Microbiome Analysis and Comparison

Documentation
Proprietary Notice

The information disclosed in this document, including all designs and related materials with SnoWMAn is the property of Graz University of Technology and/or its licensors. All credit for provided expert information lies with their respective owners.
To allow for design and specification improvements, the information in this document is subject to change at any time, without notice.
In no event shall the author be liable under any theory of contract, tort, strict liability, or other legal or equitable theory for any indirect, consequential or incidental damages with the use of SnoWMAn or this documentation.

Credits

The SnoWMAn web-application was developed by Rene Snajder, Dr. Bettina Halwachs, DI Johann Höftberger, Lornez Lilg, Dr. Gernot Stocker and Dr. Gerhard Thallinger.
This manual was written by Julian Krappinger.
1 The human microbiome

1.1 About microbiome and their relevance

Planet Earth and the human body houses a large number of microorganisms. Microorganisms exceed the number of human population by a factor of $10^{21}$ \cite{1}. Microorganisms not only surround the human body but also live and coexist with it. Microbiota are symbiotic and also pathogenic microorganisms that share our body space (The microbiome is defined as the total number of microbial genomes in a defined environment). The human microbiome plays an important role in health and disease, but structure and interaction of these bacterial communities is still poorly understood.

The human body contains 10 times more microbial cells than human cells \cite{2}. These bacteria do not cause disease (are classified as non-pathogenic) except they grow abnormally; they exist in harmony and symbiotically with their hosts \cite{3,4,5,6}. Evolutionary, host and microbiota might be seen as a unit.

![Figure 1: Human microbiota (adapted from [7])](image)

This symbiotic relationship sometimes changes to a pathogenic one. The intestinal microflora protects itself against other, more harmful microorganisms. Commensal bacteria occupy niches, which can not be inhabited by pathogenic strains. This mechanism is called "colonization resistance \cite{7,8}. Through certain environmental triggers (e.g. antibiotic treatment \cite{9}) the balance is disturbed. Commensal bacteria are depleted and allow the settlement of harmful pathogenic taxa, that can cause diseases. Most significant is the variety of organisms of the digestive tract. The human gastrointestinal (GI) tract is home to the largest microbial community which effect the hosts metabolism and immune system. The physiology, health as well as disease of the human body can not be entirely understood by the sole analysis of human genes and their
products. Also the comparison of each organism microbiotal counterparts is essential in this regard. Recent studies have shown a central role of the gut microbiota related to nutrition and many GI diseases ranging from inflammations to cancer [?]. Therefore comprehensive studies and a metagenomic analysis of the human microbiome was initiated to unravel our so-called “second genome” (e.g. The Human Microbiome Project - HMP [?, ?] - was a United States National Institutes of Health initiative with the goal of identifying and characterizing the microorganisms which are found in association with both healthy and diseased humans - conducted 2008).

The superorganism that represents the human body is therefore composed of human and microbial cells. The analysis of this microbial community which includes bacteria, eukaryotes, and viruses is primarily done through DNA-based studies. The 16S ribosomal RNA (rRNA) plays a key role in the culture-independent characterization of microbiota. According to its structure formed by alternating variable and highly conserved regions, the 16S ribosomal RNA serves as an evolutionary chronometer allowing for the identification and differentiation of subbacterial (a large group of bacteria having rigid cell walls; motile types have flagella) and archaeal (these microbes are prokaryotes, meaning that they have no cell nucleus or any other membrane-bound organelles in their cells) taxa. Sequencing variable regions of 16S rRNA genes is widely used to characterize complex microbial communities. The benefit of this genetic marker is based on the fact that it is present in all eubacteria and archaea.

Ribosomes are the gene-transcribing machines in all living cells. RNA polymerase transcribes a gene from the DNA to Messenger RNA. This mRNA is transported out of the cell nucleus into the cell fluid, where ribosomes latch onto this mRNA. The ribosomes move along the mRNA, reading the code contained in its sequence of nucleotide bases and stringing the right amino acids together based on the code to build protein chains.

Ribosomes consist of two components: the small ribosomal unit which reads the DNA and the large subunit that joins amino acids together. Each unit is composed of one or more ribosomal RNA (rRNA) molecules. Ribosomes can bind to mRNA and determine a chain of amino acids out of its sequence (mRNA is sandwiched between the two subunits). According to the code tRNA (transfer RNA) carries the required acids to the ribosome and completes one part of the code by binding to the mRNA strain. For each nucleotide triplet exists a distinct tRNA molecule with a connected amino acid who align in a chain to form a protein [?].

A ribosome has three binding sites (A,P,E):

- The A site binds to tRNA.
- The P site (stands for peptidyl-tRNA) contains the last amino acid of a growing chain.
- The E site holds the last amino acid of a chain.

The subunits contain different respective rRNA molecules in prokaryotes and eukaryotes (noted in Svedberg units S and base pairs bp):

<table>
<thead>
<tr>
<th>ribosome</th>
<th>size (S)</th>
<th>large subunit</th>
<th>small subunit</th>
</tr>
</thead>
<tbody>
<tr>
<td>prokaryotic</td>
<td>70S</td>
<td>50S (5S: 120 bp, 23S: 2906 bp)</td>
<td>30S (16S: 1542 bp)</td>
</tr>
<tr>
<td>eukaryotic</td>
<td>80S</td>
<td>60S (5S: 121 bp, 5.8S: 156 bp, 28S: 5070 bp)</td>
<td>40S (18S: 1869 bp)</td>
</tr>
</tbody>
</table>

Table 1: Prokariotic and eukariotic rRNA molecules

The genes for ribosomal RNA have changed little over millions of years as organisms evolved. Variable regions are subjected to mutation during evolution. This slight changes that have occurred provide clues as to how closely or distantly various organisms are related [?, ?, ?].
Figure 2: Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences (adapted from [?]) figure 1).
Table 2 represents each hypervariable region of *E. coli* 16S rRNA, its length and position.

<table>
<thead>
<tr>
<th>Region</th>
<th>length(bp)</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>30</td>
<td>69-99</td>
</tr>
<tr>
<td>V2</td>
<td>105</td>
<td>137-242</td>
</tr>
<tr>
<td>V3</td>
<td>195</td>
<td>338-533</td>
</tr>
<tr>
<td>V4</td>
<td>106</td>
<td>576-682</td>
</tr>
<tr>
<td>V5</td>
<td>57</td>
<td>822-879</td>
</tr>
<tr>
<td>V6</td>
<td>79</td>
<td>967-1046</td>
</tr>
<tr>
<td>V7</td>
<td>56</td>
<td>1117-1173</td>
</tr>
<tr>
<td>V8</td>
<td>51</td>
<td>1243-1294</td>
</tr>
<tr>
<td>V9</td>
<td>30</td>
<td>1435-1465</td>
</tr>
</tbody>
</table>

Table 2: *E. coli* 16S rRNA hypervariable regions [?]

Table ?? represents conserved regions and shows each primer per name, primer sequence and amplicon length:

<table>
<thead>
<tr>
<th>primer name</th>
<th>sequence (5’-3’)</th>
<th>start-position</th>
<th>end-position</th>
<th>region</th>
<th>amplicon length</th>
</tr>
</thead>
<tbody>
<tr>
<td>27F (8F)</td>
<td>AGA GTT TGA TYM TGG CTC AG</td>
<td>8</td>
<td>27</td>
<td>V1-V2</td>
<td>311</td>
</tr>
<tr>
<td>CC [F]</td>
<td>CCA GAC TCC TAC GGG AGG CAG C</td>
<td>312</td>
<td>334</td>
<td>V3</td>
<td>185</td>
</tr>
<tr>
<td>338F</td>
<td>ACT CCT ACG GGA GGC AGC</td>
<td>338</td>
<td>355</td>
<td>V3</td>
<td>164</td>
</tr>
<tr>
<td>338R</td>
<td>GCT GCC TCC CGT AGT</td>
<td>355</td>
<td>338</td>
<td>V3</td>
<td>164</td>
</tr>
<tr>
<td>357F</td>
<td>CTC CTA CGG GAG GCA GCA G</td>
<td>357</td>
<td>376</td>
<td>V3</td>
<td>143</td>
</tr>
<tr>
<td>519F</td>
<td>CAG CMG CCG CCG TAA TAC</td>
<td>519</td>
<td>536</td>
<td>V4</td>
<td>249</td>
</tr>
<tr>
<td>519R (536R)</td>
<td>GCA TTA CCG CGG CKG CTG</td>
<td>536</td>
<td>519</td>
<td>V4</td>
<td>249</td>
</tr>
<tr>
<td>515F</td>
<td>GTG CCA GCM GCC CGG GTA A</td>
<td>515</td>
<td>533</td>
<td>V4</td>
<td>252</td>
</tr>
<tr>
<td>785F</td>
<td>GGA TTA GAT ACC CTG GTA</td>
<td>785</td>
<td>805</td>
<td>V5-V6</td>
<td>102</td>
</tr>
<tr>
<td>805R</td>
<td>GAC TAC CAG GGT ATC TAA TC</td>
<td>805</td>
<td>785</td>
<td>V5-V6</td>
<td>102</td>
</tr>
<tr>
<td>907R (926R)</td>
<td>CCG TCA ATT CMT TTG AGT TT</td>
<td>926</td>
<td>907</td>
<td>V4</td>
<td>252</td>
</tr>
<tr>
<td>CD [R]</td>
<td>CTT GTG CGG GCC CCC GTC AAT TC</td>
<td>952</td>
<td>939</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100F</td>
<td>CAA CGA GCG CAA CCC</td>
<td>1100</td>
<td>1114</td>
<td>V7</td>
<td>276</td>
</tr>
<tr>
<td>1100R</td>
<td>GGG TTTG CGG TCG TTG</td>
<td>1114</td>
<td>1100</td>
<td>V7</td>
<td>276</td>
</tr>
<tr>
<td>1237F</td>
<td>GGG CTA CAC ACG YGC WAC</td>
<td>1237</td>
<td>1255</td>
<td>V8-V9</td>
<td>135</td>
</tr>
<tr>
<td>1390R (1406R)</td>
<td>ACG GCC GGT GTG TRC AA</td>
<td>1406</td>
<td>1390</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1391R</td>
<td>GAC GGG CGG TGT GTR CA</td>
<td>1403</td>
<td>1391</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1492R</td>
<td>TAC CTT GTT AYG ACT T</td>
<td>1507</td>
<td>1492</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1492R (l)</td>
<td>GGT TAC CTT GTT ACG ACT T</td>
<td>1507</td>
<td>1489</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1492R (s)</td>
<td>ACC TTG TTA CGA CTT</td>
<td>1507</td>
<td>1493</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Depending on the application, primer pairs with an estimated longer/shorter amplicon length are selected. The most common primer pair is currently referred to as 27F and 1492R for full length sequencing [?]. The difference in construction of variable and conserved regions determines their function and further area of application. Like the name suggests variable regions mutate during evolutionary processes and can serve as a kind of evolutionary clock specific to the respective taxon. Bacterial 16S rRNA consists of nine hypervariable regions (V1 up to V9) with considerably sequence diversity among different species. However, there are structural similarities between different bacteria. This fact builds the foundation of systematic studies for specific diagnostic goals.

These variable regions are flanked by conserved parts (which makes the PCR amplification of
sequences using universal primers possible), regions which are more important for ribosome function. Mutations in these regions can rarely be transferred to the offspring. Such mutations affect the bacterial cell heavily and so these regions remain unchanged [?]. The 16S rRNA gene is ideal for this processes because it is very short (a total length of 1542 nucleotides as visible in Figure ??) and offers quick and standardized amplification and sequencing. The composition of rRNA, in particular the sequences of the variable regions, depends on the type of microorganism. A different sequence of nucleotides is characteristic for each organism but all of these microbiota have identical ones on their very beginning and end of a variable region. This start- and end-sequences are used as universal and group-specific primer and essential for amplification of the 16S rRNA through PCR. The quality of a primer is directly responsible for the quality of the classification results after sequencing. Furthermore the majority of primers is specific for a certain range of bacteria with the same taxonomic rank. This means it is not possible for these specific primers to be applied for amplification of all bacteria in a microbial sample. Classification accuracy depends on sequencing technology and the chosen variable region. The V4/V5 region exhibits the highest accuracy regardless from the used sequencing technology and is therefore recommended. In contrast the V3/V4 region showed the worst classification efficiency [?].

The different colored R-fragments (R1 to R6) in figure ?? represent 6 regions of the 16S rRNA gene with distinctive length according to the V region. Together they comprise the full gene sequence of 1542 bp.

Table 4: E. coli 16S rRNA R-fragments and respective length [?]

<table>
<thead>
<tr>
<th>R-fragment</th>
<th>length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>1-250</td>
</tr>
<tr>
<td>R2</td>
<td>251-500</td>
</tr>
<tr>
<td>R3</td>
<td>501-750</td>
</tr>
<tr>
<td>R4</td>
<td>751-1050</td>
</tr>
<tr>
<td>R5</td>
<td>1051-1300</td>
</tr>
<tr>
<td>R6</td>
<td>1301-1542</td>
</tr>
</tbody>
</table>

These R-fragments can for example form a base for a prediction of taxa richness using these partial 16S ribosomal RNA sequences. Microbiota are for the most part identified and characterized through sequencing processes in large microbiome studies. A quick overview of microbiome studies is provided in a later chapter.

As the compound of highly conserved primer binding sites and hypervariable regions that can provide species-specific signature sequences useful for identification of bacteria, 16S rRNA gene sequencing has become predominant in medical microbiology because it provides a quick and cheap alternative to other methods of bacterial identification. Originally only used to identify bacteria 16S rRNA, sequencing methods have provided the capability to even reclassify bacteria into new species and classes. It has also been used to describe new species that have never been successfully cultured. The 16S rRNA gene in microbes [?] (all known prokaryotes, 18S rRNA in eukariotes) and its proper reaction to changes makes it an ideal standard for the identification and classification of microbes. Type strains of 16S rRNA gene sequences for most bacteria and archaea are available on public databases such as NCBI. However, the quality of the sequences found on these databases are often not validated. Therefore, secondary databases that collect only 16S rRNA sequences are widely used [?, ?, ?, ?, ?, ?].

1.2 Scope of microbiome studies

Advances in human microbiome research depend on carefully executed, controlled and reproducible studies. Many studies have documented a difference in the composition of host-
associated microbial communities between healthy and disease states [? , ? , ?]. For a growing
number of ailments, an altered microbiome is not just a marker of disease, but also actively
contributes to a medical condition. This is especially true for the interaction of a host and its
gut microbiome. The human gastrointestinal tract is a unique ecosystem and harbors multitudes of
diverse microorganisms. Right from birth various factors shape the GI microbiome
and it depends on a variety of factors including the age, diet, genetic composition of the host,
gender, geographic location, and health status of an individual [?]. An individual is to equal
parts affected through its microbiome as itself affects the former.

Studies have directly demonstrated that the composition of gut microbiota can alter a host's
metabolism. A microbiota of a diseased donor that is transplanted into a healthy host causes
the recipient to display the same phenotype. The range of conditions with a host-microbiome
interaction component continues to grow and has recently started to include neurological con-
ditions. Especially gut microbes are associated with various pathologies. A large number of
projects will focus on 16S rRNA genes for economical reasons as stated above. Such stud-
ies focus on different kind of microbiome and disease connection. Studies like the MetaHIT
(http://www.metahit.eu/) project chose the GI tract for detailed investigation and focused
on inflammatory bowel disease and obesity. Obesity for example, is related to a variety of
comorbidities including type II diabetes and cardiovascular diseases. It was shown that GI
microbiota influences the body weight using germ-free mice. The transplantation of an obesity
related microbiome to a lean animal leads to a significantly increased weight gain.

Like the GI microorganisms, the vaginal microbiota is very dynamic and influenced through
age, ethnicity, methods of birth control, sexual activity, personal care and other environmental
factors. One of the largest human microbial habitats is represented by the skin with an area
of about 1.8m\(^2\). Because of different pH, moisture, temperature and different structures, skin
is divided in several niches and not a uniform habitat. Disturbed microbiota structures in this
niches were linked with certain skin diseases.

