | JQ694095		Calonectria (Calonectria sp WH…	China		
more | DQ132847		Cylindrocladium (Cylindrocladi…	Canada	Plant root	Picea mariana
more | AY705981		Cylindrocladium (Cylindrocladi…			
more | DQ132844		Cylindrocladium (Cylindrocladi…	Canada	Plant root	Picea mariana
more | DQ132845		Cylindrocladium (Cylindrocladi…	Canada	Plant root	Picea mariana
more | DQ132822		Cylindrocladium (Cylindrocladi…	Canada	Plant root	Picea mariana
more | AY705980		Cylindrocladium (Cylindrocladi…			
more | DQ132848		Cylindrocladium (Cylindrocladi…	Canada	Plant root	Picea mariana
more | AY273306	**Nectriaceae**	Ascomycota (uncultured Ascomyc…	Gabon	Soil fungal DNA	

Fig. 1 A screenshot from the web-based PlutoF sequence management environment showing a *Nectriaceae* cluster, with the individual species hypotheses at different similarity levels indicated by the coloured vertical bars. Country of collection and host/interacting taxa are specified together with taxonomic re-annotations. Sequences from type material are indicated. For species hypotheses where no user has designated a reference sequence, the clustering program chooses a sequence from the most common sequence type to represent that species hypothesis (*shown in green font*). The species hypotheses are mirrored by GenBank through a LinkOut function, making it possible to go from a BLAST search in GenBank to the corresponding species hypothesis in UNITE through a single click

through ITS sequences by the scientific community. However, there are situations where one would like to influence which sequence is chosen to represent a species hypothesis. In ideal cases, the type specimen or an ex-type culture has been sequenced. Such "type sequences" form the best possible proxy for the species hypothesis, as long as they are sufficiently long and of high technical quality.

To increase the proportion of plant pathology-related fungal taxa represented by sequences from types, we scanned the 27 largest journals in plant pathology (and 12 mycological journals known for an inclination towards plant pathology or fungi otherwise associated with plants) for descriptions of new (or typifications of existing) plant pathogenic or plant-associated species of fungi (Supplementary Item 1). For all descriptions where an ITS sequence was generated from the type specimen/ex-type culture by the original authors, we examined the sequence in the corresponding UNITE cluster for read quality and length. All type sequences deemed to be of high technical quality and sufficient length were designated as reference sequences for their respective species hypothesis.

(ii) Correction of taxonomic affiliations

Taxonomic misidentifications are rife in the public nucleotide sequence databases. Similarly, more than half of all public fungal ITS sequences are not annotated to the level of species, and most of these carry little or no taxonomic annotation save, e.g., "Uncultured fungus" (cf. Hibbett et al. 2011). This makes molecular identification difficult and can lead to an incorrect name or no name at all, even when full (e.g., *Colletotrichum melonis*) or partial (e.g., *Colletotrichum* sp. or Glomerellales) naming would have been possible. Clearly it is important to avoid the common mistake of over-estimating taxonomic certainty based solely on BLAST searches, which often yield many top hits with similar quality scores and can obscure sister-level relationships to the taxa represented in the top matches. BLAST results may also differ over time according to database content, and differ markedly when, e.g., the full ITS vs. partial ITS sequences or ITS sequences with non-trivial lengths of the ribosomal small and/or large subunits for the same strain are submitted to searches (U'Ren et al. 2009). Indeed, a substantive portion of misidentified sequences in public databases appear to have resulted from spurious applications of taxonomic names to sterile mycelia, environmental samples, or otherwise unknown strains, often being studied by non-taxonomists. However, careful evaluation of database matches can provide additional information about taxonomic placement that can be applied judiciously by experts to better serve the scientific community. In addition, sequences without taxonomic annotations (e.g., "Uncultured fungus") are often unfairly disregarded in phylogenetic studies (Nilsson et al. 2011). Another reason to improve the taxonomic annotation of public ITS sequences is therefore to highlight their existence and availability for use in phylogenetic and systematic studies. Such enhanced taxon sampling carries many advantages (Heath et al. 2008). We scanned our fungal lineages of expertise in UNITE to make sure the sequences carried the most accurate name possible, viz. the full species name for fully identified sequences, and the genus, family, order, class, or phylum name for sequences that could not be fully assigned.

(iii) Addition of geographical and ecological metadata

Although DNA sequences form the core of molecular identification of fungi, additional data are often needed for final, informed decisions on the taxonomic affiliation of newly generated sequences. For plant pathogenic fungi, the identity of the host and the geographical origin of the sequences are often critical information (Britton and Liebhold 2013). Yet these metadata are usually not included with sequence data in public sequence databases; Tedersoo et al. (2011) showed, for instance, that a modest 43 % of the public fungal ITS sequences were annotated with the country of origin. To the same effect, Ryberg et al. (2009) found that host of collection was reported for less than 25 % of all public fungal ITS sequences (although not all fungi necessarily have a host). We made sure that the sequences of our core expertise were as richly annotated as possible in UNITE through recursions to the original publications.

(iv) Technical quality of sequences

Detecting sequences of substandard quality in public databases is difficult because sequence chromatograms or other original data are not present for verification of nucleotide identity, and sequencing technologies have different error rates and types of errors (e.g., 454 pyrosequencing vs. Sanger sequencing). Standards also differ among researchers and computer programs with regard to quality thresholds and what is deemed acceptable for individual nucleotides or whole-sequence reads. The extent to which sequence depositors take measures to ensure that their sequence data are of satisfactory integrity also seems to differ markedly. To discriminate with full certainty among publicly deposited sequences of high and substandard quality is simply not possible in all situations (Nilsson et al. 2012). To remove all sequences that are putatively substandard is certain to lead to many instances of false-positive removals (i.e., removal of authentic albeit poorly known biodiversity), and in this study we settled for removing entries we could prove were compromised. We evaluated sequence quality on the basis of length, evidence of chimera formations or poor read quality, and mislabelling of the genetic marker that the data represent.

Results

The participants implemented a total of 31,954 changes, including 5,135 taxonomic re-annotations, 25,028 specifications of geographical and ecological metadata, 1,368 designations of reference sequences, and 401 exclusions of substandard sequences, distributed over some 48 fungal orders. The results were incorporated in UNITE for all its users. In addition, they are made publicly available through the UNITE release of all public fungal ITS sequences (http://unite.ut.ee/repository.php) for use in, e.g., local sequence similarity searches and sequence processing pipelines such as QIIME (Caporaso et al. 2010; Bates et al. 2013), mothur (Schloss et al. 2009), SCATA (http://scata.mykopat.slu.se/), CREST (Lanzén et al. 2012), and other downstream applications. UNITE also serves as one of the data providers for BLAST (Altschul et al. 1997) searches in the EUBOLD fungal barcoding database (http://www.cbs.knaw.nl/eubold/).

(i) Selection of representative sequences for species

The extraction of sequences from type material from the literature resulted in 965 designations of reference sequences (for as many species hypotheses and a total of 194 genera of fungi; Table 1). We also designated 403 additional reference sequences based on our expertise; 174 of these stemmed from type material and 229 were from other authentic material. The latter cases involved fungal taxa of our core expertise where we knew the type material was missing or too old for DNA sequencing and where we knew that the selected sequences were as close to the type as possible in terms of morphology, country, and/or substrate of collection. A total of 202 genera were designated with at least one reference sequence.

(ii) Correction of taxonomic affiliations

The process of verifying taxonomic names given to sequences resulted in a total of 5,135 changes (Table 1), notably for the orders Hypocreales (459 changes), Glomerellales (404 changes), and Botryosphaeriales (393 changes). In addition, 22 ITS sequences were found to stem from kingdoms other than Fungi and were re-annotated accordingly.

(iii) Addition of geographical and ecological metadata

Our effort to complement the sequences with metadata from the literature resulted in a total of 14,478 specifications of host and 10,550 specifications of country of origin (Table 1).

Table 1 Summary of the changes made in the UNITE database. The 15 orders that saw the largest number of changes are specified separately; all other lineages are amalgamated into the "Others" category

Order	Taxonomic re-annotations	Country	Host	Reference sequences	Count
Hypocreales	459	3,751	2,960	118 (116)	7,288
Pleosporales	129	860	4,344	76 (76)	5,409
Capnodiales	200	960	1,696	181 (181)	3,037
Diaporthales	79	1,374	855	28 (28)	2,336
Glomerellales	404	814	824	148 (148)	2,190
Botryosphaeriales	393	428	626	70 (67)	1,517
Mucorales	90	630	631	87 (63)	1,438
Eurotiales	420	411	226	168 (168)	1,225
Xylariales	90	225	823	19 (19)	1,157
Helotiales	333	301	290	108 (46)	1,032
Chaetothyriales	22	121	521	17 (17)	681
Puccinales	134	313	194	9 (1)	650
Agaricales	442	31	8	21 (21)	502
Pezizales	297	0	97	1 (1)	395
Erysiphales	143	55	66	129 (4)	393
Others	1,500	276	317	188 (183)	2,281

Taxonomic re-annotations = The number of taxonomic (re)annotations implemented. Country = The number of specifications of country of collection. A total of 94 different countries were added. Host = The number of host specifications added in the system. Reference sequences = The number of reference sequences designated through manual inspection (of which sequences from type material are indicated in parentheses). Count = Total number of changes

(iv) Technical quality of sequences

We detected a total of 363 sequences of substandard technical quality. These were marked as compromised, which precludes them from being used in molecular identification procedures while still keeping them open to direct searches in the system. This included 84 cases of chimeric sequences and 279 cases of low read quality. Another 38 sequences were annotated as ITS sequences by their submitters but were found to represent other genes and markers (notably the ribosomal small and large subunits) and were re-annotated accordingly.

Discussion

Fungal pathogens of agricultural, silvicultural, horticultural, and wild plants can compromise ecosystem health and cause considerable economic loss globally. Correct identification of these fungi and subsequent understanding of their biology and ecology are key elements in protecting their host plants (Rossman and Palm-Hernández 2008). However, identification of plant pathogenic fungi to the species level is relevant to more than just studies of plant pathology. Because of the ease and moderate cost at which large amounts of sequence data can be generated, fungi and fungal communities are now being studied by an increasing number of non-mycologists, notably soil biologists, molecular ecologists, and researchers in the medical sciences (e.g., Ghannoum et al. 2010; La Duc et al. 2012; Pautasso 2013). Phytopathogenic fungi also occur in these substrates and ecosystems in various life stages, including sterile mycelia, resting stages, and propagules. Although some plant pathogenic fungi have been studied in great detail, the biology of the majority of phytopathogenic fungi remains poorly known. Therefore, information stemming from non-mycological or non-pathological research efforts may increase our understanding of these taxa. As a consequence, it is important that all researchers, regardless of expertise and extent of mycological knowledge, can obtain reliable estimates of the taxonomic identity of plant pathogenic – and all other – fungi in whatever form they are recovered.

