ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 1373 -1378
doi: 10.18178/wcse.2017.06.238

Next-generation Sequencing Generated Discrepancy in Abundance Characterization of Complex Microbial Community Compositions: an Error of Bioinformatics Pipeline

Huimin Zhang[1], Hongkui He[2], Runjie Cao[2], Huizhi Tang[2], Zhizhou Zhang[1]*[+], Anjun Li[2] and Jie Jiang[1]*

[1]School of Marine Science and Technology, Marine anti-fouling Engineering Technology Center of Shandong Province, Harbin Institute of Technology, Harbin, China 150006
[2]The Anhui GuJingTribute Liquor Ltd, Bozhou, Anhui, China 236800

Abstract. Next generation sequencing on metagenomes produces a lot of valuable biological and biomedical data but still with some errors. For examples, chimeras are basically originated from biological reactions, while taxonomic classification errors are easily resulted from bioinformatics pipelines. In this study the microbial compositions in the starter (Daqu) of Chinese GujingTribute liquor, especially the dominant species or OTUs (operational taxonomic unit), were determined by two approaches, one is the near full length ribosome gene (16S rDNA plus the internal transcribed spacer (ITS)) library sequencing, and another is 16S rDNA V4-V5 region based next generation sequencing approach. The two approaches gave discrepant results for both the prokaryotic microbes and eukaryotic ones. Especially, the results for prokaryotic microbes showed apparent differences in that (1)The most dominant species or OTU belong to different phyla; (2) The 20 most dominant species or OTUs overlapped only partially. Further investigation indicated that the bioinformatics analysis pipeline itself was sometimes an important source for discrepancy generation.

Keywords: next generation sequencing, bioinformatics pipeline, metagenome, discrepancy

1. Introduction

Progresses in microfluidic machinery and nanotechnology brought tremendous improvement in high throughput biotechnologies, including different platforms of next generation DNA sequencing. Such platforms are both technology pipelines and bioinformatics pipelines. In life sciences and environmental engineering fields, one hot spot is metagenome-based DNA sequencing and related functional studies, in which bioinformatics pipelines employ simple statistical algorithms and a series of DNA databases to classify a large amount of DNA sequences deciphered from thousands of microbial community species. Apparently, the classification accuracy is very important because it may affect a lot on the final biological conclusion or medical detection decision.

There are a batch of typical metagenome sequencing study cases, including soil [1], gut [2], marine water [3], fermentation pits [4], waste treatment microbes [5], sea bed [6], and some specific niche such as hospital hallway air microbes. At present, there are several methods to decode microbial population structures, including traditional pure culture [7], library sequencing [8-9], PCR-DGGE [10], and several types of next-generation sequencing (NGS) [11-12]. NGS technology is rapidly employed in hundreds of laboratories with lower and lower cost, but the short read of target genes (such as 16S rDNA variable regions) brings a great limitation on sequencing data implications. NGS technologies using full-length ribosome genes are

[+] Corresponding author. Zhizhou Zhang Tel.: + 86-631-5683176; fax: +86-631-5687230.
E-mail address: zhangzzbiox@hitwh.edu.cn.

considered still inmature [13] and may need to take several years to be fully developed. Meanwhile, the de novo sequencing technology [14] can assemble long genome fragments for dominant species in a metagenome sample with, unfortunately, still very high cost.

Hahn et al [16] employed two NGS platforms, MiSeq and PacBio RSII, to characterize the cystic fibrosis lung microbiome, and found that MiSeq allowed for the observation of many more operational taxonomic units (OTUs) and higher Chao1 and Shannon indices than the PacBio RSII, while only PacBio RSII was able to identify *Burkholderia*, an important cystic fibrosis pathogen. Such results, if used in supplementary diagnosis, would lead to different decisions because only one platform was able to detect the pathogen.

This study used a complex microbial community sample, the fermentation starter (Daqu), in the ethanol industry. GujingTribute [15] is one of the representative strong-aroma types of Chinese liquor. However, knowledge of the relationship between its flavor and fermentative microbes has been little, so detailed deciphering of Daqu microbial compositions shall be the first step to investigate the above relationship.