The characterization of human microbiomes under different conditions will help to answer a
variety of questions [?]:

• How are microbial communities formed and how do they regenerate?
• What mechanisms 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 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 most of these questions
would remain unsolved without the application of next generation sequencing technologies for
the characterization process [? , ?].

2 The sequencing process

2.1 Before the sequencing process

Before sequencing a DNA sample, preliminary steps like
DNA extraction

is the purification of a DNA sample through physical and chemical methods. This process is comprised of three basic steps (some optional steps exist). The cell membranes (of collected cells) are opened to expose DNA along with the cytoplasm of the cells. Lipids from the cell membrane and the nucleus are broken down with detergents and surfactants. After treating the solution with concentrated salt the debris (broken proteins and lipids) clumps together and can be separated from DNA through centrifugation [?].

PCR (polymerase chain reaction)

is a process to amplify a single copy or a few copies of a piece of DNA across several orders of magnitude. PCR consists of a series of repeated temperature changes. Length of applied time and temperature used, depend on a variety of parameter (like used enzymes or primers used). After the first heating cycles the hydrogen bonds between complementary bases of the template DNA are disrupted, yielding single-stranded DNA molecules. At lower temperature primer are bound to the single-stranded DNA molecules through polymerase, creating a primer-template hybrid. Next, while applying constant temperature (depending on used DNA-polymerase), nucleotide bases are attached to the strand beginning at the primer. This steps are called denaturation step, annealing step as well as elongation step and are repeated until the needed amount of DNA copies is reached [?].

barcoding

is a taxonomic method that uses a short genetic marker in an organism’s DNA to identify it as belonging to a particular species [?]. The sequence of base pairs (which functions like a barcode) of a unknown sample is thereby compared with preexisting classifications associated with a species.

have to be executed.

2.2 The sequencing process

After DNA extraction and amplification by PCR, the resulting DNA is sequenced. There are different generations of sequencing methods. The initial one is Sanger sequencing, also known as “first-generation”-sequencing. Sanger sequencing is mostly regarded as outdated but has still the advantage of long read length (>800 bp) and high accuracy (> 99.999%) over its successor [?]. It has mostly been replaced by “next-generation sequencing“ (NGS) techniques. This NGS produces an enormous amount of data in a single run within a short period of time at low costs. Additionally it makes the determination from amplified single DNA fragments without cloning possible. Different kinds of NGS techniques can be distinguished either by template preparations, chemistry, detection approaches and base calling methods. These differences result in benefits and disadvantages of each of the different techniques [?].

At this point Roche 454 should be mentioned, who played a major role in the structural analysis of the DNA and developed a number of now mostly outdated sequencing applications [?]. This sequencer uses pyrosequencing technology [?]. Instead of using dideoxynucleotides to terminate the chain amplification, pyrosequencing technology relies on the detection of pyrophosphate released during nucleotide incorporation. The initial read length was about 100-150 bp and a output of 20 MB per run. Today its read length can reach 700+ bp and an output of 14 G per run in a short amount of time. A run finishes in about 10 hours. The considerable read length and the short amount of time needed from start to finish stand as biggest advantage of this method while high costs and relatively
high error rate in homopolymer stretches longer than 6 bp are part of the reason it is mostly outdated today [?].

Further Illumina/Solexa should be mentioned with acquiring significantly more sequence data than prior processes. With their current technology an enormous amount of data can be produced in a very short time with improved Sanger sequencing methods [?]. Depending on the product initial read length amounts to 1 x 36bp up to 2 x 300bp and resulting output from about 500Mb up to 1Tb. Because of the rather small read length the number of reads is drastically increased. Via a added terminator with a fluorescent label, that blocks further PCR, and results in only single added bases to a growing DNA copy strand. The sequencing process is applied parallel on a large number of template molecules on a solid surface, which is a major innovation in the amplification of template molecules. The fluorescent label allows the detection by camera and produces a set of DNA sequence reads of uniform length.

With their MiSeq Desktop Sequencer 2 x 300 paired-end reads are produced in a single run which allows small genome sequencing and assembly as well as high detection accuracy [?]. Other features include:

<table>
<thead>
<tr>
<th>Read Length</th>
<th>total time</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x 75 bp</td>
<td>21 hrs</td>
<td>3.3-3.8 Gb</td>
</tr>
<tr>
<td>2 x 300 bp</td>
<td>56 hrs</td>
<td>13.2-15 Gb</td>
</tr>
</tbody>
</table>

3 Analysis

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 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 OTU’s: OTU’s can be either generated by unsupervised clustering of the sequences or OTUs can be formed by comparative classification using a reference database.

1. Unsupervised clustering

- preprocessing
- multiple sequence alignment
- calculation of sequence distances
- clustering of sequences into OTU’s (at a certain similarity)
- dereplication
- classification of each of the representative sequences either by similarity search against a reference database or with a classifier
- statistical analysis and visualization

In order to calculate distances between sequences all sequences have to be aligned due to their difference in length as a prerequisite to OTU generation. During the clustering step all sequences are grouped according to their distances without taking phylogenetic
information into account. OTUs are formed according to furthest, average or nearest neighbor algorithms. A variety of such metrics exist (e.g. UCLUST [?]). After the clustering step information about the OTUs and their differences or abundances is gathered and sequences assigned. A representative sequence is selected (usually the longest) for classification [?].

2. Comparative classification

- preprocessing
- dereplication
- similarity search against reference database
- combine reads based on classification
- statistical analysis and visualization

Comparative classification reversed the above mentioned steps. Sequences are assigned to a taxon to provide the basis for OTU generation. Each sequence is classified through its similarity to a known reference sequence. For this step local alignment search tools like BLAT [?] or JGAST [?] are used. OTUs are formed by pooling sequences with the same taxonomic classification [?].

Although both approaches generate OTU’s, the key difference is the homogeneity of an OTU. Sequences in cluster based OTU’s have a predefined maximum distance (sequence dissimilarity), whereas the sequence distance in OTU’s formed by comparative classification depends on the distance to the reference sequence [?].

To improve the quality of the downstream analysis sequence data is also preprocessed. Rarefaction should be mentioned at this point. It is a technique to assess species richness from the results of sampling. Rarefaction can be expressed in carefaction curves which plot the number of species as a function of the number of samples. When sampling various species in a community, the larger the number of individuals sampled, the more species will be found. Rarefaction curves are created by randomly re-sampling the pool of N samples multiple times and then plotting the average number of species found in each sample [?].

4 SnoWMAn

The number and scope of microbial community studies from diverse environments have increased exponentially with the recent introduction of next generation sequencing technologies. However, the development of analysis tools did not keep up with the advancements in data generation. Existing tools limit the number of sequences to be analyzed considerably or cover only part of the analysis workflow. Therefore, we developed SnoWMAn, a powerful web server for analysis of amplicon sequence data generated in microbiome studies [?]. It can cope with the increased data volume and integrates the complete analysis workflow covering sample splitting, sequence filtering and alignment, clustering, taxonomic classification, diversity estimation, sample comparison and visualization of the results. To accommodate the diverse analysis approaches, different analysis pipelines are available, which allow both sequence independent and taxonomy independent phylotyping. The diversity of microbial populations is estimated by rarefaction analysis and microbiomes of different samples can be compared by principal component analysis, Venn diagrams and charts representing the sequence distribution on taxonomic classification. The web server offers unique capabilities for microbiome characterizations and comparisons and facilitates fast and convenient analysis of the large and complex datasets. It can be accessed via http://snowman.genome.tugraz.at [?].
4.1 What is Snowman?

SnoWMAn, the Straightforward Novel Webinterface for Microbiome Analysis, covers the entire microbiome analysis workflow from sequence preprocessing to the visualization of the results. A typical microbial community analysis with SnoWMAn compromises three simple steps; first, the sequence and metadata are uploaded to a data repository. Second, the user can choose between five currently available analysis pipelines and define the respective parameters. Finally the user can perform statistical analysis and visualization of 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 [?] and JGAST [?] can be chosen for OTU generation by comparative classification. mothur [?], RDP [?] and UCLUST [?] 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 visualize 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 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 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 [?].