Molecular identification of plant pathogenic fungi can be challenging due to differing sequence and annotation quality of the available reference sequences. We have gone through a large number of plant pathogenic fungal groups within our collective expertise. A total of 31,954 changes in 48 fungal orders were implemented in UNITE for these groups (Table 1). However, not all plant pathogenic lineages of fungi – or, indeed, even the groups covered by the present effort – are satisfactorily resolved in UNITE. In addition, new sequences (of both known and unknown species) are continuously generated and deposited in the INSDC by the scientific community, such that a limited group of people can never stay abreast of the data deposition. A community effort is clearly required. UNITE offers third-party annotation capacities to all its registered users. Registration is free, and contributions from all relevant scientific communities are most welcome. Even small edits – such as designating a reference sequence for a single species hypothesis, correcting and improving a handful of taxonomic annotations, or adding metadata that can be used for comparative studies (Supplementary Item 2) – will improve the database significantly and may be of substantial importance to other researchers. Going through the alignments and metadata for one's fungi of expertise in the web-based system is furthermore a good way to visualize and explore patterns in the data and identify new research questions.

Many of the corrections brought about by the present effort would have been unnecessary if the original sequence authors had taken the time to examine and annotate their sequences properly prior to submission. Lack of time and awareness of these issues are the presumed culprits. Guidelines on how to process newly generated sequences in a way to establish their integrity and maximize their usefulness to the scientific community are given in Seifert and Rossman (2010), Nilsson et al. (2012), Hyde et al. (2013), and Robbertse et al. (2014). In addition, to facilitate future assessments of sequence quality and other pursuits, we urge sequence depositors in INSDC to archive chromatograms and other relevant data in UNITE or in other resources that support long-term data storage and availability. The present initiative will contribute to more accurate molecular identification of plant pathogenic fungi for three sets of users: UNITE users, anyone using the ~350,000-sequence downloadable FASTA file of the UNITE/INSDC fungal ITS sequences (http://unite.ut.ee/repository.php) for local BLAST searches or similar, and researchers using any of the major next-generation sequencing analysis pipelines or the EUBOLD database to process newly generated fungal ITS datasets. In addition, following the data sharing history between UNITE and the INSDC, the results were made available to the INSDC to reach the widest possible audience. Fungal barcoding is in a state of constant development, but it should be clear that collaboration and data sharing among resources are necessary for the future development of the field. Mycology struggles for funding in competition with fields that are often deemed larger or more fashionable, and we simply cannot afford public fungal DNA sequences to remain in a suboptimal state. On the contrary, we hope mycologists will work together to make fungal sequence data as richly annotated and as easily interpreted as possible because, after all, many of the end users of those data will not be mycologists. The present study is a small step in that direction, and we hope that others will follow.

Acknowledgments RHN acknowledges financial support from Swedish Research Council of Environment, Agricultural Sciences, and Spatial Planning (FORMAS, 215-2011-498). ArA acknowledges financial support from European Funds through COMPETE and by National Funds through the Portuguese Foundation for Science and Technology (FCT) within projects PTDC/AGR-FOR/3807/2012 - FCOMP-01-0124-FEDER-027979 and PEst-C/MAR/LA0017/2013. SB is supported by National Science Foundation Grant DBI 1046115. The Austrian Centre of Industrial Biotechnology (ACIB) contribution (BH) was supported by FFG, BMWFJ, BMVIT, ZIT, Zukunftsstiftung Tirol, and Land Steiermark within the Austrian COMET program FFG Grant 824186. Financial support to JP was partially provided by the Polish Ministry of Science and Higher Education (MNiSW), grant no. NN303_548839. OLP acknowledges financial support from FAPEMIG and CNPq. TMP was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050). The GenBank staff is acknowledged for helpful discussions and data sharing. The NEFOM network is acknowledged for infrastructural support. The authors have no conflict of interests to report.

References

Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Ursing BM, Vrålstad T, Liimatainen K, Peintner U, Kõljalg U (2010a) The UNITE database for molecular identification of fungi - recent updates and future perspectives. New Phytol 186:281–285

Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Prous M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K-H, Kõljalg U (2010b) PlutoF - a web-based workbench for ecological and taxonomic research, with an online implementation for fungal ITS sequences. Evol Bioinform 6:189–196

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Barnes I, Crous PW, Wingfield BD, Wingfield MJ (2004) Multigene phylogenies reveal that red band needle blight is caused by two distinct species of *Dothistroma*, *D. septosporum* and *D. pini*. Stud Mycol 50:551–565

Bates ST, Ahrendt S, Bik HM, Bruns TD, Caporaso JG, Cole J, Dwan M, Fierer N, Gu D, Houston S, Knight R, Leff J, Lewis C, Maestre JP, McDonald D, Nilsson RH, Porras-Alfaro A, Robert V, Schoch C, Scott J, Taylor DL, Wegener Parfrey L, Stajich JE (2013) Meeting report: fungal ITS workshop (October 2012). Stand Genomic Sci 8: 118–123

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger M, de Sousa F, Amend A, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson MK, Vik U, Veldre V, Nilsson RH (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol Evol 4:914–919

Bidartondo M, Bruns TD, Blackwell M et al (2008) Preserving accuracy in GenBank. Science 319:5870

Britton KO, Liebhold AM (2013) One world, many pathogens! New Phytol 197:9–10

Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336

Cunha KCD, Sutton DA, Fothergill AW, Gené GJ, Cano J, Madrid H, Hoog SD, Crous PW, Guarro J (2013) In vitro antifungal susceptibility and molecular identity of 99 clinical isolates of the opportunistic fungal genus *Curvularia*. Diagn Microbiol Infect Dis 76:168–174

Dean R, Van Kan JA, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, Rudd JJ, Dickman M, Kahmann R, Ellis J, Foster GD (2012) The top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol 13:414–430

Delaye L, García-Guzmán G, Heil M (2013) Endophytes versus biotrophic and necrotrophic pathogens – are fungal lifestyles evolutionarily stable traits? Fungal Divers 60:125–135

Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ (2012) Emerging fungal threats to animal, plant and ecosystem health. Nature 484:186–194

Gauthier G, Keller N (2013) Crossover fungal pathogens: the biology and pathogenesis of fungi capable of crossing kingdoms to infect plants and humans. Fungal Genet Biol 61:146–57

Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, Gillevet PM (2010) Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathog 6: e1000713

Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–257

Hibbett DS, Ohman A, Glotzer D, Nuhn M, Kirk P, Nilsson RH (2011) Progress in molecular and morphological taxon discovery in fungi and options for formal classification of environmental sequences. Fungal Biol Rev 25:38–47

Hyde KD, Udayanga D, Manamgoda DS, Tedersoo L, Larsson E, Abarenkov K, Bertrand YJK, Oxelman B, Hartmann M, Kauserud H, Ryberg M, Kristiansson E, Nilsson RH (2013) Incorporating molecular data in fungal systematics: a guide for aspiring researchers. Curr Res Environ Appl Mycol 3:1–32

Kang S, Mansfield MAM, Park B, Geiser DM, Ivors KL, Coffey MD, Grünwald NJ, Martin FN, Lévesque CA, Blair JE (2010) The promise and pitfalls of sequence-based identification of plant pathogenic fungi and oomycetes. Phytopathology 100:732–737

Ko Ko TWK, Stephenson SL, Bahkali AH, Hyde KD (2011) From morphology to molecular biology: can we use sequence data to identify fungal endophytes? Fungal Divers 50:113–120

Kõljalg U, Nilsson RH, Abarenkov K et al (2013) Towards a unified paradigm for sequence-based identification of Fungi. Mol Ecol 22: 5271–5277

Kowalski T, Holdenrieder O (2009) The teleomorph of *Chalara fraxinea*, the causal agent of ash dieback. For Pathol 39:304–308

La Duc MT, Vaishampayan P, Nilsson RH, Torok T, Venkateswaran K (2012) Pyrosequencing-derived bacterial, archaeal, and fungal diversity of spacecraft hardware destined for Mars. Appl Environ Microbiol 78:5912–5922

Lanzén A, Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Øvreås L, Urich T (2012) CREST – classification resources for environmental sequence tags. PLoS One 7:e49334

Maharachchikumbura SSN, Guo LD, Cai L, Chukeatirote E, Wu WP, Sun X, Crous PW, Bhat DJ, McKenzie EHC, Bahkali AH, Hyde KD (2012) A multi-locus backbone tree for *Pestalotiopsis*, with a polyphasic characterization of 14 new species. Fungal Divers 56:95–129

Manamgoda DS, Cai L, McKenzie EHC, Crous PW, Madrid H, Chukeatirote E, Shivas RG, Tan YP, Hyde KD (2012) A phylogenetic and taxonomic re-evaluation of the *Bipolaris*, *Cochliobolus*, *Curvularia* complex. Fungal Divers 56:131–144

Nakamura Y, Cochrane G, Karsch-Mizrachi I (2013) The international nucleotide sequence database collaboration. Nucleic Acids Res 41: D21–D24

Nilsson RH, Ryberg M, Sjökvist E, Abarenkov K (2011) Rethinking taxon sampling in the light of environmental sequencing. Cladistics 27:197–203

Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JAA, Bergsten J, Porter TM, Jumpponen A, Vaishampayan P, Ovaskainen O, Hallenberg N, Bengtsson-Palme J, Eriksson KM, Larsson K-H, Larsson E (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4:37–63

Pautasso M (2013) Fungal under-representation is (slowly) diminishing in the life sciences. Fungal Ecol 6:129–135

Pennisi E (2008) "Proposal to 'wikify' GenBank meets stiff resistance". Science 319:1598–1599

Queloz V, Grunig CR, Berndt R, Kowalski T, Sieber TN, Holdenrieder O (2011) Cryptic speciation in *Hymenoscyphus albidus*. For Pathol 41: 133–142

Robbertse B, Schoch CL, Robert V et al. (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for *Fungi*. Database, in press

Rossman AY, Palm-Hernández ME (2008) Systematics of plant pathogenic fungi: why it matters. Plant Dis 10:1376–1386

Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH (2009) An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. New Phytol 181:471–477

Samerpitak K, Van der Linde E, Choi HJ, Gerrits van den Ende AHG, Machouart M, Gueidan C, de Hoog GS (2014) Taxonomy of *Ochroconis*, a genus including opportunistic pathogens on humans and animals. Fungal Divers 65:89–126. doi:10.1007/s13225-013-0253-6

Schloss PD, Westcott SL, Ryabin T et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541

Schoch CL, Seifert KA, Huhndorf S et al (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. Proc Natl Acad Sci U S A 109:6241–6246

Seifert K, Rossman AY (2010) How to describe a new fungal species. IMA Fungus 1:109–116

Sharma G, Kumar N, Weir BS, Hyde KD, Shenoy BD (2013) Apmat gene can resolve *Colletotrichum* species: a case study with *Mangifera indica*. Fungal Divers 61:117–138

Shenoy BD, Rajesh J, Hyde KD (2007) Impact of DNA sequence-data on the taxonomy of anamorphic fungi. Fungal Divers 26:1–54

Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW (2009) The fungi. Curr Biol 19:R840–R845

