eurofins | Genomics

Data Analysis Report: Microbiome Profiling

Project / Study: NG-A2417

Date: July 17, 2024

1 Microbiome Analysis Pipeline

The microbiome analysis pipeline consists of three major steps and some intermediate filtering steps. Each major pipeline step is described in more detail in its respective report section. The following list provides an overview of the full pipeline, while the **main results** of the microbiome analysis are presented in section *Microbiome Profiling*.

Demultiplexing All reads passing the standard Illumina chastity filter (PF reads) are demultiplexed according to their index sequences.

Primer clipping The target region specific forward and reverse primer sequences are identified and clipped from the starts of the raw forward and reverse reads. If primer sequences could not be perfectly matched (no mismatches allowed), read pairs are removed at this step to retain only high-quality reads. The information on the remaining read pairs are provided in section *FASTQ Read Statistics*. The files with clipped reads are provided in the FASTQ directory and are named `*trimmed_1.fastq.gz` and `*trimmed_2.fastq.gz`. These files are not directly used as inputs for the final microbiome profiling, but are further processed as described in the following steps.

Merging If the ends of forward and reverse reads overlap, the reads are merged (assembled) to obtain a single, longer read that covers the full target region. If the target region is longer than two times the read length, merging should be impossible. If in such a case a read pair can still be merged, it is considered as an artifact and will be removed in the following quality filtering step. If the target region is only slightly shorter than two times the read length, merging my fail due to an insufficiently long high-quality overlap of the read ends. In such a case, typically only a fraction of the read pairs can be merged. In all abovementioned cases where some read pairs can't be merged, the forward read is retained and processed in the following steps instead.
In short, reads are merged if possible, and as a fallback the high quality forward read is used. No read pair is completely discarded in this step. See section *Read Merging* for additional details.

Quality filtering Merged reads are length filtered according to the expected length and known length variations of the target region (see table 1). Merged reads that are significantly shorter than the expected minimal target region length, or that are significantly longer than the expected maximal target region length, are discarded at this step. Merged and retained reads containing ambiguous bases ("N") are discarded.
The files with filtered reads are provided in the FASTQ directory and are named `*_merged_for_profiling_1.fastq.gz`. These files are used as inputs for the following microbiome profiling.

Microbiome profiling The length filtered merged reads and the quality clipped retained forward reads are used as input for the microbiome profiling, where as a first step chimeric reads are identified and removed. All details of the microbiome step can be found in section *Microbiome Profiling*:

- Methods description of chimera removal, OTU picking, taxonomic assignment, etc.
- Tables with statistics describing the results of microbiome profiling
- Overview of the taxonomic composition of samples
- Detailed descriptions of delivered result files

Region code	Expected length	Merging efficiency
MI16Sa	ca. 395 bp	high
COIa	ca. 650 bp	not expected
CYTBa	(highly variable)	(highly variable)
Fu18Sa	ca. 290 bp	high
ITS1b	(highly variable)	high
PITS1a	ca. 445 bp	high
ITS2a	ca. 350 bp	high
TRNLa	(highly variable)	high
V1V3a	ca. 490 bp	moderate
V3V4a	ca. 445 bp	high
V3V5b	ca. 535 bp	high

Table 1: Standard target regions, expected lengths (rough average), and expected merging efficiency.

2 Microbiome Profiling

2.1 Results

This section summarizes the results of read preprocessing, OTU picking, and taxonomic assignment. A description of the applied methodology and according literature references are provided in the section *Methods*. Descriptions of result files and visualizations are provided in the section *Output Files and Descriptions*.

2.1.1 Statistics

Total number of input sequences	254 524	100.0%
Remaining sequences after preprocessing and quality filtering	254 414	100.0%
Remaining sequences after chimera detection and filtering	254 169	99.9%
Total number of sequences assigned to OTUs	221 472	87.0%
Total number of sequences assigned to taxa	208 356	81.9%
Total number of OTUs	568	100.0%
Number of OTUs assigend to taxa	464	81.7%

Table 2: Summarized statistics

The number of OTUs correlates with the diversity of the data set. Sequences that were considered as noise by the OTU picking algorithm were not assigned to an OTU. The fraction of OTUs that could be assigned to taxa indicates how well the microbiome is represented in the used reference database.