In this study, the microbial composition was decoded by two approaches, one is the Illumina Miseq for 16S rDNA V4-V5 plus ITS region (NGS approach), and another is full-length 16S rDNA \ITS amplification-TA cloning-Sanger sequencing (TA-clone approach). The two approaches were expected to generate the highly similar namelists of dominant species (or OTUs) for the same Daqu sample, but the results showed that there were apparent discrepancy between the namelists, and especially, the most dominant species in TA-clone approach was not the most dominant OUT in the NGS approach. Further analysis indicated that the error was largely derived from the database quality, a key element in the NGS bioinformatics pipeline.

2. Materials and Methods

2.1. Sampling

Randomly selected twelve Daqu bricks were smashed and well-mixed into one mix-sample. Metagenome DNA was extracted as previously published protocol [15], then subjected to Illumina Miseq next generation sequencing platform (16S rDNA V4-5) and full-length 16S /ITS rDNA TA-cloning to decipher the microbial compositions.

2.2. Decoding microbial population composition by next generation sequencing (NGS)

Metagenomic DNA was extracted from the above prepared production Daqu sample using Solarbio D2600 kit for genome purification. Each 200mg sample generated 100ul metagenome DNA. The 390bp V4-V5 region of the 16S rRNA gene was amplified using the primer set 520F (5'-7bp barcode+ GCA CCT AAY TGG GYD TAA AGNG-3') and 904R (5'- CCG TCA ATT CCT TTR AGT TT -3'). For ITS, two primers were use as follow: ITS1 (5'-TCCGTA GGT GAA CCT GCG G-3') and ITS4 (5'-TCC TCC GCT TAT TGA TAT GC-3'). PCR was set up with high fidelity system (0.25ul Q5 high-fidelity DNA polymerase, 5ul 5×Reaction Buffer, 5ul 5×High GC Buffer , 0.5ul dNTP (10mM), 1 ul each primer (10uM), dH2O 11.25ul) and performed according to the following: 98°C-30s, (98°C-15s, 50°C-30s, 72°C-30s) for 26 cycles, 72°C-30s plus 72°C-5min. PCR amplicons were purified, further processed and subjected to Illumina Miseq platform. All the raw sequence data were processed in the QIIME pipeline [17]. PCR chimeras were checked and removed using the UCHIME software. The remaining good-quality sequences were clustered into operational taxonomic units (OTUs) using a 97% identity threshold with QIIME's UCLUST tool. The most abundant sequence of each OTU was picked as the representative sequence. The taxonomic information of each representative sequence, also the taxonomic information of each OTU, was annotated using Greengenes database Release 13.8 classifier [18].

2.3. Deciphering microbial community structures with full-length 16S /ITS rDNA TA-cloning (TA-clone)

The whole experimental process can be seen in our lab's publication using the universal primers 27F and 1492R [15]. For ITS amplification, primer ITS1 (5'-TCCGTA GGT GAA CCT GCG G-3') and ITS4 (5'-TCC TCC GCT TAT TGA TAT GC-3') were used. Positive colonies were subjected to Sanger sequencing using the same PCR primers. Each pair of bidirectional sequences was assembled as one single sequence with correct direction. Short sequences without primers on both ends were removed from the data and the

residual vector bases and primer bases on both ends were deleted (LaserGene, DNAStar). Then the sequences were subjected to chimera checking using UCHIME [19] in mothur software [20]. The sequences with chimeric parts were removed from the data. Finally, the remained sequence was used for further taxonomic analysis. For ITS sequences the UNITE Database was employed using User-friendly Nordic ITS Ectomycorrhiza Database (https://unite.ut.ee/index.php) [21]. The obtained taxonomic information was inspected and corrected manually based on known microbial community knowledge and BLAST (Basic Local Alignment Search Tool) plus RDP (rdp.cme.msu.edu) results. After chimera removal, refined sequences were deposited in the GenBank under the accession numbers KX603403-KX603652 (bacterial 16S rRNA) and KX911990-KX912161 (fungal ITS).

3. Results and Discussion

3.1. Bacterial compositions

For NGS approach, a total of 30353 good quality prokaryotic sequences with an average length of 390 bp were obtained after quality filtering. At the 97% identity level, 515 OTUs were classified (data not shown). The most dominant twenty OTUs were listed in Table 1. For TA-clone approach, using a 97% cutoff, 215 among 249 sequenced clones showed species-level information (data not shown). The most dominant ten species were listed in Table 2. The prefixes "p_", "g_", "s_" indicated OTUs were annotated to the level of phylum, genus, or species, respectively.