Tedersoo L, Abarenkov K, Nilsson RH, Schussler A, Grelet G-A, Kohout P, Oja J, Bonito GM, Veldre V, Jairus T, Ryberg M, Larsson K-H, Kõljalg U (2011) Tidying up international nucleotide sequence databases: ecological, geographical, and sequence quality annotation of ITS sequences of mycorrhizal fungi. PLoS One 6:e24940

Udayanga D, Liu XX, Crous PW, McKenzie EHC, Chukeatirote E, Hyde KD (2012) A multi-locus phylogenetic evaluation of *Diaporthe* (*Phomopsis*). Fungal Divers 56:157–171

Unterseher M, Peršoh D, Schnittler M (2013) Leaf-inhabiting endophytic fungi of European Beech (*Fagus sylvatica* L.) co-occur in leaf litter but are rare on decaying wood of the same host. Fungal Divers 60: 43–54

U'Ren JM, Dalling JW, Gallery RE, Maddison DR, Davis EC, Gibson CM, Arnold EA (2009) Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analyzing fungal environmental samples. Mycol Res 113:432–449

Woudenberg JHC, Groenewald JZ, Binder M, Crous PW (2013) *Alternaria* redefined. Stud Mycol 75:171–212

High-Throughput Characterization and Comparison of Microbial Communities

3

Bettina Halwachs, Gregor Gorkiewicz,
and Gerhard G. Thallinger

Abstract

The human microbiome plays an important role in health and disease, but the structure of the bacterial communities and their interaction with the human body are still poorly understood. The recent introduction of next-generation sequencing technologies allows for the first time an unbiased and in-depth characterization of a microbiome based on the sequence of certain marker genes. However, analysis of the huge amount of sequence data generated in microbiome studies poses a considerable challenge to the individual researcher. Here we provide an overview of the steps involved in the characterization and comparison of complex microbial communities starting with sequence preprocessing on to taxonomic classification ending in statistical evaluation and visualization of the analysis results. A selection of different tools and techniques of each working step is introduced and discussed. Additionally, different sequencing approaches ahead of the bioinformatics analysis are considered. Furthermore, the application of microbiome analysis in medical research is shown by selected medical studies.

The chapter is addressed to microbial ecologists or medical researchers without or little bioinformatics background as well as to bioinformatics scientists who are interested in the overall microbiome workflow, and its tools and techniques.

3.1 Introduction

Humans inhabit an earth dominated by microorganisms. This is illustrated by the fact that the number of microorganisms exceeds the number of human beings by a factor of 10^{21} (Kyrpides 2009). Humans are not just surrounded by microorganisms, but microorganisms also live on and inside the human body. The relationship with these microbes colonizing different body habitats

G.G. Thallinger (✉)
Institute for Genomics and Bioinformatics,
Graz University of Technology,
Petersgasse 14/V, 8010 Graz, Austria

Core Facility Bioinformatics,
Austrian Centre of Industrial Biotechnology
(ACIB GmbH),
Petersgasse 14, 8010 Graz, Austria
e-mail: Gerhard.Thallinger@tugraz.at

Z. Trajanoski (ed.), *Computational Medicine*,
DOI 10.1007/978-3-7091-0947-2_3, © Springer-Verlag Wien 2012

is mostly beneficial to our health. For reasons which are still poorly understood, this mutualistic ("commensal") relationship sometimes switches into a pathogenic one (Avila et al. 2009).

Commensal bacteria occupy niches, which can then not be inhabited by pathogenic strains ("colonization resistance"). However, under certain environmental triggers (e.g., antibiotic treatment) the mutualistic balance is disturbed, commensal bacteria are depleted, and certain pathogenic taxa can proliferate and subsequently harm the body.

Metagenomics enables the culture-independent study of the whole genetic information of complex microbial communities, providing information about structure, function, and interactions of the microbial community with its habitat (Eisen 2007). Comprehensive metagenomic studies have been made possible on one hand by the recent introduction of high-throughput molecular technologies, such as cheaper and faster sequencing techniques developed by Roche (Margulies et al. 2005), Illumina (Bentley et al. 2008), or Life Technologies (McKernan et al. 2009), and on the other hand by the parallel evolving sequence analysis tools such as RDP (Cole et al. 2009), FastUniFrac (Hamady et al. 2010), SnoWMAn (Stocker et al. 2011), and mothur (Schloss et al. 2009).

These tools and technologies are able to characterize microbial communities at high resolution even in bacterium-dense environments such as the mammalian gastrointestinal tract (GI). The human GI microbiota is a focus of recent research not only because it is home to the largest microbial community within individuals but also because of its effects on the host especially the host's metabolism and immune system. Recent investigations showed a central role of the gut microbiota related to nutrition and many gastrointestinal diseases ranging from inflammations to cancer (Garrett et al. 2010).

The 16S ribosomal RNA (rRNA) gene plays a key role in the culture-independent characterization of a microbial community. According to its structure formed by alternating variable and highly conserved regions, the 16S rRNA serves as an evolutional chronometer allowing for the identification and differentiation of eubacterial and archaeal taxa (Tringe and Hugenholtz 2008).

The characterization of human microbiomes under different conditions will help to answer a variety of questions, such as how are microbial communities formed and how do they regenerate? What are the mechanisms that regulate microbial composition? Which microbes are involved in health and disease? To what extent do microbial communities differ between unrelated healthy individuals? Is there a core microbiome in a habitat shared among all humans? How does microbial composition vary over time, between environments or body habitats? How can microbial composition be manipulated in respect to medical treatment?

Since the majority of microorganisms cannot be grown in laboratory (Streit and Schmitz 2004), most of these questions would remain unsolved without the application of next generation sequencing technologies for the characterization.

3.2 Human Microbial Diversity

The microbiome is defined as the total number of microbial genomes in a defined environment (National Research Council 2007).

Microorganisms colonize our body surface as well as surfaces inside our body. The vast majority of microbes is found in the human GI tract (Gill et al. 2006). Microbial cells in the human GI tract outnumber human cells by a factor of 10 (Kyrpides 2009). Human physiology, health, and disease cannot be entirely understood by the sole analysis of human genes and their products. Also the microbial counterparts scare essential in this regard. Therefore, a metagenomic analysis of the human microbiome was initiated to unravel our so-called "second genome" (Qin et al. 2010). Humans are considered as superorganisms composed of human and microbial cells (Gill et al. 2006). To understand the mutualistic relationship between humans and their associated microbes as well as to create a framework for future research, the National Institutes of Health (NIH) funded the Human Microbiome Project (HMP, http://hmpdacc.org/, Turnbaugh et al. 2007). The aim of the HMP is to characterize microbial communities found at multiple human body sites and to look for correlations between changes in the microbiome and human health. In the beginning this project focused on the sequencing of reference genomes (Nelson et al. 2010) of human associated microbes to provide the basis for subsequent metagenomic and functional studies (Turnbaugh et al. 2007).

The *MetaHIT* project (http://www.metahit.eu/), which is funded by the European Commission, chose the GI tract for detailed investigation. The prime objective of this project is to demonstrate associations between the bacterial genes of the human GI microbiome with human health and disease. The MetaHIT program is particularly focused on inflammatory bowel disease (IBD, a chronic gut inflammation) and obesity, which both become more and more prevalent in Europe. Besides the publication of the first human gut metagenome (Qin et al. 2010) the project recently described the identification of three main microbial community types, so-called enterotypes, of the human GI tract (Arumugam et al. 2011). The enterotypes represent the 3 robust clusters built of 39 stool samples from 4 countries. Each of the three types is dominated by *Bacteroides*, *Prevotella*, or *Ruminococcus*. Based on correlation analyses of the genera in the respective enterotypes, it is evident that these enterotypes were formed due to preferred community composition. Interestingly, abundant molecular functions encoded by the metagenome do not correlate with abundant species. This finding underscores earlier reports stating that the functionality of the human GI microbiome is represented by the presence or absence of genes and gene families, and not on a taxonomic level. Thus, different microbial community structures can fulfill the same functionality (Tschop et al. 2009).

Obesity is related to a variety of comorbidities including type II diabetes and cardiovascular diseases (Ahima 2011). GI bacteria are highly proficient in the degradation of complex polysaccharides providing short-chain fatty acids, the end product of bacterial fermentation to the gut (Gill et al. 2006). About 10 % of our daily calorie intake originates from this process. By studying feces samples from lean and obese mice as well as from humans, it was shown that the composition of the GI microbiota influences the body weight (Turnbaugh et al. 2006; Ley et al. 2006). In these investigations obese humans showed a higher proportion of *Firmicutes* compared to *Bacteroidetes* in their GI tract than lean individuals. Furthermore, this proportion decreases during weight loss over a period of several months reaching levels comparable to lean individuals (Ley et al. 2006). By using germ-free mice the investigators showed that transplantation of an obesity-related microbiome leads to a significantly increased weight gain in these animals compared to transplantation of a lean-associated microbiota (Turnbaugh et al. 2006). These experiments highlight that specific manipulation of the gut microbiota is an interesting rationale to combat obesity (Bajzer and Seeley 2006).

Like the GI microbiome the vaginal microbiota is very dynamic, and individual attributes such as ethnicity, age, methods of birth control, sexual activity, personal care, and environmental conditions influence the vaginal microbiota (Wilson 2005). Although recent studies focused on the vaginal microbiome, its role in women's health and disease is still poorly understood (Ravel et al. 2010). These recent studies suggest that there is not a single core vaginal microbiome prevalent, but several microbial types, which are correlated with the ethnic background of the women (Zhou et al. 2007). Disturbance of the "normal" vaginal microbiota by the increase of opportunistic pathogens leads to a frequent disease called bacterial vaginosis (BV, Thies et al. 2007). To understand the development of BV, it is important to consider the whole vaginal community, as even less abundant taxa can be important to counteract colonization with opportunistic pathogens (Thies et al. 2007).

One of the largest human microbial habitats is represented by the skin with an area of about 1.8 m^2. But skin represents not a uniform microbial habitat, and it is divided into different niches displaying different levels of pH, moisture, temperature, and also different structures such as hair and sebaceous or apocrine glands (Gill et al. 2006; Turnbaugh and Maurice 2011; Grice and Segre 2011). Like the GI microbiota the skin microbiota is a dynamic microbial community, and disturbed microbiota structures in skin diseases were observed (Gao et al. 2007, 2010; Cogen et al. 2008). Moreover, a high level of intra- and interpersonal variations in the community structures were noted (Costello et al. 2009). A physiological skin microbiota (e.g., containing lactic acid bacteria) is a safeguard against potential harmful microbes, and this fact can be exploited by therapeutic strategies in case of skin diseases (Grice and Segre 2011).

Investigation of the oral microbiome was initiated already in 1708 by Antonie van Leeuwenhoek (Parker 1965). Hundreds of years later methods and possibilities evolved, but many questions about the human oral microbiota remained unresolved. Recent studies revealed the existence of a common oral microbial composition across unrelated healthy individuals, and also identified highly complex patterns of individual niches colonized by different communities in the oral cavity (Zaura et al. 2009; Bik et al. 2010). Although the oral cavity harbors a variety of microorganisms, only six bacterial phyla, *Firmicutes*, *Actinobacteria*, *Proteobacteria*, *Bacteroidetes*, *Fusobacteria*, and *TM7*, are considered dominant with relative abundances ranging from 1 % to 36 % (Aas et al. 2005).