Sample	**1)**	**2)**	**3)**	**4)**	**5)**
952043.ITS2a	59 759	99.9%	83.4%	74.0%	317
952044.ITS2a	65 803	99.8%	87.4%	84.1%	299
952045.ITS2a	62 994	99.9%	83.6%	76.4%	319
952046.ITS2a	65 968	99.8%	93.1%	91.9%	299

Table 3: **1)** Input sequences. **2)** Sequences after preprocessing and chimera removal. **3)** Sequences assigned to OTUs. **4)** Sequences assigned to taxa. **5)** Median sequence length after preprocessing.

The tables can be found as files in the results directory. Please see the according section for details about result files.

2.1.2 Taxonomic Composition of Samples

The following table provides an overview of the identified taxonomic units in each sample. The most specific taxonomic units are listed with their taxonomy level and fraction (k...kingdom, p...phylum, c...class, o...order, f...family, g...genus, s...species). The most specific taxonomic unit is the lowest common taxonomic unit of the listed species (small font). These species came up as best hits of the OTUs representative sequences during the database comparison.
Next to each sample name, the total number of reads of this sample that were assigned to OTUs is given. All taxonomic units with less than 0.1% of reads are collapsed in the category "Other". If the representative sequence of an OTU had no significant database match, no taxonomic unit could be assigned. The total number of reads of these unclassified OTUs is stated as category "Unclassified".
Depending on the type of analysis, some taxonomic units might be removed as they to not match the expected clade, e.g. eukaryotes in a bacterial microbiome analysis. The number of removed reads is stated as category "Filtered". If this category is not listed, no filtering was performed.

Sample Name (read counts)

Taxonomic Level	Taxonomic Unit	Fraction
952043.ITS2a (44 222 reads)		
s	**Paramecium tetraurelia** (59 OTUs with 99-100% identity in 289-315bp to: Paramecium tetraurelia)	41.0%
s	**Cercomonas wylezichi** (13 OTUs with 98% identity in 375bp to: Cercomonas wylezichi)	11.9%
s	**Thaumatomonas constricta** (9 OTUs with 98% identity in 428-436bp to: Thaumatomonas constricta)	10.4%
s	**Pseudogastrostyla sp. YH-2018a** (6 OTUs with 94-95% identity in 340bp to: Pseudogastrostyla sp. YH-2018a)	7.8%
g	**Saccharomyces** (4 OTUs with 99-100% identity in 380-381bp to: Saccharomyces cerevisiae, Saccharomyces cf. cerevisiae/paradoxus, Saccharomyces sp.)	4.5%
g	**Plectosphaerella** (3 OTUs with 99-100% identity in 317bp to: Plectosphaerella cucumerina, Plectosphaerella oligotrophica, Plectosphaerella pauciseptata, Plectosphaerella sp. C21)	4.5%
s	**Paramecium multimicronucleatum** (3 OTUs with 99-100% identity in 296bp to: Paramecium multimicronucleatum)	2.9%
s	**Paramecium polycaryum** (2 OTUs with 99-100% identity in 303bp to: Paramecium polycaryum)	2.4%
s	**Cyclidium glaucoma** (11 OTUs with 98-100% identity in 321-327bp to: Cyclidium glaucoma)	2.2%
s	**Hemiurosomoida longa** (2 OTUs with 99% identity in 339bp to: Hemiurosomoida longa)	2.0%
s	**Strombidinopsis batos** (2 OTUs with 81% identity in 320-321bp to: Strombidinopsis batos)	1.7%
g	**Vorticella** (1 OTU with 97% identity in 296bp to: Vorticella convallaria, Vorticella sp. 23 PS-2013)	1.3%
o	**Saccharomycetales** (1 OTU with 100% identity in 380bp to: Hanseniaspora sp., Saccharomyces boulardii (nom. inval.), Saccharomyces cerevisiae, Saccharomyces cf. cerevisiae/paradoxus, Saccharomyces paradoxus, Saccharomyces sp.)	1.1%
s	**Opercularia sp. ZhW-2019a** (1 OTU with 100% identity in 299bp to: Opercularia sp. ZhW-2019a)	0.8%
s	**Parafurgasonia sp. FG-2015** (1 OTU with 71% identity in 329bp to: Parafurgasonia sp. FG-2015)	0.7%
s	**Chilodonella piscicola** (1 OTU with 99% identity in 304bp to: Chilodonella piscicola)	0.7%
s	**Capsicum annuum** (2 OTUs with 98-99% identity in 374bp to: Capsicum annuum)	0.6%
s	**Fuscheria nodosa** (3 OTUs with 87-88% identity in 244bp to: Fuscheria nodosa)	0.5%
s	**Rhabdostyla sp. 5 PPS-2010** (1 OTU with 96% identity in 294bp to: Rhabdostyla sp. 5 PPS-2010)	0.4%
s	**Saccharomyces cerevisiae** (2 OTUs with 99-100% identity in 382-383bp to: Saccharomyces cerevisiae)	0.4%
g	**Fusarium** (1 OTU with 100% identity in 297bp to: 3 unclassified Fusarium strains, Fusarium cf. oxysporum, Fusarium falciforme, Fusarium oxysporum)	0.4%
s	**Paracercomonas vonderheydeni** (1 OTU with 87% identity in 332bp to: Paracercomonas vonderheydeni)	0.3%

