Amplicon-based metagenomics analysis using next generation sequencing

on behalf of: Prof. Reni Kalfin and Dr. Stefan Panaiotov
Bulgarian Academy of Sciences
Institute of Neurobiology
Ste. Acad. G. Bonchev 23
1113 Sofia
Bulgaria

Project 17095

October 27, 2017

Contact:
Dr. Christin Mieth (Project Manager)
christin.mieth@imgm.com

IMGM Laboratories GmbH
Bunsenstr. 7a
82152 Martinsried, Germany

Tel:	+49.89.4524667.0
Fax:	+49.89.4524667.410
Web:	www.imgm.com

Table of contents

1 Study overview

In the present study, phylogenetic 16S analysis was carried out in human clinical samples. For this purpose, isolated DNA from 60 human clinical samples were submitted to IMGM for phylogenetic classification by next generation sequencing-based quantification of bacterial and fungal load.

For sequencing library preparation, amplicons covering V3-V4 hypervariable regions of the bacterial 16S rRNA and ITS-2 hypervariable region in between the fungal 5.8S and 26S rRNA were generated, respectively. Two amplicon libraries were prepared from all PCR products of the 16S and ITS samples, respectively and sequenced together on the Illumina MiSeq® next generation sequencing system (Illumina Inc.).

The resulting 2 x 300 bp reads were demultiplexed, quality controlled and merged into continuous reads. Further bioinformatics analysis including clustering, phylogenetic analysis and alpha and beta diversity calculation was performed with the CLC genomics workbench and its microbial genomics module.

Processed data as well as raw data were transferred to the customer after data analysis.

2 Abbreviations and nomenclature

bp	Base pairs
dsDNA	Double stranded DNA
Gb	Giga bases
gDNA	Genomic DNA
NTC	No template control
OTU	Operational taxonomical unit
PCR	Polymerase chain reaction
PE	Paired end
PF	Passed filter
PosC	Positive Control
QC	Quality control
SAV	Sequence Analysis Viewer
SBS	Sequencing by synthesis
ssDNA	Single stranded DNA
SPRI	Solid phase reversible immobilization
TE	Tris EDTA
TS	target specific

3 Material and methods

3.1 Samples

Isolated DNA from 60 human clinical samples were entrusted to IMGM Laboratories. They were delivered at IMGM Laboratories on Sptember, 05th and 13th 2017. The samples were stored at -20 °C upon arrival.

In accordance with IMGM's quality management system, IMGM internal IDs were assigned to all study samples, in order to guarantee unambiguous sample identification throughout the analysis process. Table 1 lists these IDs together with the sample name and additional sample information provided by the customer. Within the table, the order of the single samples follows their chronological order indicated by the customer.

Table 1: Sample information

IMGM-internal sample ID	customer sample name
17095_0001*	O33*
17095_0002*	O33*
17095_0003	A1
17095_0004	A11
17095_0005	A2
17095_0006	A22
17095_0007	A3
17095_0008	A33
17095_0009	A4
17095_0010	A44
17095_0011	A5
17095_0012	A55
17095_0013	A6
17095_0014	A66
17095_0015	A9
17095_0016	A99
17095_0017	B1
17095_0018	B11
17095_0019	B2
17095_0020	B22
17095_0021	B3
17095_0022	B33
17095_0023	B4
17095_0024	B44
17095_0034	AB22
17095_0035	AB3
17095_0036	AB33
17095_0037	AB4
17095_0038	AB44
17095_0039	AB5
17095_0040	AB55
17095_0041	AB6
17095_0042	AB66
17095_0043	AB7
17095_0044	AB77
17095_0045	O1
17095_0046	O11
17095_0047	O2
17095_0048	O22
17095_0049	O3
17095_0050	O33
17095_0051	O4
17095_0052	O44
17095_0053	O5
17095_0054	O55
17095_0055	O6
17095_0056	O66
17095_0057	O7

IMGM-internal sample ID	customer sample name
17095_0025	B5
17095_0026	B55
17095_0027	B6
17095_0028	B66
17095_0029	B7
17095_0030	B77
17095_0031	AB1
17095_0032	AB11
17095_0033	AB2
17095_0058	O77
17095_0059	GMG
17095_0060	VSV
17095_0061	VHA
17095_0062	COST
17095_0063	K – water
17095_0064	K xp - control
	A8**
	A88**

* Samples excluded due to unclear sample identity.
** Samples excluded due to missing matching sample. No IMGM Sample ID assigned.
*** Sample announced, but not received. No IMGM Sample ID assigned.

3.2 Sequencing analysis for phylogenetic classification

3.2.1 Amplification strategy

The amplification strategy combines amplicon generation with library preparation for Illumina sequencing. The amplicon tagging scheme, shown in Figure 1, is based on a combination of an inner target-specific (TS) primer pair and an outer index primer pair. The inner TS primers bind to a template, such as the native genomic target region or a PCR product amplified from it. It includes a universal tag sequence which anneals to the partly complementary sequence of the outer index primer pair during a second, limited cycle PCR. Besides the complementary tag the outer primer pair comprises sequencing primer binding sites, indices for multiplexing and sequencing adapters. By incorporating unique sample-specific index combinations, all singleplex PCR products generated by PCR can be multiplexed to be sequenced in a single sequencing experiment.

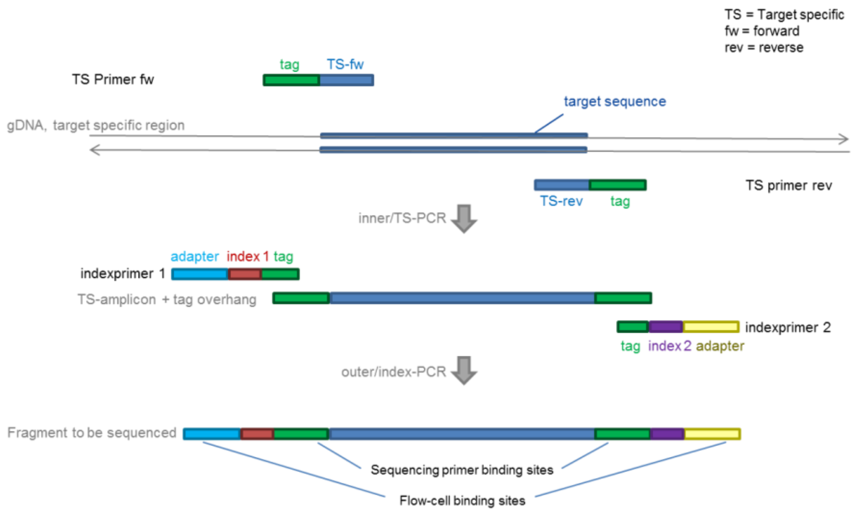


Figure 1: Scheme of the 16S amplification strategy

3.3 16S and ITS-2 rRNA amplification

In accordance with the customer, the primer pair 314F and 805R (1) was chosen to cover variable regions 3 to 4 of the 16S gene with a 465 bp target specific fragment. Additionally, the ITS3 and ITS4 (2) primer pair was used to amplify the ITS-2 region in between the 5.8S and 26S rRNA. The detailed information of the chosen primer pairs is shown in Table 2.

Table 2: 16S rRNA primer information

Primer name	Primer sequence (5’ -> 3’)	Primer Length (bp)	Reference	Amplicon size
341F	CCTACGGGRSGCAGCAG	17	(1)	465 bp
805R	GACTACHVGGGTATCTAATCC	21	(1)	
ITS3	GCATCGATGAAGAACGCAGC	20	(2)	400 bp
ITS4	TCCTCCGCTTATTGATATGC	20	(2)	

The unique sample-specific index combinations, including index sequences used for multiplexed sequencing of all samples in one sequencing run, are provided in file ***17095_Indices.xlsx*** (stored in folder: ***17095_sequencing***)

3.3.1 Amplicon quality check using agarose gel electrophoresis

An aliquot of each final PCR product including a NTC and a positive control was run on a 2% agarose gel (Midori Green-stained) to analyze the quality of the generated amplicons and to evaluate the expected amplicon size.

3.4 Library generation

3.4.1 Amplicon purification

The amplicons were separately purified using solid phase reversible immobilization (SPRI) paramagnetic bead-based technology (AMPure XP beads, Beckman Coulter) with a Bead:DNA ratio of 0.7:1 (v/v) for 16S amplicons and 0.65:1 (v/v) for ITS amplicons.

3.4.2 Library normalization and pooling

The purified PCR products were normalized to equimolar concentration with the SqualPrep™ Normalization Plate Kit (Life Technologies). An amount of at least 250 ng DNA is needed for every amplicon in order to achieve the best result during this normalization step.

Subsequently, one library pool was generated from all samples by mixing the same volume of each of the equimolar normalized amplicon samples. The pool was again purified and simultaneously concentrated using the SPRI method described in chapter 3.4.1 with a Bead:DNA ratio of 0.7:1 (v/v)

3.4.3 Integrity control and quantification of the library pool

The High Sensitivity DNA LabChip Kit (Agilent Technologies) was used on the 2100 Bioanalyzer (Agilent Technologies) to analyze the integrity and peak distribution of the library pool of purified amplicons before and after pool purification.

The library pool was quantified using the highly sensitive fluorescent dye-based Qubit® dsDNA HS Assay Kit (Invitrogen). In brief, 1 µl of each sample was used to determine dsDNA concentration (ng/µl) in comparison to a given standard provided with the kit. The DNA concentration was determined by creating a linear trend line and applying the mathematical equation of linear regression.

Concentrations were calculated according to the following equation:

$$\text{conc. [nM]} = (\text{conc. [ng/µl]} * 10^6) / (660 \text{ g/mol} * \text{average library size [bp]})$$

Where 660 g/mol is the average molecular weight of nucleotide pairs [g/mol], and average library size [bp] is the average length of the fragments generated in the experiment.

3.4.4 Library denaturation and preparation for sequencing

The final library pool was subjected to a denaturation step using NaOH, to ensure the presence of single stranded DNA (ssDNA) fragments for cluster generation. Thus, the library consisted of ssDNA fragments with sequencing adapters and indices. The library pool was diluted and used for loading on the MiSeq® system for cluster generation and sequencing.

Sequencing was performed at final concentrations of 5 pM and with a 10% PhiX v3 control library spike-in (Illumina Inc.)

3.5 Cluster generation and sequencing

For cluster generation and sequencing, a MiSeq reagents kit 500 cycles v3 (Illumina Inc.) was used. Cluster generation and sequencing were carried out under the control of the MiSeq®-inherited **MiSeq Control Software (MCS) v2.5.0.5**.