3.3 The Barcode of Life

Sequencing variable regions of 16S rRNA genes (16S rDNA) is widely used to characterize complex microbial communities (Venter et al. 2004). The benefit of this genetic marker is based on the fact that it is present in all eubacteria and archaea, and it consists of conserved and variable regions. Variable regions are subjected to mutation during evolution and can therefore serve as an evolutionary clock specific to the respective taxon. Conserved regions are often important for ribosome function (i.e., translation), and mutations in these regions can rarely be propagated to the offspring. As such mutations affect the bacterial cell heavily, these regions remain unchanged (Patel 2001).

The 16S rRNA gene comprises nine variable regions which separate regions of high conservation (Neefs et al. 1993). The variable regions can be amplified using universal or group-specific primers (Lane et al. 1985). Quality of the classification results after sequencing the amplicons highly depends on the quality of the sequenced reads. Therefore, it is necessary to minimize any kind of amplification bias. Wang and Qian (2009) studied the impact of 16S rDNA primer choice onto the resulting taxonomic classification. They found that the taxon coverage varies between 80 % and 98 % depending on the used primers. Furthermore, it was also confirmed that the majority of primers is specific for a certain range of bacterial phyla and cannot be applied for amplification of all bacteria in a microbial sample (Wang and Qian 2009). Additionally, the choice of the variable region and the used sequencing technology influence the classification accuracy (Hamp et al. 2009). With the objective to assign as many reads as possible to a certain taxonomic level, the V4/V5 region of the 16S gene is recommended, as its use exhibits the highest accuracy regardless from the used sequencing technology (Claesson et al. 2010). In contrast the V3/V4 region showed the worst classification efficiency (Liu et al. 2008).

3.4 Sequencing

The methods introduced in 1977 by Sanger et al. (1977) and by Maxam and Gilbert (1992) paved the way for a new area in DNA sequencing. The automated Sanger sequencing, later considered as "*first-generation sequencing*" (Metzker 2005), was the state-of-the-art sequencing technique over the last four decades. Although this technology still has the advantage of long read lengths (>800 bp) and high accuracy (>99.999 %), it has been largely replaced by newer methods. These methods, so-called "*next-generation sequencing*" (NGS) techniques, can be further grouped into "*second-*" and "*third-generation techniques*" (Pareek et al. 2011). One major advantage of NGS over Sanger sequencing is the ability to produce enormous amount of data in a single run within a short period of time at low costs. Moreover, NGS enables sequence determination from amplified single DNA fragments without cloning (Ansorge 2009). Different kinds of NGS techniques can be distinguished either by template preparations, chemistry, detection approaches, and base calling methods (Metzker 2010). These differences result in benefits and disadvantages of each of the different techniques, which are discussed below. The group of second-generation sequencing methods comprises three different systems.

First, the *454 Genome Sequencer FLX instrument* (Margulies et al. 2005) is based on the detection of luminescence created during conversion of pyrophosphate. It comprises four major working steps: (1) ligation of adapters to DNA fragments; (2) emulsion polymerase chain reaction (PCR, amplification); (3) distribution of beads among a picotiter plate; and (4) pyrosequencing (Voelkerding et al. 2009). With the current chemistry (FLX Titanium+) read lengths of 700 bp can be achieved. A single run produces up to 900 Mb at a raw accuracy of 99.5 % (Pareek et al. 2011).

Second, the *Illumina (Solexa) Genome Analyzer* (Bentley et al. 2008) relies on sequencing by synthesis. It consists of three major working steps: (1) library preparation; (2) cluster generation; and (3) sequencing. Illumina produces short reads with a length between 36 and 150 bp. The total throughput of a run adds up to 300–600 Gb at a raw accuracy of more than 98.5 % (Pareek et al. 2011).

Third, the *Life Technologies SOLiD system* (McKernan et al. 2009) is based on the principle of sequence ligation. It comprises six major working steps: (1) library preparation; (2) emulsion PCR and bead enrichment; (3) bead deposition onto a glass slide; (4) sequencing by ligation; (5) primer reset; and (6) exaction of call chemistry. The SOLiD system produces reads with a length between 35 and 100 bp. During a single run a total throughput of up to 180 Gb can be achieved at a raw accuracy of 99.94 % (Pareek et al. 2011).

Since the taxonomic classification of 16S rDNA fragments is influenced by sequence length, the read lengths of the different sequencing technologies have to be considered. Reads produced by Illumina and SOLiD are much shorter (~100 bp) than the 454 reads (~700 bp). Former technologies provide in turn a much higher coverage per sample or allow for analysis of significantly more samples in a single run. This can help to investigate rare species of the microbial community. The short reads on the other hand can lead to misclassifications, especially among taxa with high sequence homology. Furthermore, the increased error rate of shorter reads results in a loss of taxonomic depth. In contrast, 454 runs provide a lower coverage per sample but its longer reads can be classified down to genus or even species level (Hamady and Knight 2009; Claesson et al. 2010). Short read lengths negatively influence diversity measures and taxonomic classification. Claesson et al. (2010) argue that pyrosequencing errors can be neglected because of their little influence on the taxonomic classification and diversity measures. In contrast, tools and techniques for reducing these errors ("*denoising*") become more and more state of the art in analyzing pyrosequenced data (Quince et al. 2009; Reeder and Knight 2010).

To allow sequencing of multiple samples in a single run, primers can be labeled with unique tags before PCR amplification. This so-called *barcoding* technique enables sequencing of multiple samples within a single sequencing run. The number of samples which can be sequenced in parallel is limited by the length of the used barcode. This kind of *multiplexing* decreases sequencing time as well as costs per sample. Furthermore, it overcomes sequence loss in splitting a single plate into multiple areas (Hamady et al. 2008).

The choice of which sequencing platform to use is influenced by a variety of parameters, such as reagent costs, processing time, error rates, or read lengths. The challenge is to find a platform which is able to deliver best results under a certain level of tolerance. Glenn (2011) pointed out the lack of a standard for sequencing platforms and the resulting difficulties in comparing platform specifications. All currently available platforms have their advantages and disadvantages concerning costs and error rates (Glenn 2011). Illumina shows the broadest utility at lowest cost per reads and low error rates, whereas 454 yields the highest classification accuracy in

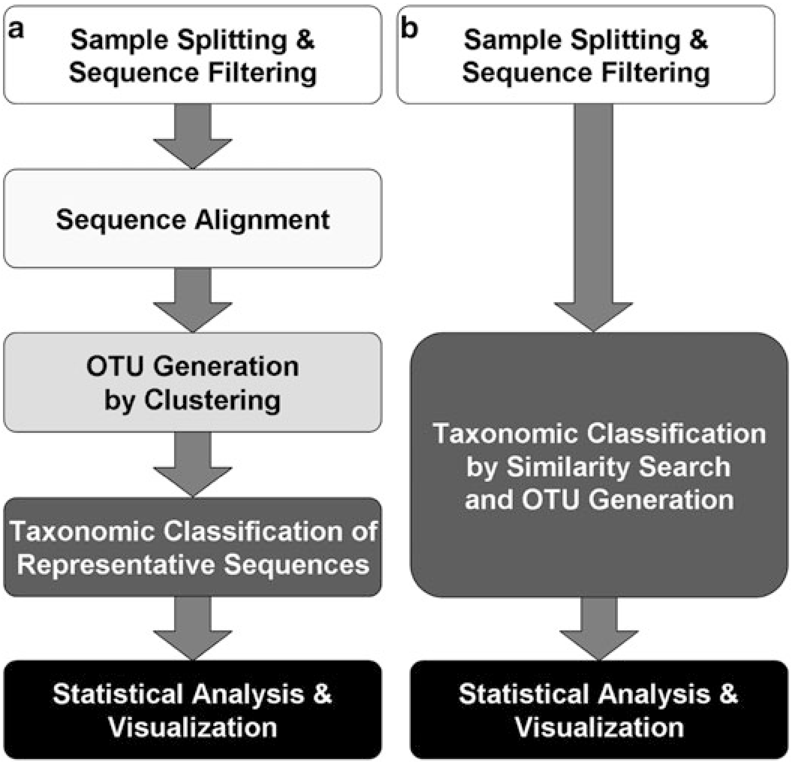

a
Sample Splitting & Sequence Filtering
Sequence Alignment
OTU Generation by Clustering
Taxonomic Classification of Representative Sequences
Statistical Analysis & Visualization
b
Sample Splitting & Sequence Filtering
Taxonomic Classification by Similarity Search and OTU Generation
Statistical Analysis & Visualization

Fig. 3.1 Microbiome analysis workflow: (**a**) working steps in case of unsupervised clustering; (**b**) comparative classification

consequence of its longer read length. However, shorter reads are more prone to classification errors.

3.5 Microbiome Analysis Workflow

The major question in microbiome projects is to figure out "what's in the mix." To characterize and classify complex microbial communities, a marker gene, in general a certain variable region of the 16S rRNA gene, is amplified from DNA, which is directly extracted from the environmental sample followed by sequencing of the amplicons. This results in thousands of sequences for a given sample which originate from hundreds of different species. To facilitate the analysis, the individual sequences are assigned to operational taxonomic units (OTUs). These OTUs represent a specific taxonomic group at a particular phylogenetic level, commonly genus or species. Each OTU consists of a taxonomic classification and an abundance, which is the number of sequence reads comprising the OTU.

A variety of tools have been developed to analyze microbiome samples. They can be divided into two main groups based on the approach to assign sequences to OTUs: OTUs can be either generated by unsupervised clustering of the sequences (Fig. 3.1a); or OTUs can be formed by comparative classification using a reference database (Fig. 3.1b) (Ghodsi et al. 2011).

OTU formation by unsupervised clustering comprises the following core working steps: (1) preprocessing (sample splitting, trimming, quality filtering, chimera removal); (2) multiple sequence alignment; (3) calculation of sequence distances; (4) clustering of sequences into OTUs; (5) dereplication (selection of a representative sequence for each OTU); (6) classification of each of the representative sequences either by similarity search against a reference database or with a classifier; and (7) statistical analysis and visualization.

OTU formation by comparative classification comprises three major working steps: (1) preprocessing (sample splitting, trimming, quality filtering, chimera removal); (2) similarity search against a reference database; and (3) statistical analysis and visualization.

Although both approaches generate OTUs, the key difference is the homogeneity of an OTU. Sequences in cluster-based OTUs have a predefined maximum distance (sequence dissimilarity), whereas the sequence distance in OTUs formed by comparative classification depends on the distance to the reference sequence.

The following sections will guide through the microbiome working steps including preprocessing of the samples, visualization, and statistical analysis of the results.

3.5.1 Preprocessing

The major goal of preprocessing sequence data is to improve the quality of the downstream analysis.