g	**Penicillium** (3 OTUs with 99-100% identity in 310bp to: 6 unclassified Penicillium strains, Penicillium albocoremium, Penicillium aurantiocandidum, Penicillium aurantiogriseum, Penicillium brevistipitatum, Penicillium cellarum, Penicillium chrysogenum, Penicillium concentricum, Penicillium cordubense, Penicillium cyclopium, Penicillium expansum, Penicillium freii, Penicillium hirsutum, Penicillium hordei, Penicillium italicum, Penicillium lapidosum, Penicillium martensii, Penicillium neoechinulatum, Penicillium nordicum, Penicillium polonicum, Penicillium resticulosum, Penicillium robsamsonii, Penicillium solitum, Penicillium thomii, Penicillium thymicola, Penicillium tricolor, Penicillium ulaiense, Penicillium verrucosum, Penicillium viridicatum)	0.3%
f	**Plectosphaerellaceae** (1 OTU with 100% identity in 318bp to: Gibellulopsis nigrescens, Gibellulopsis sp., Verticillium sp. DU18)	0.2%
s	**Coniochaeta hoffmannii** (1 OTU with 100% identity in 303bp to: Coniochaeta hoffmannii)	0.2%
f	**Pythiaceae** (1 OTU with 100% identity in 281bp to: Phytopythium citrinum, Pythium aff. dissotocum, Pythium aff. pachycaule, Pythium coloratum, Pythium diclinum, Pythium dissotocum, Pythium lutarium, Pythium sp.)	0.2%
g	**Tetrahymena** (1 OTU with 100% identity in 315bp to: Tetrahymena hegewischi, Tetrahymena tropicalis)	0.2%
	Other (6 OTUs with 0.4%)	0.4%
	Unclassified (5 601 reads)	
	Filtered (0 reads)	

952044.ITS2a (55 367 reads)