Before sequencing, cluster generation was performed. Clusters represent areas on the MiSeq® flow cell with clonally amplified DNA library fragments. The cluster generation of loaded fragments is carried out by a two-dimensional amplification of bound fragments (bridge amplification). Fragments are immobilized on the flow cell, involving two adapters attached at opposing ends of the fragment and complementary slide-anchored primers. By several cycles of flipping and binding of the immobilized fragments to the surrounding primer lawn, followed by amplification, approximately 1’000 copies of the original fragments are produced and localized in a tight cluster. A scheme of the bridge amplification reaction on the MiSeq® flow cell is shown in Figure 2.

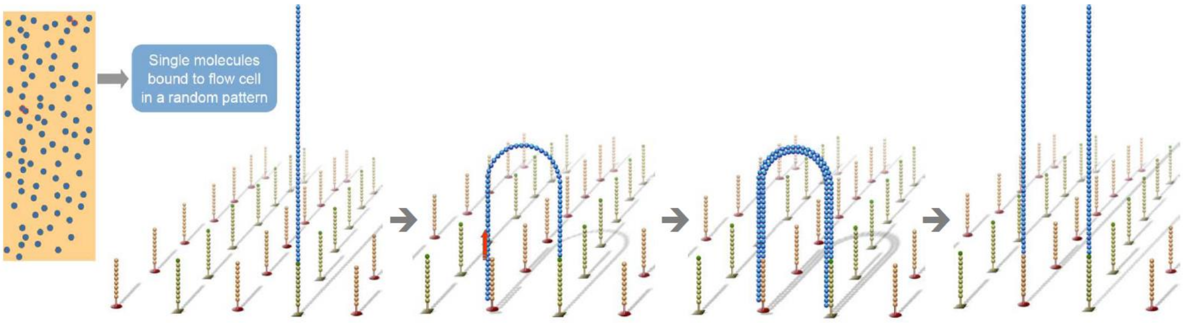


Figure 2: Schematic overview of the cluster generation by bridge amplification (adapted from Illumina Inc.)

After cluster generation, sequencing primers hybridize to the adapter sequences at the end of the fragments and sequencing is carried out.

The sequencing principle is based on the sequencing by synthesis (SBS) approach. During each sequencing cycle, the flow cell is flooded with all four nucleotides (A, T, G, C). Those are provided as fluorescently labeled terminator dNTPs, which are labeled with different fluorophores. Thereby, one base per cycle is attached to the growing antisense-strand during strand elongation by the DNA polymerase, starting from the sequencing primer. Washing away of unbound nucleotides and signal collection from the bound fluorescing nucleotides from the different clusters after each cycle allows identification of incorporated bases.

In the present study, bidirectional sequencing was performed, starting first at the end of the sense strand (read 1) and subsequently at the end of the complementary strand (read 2). Both reads have a length of 300 bases, finally producing 600 bases of sequence information in 2 x 300 bp paired-end (PE) reads.

3.6 Bioinformatic data analysis of phylogenetic classification

3.6.1 Software applied for bioinformatic data analysis

The Illumina software **MiSeq® Reporter (MSR) v 2.5.1.3** on the MiSeq® system and the **Illumina Sequence Analysis Viewer (SAV) v 2.1.8** were used for imaging and evaluation of the sequencing run performance.

Primary data analysis and QC, including signal processing and de-multiplexing, was performed using the MiSeq® inherited MSR software and the **CLC Genomics Workbench v9.5.3**.

In-depth data analysis, merging and trimming of reads as well as phylogenetic clustering, and calculation of alph diversity as well as fold change calculation and graphical representation was performed using the **CLC Genomics Workbench v9.5.3** inherited microbial genomics module, **Microsoft Excel 2010** and inhouse-developed command line tools.

3.6.2 Quality control of sequencing run performance

After completion of the sequencing run, technical quality parameters were evaluated using the **SAV** software.

The main technical quality specifications for a successful sequencing run, according to Illumina Inc., are given in Table 3.

Table 3: MiSeq® technical sequencing run quality specifications according to Illumina Inc.*

Parameter	Q 30 bases	Optimal cluster density	Sequence yield	Filter passed clusters
Illumina Specification	> 70%	1200–1400 K/ mm2	13.2–15 Gb 44–50 Mio. Reads	< 44–50 Mio. (< 100%)

*specifications are effective for an Illumina PhiX library and cluster densities between 1200 - 1400 k/mm^2 that pass filtering. Actual performance may vary based on sample characteristics.

The main quality parameter is the percentage of Q 30 bases. Each base is assigned with a quality Phred score (Q) which is defined as a property logarithmically related to the base-calling error probability P as follows:

$$Q = -10 \log_{10} P$$

For example a base with a Q score of 30 has a chance of 1 in 1’000 (accuracy 99.9%) to have been called incorrectly and a quality score of 40 means a base call accuracy of 99.99%.

The optimal cluster density resulting from an optimal flow cell loading yields an optimal sequence output. A low cluster density leads to a low sequencing yield, whereas a too high cluster density with narrow spacing of clusters leads to a loss of reads due to filtering out of mixed cluster reads. Furthermore, sample characteristics can affect cluster density, and different types of libraries lead to different cluster densities even if the same amount of DNA is loaded.

The sequencing yield is likewise dependent on the cluster density and thus on the loading of the MiSeq® flow cell and sample characteristics. A sequencing yield, given in Table 3, is expected with an excellent performing sample (PhiX control library) and an optimal cluster density in a sequencing run using MiSeq v2 reagents for 2 x 300 bp reads. Different sample types may lead to lower or higher cluster densities and thereby to lower sequencing output.

Filter passed clusters show the number of reads finally available for in-depth bioinformatic analysis, after filtering and exclusion of clusters with mixed fragment content or with non-recognized indices. Again, this parameter is dependent on the sequencing yield and thereby on the cluster density caused by sample characteristics.

3.6.3 Primary data analysis and QC

Image and signal processing was carried out by the Illumina MiSeq® inherited **MSR** software packages applying the FastQ only processing pipeline. This processing pipeline allows for 3’-end trimming of adapter sequences which is recommended by Illumina Inc. De-multiplexing of all passed filter reads was performed using indices and corresponding sample IDs. Single read 1 and read 2 ***.fastq*** files, containing quality values and sequence information were generated.

Demultiplexed raw data (read 1 and read 2 per sample) were imported into **CLC Genomics Workbench** as paired-end (forward-reverse) reads. A possible distance between 1 and 10’000 bases was allowed. Failed reads, indicated by a flag within the quality score header information inside the ***.fastq*** files, specifying if a read has passed the MiSeq®-inherited quality filters or not, were removed from the data set during this process.

Read QC was performed with the **CLC Genomics Workbench** “Create Sequencing QC report” tool.

3.6.4 In depth bioinformatics analysis for phylogenetic classification

In depth bioinformatics analysis was performed using the **CLC Genomics Workbench** and the implemented microbial genomics module.

As a first step read 1 and read 2 sequences from every sample were merged using the overlapping sequence information and taking the minimum overlap, potential mismatches, gaps and unaligned ends into account. The two reads were merged including overlaps of minimum 20 bp without any mismatch and maximum unaligned end mismatches of 2 bp.

Trimming of reads was performed according to TS primer sequences, base quality and read length, whereby a probability quality limit of 0.05 was applied to ensure high quality data for subsequent analysis. To guarantee similarity and a sufficiently high level of sequence information for phylogenetic classification, sequences <320 bp were discarded for 16S analysis and longer sequences were trimmed down to this length. For ITS analysis, a fixed length trimming of 150 bp was applied to the reversed sequences to make sure that the highly variable ITS2 region was kept for further analysis, while the homologous 5.8 rRNA region was discarded.

3.6.4.1 OTU Clustering

The remaining sequences were clustered at a 97% identity threshold defining operational taxonomic units (OTU), according to the taxonomy of the SILVA 128 16S rRNA sequence database (www.arb-silva.de) and the Unite database version 7.2 with singletons (https://unite.ut.ee/repository.php) for 16S and ITS analyses respectively. Chimeric sequences, representing PCR and sequencing artefacts, were filtered out and discarded during this step.

Out of each cluster one reference sequence of an OTU was defined and is represented with one line each in the full result table.

For *de novo* OTUs with the highest combined abundance, an additional BLAST search was applied to identify bacterial and fungal species that might have been missed based on the database annotation used. Thereby, the NCBI BLASTN application was used comparing query sequences from *de novo* OTUs detected with the nucleotide collection (nr) database at NCBI considering all organisms.

A graphical overview on the taxonomy results across all samples was generated as a pie chart. The taxonomic assignment path is shown from the highest phylogenetic level (kingdom bacteria) in the middle of the circle via order, class, family, genus down to the species level at the outer end of the circle.

3.6.4.2 Alpha and Beta Diversity

Two levels of diversity are typically considered in microbial ecology: alpha- and beta-diversity. Alpha-diversity describes the number of species (or similar metrics) in a single sample, whereas beta-diversity compares the number of species (or similar metrics) across samples (3).

For alpha and beta diversity measurements, first a multiple sequence alignment was performed including all sequences of the sample using the MUSCLE (multiple sequence comparison by log-expectation) algorithm and based on this alignment a phylogenetic tree was calculated based on a maximum likelihood approach.

Alpha Diversity

Rarefaction curves represent the species richness for a given number of individual samples and can be found as ***.png*** image files in the subfolder ***17095_Alpha-div***.

The number of different OTUs is plotted against the number of tags sampled, in our case the amount of reads sequenced for each sample. If the curve becomes flatter with increasing amount of reads, then only few additional OTUs are likely to appear after deeper sequencing.

The rarefaction analysis is done by sub-sampling the OTU abundances in the different samples at different depths. The maximum depth is set to the number of reads of the most abundant sample. The number of different depths to be sampled was specified to 5’000. At each depth, the algorithm was set to subsample the data 100 times without replacement.

The Chao-1 estimator (S) provides insight in the number of unsampled species due to undersampling. It can be calculated according to the equation:

$$S_{estimate} = D + F_1^2 / 2F_2$$

Additionally, the Shannon entropy (H) quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset. It is calculated as follows:

$$H = - \sum_{(i=1-n)} (p_i * \log_2 p_i)$$

Where F_1 is the number of singletons sampled, and F_2 is the number of doublets; n is the number of OTUs; D is the number of distinct OTUs observed in the sample and p_i is the fraction of reads that belong to OTU i.

The more unequal the taxa abundances of the different samples the smaller becomes the corresponding Shannon entropy. If practically all abundance is concentrated to one type, and the other types are very rare (even if there are many of them), Shannon entropy approaches zero.