Sample splitting is included during preprocessing. In this step barcodes and primers are separated from sequences. The barcodes serve as identifiers for a particular sample in the sequencing run. At the beginning of the analysis, the user has to specify the barcodes as well as the primer sequence, so that they can be used during the preprocessing step. Barcodes are separated from the sequences either strictly by their sequence or by using different kinds of error correction methods (Hamady et al. 2008).

Filtering of the sequences based on certain criteria is widely used. The most important approach is to discard sequences depending on their *length*. Sequences markedly longer than the average tend to be chimeric, whereas very short sequences (~20 bp) lead to misalignments. Additionally sequences can be filtered using *quality scores*, the amount of *ambiguous bases* (number of Ns), *multiplicity*, or the *sequence complexity*.

At the 5′ or 3′ end of a sequence, artifacts such as poly-A/T tails or adapters, primers might have been ligated to the sequence. *Sequence trimming* to a certain length or according to a quality score can help to get rid of these artifacts (Schmieder and Edwards 2011).

Denoising combines methods and techniques for treating and eliminating different kinds of sequencing noise. Depending on the used sequencing technique, artificial sequence differences (noise) decrease sequencing quality, and thus the downstream analysis. Sequencing noise caused by pyrosequencing results for example in an overestimated number of OTUs, the so-called *OTU inflation* (Kunin et al. 2010). The major source of pyrosequencing noise is caused by uncertainties in the base calling of long homopolymer stretches (Quince et al. 2009). Additionally, PCR errors occurring during the amplification process have to be considered, since they increase the per-base sequencing error rate. Tools such as *PyroNoise* (Quince et al. 2009), *Denoiser* (Reeder and Knight 2010), or *AmpliconNoise* (Quince et al. 2011) can be applied during preprocessing to control sequencing errors and PCR single base substitutions.

Chimera removal: Chimeras, which result from a combination of two or more sequence templates amplified during PCR, have to be considered since they distort diversity truth (Quince et al. 2009). Thus, quality of the PCR has to be taken into account and parameters such as cycle number, extension time, used primers, and polymerase type have to be considered as they directly influence PCR quality (Quince et al. 2011). The impact of chimeras can be very critical in particular when they occur at high frequencies. Tools such as *Bellerophon* (Huber et al. 2004), *Ccode* (Gonzalez et al. 2005), *Pintail* (Ashelford et al. 2005), *Chimera Slayer* (Ashelford et al. 2005), *UCHIME* (Edgar et al. 2011), or *Perseus* (Quince et al. 2011) support the detection and often also the removal of chimeric sequences.

Apart from more accurate OTU estimations, denoising and chimera checking resulting in fewer sequences for downstream analysis which in turn reduces processing time. The core step of microbiome analysis is represented by the taxonomic classification of the 16S rDNA sequences. The following sections highlight a selection of tools and techniques for each of the two major approaches (OTU generation by clustering and OTU generation by comparative classification).

3.5.2 OTU Generation by Clustering

OTU generation by clustering comprises three major working steps: Before OTUs can be defined the sequences have to be aligned in order to compensate for differences in length. Subsequently, the second step is OTU generation by distance calculation followed by clustering. Finally, the classification of OTUs is performed by assigning a single representative selected from each previously created cluster to its phylogenetic group. Basic principles of this approach as well as tools and techniques are discussed in the following sections.

3.5.2.1 Sequence Alignment

Aligning sequences is a prerequisite for the subsequent OTU generation where distances (i. e., the percentage of base changes) between sequences are calculated. Since sequences have different lengths, they have to be aligned prior to distance calculation. Therefore, either multiple sequence alignments (MSAs) of all target sequences or pair-wise alignments are created.

Tools such as *Phylip* (Felsenstein 1989), *MUSCLE* (Edgar 2004), *NAST* (DeSantis et al. 2006a), or *Infernal* (Nawrocki et al. 2009) are commonly used for sequence alignments. The major difference between these tools is the amount of structure information used for identification of the putative targets (Schloss 2009; Huse et al. 2010). In contrast to traditional sequence alignment tools, Infernal builds secondary structure profiles of the 16S rDNA sequences, which are then used to create new structure based MSAs (Nawrocki et al. 2009). The secondary structure of a sequence provides powerful information for sequence alignments, because it directs the accurate alignment of conserved sequence regions. Furthermore, user-defined parameters such as gap and extension penalties do not distort the alignment. This allows a more intuitive handling of sequencing errors and overcomes problems with aligning short partial sequences.

3.5.2.2 Clustering

The clustering step generates OTUs without taking phylogenetic information into account, as sequences are grouped according to their distances (similarities) only. Clusters/OTUs are formed according to furthest, average or nearest neighbor metrics. Examples for commonly used clustering tools are *Phylip* (Felsenstein 1989), *DOTUR* (Schloss and Handelsman 2005), *quickdist* (Sogin et al. 2006), *CD-HIT* (Li and Godzik 2006), *mothur* (Schloss et al. 2009), *UCLUST* (Edgar 2010), or *DNACLUST* (Ghodsi et al. 2011).

UCLUST is based on USEARCH and allows efficient and accurate clustering of high-throughput biological sequences. USEARCH uses a heuristic, which allows fast identification of a single or a few good hits out of all possible homologous sequences. According to the clustering method, the outcome is highly influenced by the sequence order. Sequences can be either sorted by their length or according to their abundance. In the latter case sequences have to be matched according to their prefix to keep track of misalignments of short sequences (Sun et al. 2012). UCLUST was shown to be faster and to produce highly similar clusters compared to CD-HIT (Li and Godzik 2006), but in recent studies it was outperformed by DNACLUST (Ghodsi et al. 2011).

DNACLUST represents a fast and accurate clustering tool, which is tailored toward clustering highly similar 16S rRNA sequences. Clustering is based on a greedy clustering strategy, a k-mer-based filtering algorithm, and a novel sequence alignment technique, which results in significantly increased speed and accuracy compared to existing tools (Ghodsi et al. 2011). To define the cluster size threshold, a radius is used to calculate the area of the cluster. Elements within an area are therefore defined as members of a particular cluster. DNACLUST provides MSA, k-mer filtering, and clustering with a few simple commands.

3.5.2.3 Taxonomic Classification of OTUs

In contrast to the comparative classification approach (described below), sequences are first grouped into OTUs using unsupervised clustering. After the grouping, information is gained about the number of different OTUs, the abundance of an OTU, and sequences assigned to a particular OTU. For each of the OTUs, a representative sequence is selected, usually the longest one to improve classification accuracy. Subsequently, this sequence is used for taxonomic classification either by the alignment to a reference database or by classification via the RDP classifier (Wang et al. 2007). The Bayesian classifier, which uses a secondary 16S rRNA model to confer accurate alignment of sequences, is part of the Ribosomal Database Project (Cole et al. 2009) and is currently trained with 16S rRNA sequences classified according to Bergy's *Taxonomic Outline of Prokaryotes* (Gascoyne et al. 2004) for 16S rRNA bacterial and archaeal sequences.

3.5.3 OTU Generation by Comparative Classification

Compared to the previous approach, the order of the analysis steps is reverted. First, sequences are assigned to a taxon, which is then the basis for OTU generation.

The basic idea of this approach is to classify each sequence based on its similarity to known, well-annotated reference sequences. Different taxonomic classification schemes for eubacteria and archaea exist. The widely used GreenGenes database (DeSantis et al. 2006b), combines the *Pace* (Pace 1997), *Hugenholz* (Hugenholtz and Pace 1996), *Ludwig* (Amann et al. 1995), *RDP* (Cole et al. 2009), and the *NCBI taxonomy* (Sayers et al. 2011). The most similar sequence in the reference database can be determined by using local alignment search tools such as BLAST (Altschul et al. 1990) or BLAT (Kent 2002). As the latter significantly improved the accuracy of the search and also proved to be ~500 times faster than the traditional BLAST, it is commonly used in microbiome characterization and classification. The taxonomic classification of the most similar reference sequence is then assigned to the query sequence. Finally, OTUs are formed by pooling sequences with the same taxonomic classification.

An example for such an approach is *JGAST* (Hamp et al. 2009). The implementation of JGAST is based on the principles of "nearest neighbor" algorithms and can be seen as an improved *Global Alignment for Sequence Taxonomy* method (GAST, Huse et al. 2008). The query sequence is mapped to full-length sequences in an *unaligned* reference database. The classification result of the highest scoring sequence is then assigned to the query sequence (Hamp et al. 2009). Again the GreenGenes database is often used as reference database.

3.5.4 Statistical Analysis and Visualization

The measurement of microbial diversity is a key method in understanding community organization and activity. Diversity depicts the amount of taxa or lineages in a sample with a given sample size, i.e., the number of different taxa within a respective sample (Whittaker 1972). There are two major approaches for diversity measures; *α-diversity* measures the diversity within a community or an ecosystem at a certain time point whereas *β-diversity* or *species turnover* is a comparative measure of diversity between different communities or the same community over different conditions (Whittaker 1972).

3.5.4.1 α-Diversity

As a measurement of diversity within a single community or ecosystem, it plays an important role in comparison of different communities. α-diversity can be either *qualitative* or *quantitative*.

Qualitative α-diversity is also called species richness (Lozupone and Knight 2008) and refers to the number of species in a sample (Whittaker 1972). In contrast qualitative species-based α-diversity only represents presence or absence of certain taxa within a microbial community (Lozupone and Knight 2008). To define the qualitative α-diversity, the *Chao index* (Chao 1984) or the *ACE index* (Chazdon et al. 1998) is often used.

Quantitative α-diversity is also known as richness and/or as evenness. In contrast to qualitative diversity measures it also accounts for the abundance of each taxon, i.e., evenness is high if each taxon is equally abundant in a community. Quantitative α-diversity is usually represented by the *Shannon* (Shannon and Weaver 1963) or *Simpson* (Simpson 1949) indices.

α-diversity measurements can be distinguished into *species-based* and *divergence-based* measures. In species-based methods relations between different phylotypes within a sample are not considered. In contrast divergence-based methods characterize a community as more diverse, if its individuals differ greatly from each other (Lozupone and Knight 2008). Depending on whether qualitative- or quantitative measures are used, *Phylogenetic Diversity* (Faith 1992) or *Theta* (Martin 2002) can be calculated.

3.5.4.2 Rarefaction Analysis

In the context of diversity measurement of a single community also a *rarefaction* analysis can be applied. Rarefaction curves (Fig. 3.2) illustrate the number of species or OTUs observed as a function of the number of individuals (sequences) sampled (Wooley et al. 2010). Thus, rarefaction analysis reveals how many phylotypes are in a sample (richness) and how many individuals have to be sampled to reach saturation of the analysis. In microbiome analysis this relates to the sequencing depth (the number of sequences) obtained from a sample. Figure 3.2 shows three typical rarefaction curves: (1) the solid black curve demonstrates the best case: the curve shows a steep increase at the beginning and flattens with increasing number of individuals sampled. Here most or all species or OTUs have been sampled and further sampling would not increase the number of species or OTUs. (2) The dashed curve shows also the steep increase at first, but does not saturate. This means that each newly analyzed sample or an increased sequencing depth would lead to more species or more OTUs. (3) The very steep increase of the gray curve indicates a species-rich habitat. The current number of individuals sampled only covers a small fraction of the given diversity and additional sequencing is necessary to characterize the community (Wooley et al. 2010).