Table 1. The most abundant 20 prokaryotic OTUs detected with 16S rDNA V4-5 (OTU level)

	#OTU ID	Read	Abundance (%)	taxonomy
1	denovo450	11970	39.43	p_Proteobacteria; g_*Erwinia*; s_
2	denovo41	2301	7.58	p_Firmicutes; g_*Staphylococcus; s_succinus*
3	denovo376	1899	6.25	p_Firmicutes; g_*Leuconostoc*; s_
4	denovo105	1581	5.20	p_Actinobacteria; g_*Streptomyces*; s_
5	denovo380	1511	4.97	p_Firmicutes; g_*Lactobacillus;* s_*paraplantarum*
6	denovo76	1434	4.72	p_Firmicutes; g_*Lactobacillus*; s_*brevis*
7	denovo309	1267	4.17	p_Firmicutes; g_*Thermoactinomyces*; s_*sanguinis*
8	denovo379	1128	3.71	p_Firmicutes; g_*Lactobacillus*; s_
9	denovo420	748	2.46	p_Proteobacteria; g_*Sarcandra*; s_*grandifolia*
10	denovo177	573	1.88	p_Firmicutes; g_*Lactobacillus*; s_
11	denovo99	529	1.74	p_Firmicutes; g_*Bacillus*; s_
12	denovo178	528	1.73	p_Cyanobacteria; g_; s_
13	denovo282	478	1.57	p_Proteobacteria; g_*Burkholderia*; s_
14	denovo312	414	1.36	p_Proteobacteria; g_*Enterobacter*; s_
15	denovo220	301	0.99	p_Proteobacteria; g_*Rhodanobacter*; s_
16	denovo144	221	0.72	p_Proteobacteria; g_*Erwinia*; s_*oleae*
17	denovo6	211	0.69	p_Firmicutes; g_*Pediococcus*; s_*acidilactici*
18	denovo476	201	0.66	p_Proteobacteria; g_*Pseudomonas*; s_
19	denovo154	169	0.55	p_Firmicutes; g_*Lactobacillus*; s_
20	denovo330	138	0.45	p_Actinobacteria; g_*Rhodococcus*; s_

Table 2. Bacterial compositions in GujingTribute Daqu determined by TA-clone approach (species level)

	Species	Abundance (%)	Taxonomy	BLAST similarity level (%)
1	*Virgibacillus halotolerans*	38.37	p_Firmicutes; g_*Virgibacillus*	99
2	*Thermoactinomyces sanguinis*	19	p_Firmicutes; g_*Thermoactinomyces*	99
3	*Virgibacillus sp.*	6.48	p_Firmicutes; g_*Virgibacillus*	100
4	*Lactobacillus plantarum*	6.34	p_Firmicutes; g_*Lactobacillus*	99
5	*Pantoea agglomerans*	5.94	p_Proteobacteria; g_*Pantoea*	99
6	*Staphylococcus sp.*	4.52	p_Firmicutes; g_*Staphylococcus*	100
7	*Pantoea vagans*	4.32	p_Proteobacteria; g_*Pantoea*	99
8	*Lactobacillus sp.*	3.24	p_Firmicutes; g_*Lactobacillus*	98
9	*Bacillus sp.*	2.16	p_Firmicutes; g_*Bacillus*	99
10	*Staphylococcus kloosii*	2.16	p_Firmicutes; g_*Staphylococcus*	100
11	*Bacillus subtilis*	1.62	p_Firmicutes; g_*Bacillus*	99
12	*Bacillus licheniformis*	0.74	p_Firmicutes; g_*Bacillus*	99
13	*Planomicrobium sp.*	0.54	p_Firmicutes; g_*Planomicrobium*	100
14	*Lactobacillus brevis*	0.54	p_Firmicutes; g_*Lactobacillus*	100
15	*Lactobacillus fermentum*	0.54	p_Firmicutes; g_*Lactobacillus*	99

16	*Lactobacillus pontis*	0.54	p_Firmicutes; g_*Lactobacillus*	99
17	*Lactobacillus rossiae*	0.54	p_Firmicutes; g_*Lactobacillus*	100
18	*Weissella sp.*	0.54	p_Firmicutes; g_*Weissella*	100
19	*Enterobacter hormaechei*	0.54	p_Proteobacteria; g_*Enterobacter*	99
20	*Pantoea ananatis*	0.54	p_Proteobacteria; g_*Pantoea*	99