f	**Cephalothecaceae** (39 OTUs with 97-100% identity in 292-299bp to: 2 unclassified Cephalotheca strains, Cephalotheca sulfurea, Phialemonium inflatum, Phialemonium limoniforme, Phialemonium sp. DI14-259)	48.1%
s	**Conlarium sacchari** (10 OTUs with 98-100% identity in 311-312bp to: Conlarium sacchari)	5.6%
f	**Chaetomiaceae** (2 OTUs with 99-100% identity in 300bp to: 3 unclassified Chaetomium strains, Chaetomium sphaerale, Humicola homopilata)	5.3%
f	**Thermoascaceae** (5 OTUs with 97-100% identity in 313-322bp to: 2 unclassified Paecilomyces strains, 3 unclassified Byssochlamys strains, Byssochlamys nivea, Paecilomyces dactylethromorphus, Paecilomyces tabacinus)	4.8%
g	**Pseudeurotium** (3 OTUs with 91-100% identity in 294bp to: 3 unclassified Pseudeurotium strains, Pseudeurotium bakeri, Pseudeurotium hygrophilum, Pseudeurotium ovale, Pseudeurotium zonatum)	4.2%
s	**Trichomonascus vanleenenius** (2 OTUs with 99-100% identity in 315-317bp to: Trichomonascus vanleenenius)	3.2%
s	**Oidiodendron rhodogenum** (3 OTUs with 99-100% identity in 287bp to: Oidiodendron rhodogenum)	2.8%
s	**Oidiodendron periconioides** (3 OTUs with 98-99% identity in 289bp to: Oidiodendron periconioides)	1.8%
p	**Ascomycota** (9 OTUs with 98-100% identity in 293-323bp to: 14 unclassified Geomyces strains, 17 unclassified Coniochaeta strains, 4 unclassified Pseudogymnoascus strains, 6 unclassified Aureobasidium strains, 7 unclassified Penicillium strains, Blumeria graminis, Chrysosporium merdarium, Chrysosporium sp. 23WI04, Coniochaeta fasciculata, Coniochaeta velutina, Geomyces auratus, Ophiocordyceps sinensis, Penicillium aeneum, Penicillium citreonigrum, Phialemonium atrogriseum, Phialocephala fortinii, Phialophora intermedia, Phialophora mustea, Pleurostoma richardsiae, Pleurostoma sp., Pseudogymnoascus pannorum, Purpureocillium lilacinum, Trichoderma aff. songyi, Trichoderma atroviride, Trichoderma sp., Trichoderma viride)	1.7%
s	**Sporothrix cf. inflata 2 PB-2018** (1 OTU with 99% identity in 335bp to: Sporothrix cf. inflata 2 PB-2018)	1.7%
g	**Trichoderma** (3 OTUs with 100% identity in 318-320bp to: 2 unclassified Trichoderma strains, Trichoderma afroharzianum, Trichoderma asperellum, Trichoderma atrobrunneum, Trichoderma atroviride, Trichoderma aureoviride, Trichoderma breve, Trichoderma cf. harzianum, Trichoderma hamatum, Trichoderma harzianum, Trichoderma lentiforme, Trichoderma lixii, Trichoderma rifaii, Trichoderma rugulosum, Trichoderma virens, Trichoderma viride)	1.4%
o	**Eurotiales** (1 OTU with 100% identity in 315bp to: 6 unclassified Penicillium strains, Paecilomyces parvisporus, Paecilomyces sp. CZ-2011c, Penicillium amphipolaria, Penicillium daleae, Penicillium simplicissimum)	1.4%
g	**Coniochaeta** (1 OTU with 100% identity in 302bp to: 4 unclassified Coniochaeta strains, Coniochaeta lignicola, Coniochaeta velutina)	1.4%
s	**Pseudeurotium bakeri** (1 OTU with 100% identity in 294bp to: Pseudeurotium bakeri)	1.3%
s	**Conlarium sp. JR11** (1 OTU with 95% identity in 311bp to: Conlarium sp. JR11)	1.3%
o	**Hypocreales** (1 OTU with 100% identity in 323bp to: 2 unclassified cf. Trichoderma strains, 6 unclassified Trichoderma strains, Fusarium sp., Trichoderma atroviride, Trichoderma viride)	1.3%
s	**Tausonia pullulans** (2 OTUs with 98% identity in 303bp to: Tausonia pullulans)	1.1%
g	**Penicillium** (5 OTUs with 100% identity in 304-313bp to: 17, Penicillium)	0.8%
s	**Oidiodendron sp. 1 2017** (1 OTU with 100% identity in 287bp to: Oidiodendron sp. 1 2017)	0.8%