In addition to checking the sampling depth, a rarefaction analysis facilitates the comparison of samples with different sample size, by comparing the number of OTUs or species at a specific number of sequences in a sample. This is in general the number of sequences in the smallest sample.

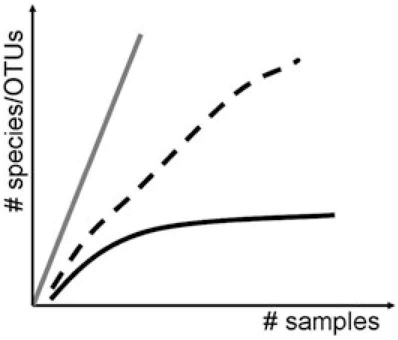


Fig. 3.2 Rarefaction curves. *Solid black*, ideal case, nearly all species/OTUs have been sampled. *Dashed black*, more sampling is needed; the habitat has not been sufficiently sampled. *Solid gray*, indicates a species rich habitat; the current number of individuals sampled only covers a small fraction of the species in the habitat

3.5.4.3 β-Diversity

β-Diversity describes the degree of variation between microbial communities according to the number of different species, and their abundance in a habitat across space and/or time or environmental condition, i.e., how many taxa or lineages are shared among samples/along a gradient (Koleff et al. 2003). Species-based approaches can be used to observe a microbial environment during different disease stages. It reveals changes in composition and diversity of a microbiome in course of a disease compared to healthy state. Additionally, species-based β-diversity measures allow evaluating whether the same environment in different ecosystems (i.e., the same body site of different individuals) share a similar or equal microbial composition (Noguez et al. 2005).

As with α-diversity, qualitative and quantitative indices of β-diversity can be discriminated. *Sörensen* (Soerensen 1948), *Bray–Curties* (Bray and Curtis 1957), and *Jaccard* (Jaccard 1901) indices are often calculated to get a qualitative measure. For quantitative diversity index calculations, the *Sörensen quantitative* index (Chao et al. 2006) or *Morisita-Horn* (Magurran 2004) measure are widely applied. Due to limitations within species-based β-diversity calculations, the divergence-based approach is preferred. The underlying principle of the divergence-based measure is that similarity/dissimilarity of the different taxa within a microbial cohort is taken into account. To calculate qualitative and quantitative divergence-based measures, *Unweighted UniFrac* (Lozupone and Knight 2005) or *Taxonomic Similarity* (Izsak and Price 2001), and *Weighted UniFrac* (Lozupone et al. 2007), F_{ST} (Martin 2002), or *DPCoA* (Pavoine et al. 2004) are used respectively (Lozupone and Knight 2008).

In measuring α- as well as β-diversity, divergence-based methods are more accepted than species-based techniques. In addition, divergence-based methods can resolve the phylogenetic membership of a given OTU even when exact matches to reference sequences are not available. Furthermore, these dramatic differences often directly correlate with phenotypic similarities, which represent fundamental features (Lozupone and Knight 2008).

3.5.4.4 Visualization

Rapid interpretation of the results can be facilitated by different types of diagrams visualizing a single sample or multiple samples.

Simple *barcharts* (Fig. 3.3a), *piecharts*, or *line plots* can be used for visualizing sequence distribution and composition of a sample at a particular taxonomic rank. They enable easy and fast comparison of differences in microbial composition and in abundance between different samples. Furthermore, line plots are a powerful tool to illustrate changes in microbial composition over time.

Heatmaps (Fig. 3.4a) are the best way to illustrate two-dimensional data. The degree of correlation of each *x*-value to its corresponding *y*-value is represented by a certain color. Heatmaps are often used for graphical representation of β-diversity measures. In this particular case the heatmap results in an upper triangular matrix with the same categories on *x*- and *y*-axis. As a consequence of the cross-correlation, the leading diagonal represents identity. Typically, heatmaps are used to visualize differences in microbial community compositions between healthy and diseased states or between states in the course of a disease. Furthermore, heatmaps are suitable to illustrate the relative abundance of each OTU between different samples.

Rank abundance plots (Fig. 3.3b) illustrate the species abundance of a certain habitat. Naturally occurring microbial communities are typically composed of a small number of high abundant phylotypes representing the majority of cells in a community and a vast amount of low abundant or rare phylotypes. This so-called long-tailed distribution of phylotypes together with incomplete sampling leads to an insufficient detection of rare taxa. It has to be noted that rare taxa could play major roles in the ecology of the microbial community; for instance they could serve as a "seed bank" for species whose numbers increase under certain conditions that favor their growth and may therefore be important for community function (Lennon and Jones 2011). Abundance ranks of the OTUs are plotted on the *x*-axis, starting with the highest rank of 1. The *y*-axis represents the logarithm of the species abundance. This kind of graphical representation allows visualizing richness and evenness of microbial communities. Richness is simply represented by the number of ranked species. Evenness can be determined according to the trend of the rank abundance curve. Low evenness is indicated by a steep gradient, since high ranking species are more abundant than low ranked species. In contrast, a flat slope means high evenness, because all ranked species are equally abundant (Magurran 2004).

Principal component analysis (PCA) scatter plots are used to visualize groupings within the data according to two principal components (two-dimensional PCA, Jolliffe 2002). Through a microbial community analysis the composition of individual microbiomes can be visually compared to PCA scatter plots. In this particular case abundance values of taxonomic groups are used for the PCA.

Venn diagrams (Fig. 3.4b, Venn 1880) represent logical relations of different cohorts as overlapping circles. These circles contain all species of a particular microbial community. Overlapping areas of different circles represent

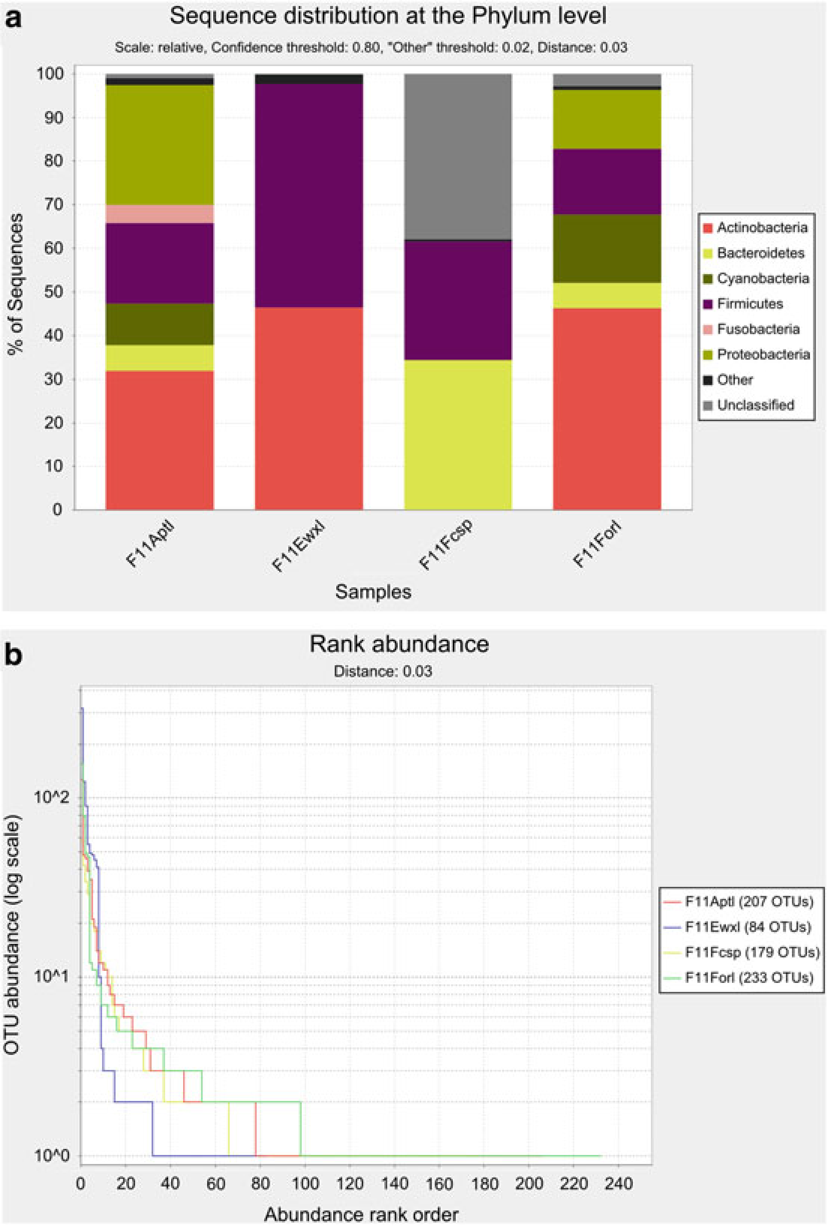


Fig. 3.3 Visualization of α-diversity: Data provided by Costello et al. (2009) have been analyzed using SnoW-MAn's integrated RDP pipeline. Samples from four different body sites (*F11Aptl: left armpit; F11Ewxl: earwax; F11Fcsp: stool; F11Forl: left forearm*) of one female individual have been selected for visualization. (**a**) The relative sequence distribution at phylum level at a classification confidence threshold of 80 % is illustrated as a barchart. Each bar comprises all phyla of particular sample colored individually. The height of a phylum relates to the relative abundance of all OTUs assigned to that specific phylum. The microbial composition of the armpit and the forearm is similar with regard to prevalent phylotypes. Moreover, the earwax sample can be treated as a skin sample and shows two prevalent phyla which are also present in forearm and armpit. As expected, the microbial compositions of the armpit and the stool have not very much in common. (**b**) The rank abundance plot illustrates that species richness is very similar for three of the four samples and significantly lower for the ear wax sample. The steep gradient of the slope in the rank abundance plot indicates low evenness for all four samples

shared species among microbial communities. This is a simple way of comparing the composition of different microbial communities.

Cytoscape networks (Shannon et al. 2003) allow visualizing species co-occurrence networks of different microbial communities (Qin et al. 2010; Arumugam et al. 2011). This kind of graphical illustration enables a clear visualization of similar community structures among a variety of habitats or ecosystems.