5 Introduction of analysis options/methods

SnoWMAn consists of two modules: (i) the web interface which offers access to its analysis pipelines and their result visualization, (ii) the computation service which is delegating computationally intensive calculations to the high performance computing infrastructure. The user interface is decoupled from the calculation back-end to ensure a responsive browser interaction.
5.1 Pipelines
SnoWMAn currently supports five different analysis pipelines - two based on comparative clustering (based on BLAT and JGast) and three clustering based (UCLUST, mothur and RDP). All of them result in a unified output format that can later be used for further statistical analyses.

5.2 Clustering based pipelines
OTU generation by clustering comprises three major working steps: Before OTUs can be defined the sequences have to be aligned. Subsequently, OTUs are generated 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. Aligning sequences is a prerequisite for the subsequent OTU generation where distances (i.e. the percentage of base changes) between sequences are calculated. Therefore, either multiple sequence alignments (MSAs) of all target sequences or pair-wise alignments are created.

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 the furthest, average or nearest neighbor metrics. SnoWMAn uses mothur, RDP and UCLUST as clustering tools. According to the clustering method, the outcome may be 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 [?].
5.2.1 mothur pipeline

The first step of the in SnoWMAn implemented version of the mothur pipeline splits the input sequence file (normally in FASTA format) into multiple smaller files based on the sample tag.

To ensure efficient further processing these small files are merged again to an ordered larger file which is the basis for following analysis steps.

Duplicates and chimeras are removed in order to keep the run time as low as possible and to avoid OTU inflation respectively.

A representative sequence serves as template for further alignment procedures.

Sequences are aligned by comparing the closest match of the first half of our sequences to the template alignment and the closest match of the other half of our sequences to the template alignment.

Representative sequences, containing the information about its distance to the closest annotated sequence, are clustered into OTUs.

A rarefaction creates samples of a certain depth, the number of species is sampled to predict the number of genera in a particular community.

Based on the classification results a output of the different samples is created.

More information on mothur can be found in the Appendix.
5.2.2 RDP pipeline

- The first step of the in SnoWMAn implemented version of the RDP pipeline splits the input sequence file (normally in FASTA format) into multiple smaller files based on the sample tag.

- These files are independent from one another filtered (multiple appliable filter), trimmed and their orientation determined.

- To ensure efficient further processing these small files are merged again to an ordered larger file which is the basis for following analysis steps.

- Duplicates and chimeras are removed in order to keep the run time as low as possible and to avoid OTU inflation respectively.

- Hierarchical clustering is used to group together reads that are similar to each other. Key to this analysis is repeated calculation of distance measures between reads [?].

- Through the complete linkage clustering (farthest neighbor method) distance between clusters in hierarchical cluster analysis is calculated which allows to create a cluster file based on the aligned sequence [?].

- Representative Sequences are extracted for each OTU.

Figure 4: RDP pipeline uses tools developed by Cole et al [?].
• Representative sequences, containing the information about its distance to the closest annotated sequence, are clustered into OTUs.

• A rarefaction creates samples of a certain depth, the number of species is sampled to predict the number of genera in a particular community.

• Based on the classification results a output of the different samples is created.

• More information on RDP can be found in the Appendix.

5.2.3 UCLUST pipeline

![UCLUST pipeline diagram]

• The first step of the in SnoWMAn implemented version of the UCLUST pipeline splits the input sequence file (normally in FASTA format) into multiple smaller files based on the sample tag.

• To ensure efficient further processing these small files are merged again to an ordered larger file which is the basis for following analysis steps.

• Duplicates and chimeras are removed in order to keep the run time as low as possible and to avoid OTU inflation respectively.

• Through the alignment and search algorithm UCLUST OTUs are generated. The details of this procedure are available in the Appendix.

• The through UCLUST generated clusters are extracted and converted to a mothur format.

• Representative sequences, containing the information about its distance to the closest annotated sequence, are clustered into OTUs.
• A rarefaction creates samples of a certain depth, the number of species is sampled to predict the number of genera in a particular community.
• Based on the classification results a output of the different samples is created.
• More information on UCLUST can be found in the Appendix.

5.3 Comparative classification based pipelines

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 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. There are a variety of different reference databases.

The most similar sequence in the reference database (Greengenes [?], SILVA [?] and the RDP reference database [?]) can be determined by using local alignment search tools like BLAST [?] or BLAT [?], 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 [?]. 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 [?]). 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 [?] under the use of different reference databases [?].
5.3.1 BLAT pipeline

- The first step of the in SnoWMAn implemented version of the BLAT pipeline splits the input sequence file (normally in FASTA format) into multiple smaller files based on the sample tag.
- Duplicates and chimeras are removed in order to keep the run time as low as possible and to avoid OTU inflation respectively.
- The local alignment search tool BLAT aligns each small sample sequence. The details of this procedure are available in the Appendix.
- Representative sequences, containing the information about its distance to the closest annotated sequence, are clustered into OTUs, depending on their taxonomic classification.
- A rarefaction creates samples of a certain depth, the number of species is sampled to predict the number of genera in a particular community.
- Based on the classification results a output of the different samples is created.
- More information on BLAT can be found in the appendix.

Figure 6: BLAT pipeline: performs BLAT alignment to the greengenes database [1].
5.3.2 JGAST pipeline

- The first step of the in SnoWMAn implemented version of the JGAST pipeline splits the input sequence file (normally in FASTA format) into multiple smaller files based on the sample tag.

- To ensure efficient further processing these small files are merged again to an ordered larger file which is the basis for following analysis steps.

- Duplicates are removed to reduce computation time.

- Chimeras (< 97% identical to a biological sequence) are removed to prevent OTU inflation.

- JGAST classifies sequences through the use of BLAST (a alignment search tool similar to BLAT) which compares all sequences to a reference.

- The best BLAST hits are selected and through RDP classification formed into OTUs.

- Based on the classification results an output for the different samples is created.

- More information on JGAST can be found in the Appendix.
5.4 Preprocessing and pipeline parameters

The major goal of preprocessing sequence data is to improve quality of downstream analysis. Filtering of 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 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, artifact such as poly-A/T trails or adapters, primers might have been ligated to the sequence. Sequence trimming to a certain length or according to quality score can help to get rid of these artifacts [?].

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 [?]. The major source of pyrosequencing noise is caused by uncertainties in the base calling of long homopolymer stretches [?]. Additionally PCR errors occurring during the amplification process have to be considered, since they increase the per-base sequencing error rate.

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 [?]. 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 [?]. The impact of chimeras can be very critical in particular when they occur at high frequencies.

Apart from more accurate OTU estimations denoising and chimera checking result 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 as mentioned before [?].

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 sequence either by their sequence or by using different kind of error correction methods [?].

6 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. 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 [?, ?]. According to this reasoning, α-diversity and β-diversity constitute independent components of γ-diversity.

γ-diversity is therefore the overall diversity for the different ecosystems within a region as shown in the Table ?? below [?, ?]:

Table 6: Comparison between different diversity values [?]
<table>
<thead>
<tr>
<th>Species</th>
<th>Woodland habitat (W)</th>
<th>Plane habitat (P)</th>
<th>Swamp habitat (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>H</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>J</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α-Diversity</td>
<td>6</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>β-Diversity</td>
<td>W VS. P → 5</td>
<td>P VS. S → 6</td>
<td>W VS. S → 9</td>
</tr>
<tr>
<td>γ-Diversity</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

### 6.1 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 [?] and refers to the number of species in a sample [?]. In contrast to qualitative species-based α-diversity only represents presence or absence of a certain taxa within a microbial community [?]. To define the qualitative α-diversity, the Chao index or the ACE index 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 [?] or Simpson [?] 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 [?, ?].

### 6.2 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 [?]. 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 a healthy state. Additionally, species based β-diversity measures allow evaluating same body space of different individuals) share a similar or equal microbial composition [?]. As with α-diversity, qualitative and quantitative indices of β-diversity can be discriminated. Sørensen [?], Bray-Curties [?], and Jaccard [?] indices are often calculated to get a qualitative measure. For quantitative diversity index calculations, the Sørensen quantitative index [?] or Morisita-Horn [?] 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 microbial cohort is taken into account.

In measuring α-diversity 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 differences often directly correlate with phenotypic similarities, which represent fundamental features [?, ?].