s	**Pseudeurotium zonatum** (1 OTU with 100% identity in 294bp to: Pseudeurotium zonatum)	0.6%
k	**Eukaryota** (2 OTUs with 99-100% identity in 437bp to: Mortierella chlamydospora, Pythium myriotylum)	0.6%
s	**Pseudogymnoascus pannorum** (4 OTUs with 93-100% identity in 293-294bp to: Pseudogymnoascus pannorum)	0.5%
s	**Sebacina sp.** (4 OTUs with 81-82% identity in 361-367bp to: Sebacina sp.)	0.5%
f	**Aspergillaceae** (1 OTU with 100% identity in 305bp to: 2 unclassified Penicillium strains, Aspergillus flavus, Penicillium aurantioviolaceum, Penicillium glabrum, Penicillium grancanariae, Penicillium palmense, Penicillium pulvis, Penicillium purpurascens, Penicillium spinulosum, Penicillium thomii)	0.5%
s	**Pezoloma ericae** (1 OTU with 99% identity in 292bp to: Pezoloma ericae)	0.5%
c	**Sordariomycetes** (4 OTUs with 98-100% identity in 293-314bp to: 2 unclassified Chloridium strains, Phaeoacremonium krajdenii, Phialocephala humicola, Pleurostoma repens)	0.4%
s	**Capsicum annuum** (2 OTUs with 98-99% identity in 374bp to: Capsicum annuum)	0.4%
g	**Oidiodendron** (2 OTUs with 100% identity in 288-289bp to: 2 unclassified Oidiodendron strains, Oidiodendron echinulatum, Oidiodendron griseum)	0.4%
s	**Umbelopsis isabellina** (1 OTU with 100% identity in 344bp to: Umbelopsis isabellina)	0.4%
s	**Umbelopsis sp.** (2 OTUs with 91-92% identity in 346-348bp to: Umbelopsis sp.)	0.3%
g	**Conlarium** (4 OTUs with 92-100% identity in 309-314bp to: Conlarium nanningense, Conlarium sp.)	0.3%
f	**Ophiostomataceae** (1 OTU with 100% identity in 335bp to: Raffaelea quercivora, Sporothrix sp.)	0.3%
f	**Pseudeurotiaceae** (3 OTUs with 91-100% identity in 294bp to: 10 unclassified Pseudogymnoascus strains, 2 unclassified Pseudeurotium strains, 36 unclassified Geomyces strains, Geomyces vinaceus, Pseudeurotium bakeri, Pseudeurotium zonatum, Pseudogymnoascus destructans, Pseudogymnoascus pannorum, Pseudogymnoascus roseus, Pseudogymnoascus verrucosus)	0.3%
s	**Olpidium brassicae** (1 OTU with 99% identity in 400bp to: Olpidium brassicae)	0.3%
s	**Candida sp. (in** (2 OTUs with 99-100% identity in 331bp to: Candida sp. (in)	0.3%
g	**Fusarium** (2 OTUs with 100% identity in 318bp to: Fusarium cf. solani, Fusarium falciforme, Fusarium keratoplasticum, Fusarium oxysporum, Fusarium phaseoli, Fusarium sedimenticola, Fusarium solani, Fusarium sp., [Nectria] haematococca)	0.3%
s	**Uroleptus pisces** (1 OTU with 93% identity in 338bp to: Uroleptus pisces)	0.2%
s	**Rhodotorula sp. AY214** (1 OTU with 81% identity in 379bp to: Rhodotorula sp. AY214)	0.2%
s	**Myrmecridium thailandicum** (1 OTU with 82% identity in 298bp to: Myrmecridium thailandicum)	0.2%
s	**Sporothrix schenckii** (1 OTU with 100% identity in 336bp to: Sporothrix schenckii)	0.2%
s	**Pholiota multicingulata** (2 OTUs with 73-74% identity in 376-377bp to: Pholiota multicingulata)	0.2%
s	**Hemiurosomoida longa** (1 OTU with 100% identity in 339bp to: Hemiurosomoida longa)	0.2%
s	**Podospora ellisiana** (1 OTU with 96% identity in 299bp to: Podospora ellisiana)	0.1%
f	**Trichosporonaceae** (1 OTU with 100% identity in 311bp to: 12 unclassified Trichosporon strains, Apiotrichum dehoogii, Apiotrichum porosum, Apiotrichum xylopini)	0.1%
s	**Phialophora intermedia** (1 OTU with 100% identity in 299bp to: Phialophora intermedia)	0.1%
g	**Sporothrix** (2 OTUs with 100% identity in 335bp to: Sporothrix chilensis, Sporothrix schenckii, Sporothrix sp., Sporothrix stylites)	0.1%
s	**Hemiamphisiella terricola** (1 OTU with 99% identity in 339bp to: Hemiamphisiella terricola)	0.1%
s	**Zopfiella lundqvistii** (1 OTU with 97% identity in 299bp to: Zopfiella lundqvistii)	0.1%
s	**Paramecium tetraurelia** (1 OTU with 100% identity in 296bp to: Paramecium tetraurelia)	0.1%
s	**Scytalidium sp.** (1 OTU with 100% identity in 307bp to: Scytalidium sp.)	0.1%
s	**Clitopilus scyphoides** (1 OTU with 95% identity in 343bp to: Clitopilus scyphoides)	0.1%
s	**Cephalotheca sp.** (1 OTU with 100% identity in 299bp to: Cephalotheca sp.)	0.1%
	Other (14 OTUs with 0.9%)	0.9%
	Unclassified (2 173 reads)	
	Filtered (0 reads)	