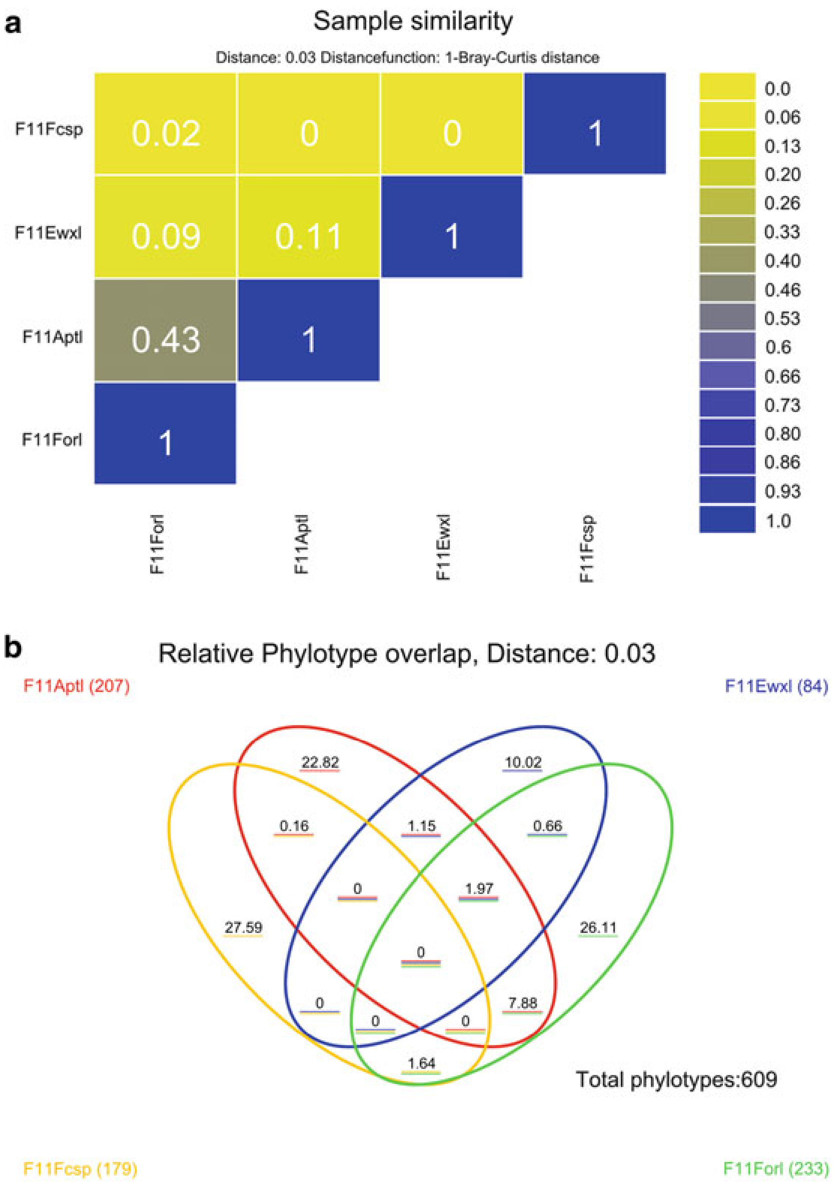


Fig. 3.4 Visualization of β-diversity: Data provided by Costello et al. (2009) have been analyzed using SnoW-MAn's integrated RDP pipeline. Samples from four different body sites (*F11Aptl: left armpit; F11Ewxl: earwax; F11Fcsp: stool; F11Forl: left forearm*) of one female individual have been selected for statistical visualization. (**a**) A heatmap illustrates sample similarity based on the Bray–Curtis distance (Bray and Curtis 1957). For easier interpretability the similarity (1-distance) is shown. The microbial composition of the armpit and the forearm samples are quite similar. In contrast, the stool and the armpit or the stool and the earwax microbiomes are very different, showing a similarity of zero. (**b**) A Venn diagram shows the relative phylotype (OTU) overlap between the four samples. From a total number of 609 distinct phylotypes in the samples, the vast majority (86.54 %) is unique to a specific sample. Moreover, no OTU is shared among all samples, caused by the distinctive composition of the stool sample

3.6 Web-Based Pipelines for Microbiome Sequence Analysis

In the following sections, three selected web-based analysis pipelines are described. These pipelines simplify microbiome data analysis considerably and cover the analysis steps aforementioned to a certain extent. As web-based analysis pipelines do not require any installation on the user's computer, they can be readily used. Furthermore, intuitive web interfaces allow analysis from any computer with an Internet connection and without detailed knowledge about underlying programming techniques and methods. Users can so start immediately with data analysis. Some of the web-based analysis pipelines also allow data storage and organization. In contrast, limitations to the maximum amount of sequence data, account space, or reproducibility have to be considered. Data analysis consists in general of four steps: (1) upload of sequence data to the web platform; (2) selection of analysis parameters and initiation of the analysis; (3) visualization of the results; and (4) download of the results for further analysis.

3.6.1 RDP Pyrosequencing Pipeline

The *RDP Pyrosequencing Pipeline* (http://pyro.cme.msu.edu/, Cole et al. 2009) provides a collection of tools for the analysis of 16S pyrosequence data. The pipeline is organized in three tiers: The first tier comprises tools for the initial processing like trimming, sorting, or quality filtering. The second tier, the so-called core tools, includes the calculation logic such as alignment, clustering, and dereplication as well as the classification of OTUs with the RDP classifier. At the top tier specialized tools for rarefaction analysis, library comparison, ecological metrics, and data export utilities for multiple output formats are combined. Each step in the analysis workflow has to be addressed, configured, and executed separately, and requires the download of intermediate results as well as their upload for the next analysis step. E-mail notifications inform the user when a job is completed. Analysis results can then be used for further processing within the pipeline as well as exported as common file formats for further analysis with statistical and ecological packages like *EstimateS* (Colwell 1997), R (R Core Team 2012), or *Spade* (Chao A and Shen T-J, 2010).

The RDP Pyrosequencing pipeline can analyze studies with up to 350,000 raw sequences, but the input to the RDP classifier is limited to 100,000 sequences.

3.6.2 SnoWMAn

SnoWMAn, the Straightforward Novel Webinterface for Microbiome Analysis (http://SnoWMAn.genome.tugraz.at, Stocker et al. 2011), covers the entire microbiome analysis workflow from sequence preprocessing to the visualization of the results. A typical microbial community analysis with SnoWMAn comprises three simple steps: first, the sequence and metadata are uploaded to a data repository. Second, the user can chose between five currently available analysis pipelines and define the respective parameters. Finally, the user can perform statistical analysis and visualization on the results.

An intuitive and user-friendly web interface guides the user through the analysis. Data can be uploaded into the repository as a compressed archive or as single files. Files containing sequence data need to be submitted in FASTA format and can be accompanied by their respective quality files. Metadata files are plain text files and comprise primer- and sample description files. The sample description file keeps information about sample barcodes, sample names, and sample grouping. The latter information is important for subsequent statistical analysis and visualization. Data files are organized in the repository of the user allowing the analysis of a data set with multiple pipelines and parameter settings. Additionally, data files and analysis results can be shared with other SnoWMAn users working on the same study.

Currently, five different pipelines are supported: *BLAT* (Kent 2002) and *JGAST* (Hamp et al. 2009) can be chosen for OTU generation by comparative classification. *mothur* (Schloss et al. 2009), *RDP* (Cole et al. 2009), and *UCLUST* (Edgar 2010) are available for OTU formation by clustering. According to the chosen analysis pipeline a set of preprocessing or pipeline parameters are available. For example, the user can define the reference databases used for comparative classification or alignment. This gives the user control over the database used and allows for the reproduction of analysis results at a later time.

Based on the amount of sequences in the data set and on the selected pipeline, the calculation time varies considerably. Current analysis status and time estimation are available via the web interface. If an e-mail address was provided, the user is notified when the analysis has been completed.

For statistical analysis and visualization, various possibilities are offered depending on the selected samples. α-Diversity and β-diversity measures or rarefaction curves can be calculated for samples. Comparison of individual samples is offered by PCA. Additionally, different chart types (i.e., barchart, piechart, line plot) can be chosen to illustrate the number of sequences in the samples, the taxonomic composition of samples, or the rank abundance relationship of a given sample. OTU overlap of different samples can be easily compared using integrated Venn diagrams.

Analysis results are summarized and illustrated in user-friendly tables. Furthermore, results of distance calculation, clustering, and taxonomic classification can be exported for further statistical analysis. All generated graphical illustrations can be downloaded in either PNG or SVG format or as an Excel sheet containing the data used to generate the chart.

SnoWMAn imposes no restrictions on the number of sequences or number of samples which can be analyzed with a single run.

3.6.3 FastUniFrac

FastUniFrac (http://bmf2.colorado.edu/fastunifrac/, Hamady et al. 2010) can be assigned neither to the category of comparative classification nor to the unsupervised clustering techniques within the analysis of complex microbial communities. FastUniFrac is the web-based version of UniFrac (bmf.colorado.edu/unifrac/, Lozupone and Knight 2005) and represents a phylogenetic method for computing differences between microbial communities. The main principle is the measurement of the pair-wise distances between communities based on the lineages these communities contain. These distances are used to build a phylogenetic tree containing all taxa found either in one or in both communities. Branches of the tree are either shared or unshared, depending whether on the taxa it holds belong to one or both communities. Consequently, two similar communities would share much of the branch length. In contrast, distinct communities would be represented by a highly branched tree which contains barely any shared branches (Lozupone and Knight 2005; Lozupone et al. 2006).

FastUniFrac allows investigating the microbial community composition. In particular, samples, which have been added to the phylogenetic tree, differ significantly in microbial composition. Additionally, the impact of environmental factors can be determined as well as if the sample size was sufficient for reliable investigation. Finally, clear and easy graphical illustration of differences between samples is provided by FastUniFrac.

However, data analysis with the FastUniFrac web version is limited to 50,000 sequences and 100 samples.

3.7 Command Line-Based Pipelines for Microbiome Sequence Analysis

In contrast to web-based pipelines, command line-based pipelines do not offer a graphical user interface and have to be run from a command shell. They often require complex installation and are therefore not available for users without a bioinformatics background. Hardware requirements are quite demanding, especially for large studies. Additionally, reference databases have to be downloaded and updated regularly and stored within the local network to be available for the analysis.

Nevertheless, command line tools have several advantages. They can be integrated into individualized analysis workflows and sequence data does not have to be transferred to external servers, as well as the analysis results are directly available in the local network.

In the next sections two commonly used command line-based analysis tools (mothur, Schloss et al. 2009; QIIME, Caporaso et al. 2010) are introduced and discussed.

3.7.1 mothur

mothur (http://www.mothur.org/, Schloss et al. 2009) was designed as a platform for microbial ecologists to support their needs to analyze 16S rRNA gene sequences. The platform combines preprocessing methods, alignment tools, pairwise distance calculation, clustering sequences into OTUs, and analysis strategies for distance matrices like α- and β-diversity measures as well as rarefaction. Moreover, visualization plots such as Venn diagrams, heatmaps, and dendograms can be created. The included techniques and algorithms have been mostly modified and extended to overcome limitations including number of sequences allowed or calculation time.

mothur is a powerful, free, open source, and platform-independent command line tool. Due to its large development and user community, existing features are continuously improved as well as new tools are integrated into the platform (Schloss et al. 2009).

3.7.2 QIIME

QIIME (pronounced "chime", http://qiime.sourceforge.net/, Caporaso et al. 2010) is a pipeline designed for the analysis of high-throughput microbial community sequence data. It combines many third party tools such as options for library demultiplexing and quality filtering as well as techniques for denoising. Different clustering tools can be selected for grouping sequences to OTUs. Tools including MUSCLE or Infernal are provided for sequence alignment. Chart types such as piecharts and histograms can be selected for visualization of the sample composition. Additionally, rarefaction and diversity measures can be calculated using different metrics and they can also be graphically illustrated.