6.3 \(\gamma\)-Diversity

R. H. Whittaker introduced \(\gamma\)-diversity together with \(\alpha\)- and \(\beta\)-diversity [?]. As stated above \(\alpha\)-diversity and \(\beta\)-diversity constitute independent components of \(\gamma\)-diversity as the total species diversity in a landscape is determined by the mean species diversity in sites or habitats at a more local scale and the differentiation among those habitats. \(\gamma\)-diversity is not tied to a specific spatial scale, it can be measured for an existing dataset at any scale of interest [?].

\[ \gamma = \alpha \times \beta \]

6.4 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical procedure which uses orthogonal transformation (length and angle between vectors are preserved) to convert a set of observations of possibly correlated variables in a set of linear uncorrelated variables. These are called principal components. The first component has the highest possible variance while each following component has the highest possible variance in respect to their orthogonality to the proceeding components. This results in an uncorrelated orthogonal basis set of vectors. PCA is used for exploratory data analysis and to make predictive models about systems. It can be understood as fitting an n-dimensional ellipsoid to the data and is therefore a procedure which is sensitive to scaling of data as each principal component represents a axis of the resulting ellipse.
7 How to use SnoWMAn?

7.1 General functionality and User Interface

This chapter contains information about the general navigation and functionality of the SnoWMAn web-application.

7.1.1 Login and creating an Account

The user can choose between creating an account or using SnoWMAn as a guest. For usage as a guest enter following username and password:

user: guest password: guest

Figure 8: The Start up interface of SnoWMAn after accessing the web-application

Figure 9: The login screen
In order to create a personal account click the “Create account” link. Choose a username, password, state an E-mail address and submit. After the creation of an account, it can be accessed through the “login” link. The creation of a personal account is recommended as it allows the storage and management of data.

SnoWMAAn accounts

![Create a new SnoWMAAn account](image)

Figure 10: The account creation screen.

### 7.1.2 Home screen

![Welcome to SnoWMAAn](image)

Figure 11: The home screen.

After login as guest or with a personal account the home screen is displayed. At the left-hand side of the home screen the navigation tree is displayed. It contains all previous analysis runs, a link to the "data directory" with prior uploaded files and a link to the analysis pooling tool. At the top of the screen the main navigation bar is displayed containing a "Home-", "Upload Data-" and "Start Analysis" link. Per click on the "Start Analysis" links the Analysis View is displayed.
7.1.3 Data Directory

The "Data Directory" displays a list with all the user uploaded files. With a click on the "Choose File" button a new file can be selected and uploaded per a simple click on "upload". With a click on "Upload multiple files" several files can be uploaded at once. It is also possible to upload .zip files. Every file in the .zip container can be accessed for analyses.

Via the buttons next to a single file the respective data set can be downloaded, renamed, deleted or inspected.

The "load test files" buttons allows the test of various features by adding of some tryout files to the directory. This allows to try out the different analysis options with files containing a small number of sequences.

The magnification glass next to a file opens a "File Inspector" window which contains the first 200 lines of the selected file. Close the window by pressing the button in the upper right corner of the layer. The button allows the behavior definition of text not displayed in text area (wrap around or display horizontal scrollbar). The button displays sequence file statistics like sequence length distribution and base frequency.
7.1.4 Start Analysis

This window allows the execution of different analyses through the in Chapter 5 mentioned analysis pipelines. The execution of an analysis requires the prior upload of files to the ”Data Directory”.

To start an analysis click on the ”Start Analysis” button in the main navigation bar. This opens a window where different analysis pipelines can be selected. A yellow arrow indicates each main pipeline with its different modi listed below. While moving the mouse and hover over a pipeline or modi a flowchart of the selection is displayed on the right side of the window. If only parts of the parent pipeline are used, the skipped parts are grayed out. Per click on a link the pipeline with its modifications is selected and the input data screen is displayed.

One of the following pipelines can be selected:

- mothur
- RDP
- UCLUST
- BLAT
- JGAST
7.1.5 Input Data

At first the used input files can be specified per ”Add file” dropdown list. Not every file can be selected as files in the ”Data Directory” are filtered by their extension. All analyses require FASTA files in order to progress. Here can also files inside of .zip archives can be added to the selection. Each selected file is displayed in the ”Already added files” section above the dropdown list. Per click on the button a selected file can be removed again. A click on the ”Next” button accepts the selection and moves to the following step.

7.2 Workflow

7.2.1 Preprocessing parameters

All workflows will show the ”Select preprocessing parameters” window as first step. It is differentiated between two versions:

- IGB preprocessing
- RDP preprocessing (combines the RDP Tagsorter and RDP Trimmer)

Depending on the selected workflow there may be one or both of them available for selection while on the preprocessing screen. Each one provides different functions and parameters that include:
• Sample splitting based on sample tags (barcodes)
• Filtering based on forward primer
• Cropping tag and primer from sequence
• Filtering short sequences
• Filtering sequences that exceed the maximum number of Ns
• Filtering based on quality (Quality file required)
• Reverse complement sequences

A tag file is required to perform preprocessing. The format of such a tag file is explained in the "Requirements" section of the user manual. The tag file (ends with .txt), quality filtering and quality files (end in .qual) with the same name as the sequence file are automatically selected. As tag files are required for preprocessing their absence is indicated with the error message: "One or more of your selected sequence files is missing a tag file.". Further Information to different options:

Primer File:
A list of primer file candidates is generated by searching the files in the data directory for a file with the same name as the selected sequence files and the ending ".primers.fa". If no primer file exists, or per user preference, forward and reverse primer can be manually entered into a text field.

Perform Quality Filtering:
Quality filtering can be enabled if a quality file for each sequence file exists. The quality files are loaded and checked prior to the preprocessing steps where all sequences that fall below the defined quality threshold will be removed from further analysis steps.

Analyze runs individually:
If this option is enabled the selected workflow will run separately for each input file. In any other case the preprocessed samples will be treated as if they all originated from the same file.

Maximum number of Ns:
Sequences with a total number of characters that exceed N are removed from further analysis steps.

Minimum sequence length:
All sequences shorter than the defined minimum sequence length will be removed during preprocessing.

Forward primer sequences/Reverse primer sequences:
If the forward and reverse primer are specified all sequences that do not match the specified primer at the beginning and the end of the sequence are removed from further steps.

Quality threshold:
If "Quality filtering" is selected all sequences below the threshold are removed.
Crop primers:
Defines whether primer sequences are cropped from the sequence.

Reverse complement sequences:
Reverse complements of the original sequence before further analysis steps.

7.2.2 Full length sequences
If no primer, barcode or tag sequence is specified and only the full sequence file is present the "no splitting" option at the beginning of the workflow can be selected to skip the preprocessing steps.

After selection of the input file a tag file in the tag format must be specified. The barcode or tag sequence can be empty in this case. For each selected sequence a sample must be defined in the stated tag file in order to acquire grouping information for a sample. On the selection screen is a visual indication if the tag file contains a definition for each sequence file.

7.2.3 mothur

![Select Mothur pipeline parameters](image)

Figure 16: The mothur pipeline parameters.

Here a reference database for the alignment and settings for the clustering process can be selected as well as "Maximum cluster similarity" and "Similarity step parameter". These define on which maximum distance clustering occurs and the time period the analysis needs to be finished. Clustering method determines the algorithm that assigns sequences to OTUs (nearest neighbor, furthest neighbor and average neighbor). The classification model option provides taxonomic assignments from domain to genus (e.g. trichoderma). Rarefaction method can be set to mothur or RDP.
7.2.4 RDP

Here alignment model (reference database) and the classification (setting for the RDP classifier) can be selected as well as "Maximum similarity" and "Step parameter". These define on which maximum distance clustering occurs and the time period the analysis needs to be finished. The classification model option provides taxonomic assignments from domain to genus (e.g. trichoderma). Rarefaction method can be set to mothur or RDP.