Additionally, the phylogenetic diversity was plotted and calculated based on the following equation:

$$PD = \sum_{(i=1-n)} b_i I\ (p_i > 0)$$

whith n presenting the number of branches of the phylogenetic tree, b_i presenting the length of branch i; p_i being the proportion of taxa descending from branch i and the indicator function I($p_i > 0$*) and I(*$p_i^B > 0$*) assuming the value 1 if any taxa descending from branch i is present in the sample or 0 otherwise.*

This estimator equations are provided graphically plotted (***.png*** image files) for the data of all samples including all subsampling steps. They allow researchers to have a good idea of how their limited sampling relates to the entire sampled population.

Beta diversity

Beta diversity examines the change in species diversity between ecosystems. The analysis was done in two steps. First, the distance between each pair of samples was estimated. Therefore, the calculations according to Bray Curtis and Jaccard were used, as follows:

$$\text{Bray Cutis } (B) = \left(\sum_{(i=1-n)} | x_i^A - x_i^B |\right) / \left(\sum_{(i=1-n)\,(} x_i^A - x_i^B) \right)$$

and

$$\text{Jaccard } (J) = 1- \left(\sum_{(i=1-n)} \min(x_i^A , x_i^B)\right) / \left(\sum_{(i=1-n)} \max(x_i^A , x_i^B)\right)$$

Where n is the number of OTUs and x_i^A and x_i^B are the abundances of OTU i in samples A and B, respectively.

The Bray–Curtis dissimilarity is used to quantify the compositional dissimilarity between two different samples, including the species abundances, and is calculated based on raw counts at each sample. The Jaccard coefficient measures similarity between sample sets, based on present or absent calls for each taxon and is defined as the size of the intersection divided by the size of the union of the sample sets.

Using these calculations, matrices were generated as a measure for the similarity (Jaccard) or dissimilarity (Bray-Curtis) between each possible sample pair and Principal Coordinate Analysis (PCoA) plots were calculated on these. This multidimensional scaling technique takes an input matrix giving dissimilarities between pairs of items and outputs a coordinate matrix which can be plotted on 2-dimensional or 3 dimensional scale. The PCoA ordination technique is similar to Principal Component Analysis (PCA) but has the advantage over that any ecological distance can be investigated.

Additionally, the Unweighted UniFrac was calculated as measure of the beta diversity giving comparatively more importance to rare lineages. The following equation is underlying the calculation:

$$d^{(U)}=(\sum_{(i=1-n)} b_i|I(p_i^A > 0)-I(p_i^B > 0)|)/ \sum_{(i=1-n)} b_i$$

with n presenting the total number of branches in the phylogenetic tree, b_i being the length of branch i; p_i^A and p_i^B presenting the proportion of taxa descending from branch i for samples A and B, respectively and the indicator functions $I(p_i^A > 0)$ and $I(p_i^B > 0)$ assuming the value 1 if any taxa descending from branch i is present in samples A and B, respectively or 0 otherwise.

For the calculation of the beta diversity, metadata provided by the customer where incorporated. Thereby, a sample type grouping samples into four classes for which a similar microbial classification was expected were assigned. Furthermore, all samples were assigned with a category separating cultured and non-cultured samples from each other.

4 Results

4.1 Period of analysis

The experimental analyses of the present study were carried out in the period from September 19, 2017, through September 29, 2017.

4.2 Amplicon QC

4.2.1 16S amplicon generation

After amplification with target specific and index primers all PCR products were checked on 2% agarose gels for purity and product sizes.

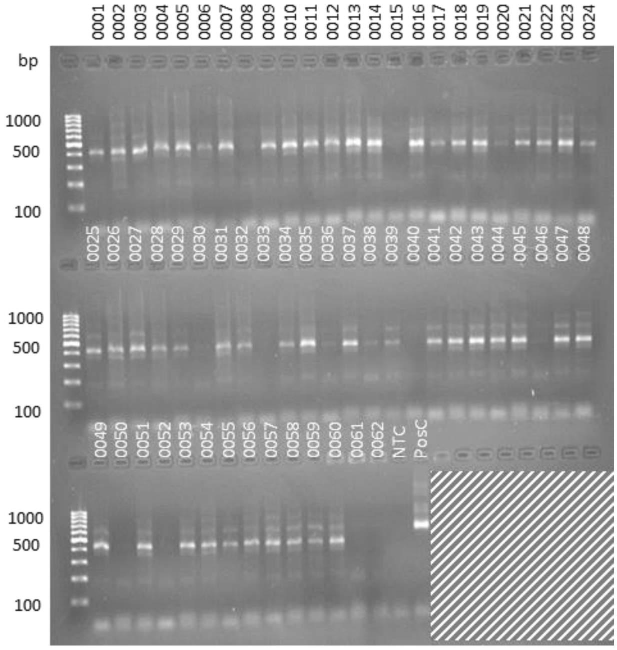


Figure 3: Agarose gel image for the 16S initial PCR reactions. Sample names are indicated with their last 4 digits only, prefix 17095_ is omitted. The GeneRuler^{TM} 100 bp DNA Ladder (Fermentas) was used as a size marker (*17095-16S-PCR01.png*).

Most samples showed good amplification results and were included in the amplicon library. Twenty samples (17095_0003, 17095_0008, 17095_0010, 17095_0017, 17095_0019, 17095_0022, 17095_0026, 17095_0030, 17095_0031, 17095_0032, 17095_0034, 17095_0035, 17095_0038, 17095_0040, 17095_0042, 17095_0048, 17095_0052, 17095_0054, 17095_0057, 17095_0061) showed no band or only very low amounts of PCR products. For these, amplification was repeated using 3 x more and 10 x less DNA input amounts. The result of the repeated PCR is shown in Figure 4.

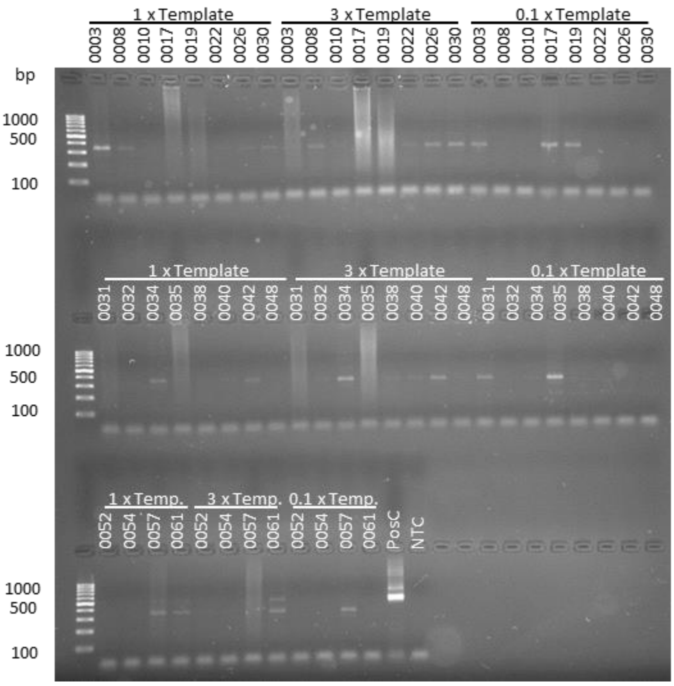


Figure 4: Agarose gel image for the repeated 16S TS PCR reactions of 20 samples. Sample names are indicated with their last 4 digits only, prefix 17095_ is omitted. The GeneRuler^{TM} 100 bp DNA Ladder (Fermentas) was used as a size marker (*17095_16S_PCR02.png*).

All of the repeated samples showed clear bands for at least one condition. As some samples showed still a low amount of PCR product on agarose gel, TS amplicons of good quality on agarose gel were pooled and purified using AMPure XP beads (Beckman Coulter) with an elution volume of 20 µl according to 3.4.1 (Table 4).

Table 4: Pooling and purification of TS PCR amplicons. Samples that showed a smear of genomic DNA on agarose gel were excluded. X) sample included; -) sample excluded

IMGM Sample ID	1 x Template PCR	3 x Template PCR	0.1 x Template PCR	µl used for Index-PCR
17095_0003	X	-	X	2
17095_0008	X	X	X	4
17095_0010	X	X	X	14.25
17095_0017	-	-	X	4
17095_0019	-	-	X	4
17095_0022	X	X	X	14.25
17095_0026	X	X	X	8
17095_0030	X	X	X	4
17095_0031	-	-	X	8
17095_0032	X	X	X	14.25
17095_0034	X	X	X	2
17095_0035	-	-	X	4

IMGM Sample ID	1 x Template PCR	3 x Template PCR	0.1 x Template PCR	µl used for Index-PCR
17095_0038	X	X	X	14.25
17095_0040	X	X	X	14.25
17095_0042	X	X	X	4
17095_0048	X	X	X	14.25
17095_0052	X	X	X	14.25
17095_0054	X	X	X	14.25
17095_0057	-	-	X	4
17095_0061	X	X	X	4

Volumes of purified TS PCR amplicons used for subsequent indexing PCR are indicated in Table 4.

The results of the repeated Index PCR are shown in Figure 5.

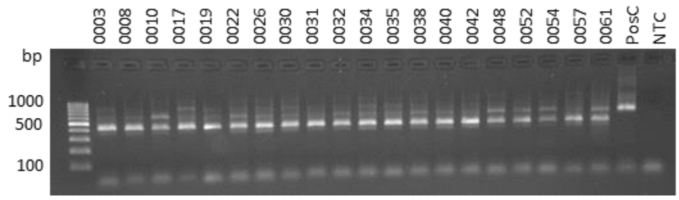


Figure 5: Agarose gel image for the repeated 16S Index PCR reactions of 20 repeated samples. Sample names are indicated with their last 4 digits only, prefix 17095_ is omitted. The Gene-Ruler^{TM} 100 bp DNA Ladder (Fermentas) was used as a size marker (*17095_16S_PCR03.png*).

All samples showed good amplification results and were included in the amplicon library.

4.2.2 ITS amplicon generation

After amplification with target specific and index primers all PCR products were checked on 2% agarose gels for purity and product sizes.