Table 3. Bacterial compositions in GujingTribute Daqu (genus level)

TA-clone		NGS	
Abundance (%)	Taxonomy	Abundance (%)	Taxonomy
44.85	p_Firmicutes; g_*Virgibacillus*;	40.15	p_Proteobacteria; g_*Erwinia*
19	p_Firmicutes; g_*Thermoactinomyces*	15.83	p_Firmicutes; g_*Lactobacillus*
11.74	p_Firmicutes; g_*Lactobacillus*	7.58	p_Firmicutes; g_*Staphylococcus*
10.8	p_Proteobacteria; g_*Pantoea*	6.25	p_Firmicutes; g_*Leuconostoc*
6.68	p_Firmicutes; g_*Staphylococcus*	5.20	p_Actinobacteria; g_*Streptomyces*
4.52	p_Firmicutes; g_*Bacillus*	4.17	p_Firmicutes; g_*Thermoactinomyces*
0.54	p_Firmicutes; g_*Planomicrobium*	2.46	p_Proteobacteria; g_*Sarcandra*
0.54	p_Firmicutes; g_*Weissella*	1.74	p_Firmicutes; g_*Bacillus*
0.54	p_Proteobacteria; g_*Enterobacter*	1.73	p_Cyanobacteria; g_

3.2. Eukaryotes compositions

For eukaryotic data, a total of 15591 good quality sequences with an average length of 265 bp were obtained after quality filtering. At the 97% identity level, 84 OTUs were classified (data not shown). The most dominant twenty OTUs were listed in Table 4. For TA-clone approach, all 172 curated sequences of ITS clones can give species-level annotation (data not shown). The most dominant ten species were listed in Table 5.

Table 4. The most abundant 20 eukaryotic OTUs detected with ITS (OTU level)

	#OTU ID	Read	Abundance (%)	taxonomy
1	denovo61	4623	29.65	p_Ascomycota; g_*Aspergillus*; s_*Aspergillus flavus*
2	denovo60	2903	18.61	p_Zygomycota; g_*Rhizopus*; s_*Rhizopus arrhizus*
3	denovo46	2627	16.84	p_Zygomycota; g_*Rhizomucor*; s_*Rhizomucor pusillus*
4	denovo66	791	5.07	p_Ascomycota; g_*Paecilomyces*; s_*Paecilomyces verrucosus*
5	denovo42	754	4.83	p_Ascomycota; g_unidentified; s_*Saccharomycetales sp*
6	denovo43	661	4.23	p_Ascomycota; g_*Aspergillus*; s_*Aspergillus cibarius*
7	denovo19	580	3.72	p_unidentified; g_unidentified; s_*Plantae sp*
8	denovo21	519	3.32	p_Ascomycota; g_*Thermomyces*; s_*Thermomyces lanuginosus*
9	denovo6	488	3.13	p_Ascomycota; g_*Candida*; s_*Candida xylopsoci*
10	denovo49	407	2.61	p_unidentified; g_unidentified; s_*Plantae sp*
11	denovo64	232	1.48	p_Ascomycota; g_unidentified; s_*Eurotiomycetes sp*
12	denovo4	159	1.01	p_Zygomycota; g_*Lichtheimia*; s_*Lichtheimia ornata*
13	denovo51	116	0.74	p_Ascomycota; g_*Thermoascus*; s_*Thermoascus aurantiacus*
14	denovo28	108	0.69	p_Ascomycota; g_*Aspergillus*; s_*Aspergillus piperis*
15	denovo8	103	0.66	p_Ascomycota; g_*Candida*; s_*Candida blattae*
16	denovo5	86	0.55	p_Ascomycota; g_*Aspergillus*; s_*Aspergillus candidus*
17	denovo53	60	0.38	p_Ascomycota; g_*Wickerhamomyces*; s_*Wickerhamomyces anomalus*
18	denovo59	48	0.30	p_Ascomycota; g_*Penicillium*; s_*Penicillium polonicum*
19	denovo77	38	0.24	p_Ascomycota; g_*Penicillium*; s_*Penicillium citrinum*
20	denovo68	27	0.17	p_Ascomycota; g_*Monascus*; s_*Monascus purpureus*