7.2.5 UCLUST

Here classification settings for "Maximum similarity" can be selected as well as the RDP Classifier and "Step parameter". These define on which maximum distance clustering occurs and the time period the analysis needs to be finished. The classification model option provides taxonomic assignments from domain to genus (e.g. trichoderma). Rarefaction method can be set to mothur or RDP.
7.2.6 BLAT

**Figure 19: The pipeline settings screen.**

Here the reference database used for the BLAT alignment used in the *BLAT* pipeline, or for the *BLAST* alignment used in the JGast pipeline is selected.
This is the last step of each workflow. A summary of selected analysis steps, selected tag files, queue position and estimated duration is provided. Completely optional a name and description of an analysis can be provided. If no further information is provided one will be generated. The analysis starts with a click on the "Start" button and is then added to the program tree where detailed information about the process is available. Depending on the type of analysis information about used sequences or number of total runs may be provided after the finished workflow.
7.3 File formats and requirements

7.3.1 Sequence file

This file contains all actual sequences to classify.

Format:
The sequence file must be provided in FASTA format. Each sequence must have a unique identifier. If the file is formatted like a Roche 454 file, then the “uaccno=” field is considered to be the identifier. It is mandatory that the all identifiers are completely unique over the whole analysis.

Allowed characters are:

- A-Z
- a-z
- 0-9
- - and _

File name:
Sequence files have to be named with the file extension .fa, .fs, .fna or .fasta.

```
>000098_0532_0995 length=236 uaccno=FH2TH0T01BK0FP
ATCATCATTGGGCTTAAAGCGTGCGCAGGCGGTTCTGTAAGATAGATGTGAAATCCCCGG
GCTCAACCTGGGAATTGCATATATGACTGCAGGACTTGAGTTTGTCAGAGGAGGGTGGAA
TTCCACGTGTAGCAGTGAAATGCGTAGATATGTGGAAGAACACCGATGGCGAAGGCAGCC
CTCTGGGACATGGACTCGCTCATGACCCGAAACGCGTGGGGAGCCGCTGCAGGATAGAT
```
7.3.2 Tag file

The tag file identifies different kind of samples within the sequence file. A tag sequence is usually 6 characters long and matched with the first characters of each sequence do determine which sample it belongs to. This also allows the definition of groups over various samples.

Format:
The tagfile has to be provided in tab delimited format.
Allowed characters are:

- A-Z
- a-z
- 0-9

File name:
Tag files have the same name as their sequence file and end with .txt. The following excerpt is a sample for a tag file:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Sample</th>
<th>Patient</th>
<th>Timepoint</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGCAGC</td>
<td>A01</td>
<td>A</td>
<td>1</td>
<td>F</td>
</tr>
<tr>
<td>CTCAGC</td>
<td>A02</td>
<td>A</td>
<td>2</td>
<td>F</td>
</tr>
<tr>
<td>AGAGAG</td>
<td>A03</td>
<td>A</td>
<td>3</td>
<td>F</td>
</tr>
<tr>
<td>AGATGC</td>
<td>A04</td>
<td>A</td>
<td>4</td>
<td>F</td>
</tr>
<tr>
<td>AGCATG</td>
<td>C01</td>
<td>C</td>
<td>1</td>
<td>F</td>
</tr>
<tr>
<td>ATCATC</td>
<td>C02</td>
<td>C</td>
<td>2</td>
<td>F</td>
</tr>
<tr>
<td>ATCTGC</td>
<td>C03</td>
<td>C</td>
<td>3</td>
<td>F</td>
</tr>
<tr>
<td>ATGAGC</td>
<td>C04</td>
<td>C</td>
<td>4</td>
<td>F</td>
</tr>
</tbody>
</table>

"Sequence" and "Sample" columns are mandatory while "Patient", "Timepoint" and "Type" etc. are optional and define tags in later analysis results.

7.3.3 Primer file

Format:
The primer file must be provided in FASTA format. Each entry is seen as a forward primer except it is stated as reverse primer in the FASTA header (header contains the word reverse).

File name:
Primer files have the same name as their sequence file and end with .primers.fa. The following excerpt is a sample for a primer file:

```
>forwardprimer1
AYTGGGYDTAAAGNG
>forwardprimer2
AGATTTTGATCTGGCTCAG
```

Primer can also be specified manually in their respective fields.
7.4 Analysis results

After an analysis is finished a window containing its results will open. It contains a ”View”- and the context sensitive ”Action” Bar.

7.4.1 Action Bar

As this bar is context sensitive its functions may vary. In general it contains actions depending on the view and state the analysis results are currently in as:

- Download .zip file.
- Sample Merging.
- Sample Deletion.

7.4.2 View Bar

Results can be displayed in five different ways which are explained in detail in the Illustrations chapter.

Table 7: View bar actions
<table>
<thead>
<tr>
<th>Icon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Icon" /></td>
<td>Displays selected settings and parameter for the current analysis.</td>
</tr>
<tr>
<td><img src="image2.png" alt="Icon" /></td>
<td>Displays analysis history.</td>
</tr>
<tr>
<td><img src="image3.png" alt="Icon" /></td>
<td>Displays separate analysis results.</td>
</tr>
<tr>
<td><img src="image4.png" alt="Icon" /></td>
<td>Displays a sample overview.</td>
</tr>
<tr>
<td><img src="image5.png" alt="Icon" /></td>
<td>Displays /alpha-diversity.</td>
</tr>
<tr>
<td><img src="image6.png" alt="Icon" /></td>
<td>Displays statistical diagrams.</td>
</tr>
</tbody>
</table>

### 7.5 Illustrations

#### 7.5.1 File Browser

After the finished Analysis the displayed information defaults to the "File Browser" where the results of each analysis step for each run is available as a table. An example how the file browser looks like is shown in the analysis results section.

Different actions can be performed while in "File Browser" view:

<table>
<thead>
<tr>
<th>Icon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7.png" alt="Icon" /></td>
<td>Download analysis results.</td>
</tr>
<tr>
<td><img src="image8.png" alt="Icon" /></td>
<td>Copy analysis results to Data Directory.</td>
</tr>
<tr>
<td><img src="image9.png" alt="Icon" /></td>
<td>Delete analysis results.</td>
</tr>
<tr>
<td><img src="image10.png" alt="Icon" /></td>
<td>Open analysis results in File Inspector.</td>
</tr>
<tr>
<td><img src="image11.png" alt="Icon" /></td>
<td>View Sequence Length Distribution.</td>
</tr>
<tr>
<td><img src="image12.png" alt="Icon" /></td>
<td>Download results as ZIP File.</td>
</tr>
<tr>
<td><img src="image13.png" alt="Icon" /></td>
<td>Download currently viewed files as ZIP File.</td>
</tr>
</tbody>
</table>
**7.5.2 Settings Overview**

**Analysis results for "vfew BLAT"**

![View options]

**Settings for run Run1**

<table>
<thead>
<tr>
<th>IGB preprocessing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fasta files</strong></td>
<td>vfew2.fa from &lt;vfew.zip&gt;, vfew1.fa from &lt;vfew.zip&gt;</td>
</tr>
<tr>
<td><strong>Tag files</strong></td>
<td>vfew2.txt from &lt;vfew.zip&gt;, vfew1.txt from &lt;vfew.zip&gt;</td>
</tr>
<tr>
<td><strong>Quality files</strong></td>
<td>vfew2.qual from &lt;vfew.zip&gt;, vfew1.qual from &lt;vfew.zip&gt;</td>
</tr>
<tr>
<td><strong>Maximum number of N's</strong></td>
<td>0</td>
</tr>
<tr>
<td><strong>Minimum sequence length</strong></td>
<td>150</td>
</tr>
<tr>
<td><strong>Perform quality filtering</strong></td>
<td>true</td>
</tr>
<tr>
<td><strong>Quality threshold</strong></td>
<td>20</td>
</tr>
<tr>
<td><strong>Forward primer sequence(s)</strong></td>
<td>AYTGGGYDTAAAGNG AGAGTTTGATCCTGCTGCTGAG</td>
</tr>
<tr>
<td><strong>Crop primers</strong></td>
<td>true</td>
</tr>
<tr>
<td><strong>Reverse complement sequences</strong></td>
<td>true</td>
</tr>
<tr>
<td><strong>Chimera filtering reference database:</strong></td>
<td>chimeraref dbs_39-Feb-2012_16S_aligned</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BLAT settings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference database:</strong></td>
</tr>
<tr>
<td><strong>Rarefaction method:</strong></td>
</tr>
<tr>
<td><strong>Alignment reference database:</strong></td>
</tr>
</tbody>
</table>

Figure 23: The settings overview screen.