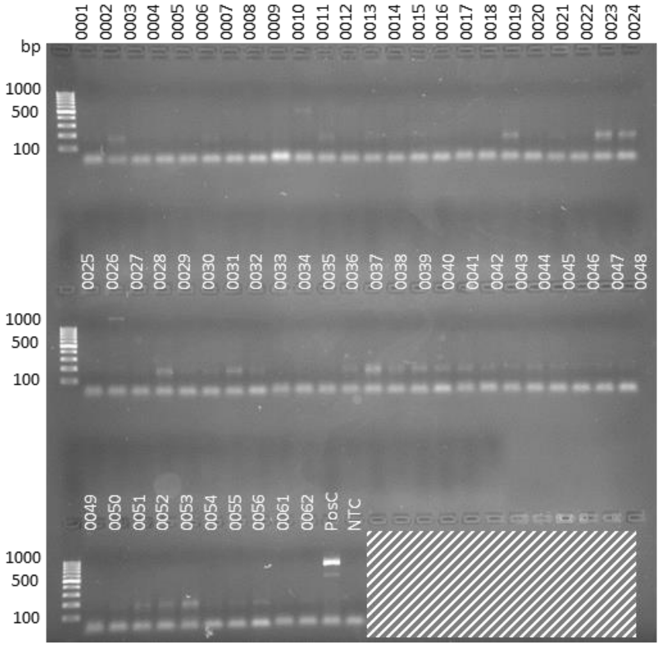


Figure 6: Agarose gel image for the initial ITS PCR reactions. Sample names are indicated with their last 4 digits only, prefix 17095_ is omitted. The GeneRuler^{TM} 100 bp DNA Ladder (Fermentas) was used as a size marker (*17095-ITS-PCR01.png*).

All samples showed poor amplification results (Figure 6). Therefore, the TS PCR was repeated with 10 more amplification cycles compared to the standard protocol (25 cycles). The result of the repeated PCR is shown in Figure 7.

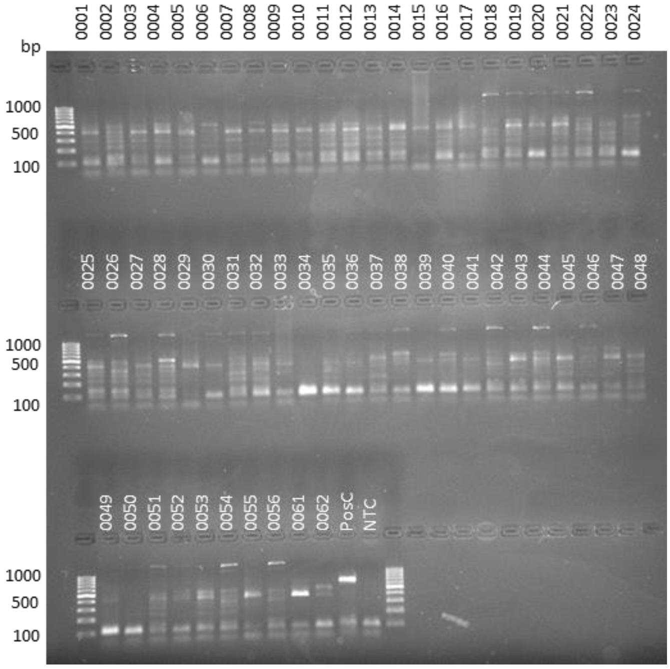


Figure 7: Agarose gel image for the repeated ITS TS PCR reactions of all samples. Sample names are indicated with their last 4 digits only, prefix 17095_ is omitted. The GeneRuler^{TM} 100 bp DNA Ladder (Fermentas) was used as a size marker (*17095_ITS_PCR02.png*).

Most samples showed multiple low molecular weight bands. Six samples (17095_0017, 17095_0019, 17095_0031, 17095_0035, 17095_0036, 17095_0051, 17095_0052) showed no band or only very low amounts of PCR products. For these, amplification was repeated using 3 x more and 10 x less DNA input amounts. The result of the repeated PCR is shown in Figure 8.

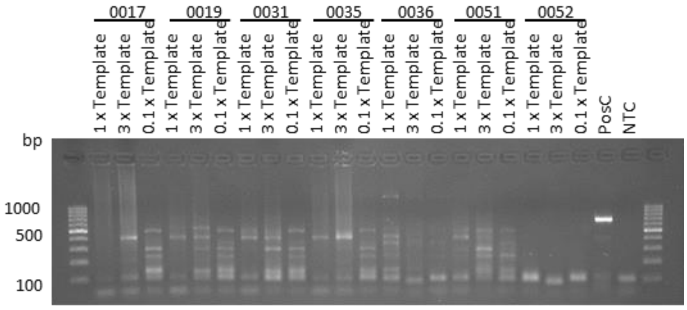


Figure 8: Agarose gel image for the repeated ITS TS PCR reactions of 20 samples. Sample names are indicated with their last 4 digits only, prefix 17095_ is omitted. The GeneRuler^{TM} 100 bp DNA Ladder (Fermentas) was used as a size marker (*17095_ITS_PCR03.png*).

A low molecular weight band pattern was obtained for all samples repeated. The pattern of bands differed depending on the amount of input DNA. This might indicate the amplification of unspecific PCR products. In accordance with the customer, it was decided to use the 0.1 x template TS PCR samples for further indexing PCR and library generation. Two exceptions were sample 17095_0036 of which the 1 x template PCR performed best and was used for indexing PCR and sample 17095_0052 of which the two 1 x template and the 3 x template PCRs were pooled. TS PCR reactions were purified using AMPure XP beads (Beckman Coulter) with an elution volume of 20 µl according to 3.4.1. The results of the indexing PCR are shown in Figure 9.

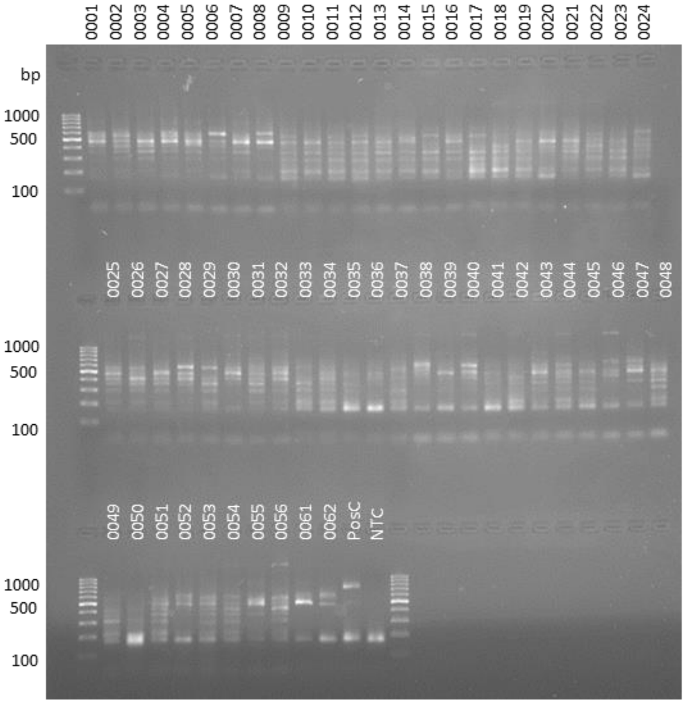


Figure 9: Agarose gel image for the repeated PCR reactions of twice failed samples. Sample names are indicated with their last 4 digits only, prefix 17095_ is omitted. The GeneRuler^{TM} 100 bp DNA Ladder (Fermentas) was used as a size marker (*17095_ITS_PCR04.png*).

Multiple low molecular weight bands were obtained for all samples. 16S and ITS amplicons were used for further library preparation.

4.3 Library generation

Amplicons were purified once using AMPure XP beads and normalized using the Sequal-Prep Kit. An amount of at least 250 ng DNA is needed for every amplicon in order to achieve the best result during this normalization step. All samples contained sufficient DNA amounts for Sequal-Prep normalization.

4.4 Integrity control of the library

One library pool was generated by combining all amplicons of each plate separately including additionally the NTC with the same volume and the positive control with 1/10 sample volume. The library pool was again purified and integrity and purity of the library pool before and after purification was checked using the HighSensitivity DNA LabChip Kit on the Bioanalyzer.

The electropherogramm of the purified amplicon library quality check is shown in Figure 10.

The library comprised clear amplicon peaks at the expected amplicon size and no signs for contamination with sequencing-compromising primer dimers in lower molecular ranges.

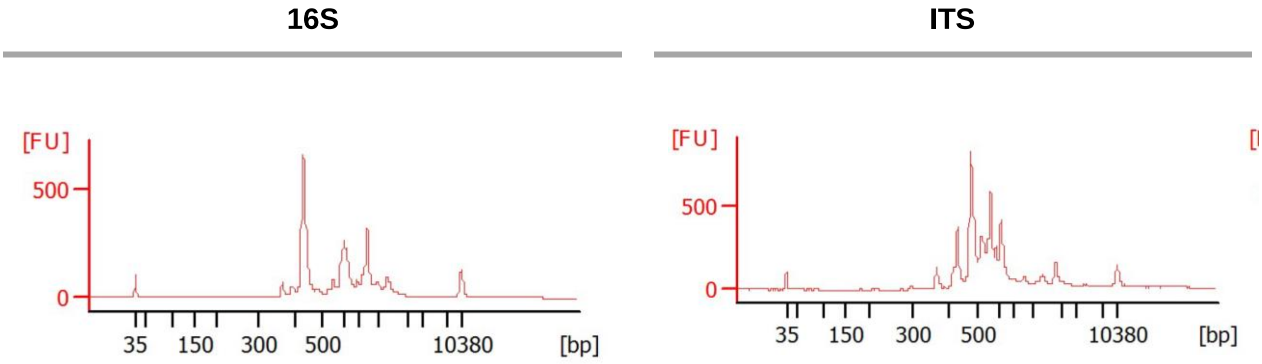


Figure 10: Integrity control of the purified library 17095 16S and ITS using the HighSensitivity LabChip Kit (*17095_High Sensitivity DNA Assay.pdf*)

Results of the library pool quality control before and after purification including electropherograms, virtual gel image, etc. were stored and are available as a separate file called ***17095_High Sensitivity DNA Assay.pdf.***

4.4.1 Quantification and finalization of the library

Library concentration and the number of molecules per µl were quantified as described in 3.4.3. The pool was diluted, denatured and used for the sequencing in a final concentration of 2 pM. A standard control library (*PhiX v3*, Illumia Inc.) spike-in of 10% was additionally used to ensure balanced nucleotide representation during the whole sequencing run.

4.5 Cluster generation and sequencing results

4.5.1 Cluster generation

Cluster generation as well as sequencing was carried out. The cluster generation resulted in a cluster density of 824 +/- 37 k/mm^2 (Table 5). This represents a little bit lower cluster density than the optimal cluster density indicated by Illumina Inc for standard libraries. (Table 3). However, for low diversity libraries as metagenomics amplicons a lower cluster density is recommended by Illumina to achieve high quality sequencing data.