Table 5. Eukaryotic compositions in GujingTribute Daqu determined by TA-clone approach (species level)

	Species	Abundance (%)	Taxonomy	BLAST similarity level (%)
1	*Rhizopus_ arrhizus*	45.09	p_Zygomycota;g_*Rhizopus*	100
2	*Aspergillus_flavus*	14.63	p_Ascomycota; g_*Aspergillus*	100
3	*Thermomyces_lanuginosus*	13.94	p_Ascomycota; g_*Thermomyces*	100
4	*Aspergillus_amstelodami*	8.02	p_Ascomycota; g_*Aspergillus*	100
5	*Thermoascus_crustaceus*	6.44	p_Ascomycota; g_*Thermoascus*	99
6	*Thermoascus_aurantiacus*	5.31	p_Ascomycota; g_*Thermoascus*	100
7	*Penicillium_chrysogenum*	1.94	p_Ascomycota; g_*Penicillium*	100
8	*Pichia_kudriavzevii*	1.94	p_Ascomycota; g_*Pichia*	99
9	*Aspergillus_candidus*	0.96	p_Ascomycota; g_*Aspergillus*	99
10	*Lichtheimia_ramosa*	0.96	p_Zygomycota; g_*Lichtheimia*	100

3.3. The discrepancy is related with data processing pipelines

Though Table 6 suggested that the two approaches generate apparent discrepancy on characterization of eukaryotic species, from the Table 2 and Table 3, however, discrepancy in prokaryotic microbial compositions may be more dramatic. It was clear that the most dominant bacterial species was a *Virgibacillus* (Firmicute) strain, while Table 1 and Table 3 demonstrated that the most dominant OTU was a Erwinia (Proteobacteria) microbe. Because they belonged to different phyla, the classification discrepancy looked unacceptable. The NGS platform in this study used GreenGenes database to annotate each filtered sequence while the TA-clone approach used BLAST and RDP as annotation tools, and that may be the reason why the most dominant OTU (denovo 450) [22] was classified as an *Erwinia* species (OTU). If the representative sequence of OTU 450 was tested on BLAST and RDP databases, it would be assigned as a *Pantoea agglomerans* strain, relatively consistent with the TA-clone result (Table 2) in which *Pantoea agglomerans* was the 5th most dominant species.

Table 6. Eukaryotic compositions in GujingTribute Daqu (genus level)

TA-clone		NGS	
Abundance (%)	Taxonomy	Abundance (%)	Taxonomy
45.09	p_Zygomycota;g_*Rhizopus*	35.12	p_Ascomycota; g_*Aspergillus*
23.61	p_Ascomycota; g_*Aspergillus*	18.61	p_Zygomycota; g_*Rhizopus*
13.94	p_Ascomycota; g_*Thermomyces*	16.84	p_Zygomycota; g_*Rhizomucor*
11.75	p_Ascomycota; g_*Thermoascus*	5.07	p_Ascomycota; g_*Paecilomyces*
1.94	p_Ascomycota; g_*Penicillium*	4.83	p_Ascomycota; g_unidentified
1.94	p_Ascomycota; g_*Pichia*	3.72	p_unidentified; g_unidentified
0.96	p_Zygomycota; g_*Lichtheimia*	3.32	p_Ascomycota; g_*Thermomyces*

Metagenome sequencing and sequence annotation rely heavily on the bioinformatics databases in which RDP [rdp.cme.msu.edu], GreenGenes [greengenes.secondgenome.com], SILVA [http://www.arb-silva.de] and NCBI [www.ncbi.nlm.nih.gov/nucleotide] are all good ones. However, for ribosome genes, GreenGenes contains most high-quality sequences and loses the general coverage range in that some different taxonomic elements are only in those low or intermediate quality sequences. For NCBI source, it has many misleading or wrong sequences (such as some chimeras) due to a historic reason and those low quality or wrong sequences have not been re-confirmed and cleaned. However, RDP has a general coverage that is between GreenGenes and NCBI in the context of ribosome gene representations. So as one important step in the bioinformatics pipeline, database quality curation is a continuous and stringent task from now on. On the other hand, choice of a specific NGS protocol for a specified research task is also critical; for example, Yang et al [23] conducted a survey for sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis and found that V4-V6 region was best to represent the optimal sub-regions of a new phylum.