QIIME is a free, open source analysis pipeline, which can be used either locally or in the "Cloud" as part of the CloVR Cloud Computing Research Project (http://clovr.org/, Angiuoli et al. 2011).

Conclusion

In this chapter we reviewed bioinformatics tools and techniques which are commonly used for characterization and classification of complex microbial communities. Furthermore, the entire workflow of a microbiome analysis was introduced and challenges of each step were discussed. Although the focus was on the analysis possibilities, their tools and techniques as well as practical examples of complex microbial communities of the human body were shown.

We conclude that the rapid progress in sequencing technologies and the continuous increasing amount of sequences they produce pose a challenge to bioinformatics analysis tools to keep up with these fast developments.

References

Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE (2005) Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 43(11):5721–5732

Ahima RS (2011) Digging deeper into obesity. J Clin Invest 121(6):2076–2079

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59 (1):143–169

Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, Arze C, White JR, White O, Fricke WF (2011) CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 12:356

Ansorge WJ (2009) Next-generation DNA sequencing techniques. N Biotechnol 25(4):195–203

Arumugam M, Raes J, Pelletier E, Le P, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Dore J, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Foerstner KU, Friss C, van de GM, Guedon E, Haimet F, Huber W, Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le R, Maguin E, Merieux A, Melo M, M'rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P (2011) Enterotypes of the human gut microbiome. Nature 473(7346):174–180

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol 71(12):7724–7736

Avila M, Ojcius DM, Yilmaz O (2009) The oral microbiota: living with a permanent guest. DNA Cell Biol 28(8):405–411

Bajzer M, Seeley RJ (2006) Physiology: obesity and gut flora. Nature 444(7122):1009–1010

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira CR, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E, Catenazzi CS, Neil CR, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott FW, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling NB, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris PD, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva RA, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna SJ, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456 (7218):53–59

Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF, Nelson KE, Gill SR, Fraser-Liggett CM, Relman DA (2010) Bacterial diversity in the oral cavity of 10 healthy individuals. ISME J 4(8):962–974

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr 27(4):325–349

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336

Chao A (1984) Nonparametric estimation of the number of classes in a population. Scand J Stat 11(1):265–270

Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62(2):361–371

Chao A, Shen T-J (2010) Program SPADE (Species Prediction and Diversity Estimation). Program and User's Guide. http://chao.stat.nthu.edu.tw/

Chazdon RL, Colwell RK, Denslow JS, Guariguata MR (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of Northeastern Costa Rica. In: Dallmeier FCJA (ed) Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies. Parthenon Publishing, France, pp 285–309

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16 S rRNA gene regions. Nucleic Acids Res 38(22):e200

Cogen AL, Nizet V, Gallo RL (2008) Skin microbiota: a source of disease or defence? Br J Dermatol 158(3):442–455

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37(Database issue):D141–D145

Colwell RK (1997) EstimateS: Statistical estimation of species richness and shared species from samples. Version 5. User's Guide and application. http://viceroy.eeb.uconn.edu/estimates

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R (2009) Bacterial community variation in human body habitats across space and time. Science 326(5960):1694–1697

DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL (2006a) NAST:

a multiple sequence alignment server for comparative analysis of 16 S rRNA genes. Nucleic Acids Res 34 (Web server issue):W394–W399

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006b) Greengenes, a chimera-checked 16 S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72(7):5069–5072

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27(16):2194–2200

Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol 5(3):e82

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61(1):1–10

Felsenstein J (1989) PHYLIP-phylogeny inference package (version 3.2). Cladistics 5(2):164–166

Gao Z, Tseng CH, Pei Z, Blaser MJ (2007) Molecular analysis of human forearm superficial skin bacterial biota. Proc Natl Acad Sci USA 104(8):2927–2932

Gao Z, Perez-Perez GI, Chen Y, Blaser MJ (2010) Quantitation of major human cutaneous bacterial and fungal populations. J Clin Microbiol 48(10):3575–3581

Garrett WS, Gordon JI, Glimcher LH (2010) Homeostasis and inflammation in the intestine. Cell 140(6):859–870

Gascoyne R, Bell JA, Lilburn TG (2004) Taxonomic outline of prokaryotes, 2nd edn. Bergey's manual of systematic bacteriology, vol 5. Springer, NewYork

Ghodsi M, Liu B, Pop M (2011) DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics 12:271

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. Science 312 (5778):1355–1359

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11(5):759–769

Gonzalez JM, Zimmermann J, Saiz-Jimenez C (2005) Evaluating putative chimeric sequences from PCR-amplified products. Bioinformatics 21(3):333–337

Grice EA, Segre JA (2011) The skin microbiome. Nat Rev Microbiol 9:244–253

Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res 19(7):1141–1152

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods 5(3):235–237

Hamady M, Lozupone C, Knight R (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J 4(1):17–27

Hamp TJ, Jones WJ, Fodor AA (2009) Effects of experimental choices and analysis noise on surveys of the "rare biosphere". Appl Environ Microbiol 75 (10):3263–3270

Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20 (14):2317–2319

Hugenholtz P, Pace NR (1996) Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol 14 (6):190–197

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 4(11):e1000255

Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol 12 (7):1889–1898

Izsak J, Price A (2001) Measuring beta-diversity using a taxonomic similarity index, and its relation to spatial scale. Mar Ecol Prog Ser 215:69–77

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37:547–579

Jolliffe I (2002) Principal component analysis. Springer, New York

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12(4):656–664

Koleff P, Gaston KJ, Lennon Jack J (2003) Measuring beta diversity for presence–absence data. J Anim Ecol 72(5):367–382

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12(1):118–123

Kyrpides NC (2009) Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. Nat Biotechnol 27(7):627–632

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR (1985) Rapid determination of 16 S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA 82(20):6955–6959

Lennon JT, Jones SE (2011) Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol 9(2):119–130

Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444(7122):1022–1023

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659

Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16 S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res 36(18):e120

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71(12):8228–8235

Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev 32(4):557–578

Lozupone C, Hamady M, Knight R (2006) UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 7(7):371

Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 73(5):1576–1585

Magurran AE (2004) Measuring biological diversity. Blackwell, Oxford

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380

Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl Environ Microbiol 68(8):3673–3682

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74(2): 560–564

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De L, V, Blanchard AP (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19(9):1527

Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Res 15(12):1767–1776

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46

National Research Council (2007) The new science of metagenomics: revealing the secrets of our microbial planet. The National Academies Press, Washington, DC, USA

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25 (10):1335–1337

Neefs JM, Van de Peer Y, De Rijk P, Chapelle S, De Wachter R (1993) Compilation of small ribosomal subunit RNA structures. Nucleic Acids Res 21 (13):3025–3049

Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, Young S, Warren WC, Badger J, Crabtree J, Markowitz VM, Orvis J, Cree A, Ferriera S, Fulton LL, Fulton RS, Gillis M, Hemphill LD, Joshi V, Kovar C, Torralba M, Wetterstrand KA, Abouellleil A, Wollam AM, Buhay CJ, Ding Y, Dugan S, FitzGerald MG, Holder M, Hostetler J, Clifton SW, Allen-Vercoe E, Earl AM, Farmer CN, Liolios K, Surette MG, Xu Q, Pohl C, Wilczek-Boney K, Zhu D (2010) A catalog of reference genomes from the human microbiome. Science 328(5981):994–999

Noguez AM, Arita HT, Escalante AE, Forney LJ, Garcia-Oliva F, Souza V (2005) Microbial macroecology: highly structured prokaryotic soil assemblages in a tropical deciduous forest. Glob Ecol Biogeogr 23:241–248

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276(5313):734–740

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52(4):413–435

Parker V (1965) Antony van Leeuwenhoek. Bull Med Libr Assoc 53:442–447

Patel JB (2001) 16 S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. Mol Diagn 6(4):313–321

Pavoine S, Dufour AB, Chessel D (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. J Theor Biol 228(4):523–537

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le P, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Artiguenave F, Blottiere H, Borruel N, Bruls T, Casellas F, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Forte M, Friss C, van de GM, Guedon E, Haimet F, Jamet A, Juste C, Kaci G, Kleerebezem M, Knol J, Kristensen M, Layec S, Le R, Leclerc M, Maguin E, Melo M, Oozeer R, Rescigno M, Sanchez N, Tims S, Torrejon T, Varela E, de Vos W, Winogradsky Y, Zoetendal E, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6(9):639–641

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12:38

R Core Team (2012) R: A Language and Environment for Statistical Computing. http://www.R-project.org/

Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ (2010) Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci USA 108 (Suppl 1):4680–4687

Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods 7(9):668–669

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(12):5463–5467

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J (2011) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39(Database issue): D38–D51

Schloss PD (2009) A high-throughput DNA sequence aligner for microbial ecology studies. PLoS One 4 (12):e8230

Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol 71(3):1501–1506

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van H, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75(23):7537–7541

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

Shannon CE, Weaver W (1963) The mathematical theory of communication. University of Illinois Press, Urbana

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504

Simpson EH (1949) Measurement of diversity. Nature 163:688

Soerensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr 5(4):1–34

Sogin ML, Morrison HG, Huber JA, Mark WD, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103 (32):12115–12120

Stocker G, Snajder R, Rainer J, Trajanoski S, Gorkiewicz G, Trajanoski Z, Thallinger GG (2011) SnoWMAn: high-throughput phylotyping, analysis and comparison of microbial communities (submitted for publication)

Streit WR, Schmitz RA (2004) Metagenomics—the key to the uncultured microbes. Curr Opin Microbiol 7:492–498

Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, Mai V (2012) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Brief Bioinformatics 13(1):107–21

Thies FL, Konig W, Konig B (2007) Rapid characterization of the normal and disturbed vaginal microbiota by application of 16S rRNA gene terminal RFLP fingerprinting. J Med Microbiol 56(Pt 6): 755–761

Tringe SG, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol 11(5):442–446

Tschop MH, Hugenholtz P, Karp CL (2009) Getting to the core of the gut microbiome. Nat Biotechnol 27(4):344–346

Turnbaugh PJ, Maurice CF (2011) The human microbiome: exploring and manipulating our microbial selves. In: Marco D (ed) Metagenomics: current innovations and future trends. Caister Academic, Norfolk, pp 179–210

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444(7122):1027–1031

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The human microbiome project. Nature 449(7164):804–810

Venn J (1880) On the diagrammatic and mechanical representation of propositions and reasonings. Dublin Philos Magn J Sci 10(59):1–18

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304(5667):66–74

Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55(4):641–658

Wang Y, Qian PY (2009) Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One 4(10):e7401

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73(16):5261–5267

Whittaker RH (1972) Evolution and measurement of species diversity. Taxon 21(2/3):213–251

Wilson M (2005) Microbial inhabitants of humans. Cambridge University Press, Cambridge

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6(2):e1000667

Zaura E, Keijser BJ, Huse SM, Crielaard W (2009) Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol 9:259

Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P, Foster JA, Forney LJ (2007) Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. ISME J 1 (2):121–133