952045.ITS2a (48 140 reads)

p	**Ascomycota** (31 OTUs with 99-100% identity in 302-310bp to: 7 unclassified Penicillium strains, Alternaria alternata, Penicillium astrolabium, Penicillium brevicompactum, Penicillium olsonii, Penicillium volgaense)	30.2%
s	**Hemiurosomoida longa** (17 OTUs with 99% identity in 339-355bp to: Hemiurosomoida longa)	11.5%
s	**Cyclidium glaucoma** (36 OTUs with 98-100% identity in 321-327bp to: Cyclidium glaucoma)	8.5%
g	**Desmodesmus** (3 OTUs with 99-100% identity in 389bp to: 6 unclassified Desmodesmus strains)	6.7%
s	**Capsicum annuum** (3 OTUs with 98-99% identity in 374bp to: Capsicum annuum)	4.9%
s	**Thaumatomonas constricta** (2 OTUs with 98% identity in 435bp to: Thaumatomonas constricta)	4.7%
s	**Paramecium multimicronucleatum** (1 OTU with 100% identity in 296bp to: Paramecium multimicronucleatum)	4.5%
g	**Plectosphaerella** (1 OTU with 100% identity in 317bp to: Plectosphaerella cucumerina, Plectosphaerella oligotrophica, Plectosphaerella pauciseptata, Plectosphaerella sp. C21)	3.7%
f	**Pythiaceae** (1 OTU with 100% identity in 281bp to: Phytopythium citrinum, Pythium aff. dissotocum, Pythium aff. pachycaule, Pythium coloratum, Pythium diclinum, Pythium dissotocum, Pythium lutarium, Pythium sp.)	3.6%
g	**Paramecium** (1 OTU with 100% identity in 296bp to: Paramecium biaurelia, Paramecium decaurelia, Paramecium dodecaurelia)	3.5%
g	**Penicillium** (1 OTU with 100% identity in 319bp to: Penicillium citrinum, Penicillium raistrickii, Penicillium sp. BMP3042, Penicillium steckii)	2.7%
s	**Cyclidium varibonneti** (7 OTUs with 76-91% identity in 326-344bp to: Cyclidium varibonneti)	2.7%
s	**Pseudogastrostyla sp. YH-2018a** (3 OTUs with 95% identity in 340bp to: Pseudogastrostyla sp. YH-2018a)	1.2%
s	**Paramecium tetraurelia** (2 OTUs with 99-100% identity in 296bp to: Paramecium tetraurelia)	1.2%
s	**Penicillium fluviserpens** (3 OTUs with 99-100% identity in 305bp to: Penicillium fluviserpens)	1.1%
g	**Cladosporium** (1 OTU with 100% identity in 297bp to: 3 unclassified Cladosporium strains, Cladosporium dominicanum, Cladosporium pseudocladosporioides, Cladosporium sphaerospermum)	1.1%
s	**Cercomonas zhukovi** (2 OTUs with 99% identity in 399-400bp to: Cercomonas zhukovi)	0.7%