Here information about all setting the analysis was started with are displayed.
7.5.3 History

Analysis results for "vfew BLAT"

Figure 24: The analysis history screen.

Here information about runtime of the analysis is displayed. The first column displays the current analysis step, followed by the number of sequences. The last two columns display a time stamp of the starting point of a respective analysis step and estimated time for future steps as well as consumed time for finished steps.
Sample view provides a table for each run containing all its samples with their respective sequences and OTUs per distance. If a workflow does not use clustering table with the number of phylotypes for each taxonomy is instead displayed. According to the specifications in the tag file, tabs can be switched to display different sortings (optional columns). The "Sample View" allows the merging of data. This is not only seen here but expands to the "Statistical Analysis View". The default "Sample View" is in read-only mode.
<table>
<thead>
<tr>
<th>Icon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="download" alt="Icon" /></td>
<td>Download table as a .txt file.</td>
</tr>
<tr>
<td><img src="lock" alt="Icon" /></td>
<td>Unlock/Lock for editing.</td>
</tr>
<tr>
<td><img src="save" alt="Icon" /></td>
<td>Save changes.</td>
</tr>
<tr>
<td><img src="discard" alt="Icon" /></td>
<td>Discard changes.</td>
</tr>
<tr>
<td><img src="merge" alt="Icon" /></td>
<td>Mark for merging.</td>
</tr>
<tr>
<td><img src="delete" alt="Icon" /></td>
<td>Mark for deletion.</td>
</tr>
</tbody>
</table>

### 7.5.5 Statistical Analysis View

The "Statistical Analysis View" the visualization of analysis results in various charts cited from [?]. These charts can be exported to an excel file or saved graphic in .png or .svg format.

<table>
<thead>
<tr>
<th>Icon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="download" alt="Icon" /></td>
<td>Download dataset and image as .xls file.</td>
</tr>
<tr>
<td><img src="png" alt="Icon" /></td>
<td>Download image as PNG file.</td>
</tr>
<tr>
<td><img src="svg" alt="Icon" /></td>
<td>Download image as SVG file.</td>
</tr>
<tr>
<td><img src="zoom-out" alt="Icon" /></td>
<td>Zoom out (smaller image).</td>
</tr>
<tr>
<td><img src="zoom-in" alt="Icon" /></td>
<td>Zoom in (larger image).</td>
</tr>
</tbody>
</table>

The "Sequence Distribution Tab" allows the application of the in Table ?? displayed settings:

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select samples:</td>
<td>Select one or more samples to be displayed.</td>
</tr>
<tr>
<td>Chart Type:</td>
<td>Select the displayed chart type (as mentioned below).</td>
</tr>
<tr>
<td>Taxonomy/Distance:</td>
<td>Depending on a OTU/not OTU based workflow the distance/taxonomy can be defined.</td>
</tr>
<tr>
<td>Rank:</td>
<td>Defines the displayed taxonomix rank.</td>
</tr>
<tr>
<td>Share threshold:</td>
<td>Taxa with relative abundance below this threshold will be displayed as &quot;other&quot;.</td>
</tr>
<tr>
<td>Probability threshold:</td>
<td>Classifications with a probability lower than this threshold will be displayed as &quot;Unclassified&quot;.</td>
</tr>
<tr>
<td>Show unclassified:</td>
<td>Determines if unclassified sequences are displayed.</td>
</tr>
<tr>
<td>Display as:</td>
<td>Display as bar or line chart.</td>
</tr>
<tr>
<td>Scale:</td>
<td>Display absolute or relative values based on the percentage of the sum.</td>
</tr>
<tr>
<td>Display abundance:</td>
<td>Used in OTU Overlap Diagram to display the total number of sequences instead of the number of clusters.</td>
</tr>
<tr>
<td>Display group size:</td>
<td>Display the number of sequences per group next to the group names.</td>
</tr>
</tbody>
</table>
In “Statistical Analysis View” various chart types can be selected.

**Sequence VS Taxa:**

![Sequence distribution at the Phylum level](image)

Figure 26: The sequence VS. taxa diagram.

Plots samples against the number of sequences, categorized by their classification at a given taxonomic rank/level.

**Sequences VS Rank:**

![Sequence distribution over all taxonomic ranks](image)

Figure 27: The sequence VS. taxonomic ranks diagram.

Plots each taxonomic rank against the number of sequences that were successfully classified.
Sequences per Sample:

![Sequence per Sample diagram](image)

Figure 28: The sequence per sample diagram.

Plots each sample against the total number of sequences.

Endpoint depth:

![Endpoint depth diagram](image)

Figure 29: The endpoint depth diagram.

Plots each taxonomic rank against the number of sequences whose classification ended at that respective rank.
Cluster size distribution:

![Cluster size distribution diagram](image)

Figure 30: The cluster size distribution diagram.

Plots each 6 preset groups of cluster sizes against the number of clusters in that group.

Phylogtype Overlap:

![Phylogenotype overlap diagram](image)

Figure 31: The phylogenotype overlap diagram.

Plots a Venn Diagram for up to 4 selected samples, displaying the overlap in Phylogenotypes. In this Venn Diagram every selected sample is displayed as set of phylogenotypes and all subsets. A list of each subset can be accessed by clicking on the subset in the diagram. Scale can be switched...
between absolute and relative, used data between phylotypes and sequences.

**Classification Confidence:**

![Classification confidence diagram](image)

Figure 32: The classification confidence diagram.

Plots the classification confidence in per cent depending on the taxonomic rank as generated by the Rarefaction analysis. Distance, threshold and legend position can be modified.

**Alpha Diversity:**

![Alpha diversity scores per sample](image)

Figure 33: $\alpha$-diversity scores per sample.

Displays $\alpha$-diversity scores per sample depending on diversity score with a rarefaction option.
Sample Similarity:

Statistical diagram used for comparing the similarity and diversity of sample sets at set distance and distance function. Applicable distance functions are: Bray-Curtis distance, Anderberg distance, Hamming distance, Jaccard distance, Kulczynski distance, Kulczynski-Cody distance, Lennon distance, Memchi2 distance, Memchord distance, Memeuclidean distance, Mempearson distance, Ochiai distance, Sorenson distance, Whittaker distance, Jest distance, Sorest distance.

Hierarchical Clustering:

Hierachy of clusters: Divisive. All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
Cluster Abundance:

Displays the number of sequences distributed over all clusters.

Rank Abundance:

Displays relative species abundance. It can also be used to visualize species richness and species evenness.
8 Appendix

8.1 Description of tools used in SnoWMAn pipelines

Here follows a detailed description of tools used in the different analysis pipeline, not only analyzing parts that find application in the SnoWMAn workflow but about their general construction. This shall give a more detailed understanding about classification processes and different possibilities in approaching an analysis.

8.1.1 mothur

mothur was initiated by Dr. Patrick Schloss and his software development team in the Department of Microbiology & Immunology at The University of Michigan with the goal of creating a single piece of expandable open source software for use in the microbial ecology community. Its flexible algorithms, calculators and visualization tools offer an advantage over similar tools.

mothur aims to be a single resource to analyze molecular data and offers fast analysis and visualization to describe $\alpha$- and $\beta$-diversity.

For prokaryotic systems the use of the 16S rRNA gene is advantageous for the use of mothur because it can be reasonable aligned and is useful for distinguishing species. Depending on which type of microbiota is the target of an analysis the use of mothur may provide certain difficulties. For example the determination of OTUs for different kind of fungi may prove difficult (clustering OTUs based on multiple sequence alignments) and other programs may provide better alternatives. How does mothur work?