4.5.2 Sequencing results

After image and signal processing for read 1 and read 2 the following two *.fastq-files for each sample were created, containing quality values and sequence information of the 2x300 bp read data:

- ***16S-17095-0###_S#_L001_R1_001.fastq.gz***

- ***16S-17095-0###_S#_L001_R2_001.fastq.gz***
- ***ITS-17095-0###_S#_L001_R1_001.fastq.gz***
- ***ITS-17095-0###_S#_L001_R2_001.fastq.gz***

The respective raw data are stored in the folder ***17095_RawData***.

Quality criteria according to Illumina Inc. were evaluated. The quality values for the performed sequencing run are summarized in Table 5.

Table 5: Sequencing run quality values

Parameter	Q30 bases	Cluster density	Sequencing yield	Filter-passed clusters
Obtained value	73.44%	824 +/- 37 k/mm^2	11.22 Gb	18.7 Mio (82.7%)

The major quality criterion of passed Q 30 bases of >70% (Table 3) was fulfilled by the sequencing run. The sequencing yield was with 11.22 Gb and 18.7 Mio reads below the Illumina specifications (Table 3). This underload was accepted for obtaining higher quality scores for each sequencing read instead.

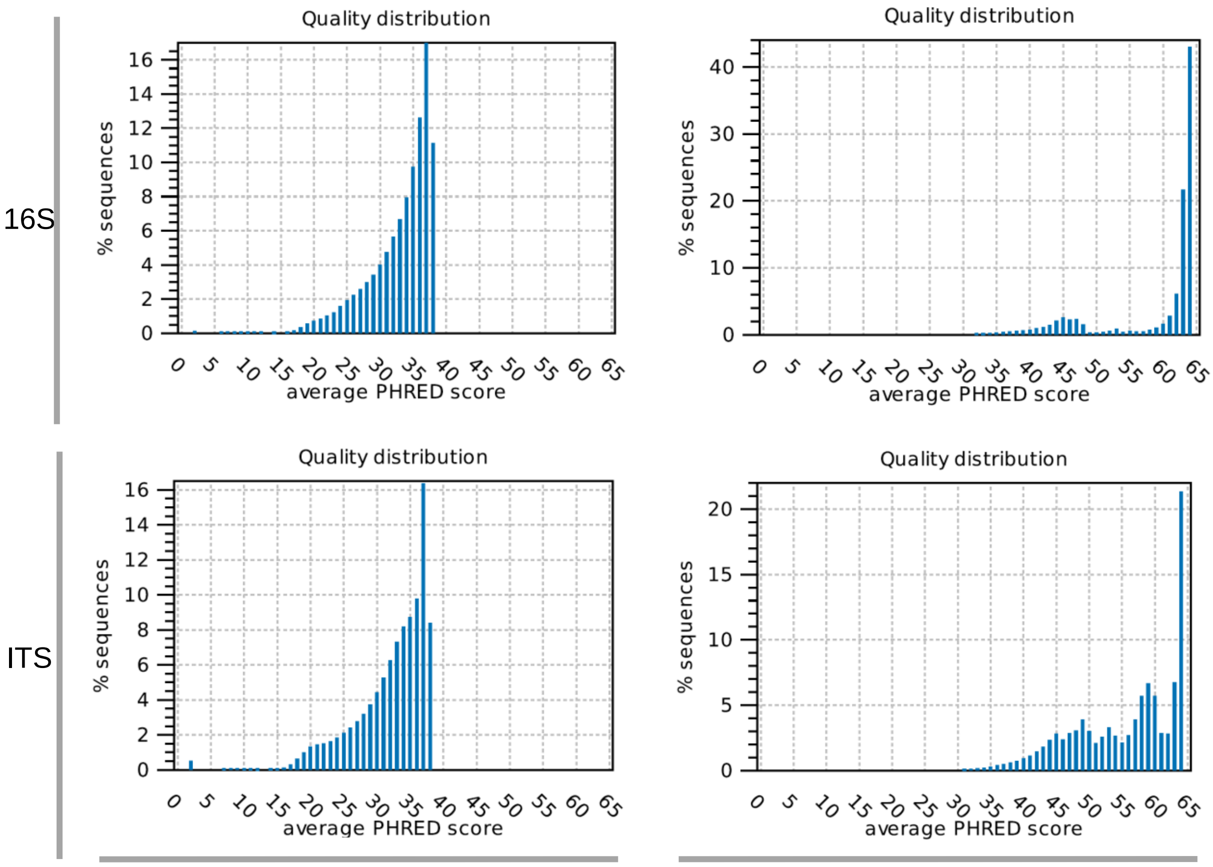
All reads from low quality clusters as well as mixed read clusters which did not pass quality criteria were discarded during the primary analysis pipeline. Filter-passed clusters were with 82.7% at a high level for the achieved cluster density.

These filter-passed clusters led to 3’029 - 291’663 reads per sample. However, the project specific controls 17095_0063 and 17095_0064 showed a low number of 5’048 and 3’029 reads in the 16S library as expected. After exclusion of these samples the minimum number of reads was 25’314 for sample 17095_0041 and therefore all other samples exceeded the expected 10’000 reads.

Read counts per sample are provided in file ***17095_PerSampleReadCount.xlsx***.

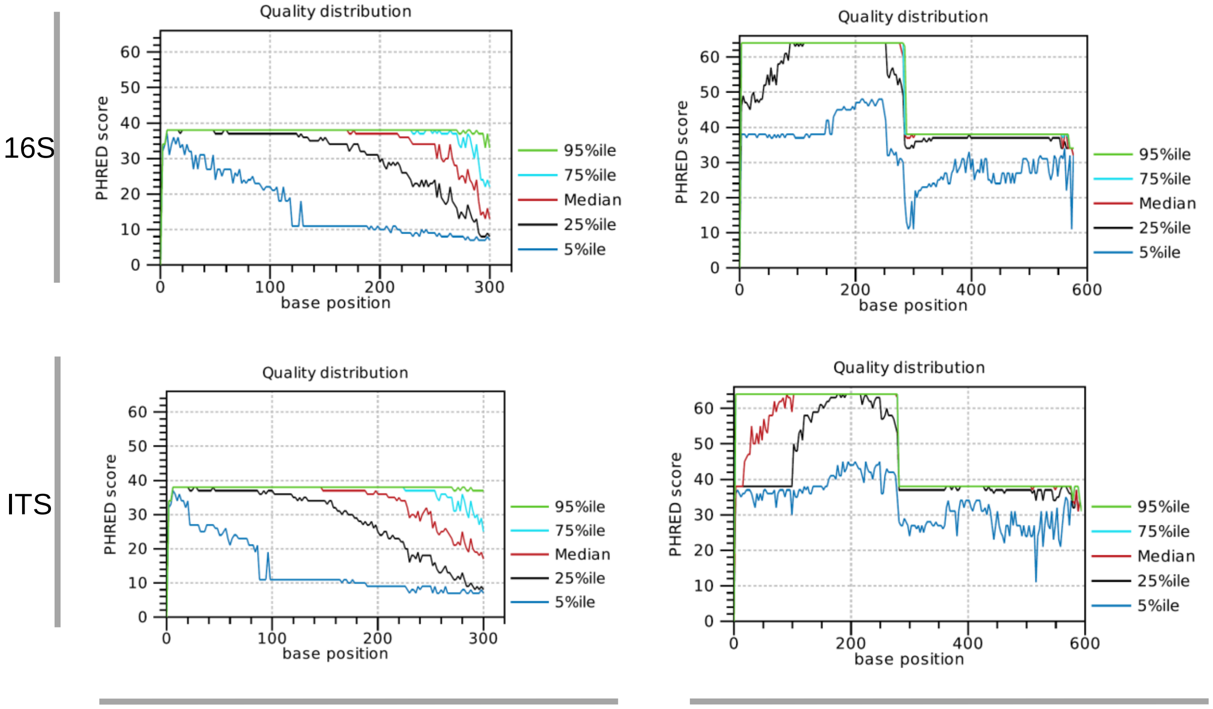
The paired-end FastQ files were imported into the **CLC Genomics Workbench** and were quality-checked before and after merging and trimming.

The sequences, available after trimming and merging show a high amount of high quality reads, suitable for downstream analysis (see Figure 11 and Figure 12).



Before merging and trimming After merging and trimming

Figure 11: Mean Q score distribution before and after merging and trimming of reads



Before merging and trimming After merging and trimming

Figure 12: Q score distribution along the reads before and after merging and trimming of reads

The summary of the QC is provided in the corresponding QC reports:

- ***17095-16S-Sequencing-QC-raw.pdf***
- ***17095-16S-Sequencing_QC_After _Merge+Trim.pdf***
- ***17095-ITS-Sequencing-QC-raw.pdf***
- ***17095-ITS-Sequencing_QC_After _Merge+Trim.pdf***

4.6 Phylogenetic classification results

4.6.1 OTU clustering

The phylogenetic analysis was performed with the merged reads per sample as described in chapter 3.6.4. The sequences which remained after alignment, quality trimming and filtering, were clustered at a 97% identity threshold. Each cluster represents one OTU, and one representative reference read was defined for each OTU. These OTU reference reads were used for further phylogenetic analysis. The results of the phylogenetic classification of the complete sample set are provided in absolute numbers in file ***17095-16S_OTU-Table-full.xlsx*** and ***17095-ITS_OTU-Table-full.xlsx*** and as relative abundances in files ***17095-16S_OTU-Table-relative.xlsx*** and ***17095-ITS_OTU-Table-relative.xlsx***. The meaning of the various column headers are given in Table 6.