4. Conclusion

In conclusion, metagenome sequencing generates large amount of DNA sequences that are subjected to possible wrong or misleading annotations due to some systematic error in bioinformatics analysis pipelines. Sometimes such an error or discrepancy may lead to big difference in the final biological research conclusion or biomedical detection decision. The results in this study further confirmed that several different approaches shall be considered at the same time for accurately determining compositions of a complex microbial community and database quality curation shall be a systematic and continuous endeavor worldwide. In the future, the authors would like to develop some novel nanoparticle-assisted DNA amplification techniques with higher fidelity than those used in the present NGS approaches so that some sequence errors can be avoided before they enter the following bioinformatics pipelines.

5. Acknowledgements

This study was supported by NSFC (No.31071170), GujingTribute fund (2016-1), GREDBIO (201401) and HIT fund (hitwh200904, 2016GSF115022).

6. References

[1] Erick C. et al. Forest harvesting reduces the soil metagenomic potential for biomass decomposition. ISME J 2015, **9**(11): 2465–2476.

[2] Zeng, B. et al. The bacterial communities associated with fecal types and body weight of rex rabbits. Sci Rep 2015, **5**, 9342.

[3] Shinichi, S. et al. Structure and function of the global ocean microbiome. Science 2015, **348**(6237), 1261359.

[4] Luo, X.M. et al. Phylum-specific primer design and implication in quantification of the microbial community structure in GuJingGong pit mud., Adv Mater Res 2014, **1051**,311-316.

[5] Shu D. et al. Metagenomic and quantitative insights into microbial communities and functional genes of nitrogen and iron cycling in twelve wastewater treatment systems. Chem Eng J 2016,**290**,21-30.

[6] Christner, B.C. et al.A microbial ecosystem beneath the West Antarctic ice sheet. Nature 2014, **512**(7514):310-313

[7] Kaeberlein, T., Lewis, K. and Epstein, S.S. Isolating" uncultivable" microorganisms in pure culture in a simulated natural environment. Science 2002, **296**, 1127-1129.

[8] Rondon, M.R. et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Micro 2000, **66**, 2541-2547.

[9] Gou, M. et al. Characterization of the microbial community in three types of fermentation starters used for Chinese liquor production. J Inst Brew 2015, **121**, 620-627.

[10] Zhang, L. et al. Characterisation of microbial communities in Chinese liquor fermentation starters Daqu using nested PCR-DGGE. World J Microbiol Biotechnol 2014, **30**, 3055-3063.

[11] Klindworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 2012, gks808.

[12] Hong, X.et al. Metagenomic sequencing reveals the relationship between microbiota composition and quality of Chinese Rice Wine. Sci Rep 2016, **6**, 26621

[13] Wagner, J. et al. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol 2016, **16**(1):274.

[14] Hess, M. et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 2011, **331**, 463-467.

[15] Zhang, H. et al. Employment of near full-length ribosome gene TA-cloning and Primer-Blast to detect multiple species in a natural complex microbial community using species-specific primers designed with their genome sequences. Mol Biotechnol 2016, **58**, 729-737.

[16] Hahn A et al. Different next generation sequencing platforms produce different microbial profiles and diversity in cystic fibrosis sputum. J Microbiol Methods 2016, **130**, 95-99.

[17] Navas-Molina, J.A. et al. Advancing our understanding of the human microbiome using QIIME. Methods Enzymol 2013, **531**, 371-444.

[18] DeSantis, T.Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microb 2006, **72**, 5069-5072.

[19] Edgar, R.C. et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011, **27**, 2194-2200.

[20] Schloss, P.D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Micro 2009, **75**, 7537-7541.

[21] Abarenkov, K. et al. The UNITE database for molecular identification of fungi – recent updates and future perspectives. New Phyto 2010, **186**, 281-285.

[22] >OTU 450
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTCTGTTAAGTCAGATGTGAAATCCCCGGGCTTAA
CCTGGGAACTGCATTTGAAACTGGCAGGCTTGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGG
AGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTTCCCTTGAGGAGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGA
GTACGGCCGCAAGGTTAAAACTTAAATGAATTGACGG

[23] Yang, B. et al. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 2016, **17**, 135.