- First the user can clean up and sort the datasets of interest via different preprocessing procedures. It is possible to trim off primer sequences and barcodes, and further to use the barcode information to generate a group file (e.g. sorting sequences by barcode).
- After sorting, trimming and scrapping the provided sequences mothur can collapse all the identical sequences and will generate a list of unique sequences. Redundant sequences must have matching nucleotides and sequence length.
- mothur has an integrated tool to align sequences rapidly and can process huge amounts of data. It needs however an template alignment to work with.
- By comparing the closest match of the first half of our sequences to the template alignment and the closest match of the other half of our sequences to the template alignment mothur provides an excellent way to check for chimeras.
- After removing chimeras, cleaning and sorting the files a distance calculation is possible through three different types of analysis:
  - Parsimony: This generates random trees from a pool of ”species” and compares the number of changes in tree topology. According to maximum parsimony the shortest possible tree that explains the data is considered best.
  - Unifrac: mothur provides a weighted Unifrac algorithm to make a quantitative $\beta$-diversity measurement. This is however not a reliable approach to measuring similarity and should be applied for hypothesis testing.
  - Libshuff: Can describe whether several communities have the same structure using the Cramer-von Mises test. This indicates the probability that the communities have the same structure by chance. A correction for multiple comparisons must be applied.
• Various options for clustering sequences into OTUs or to bin them into phylotypes are provided. Phylogenetic trees can be created also. This opens room for different OTU-based analysis options like calculating diversity estimates or $\beta$-diversity statistics [?, ?]. Additional information can be obtained from http://www.mothur.org/.

8.1.2 RDP

Due to the conserved regions of 16S rRNA it is possible to determine phylogenetic relationships between organisms. The Ribosomal Database Project provides the user with aligned and annotated rRNA gene sequences along with analysis tools and a framework for this data. This includes bacterial and archaeal 16S rRNA sequences, and fungal 28S rRNA sequences [?]. How does RDP work?

• For processing with RDP [?] at least one sequence file and one forward primer are needed. This sequence files should be in FASTA format. Optionally a tag file can be provided [?, ?].

  – Sequence File: Contains FASTA formatted nucleotide reads.
  – Primer File: Contains Primer sequences. The entered Orimer sequences are for the target region only and do not include barcode or adapter regions of the primer.
  – Tag File: Is an optional file that organizes samples based on user-defined nucleotide tag sequences. It is a tab-delimited text file with a tag sequence, followed by a sample name in each line.

• After splitting and trimming the sequences different tools and filter types can be applied:

  – Hierarchy Browser allows navigation through the sequence data through a number of applicable display hierarchies and different search tools as well as filter [?].
  – RDP Classifier places sequences in the RDP hierarchy to give a initial taxonomic placement for the selected sequences. Further filtering is possible with user defined threshold levels that only allow assignments to be displayed if the calculated confidence level is above the same [?].
  – Sequence Match finds similar sequences to the user input without prior alignment. This makes it more accurate than for example BLAST at finding closely related rRNA sequences [?].
  – Probe Match is a program used for the detection of cross hybridization [?].

• After merging the data into one file a check for duplicates and chimeras is executed and the affected parts removed followed by sequence selection and alignment processes [?].

• Hierarchical clustering is used to group together object that are close to each other. Key to this analysis is repeated calculation of distance measures between objects, and between clusters once objects begin to be grouped into clusters [?].

• Through the complete linkage clustering (farthest neighbor method) distance between clusters in hierarchical cluster analysis is calculated which allows to create a cluster file based on the aligned sequence [?].

• Various formatting options for the clustered files are available for example the Rarefaction of the obtained data to create tab delimited files for further processing in Excel [?].

Additional information can be obtained from https://rdp.cme.msu.edu/.
8.1.3 UCLUST

UCLUST [?] is an alignment and search algorithm that allows the clustering of millions of nucleotide or amino-acid sequences based on their sequence similarity and is based on two criteria.

- A similarity threshold $T$ is defined. The first criteria states that any given cluster’s centroid sequence has a similarity value smaller to another centroid of a sequence.
- The second criteria states that any member sequence in a given cluster will have a similarity equal or greater to $T$ to the cluster’s centroid sequence.

The algorithm is implemented in a program of the same name. With UCLUST the order of the provided sequences will have an impact on the resulting clusters and their quality. Because of this the sorting of sequences before executing the algorithm is advised. With this UCLUST is equipped with capabilities of OTU assignment, creating non-redundant gene catalogs, taxonomic assignment and phylogenetic analysis [?, ?, ?]. How does UCLUST work?

- At the very core of UCLUST stands the search through a set of sequences. A global identity from all letters of both sequences is calculated. A sequence matches one in memory if the identity is high enough [?].
- Per prior definition UCLUST stops searching the set when it finds a match (usually the best match first) or if needed all matches. UCLUST also stops searching if it doesn’t find a match. Here the order of the input sequences plays a important role [?].
- A cluster owns a representative sequence called seed. Each sequence in the cluster matches the seed according to the second criteria [?].
- sequences are processed in input order and assigned to cluster if a match is found. If no match is found the input sequence becomes seed to a new cluster [?].
- Only seeds are stored into memory. This results in an advantage for large datasets because of the lesser needed storage capabilities but can result in non-seed sequences in the same cluster to fall below the identity threshold [?].
8.1.4 BLAT

BLAT [?] is a pairwise sequence alignment algorithm developed to assist in the assembly and annotation of the human genome [?]. Its main purpose lies in the alignment of sequences against the human genome sequence much faster than comparable tools. The BLAT algorithm analyzes and compares biological sequences such as DNA, RNA and Proteins in order to establish homology. With this process biological functions of genetic sequences can be discovered. It aims to rapidly find short sequences which are more likely to be homologous and then extends and aligns the homologous areas [?, ?, ?]. How does BLAT work?

- As stated above BLAT is used to find homologous regions in respect to a query sequence.
- First BLAT searches for short elements in a database and the query sequence that have a certain number of matching elements. These alignment seeds are then extended in both directions of the sequences in order to form high-scoring pairs [?]. Because of its indexing approach BLAT is able to search large genomic and protein databases for similarities in a very short amount of time. This is possible through a hash-table (indexed list) of the target-database in memory which reduces the time required for comparison.
- In particular this index is created by taking the coordinates of all the non-overlapping k-mers (part of sequences with k letters) in the target database, except for highly repeated k-mers. BLAT compares a list of overlapping k-mers of the query sequence with the target database and notes the matches between sequences [?].
- There are different search strategies and stages:
  1. Requires a perfect match between query and database occurs when two k-mers are exactly the same. With this approach arise several considerations and problem because to achieve a high sensitivity a small k-mer size is necessary but also causes the ratio of false positive hits to rise [?].
  2. At least one mismatch between two k-mers is allowed which decreases the amount of false positives and allows larger k-mer sizes. A larger k-mer size means less
taxing calculations. This strategy is very effective in indentifying small homologous regions [?].

3. Requires multiple perfect matches which are in close proximity to each other. This method is capable of taking into consideration small insertions and deletions within the homologous regions [?].

- Depending on the area of application different strategies are applied by BLAT. The indexing approach and search algorithm make BLAT to a much faster option than for example the BLAST [?] algorithm.

8.1.5 JGAST

JGAST [?] is an approach based on the principles of the "nearest neighbour" algorithms and can be seen as an improved GAST (Global Alignment for Sequence Taxonomy). The GAST [?] algorithm compares small subunits of the ribosomal RNA gene (SSU rRNA) with a set of reference sequences of known taxonomy. Taxonomy is then assigned in accordance to the nearest reference sequence. The alignment distance between query and reference is recorded as a measure of similarity. Several different RNA gene reference databases can be used (in this case BLAST [?]). Depending on the technology used, the read length of the rRNA gene varies (about 100–450pb) and although next generation sequencing produces extensive reads they are short compared to the full gene. Therefore it is important to select the right rRNA region for sequencing [?].

Massively parallel pyrosequencing of the hypervariable regions allows assembly of small genomes or detection of target variants with high accuracy. How does GAST work?

- The sequence file in FASTA format is split in smaller sequence parts. Through the hypervariable-region specific primers the microbial sample is sequenced.
- Each tag has the primers trimmed and quality filters applied.
- Through the BLAST algorithm each tag is compared with a set of reference sequences and then aligned against the top 100 BLAST hits.
- The matches having a minimum pairwise distance to the tag are selected.
- For this best match all source sequences are selected.
- A consensus agreement of >= 66% of the selected sources is calculated.
- The consensus taxonomy is applied to the tag.