Table 6: Colum header description of the phylogenetic classification result table

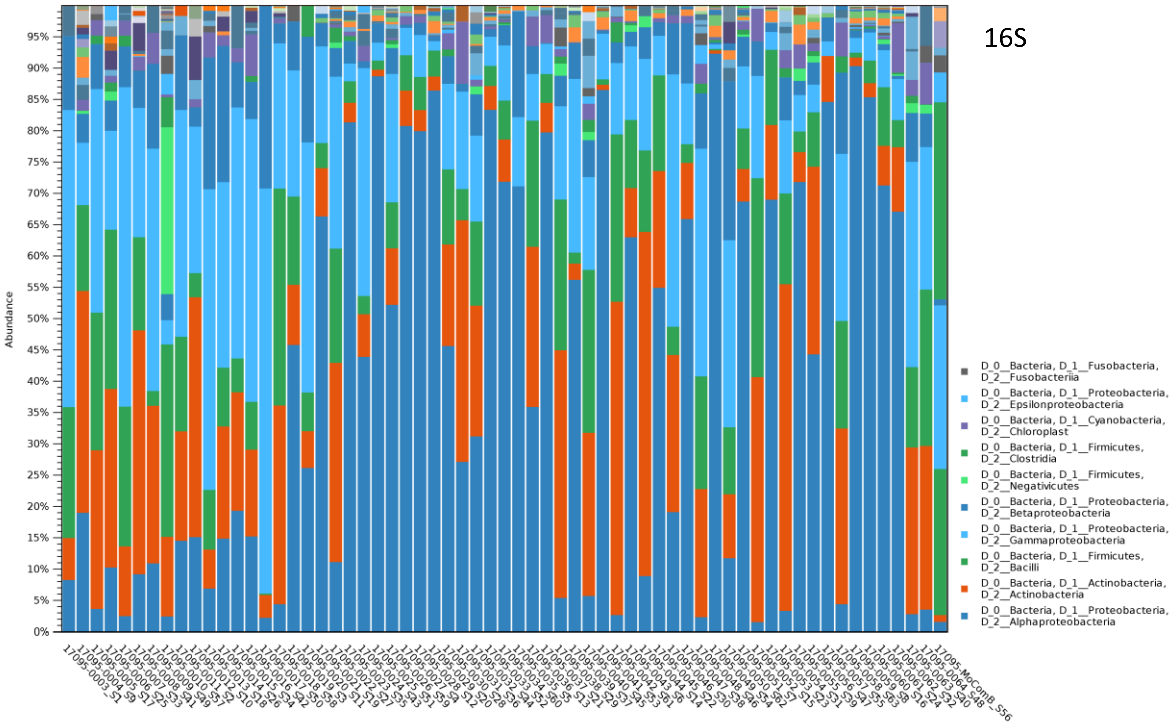
Column header	Description
Name	Name of the representative sequence/OTU
Taxonomy	Full taxonomic path, in descending order (D_0:kingdom, D_1: phylum, D_2: order; D_3: class, D_4: family, D_4: genus, D_6: species, D_7: subspecies,…)
Combined Abundance	Summarized amount of reads for the specific taxon over the complete sample set
Min	Minimum number of reads for the specific taxon over the complete sample set
Max	Maximum number of reads for the specific taxon over the complete sample set
Mean	Mean number of reads for the specific taxon over the complete sample set
Median	Median number of reads for the specific taxon over the complete sample set
Std	Standard deviation of number of reads for the specific taxon over the complete sample set
17095-0###_S##	IMGM-internal sample ID and identifier implemented by the MiSeq sequencer
Sequence	Representative sequence for the specific taxon

The statistics of OTU clustering input sequences and results, including the number of representative sequences per sample, their percentage compared to all reads per sample as well as number of unique and total chimeras in the OTU clustering step are provided in file

17095-16S_OTU_ClustStats.pdf and ***17095-ITS_OTU_ClustStats.pdf***.

In total 18’093 different 16S OTUs and 34’987 ITS OTUs were found in the complete sample set. Out of these, 1’670 16S OTUs and 200 ITS OTUs were annotated in the respective database used for analysis (see section 3.6.4.1). As a large number of *de novo* OTUs were predicted by the OTU clustering analysis, the 12 OTUs with the highest combined abundance across all samples were blasted at NCBI to the nr nucleotide collection considering all organisms. Results from the BLAST search are reported in table ***17095_BLASTN.xlsx***. While all 12 most abundant *de novo* OTUs in the 16S dataset were of human origin, four of the 12 most abundant *de novo* OTUs in the ITS dataset were classified as fungal sequences. To cover all highly abundant OTUs as good as possible, 25 more de novo OTUs were used for an NCBI BLAST search, giving 16 more hints of fungal origin.

For each sample a bar chart presenting the relative abundances of OTU classes was generated (Figure 13). Only OTUs exceeding a combined abundance of 10 were included.



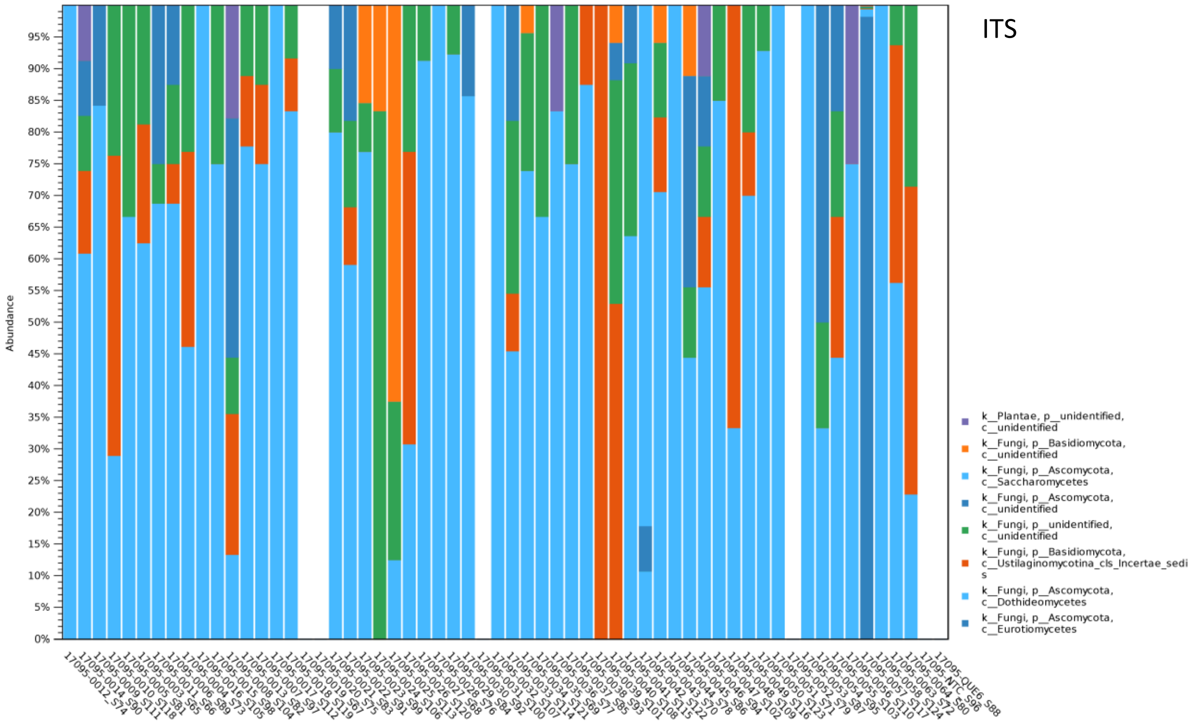
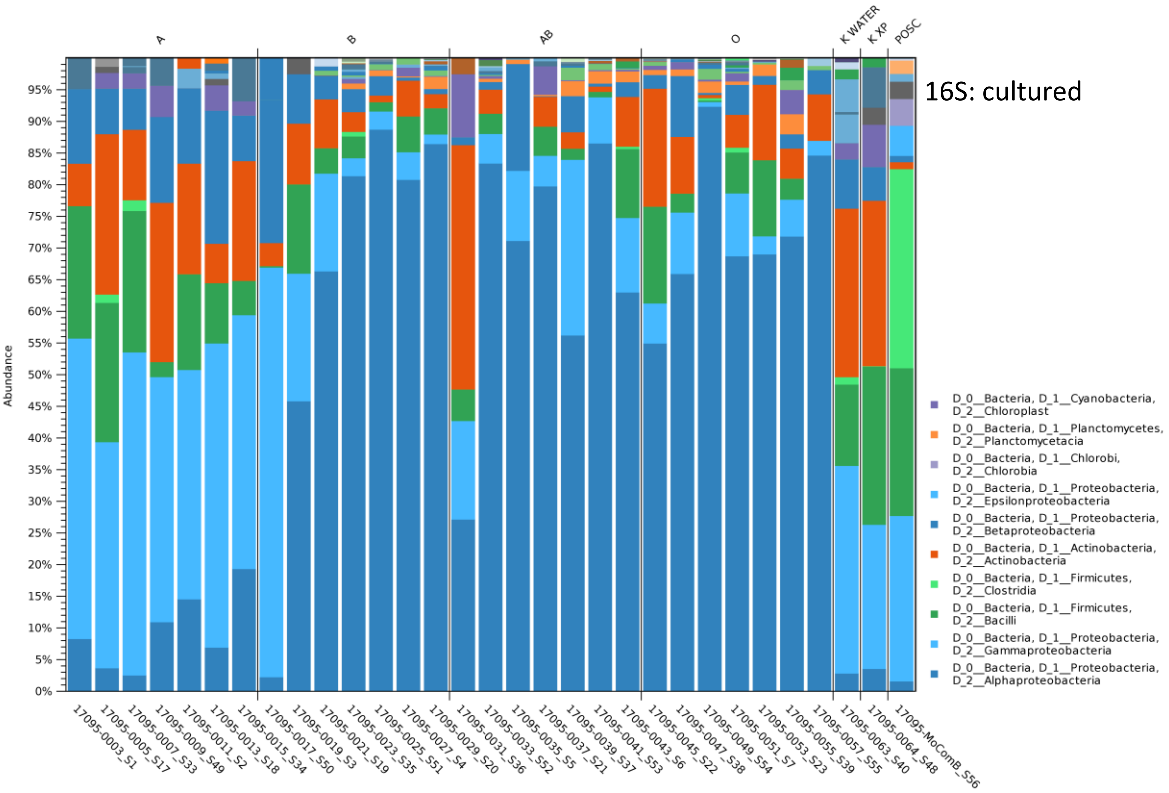