s	**Desmodesmus intermedius** (1 OTU with 100% identity in 398bp to: Desmodesmus intermedius)	0.6%
s	**Paracercomonas vonderheydeni** (3 OTUs with 86-87% identity in 330-332bp to: Paracercomonas vonderheydeni)	0.6%
g	**Fusarium** (2 OTUs with 100% identity in 297-299bp to: 4 unclassified Fusarium strains, Fusarium cf. oxysporum, Fusarium falciforme, Fusarium oxysporum, Fusarium proliferatum)	0.6%
f	**Chlorellaceae** (1 OTU with 94% identity in 384bp to: Dictyosphaerium sp. UTEX 731, Mucidosphaerium pulchellum)	0.6%
g	**Saccharomyces** (2 OTUs with 100% identity in 380-381bp to: Saccharomyces cerevisiae, Saccharomyces cf. cerevisiae/paradoxus, Saccharomyces sp.)	0.5%
s	**Tolypocladium lignicola** (1 OTU with 95% identity in 335bp to: Tolypocladium lignicola)	0.5%
s	**Wallemia mellicola** (2 OTUs with 99% identity in 357-358bp to: Wallemia mellicola)	0.5%
s	**Westella botryoides** (1 OTU with 95% identity in 391bp to: Westella botryoides)	0.5%
s	**Litonotus crystallinus** (2 OTUs with 98-99% identity in 240bp to: Litonotus crystallinus)	0.5%
g	**Thaumatomonas** (1 OTU with 96% identity in 434bp to: 2 unclassified Thaumatomonas strains)	0.4%
s	**Candida sp. C22** (1 OTU with 100% identity in 341bp to: Candida sp. C22)	0.3%
g	**Capsicum** (1 OTU with 98% identity in 374bp to: Capsicum annuum, Capsicum frutescens)	0.3%
s	**Chaetonotus antrumus** (2 OTUs with 78-79% identity in 297bp to: Chaetonotus antrumus)	0.2%
f	**Trichosporonaceae** (1 OTU with 100% identity in 311bp to: 12 unclassified Trichosporon strains, Apiotrichum dehoogii, Apiotrichum porosum, Apiotrichum xylopini)	0.2%
s	**Micronuclearia podoventralis** (2 OTUs with 98-99% identity in 281bp to: Micronuclearia podoventralis)	0.2%
s	**Paramecium polycaryum** (1 OTU with 99% identity in 303bp to: Paramecium polycaryum)	0.2%
g	**Tetrahymena** (1 OTU with 99% identity in 315bp to: Tetrahymena hegewischi, Tetrahymena tropicalis)	0.1%
s	**Micractinium pusillum** (1 OTU with 98% identity in 389bp to: Micractinium pusillum)	0.1%
o	**Saccharomycetales** (1 OTU with 100% identity in 380bp to: Hanseniaspora sp., Saccharomyces boulardii (nom. inval.), Saccharomyces cerevisiae, Saccharomyces cf. cerevisiae/paradoxus, Saccharomyces paradoxus, Saccharomyces sp.)	0.1%