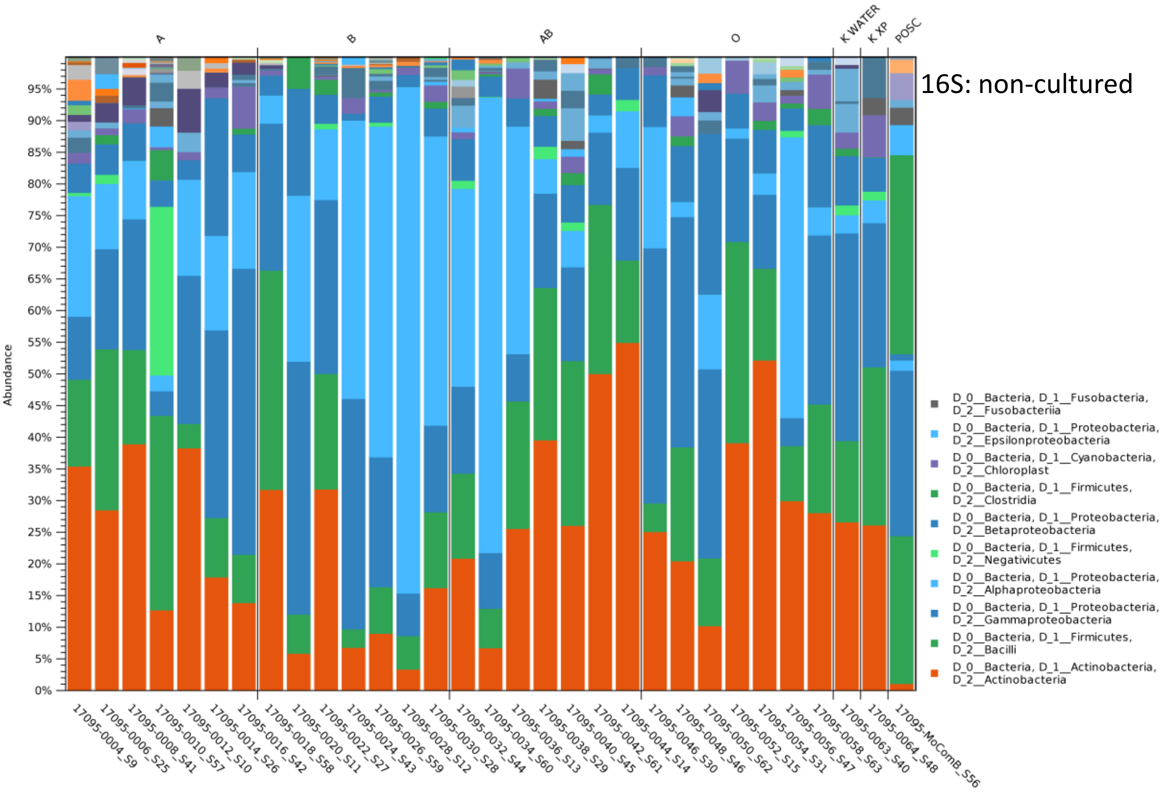
Figure 13: Distribution of OTU abundances on class level for each sample for 16S and ITS analysis respectively

Metadata containing specific sample information regarding sample type (A, B, AB and O) and category (cultured vs. non-cultured) were included for visualization of differences in OTU abundances between sample type and category. Thereby, it was shown that OTU classes differed depending on whether the sample was cultured or not (Figure 14). All bar charts can be found in the folder ***17095_Sequencing\17095_phylogenetic-abundance\17095_OTUs\Bar-charts***.

17095



16S: cultured



16S: non-cultured

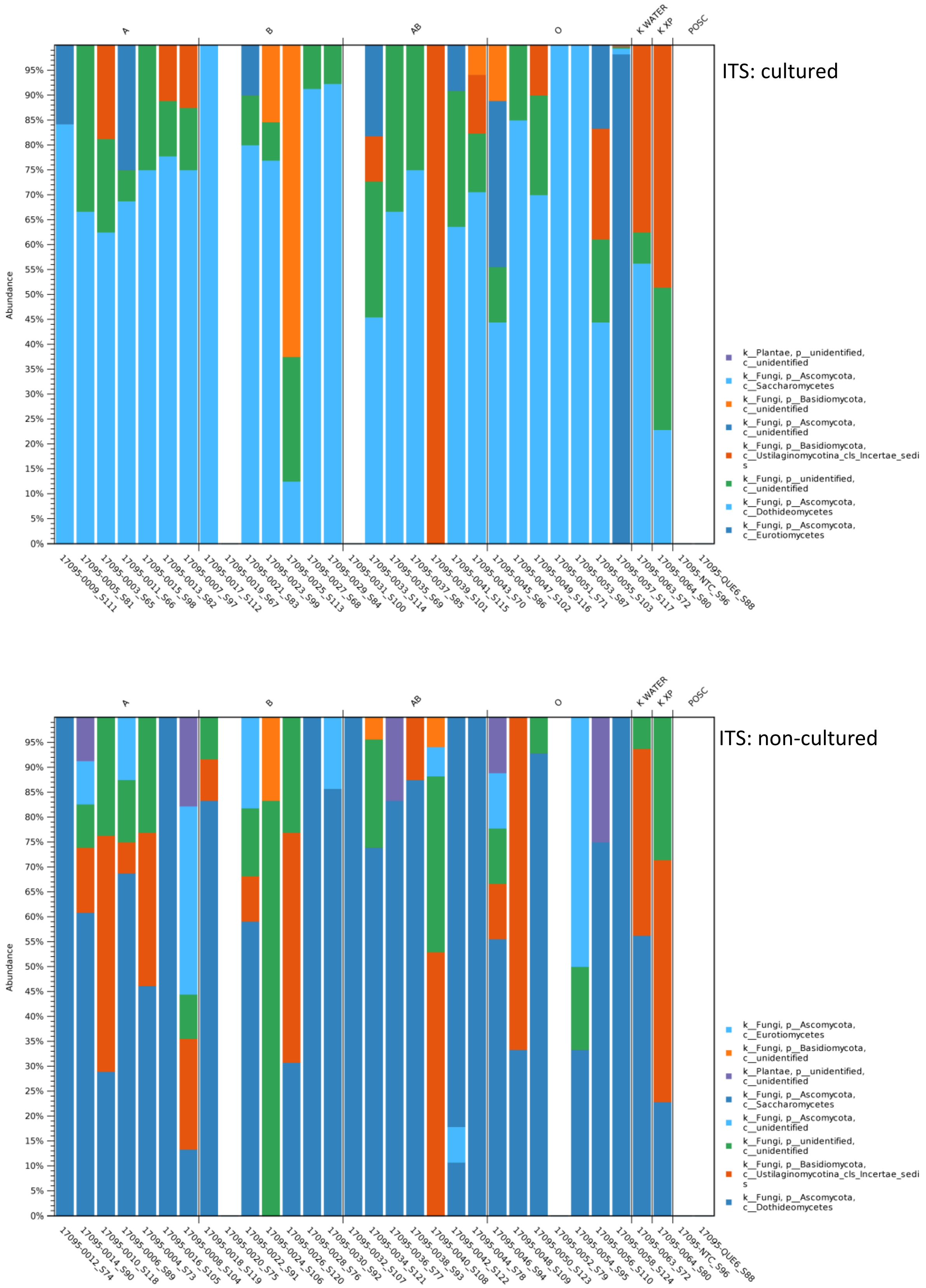


Figure 14: Distribution of OTU abundances on class level for cultured and non-cultured 16S and ITS samples separately

To visualize the heterogeneity of OTUs found in the 16S and ITS analyses, pie charts presenting the taxonomic classification of OTUs across all samples were generated. Only OTUs exceeding a combined abundance of 10 were included. As a large fraction of OTUs were not annotated in the respective databases, a separate pie chart presenting the annotated OTUs only was included (Figure 15). The taxonomic assignment path is shown from the highest phylogenetic level (kingdom bacteria) in the middle of the circle via order, class, family, genus down to the species level at the outer end of the circle.

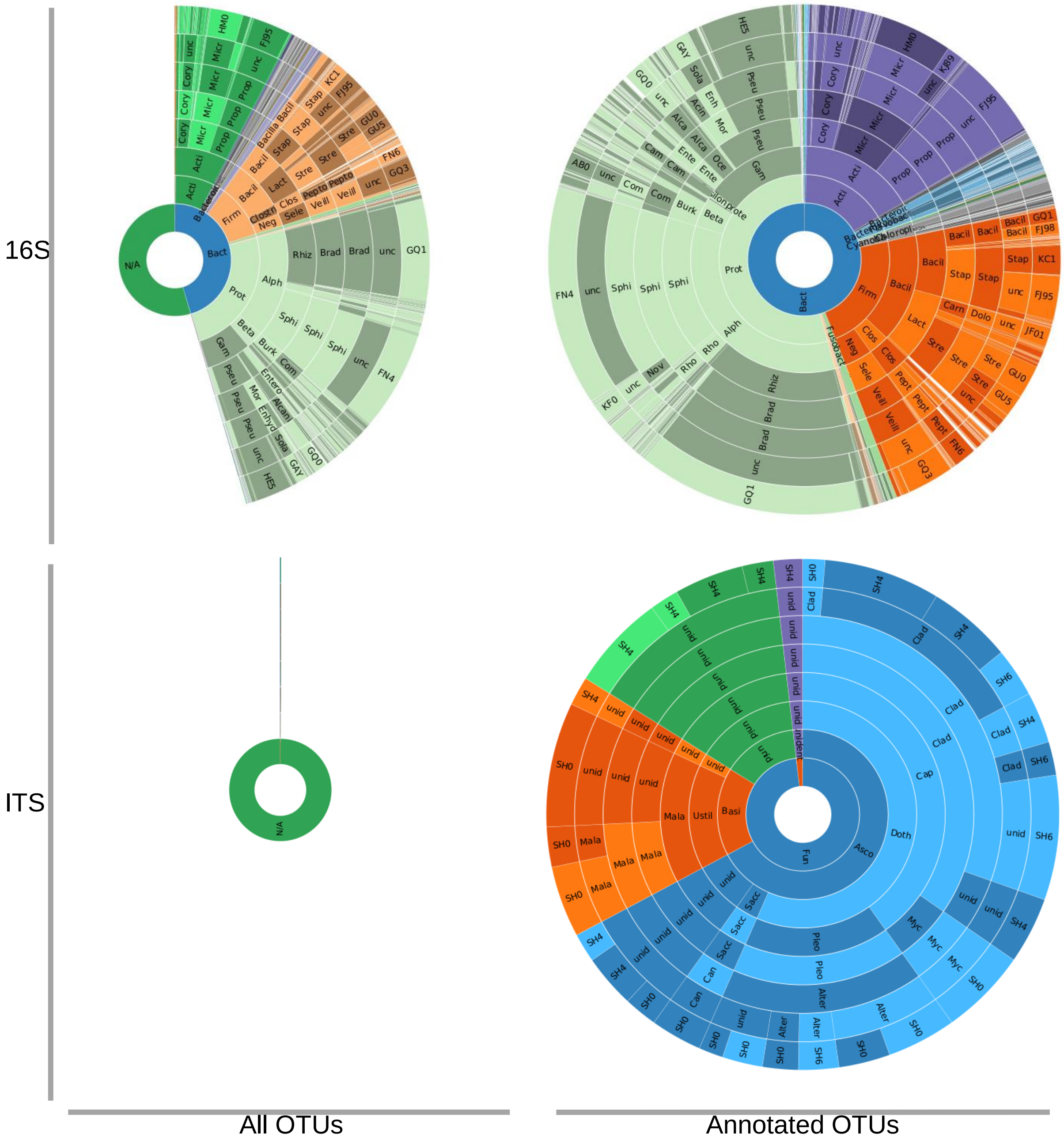


Figure 15: Taxonomic classification of OTUs identified in 16S and ITS samples

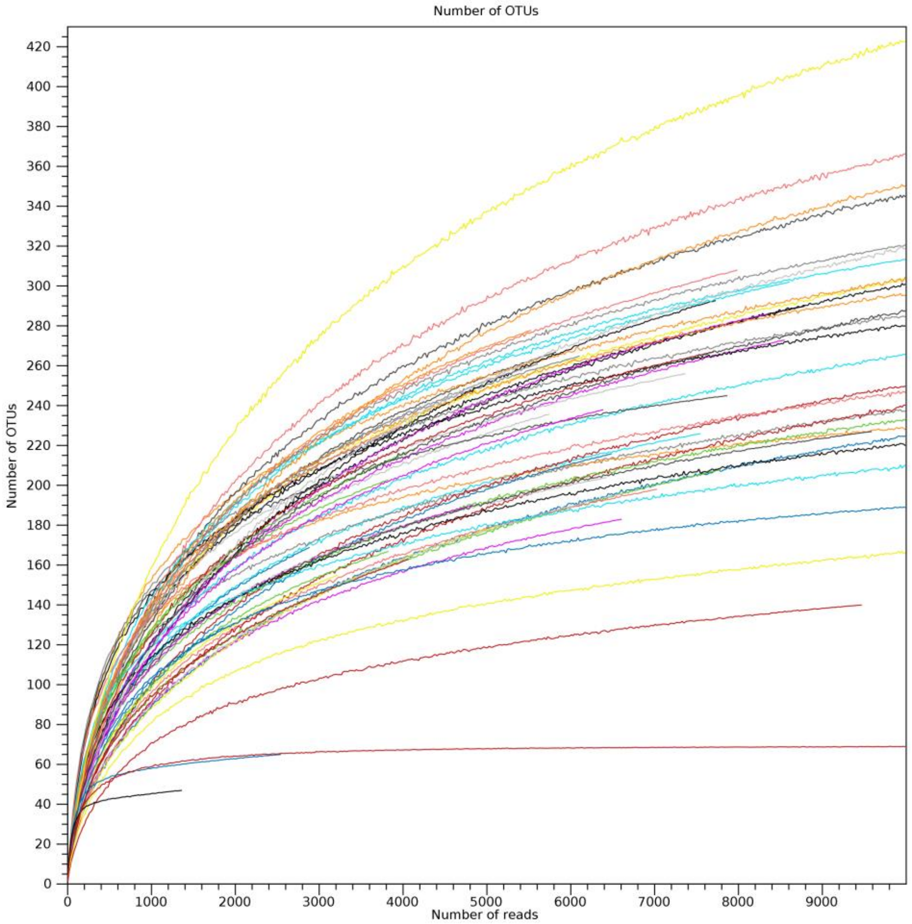
All pie charts are provided in folder:

17095_Sequencing\17095_phylogenetic-abundance\17095_OTUs\Pie-charts.

4.6.2 Alpha diversity

The alpha-diversity describes the number of species (or similar metrics) in a single sample. To calculate the alpha diversity OTUs with a combined abundance of lower than 10 were excluded due to minor relevance. The rarefaction curves showing the OTU vs. read plot for the 16S and ITS samples reach a saturation level for most samples (Figure 16). Hence, the amount of analyzed reads gives a good representation of the entire sampled population and only a low amount of new OTUs will appear with the analysis of more sequence reads.

16S



ITS

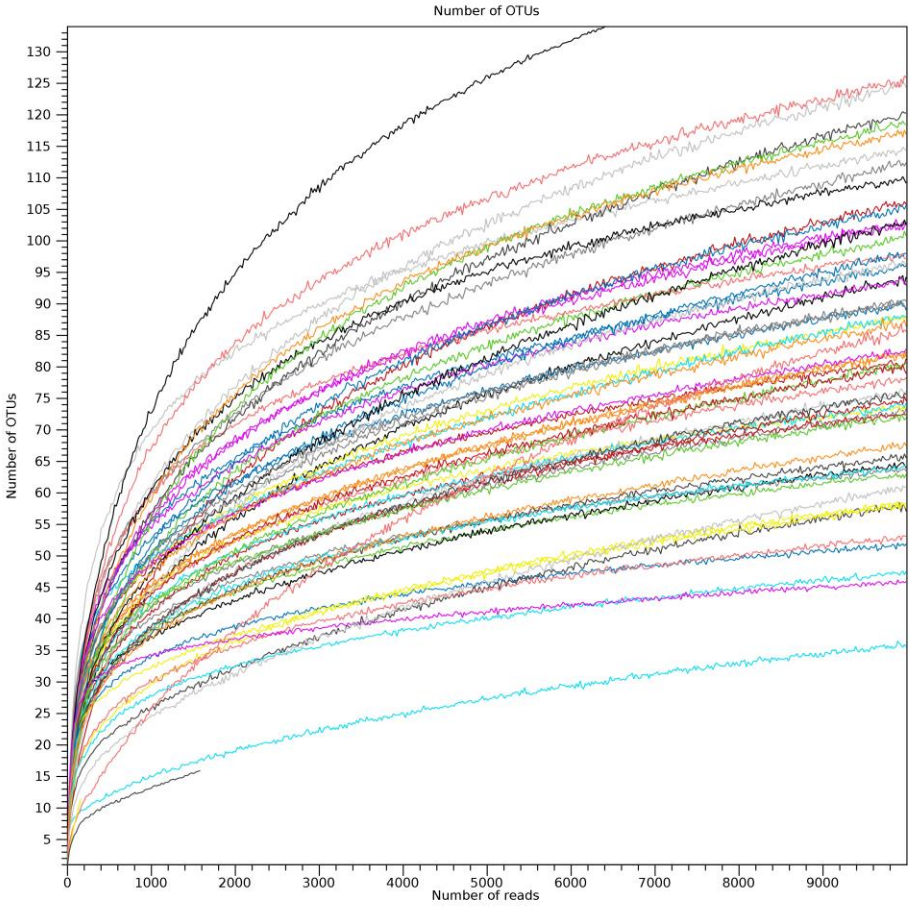


Figure 16: 16S Rarefaction curves plotting number of OTUs vs. number of reads for the sample set (*17095-16S_AlphaDiv_OTUs.png; 17095-ITS_AlphaDiv_OTUs.png*)

Further alpha diversity estimators were calculated for all subsampling steps as described in chapter 3.6.4.2. All alpha diversity results can be found in the files:

- ***17095-##S_AlphaDiv_OTUs.png*** (graphical OTU vs. reads plot)
- ***17095-##S_AlphaDiv_Chao1.png*** (graphical plot of Chao-1 measure)
- ***17095-##S_AlphaDiv_Shannon.png*** (graphical plot of Shannon-entropy measure)
- ***17095-##S_AlphaDiv_PhylogeneticDiversity.png*** (graphical plot of phylogenetic diversity measure)

4.6.3 Beta diversity

The beta-diversity compares the number of species across samples. Beta diversity measures according to Bray-Curtis, Jaccard and the unweighted UniFrac were calculated and used to generate PCoA matrices. For better visualization PCoA plots were generated for all samples colored based on sample name, type (A, B, AB, O) and category (cultured vs. non-cultured) separately. The 3-dimensional PCOA plots visualized in the **CLC Genomics Workbench** screenshots are provided as:

- ***17095-##S_BetaDiv-category-PCoA-Bray-Curtis.png***
- ***17095-##S_BetaDiv-samples-PCoA-Bray-Curtis.png***
- ***17095-##S_BetaDiv-type-PCoA-Bray-Curtis.png***
- ***17095-##S_BetaDiv-category-PCoA-Jaccard.png***
- ***17095-##S_BetaDiv-samples-PCoA-Jaccard.png***
- ***17095-##S_BetaDiv-type-PCoA-Jaccard.png***
- ***17095-##S_BetaDiv-category-PCoA-Unweighted UniFrac.png***
- ***17095-##S_BetaDiv-samples-PCoA-Unweighted UniFrac.png***
- ***17095-##S_BetaDiv-type-PCoA-Unweighted UniFrac.png***

Furthermore, the underlying data is provided in table format to allow procession and visualization of the data with other, user accessible, PCoA viewer tools:

- ***17095_BetaDiv_Distance-matrix.xlsx***

5 Summary

In study 17095 phylogenetic 16S and ITS analyses were carried out in gDNA samples of human blood samples infected with bacteria and fungi. 62 gDNA samples and two negative controls were submitted to IMGM for 16S and ITS library generation and phylogenetic classification by next generation sequencing.

During sample registration, four samples were excluded due to ambiguous sample names or missing samples and replaced by the customer later on.

For phylogenetic amplification, amplicons covering V3-V4 hypervariable regions of the 16S rRNA and ITS2 hypervariable region of ITS rRNA were generated and two amplicon libraries were prepared from all PCR products of 64 different 16S and 62 different ITS samples, respectively. Sequencing was performed on the MiSeq sequencing system with 2x300 bp PE reads. Image and signal processing of the recorded signals as well as de-multiplexing of reads according to their respective index was performed. Quality was evaluated and quality criteria were fulfilled in the sequencing run. The resulting 2 x 300 bp reads were quality controlled, merged into continuous 600 bp reads and trimmed according to primer sequences, quality and length. The two negative control samples only showed negligible amplification and thus a very low amount of reads and in the 16S analysis. For the other samples a high number of high quality reads was generated.

Further bioinformatics analysis including clustering, phylogenetic analysis and alpha and beta diversity calculation was performed with the **CLC Genomics Workbench** and its microbial genomics module.

Metadata including information regarding expected similarities of microbial load across samples as well as culture conditions were included after OTU clustering and used for OTU plotting and beta diversity calculations.

Martinsried, October 27, 2017

…...

Dr. Christin Mieth (Project Manager)

6 References

1. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, u. a. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res [Internet]. 28. August 2012 [zitiert 4. April 2013]; Verfügbar unter: http://nar.oxfordjournals.org/content/early/2012/10/31/nar.gks808

2. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, u. a. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A. 17. April 2012;109(16):6241–6.

3. Whittaker RH. Evolution and Measurement of Species Diversity. Taxon. 1. Mai 1972;21(2/3):213–51.