g	**Trichoderma** (1 OTU with 100% identity in 320bp to: Trichoderma asperelloides, Trichoderma asperellum, Trichoderma sp., Trichoderma viride, Trichoderma yunnanense)	0.1%
	Other (2 OTUs with 0.1%)	0.1%
	Unclassified (4 531 reads)	
	Filtered (0 reads)	
952046.ITS2a (60 627 reads)		
f	**Cephalothecaceae** (28 OTUs with 99-100% identity in 292-300bp to: Cephalotheca sulfurea, Phialemonium inflatum)	84.7%
s	**Trichocladium brosimi** (8 OTUs with 98-100% identity in 293-294bp to: Trichocladium brosimi)	7.7%
f	**Chaetomiaceae** (3 OTUs with 100% identity in 298-309bp to: 2 unclassified Thielavia strains, 6 unclassified Chaetomium strains, Chaetomium thermophilum, Mycothermus thermophilus, Ovatospora brasiliensis, Ovatospora mollicella, Ovatospora pseudomollicella, Thermothielavioides terrestris)	1.3%
s	**Conlarium sacchari** (4 OTUs with 99-100% identity in 311-312bp to: Conlarium sacchari)	0.9%
s	**Phialemonium inflatum** (4 OTUs with 99-100% identity in 318-319bp to: Phialemonium inflatum)	0.6%
s	**Achroceratosphaeria potamia** (2 OTUs with 82-83% identity in 303-310bp to: Achroceratosphaeria potamia)	0.5%
s	**Capsicum annuum** (3 OTUs with 98-99% identity in 374bp to: Capsicum annuum)	0.4%
s	**Hemiamphisiella terricola** (1 OTU with 99% identity in 339bp to: Hemiamphisiella terricola)	0.4%
g	**Aspergillus** (1 OTU with 100% identity in 313bp to: Aspergillus fumigatus, Aspergillus neoellipticus, Aspergillus niger, Aspergillus sp.)	0.4%
s	**Trichomonascus vanleenenius** (1 OTU with 100% identity in 315bp to: Trichomonascus vanleenenius)	0.4%
g	**Rasamsonia** (2 OTUs with 100% identity in 316-317bp to: Rasamsonia argillacea, Rasamsonia cylindrospora, Rasamsonia emersonii, Rasamsonia piperina)	0.4%
s	**Hemiurosomoida longa** (2 OTUs with 99-100% identity in 339bp to: Hemiurosomoida longa)	0.3%
s	**Tausonia pullulans** (2 OTUs with 98% identity in 303bp to: Tausonia pullulans)	0.3%
s	**Acrobeloides cf. nanus WB-2017** (3 OTUs with 99-100% identity in 407-409bp to: Acrobeloides cf. nanus WB-2017)	0.2%
g	**Mortierella** (1 OTU with 100% identity in 398bp to: 3 unclassified Mortierella strains, Mortierella alpina, Mortierella amoeboidea)	0.2%
s	**Sebacina sp.** (1 OTU with 82% identity in 367bp to: Sebacina sp.)	0.2%
s	**Anteholosticha marimonilata** (1 OTU with 94% identity in 338bp to: Anteholosticha marimonilata)	0.2%
o	**Eurotiales** (1 OTU with 100% identity in 315bp to: 6 unclassified Penicillium strains, Paecilomyces parvisporus, Paecilomyces sp. CZ-2011c, Penicillium amphipolaria, Penicillium daleae, Penicillium simplicissimum)	0.2%
s	**Cirrenalia sp.** (1 OTU with 84% identity in 340bp to: Cirrenalia sp.)	0.1%
s	**Umbelopsis sp.** (2 OTUs with 91-92% identity in 346-348bp to: Umbelopsis sp.)	0.1%
	Other (9 OTUs with 0.6%)	0.6%
	Unclassified (811 reads)	
	Filtered (0 reads)	

Table 4: Condensed overview of the taxonomic composition of samples.

This table can be found as a file in the results directory. Please see the according section for details about result files.

2.2 Methods

As a first step of the microbiome analysis, all reads with ambiguous bases ("N") were removed. Chimeric reads were identified and removed based on the de-novo algorithm of UCHIME (Edgar RC et al., 2011) as implemented in the VSEARCH package (Rognes T et al., 2016).
The remaining set of high-quality reads was processed using minimum entropy decomposition (Eren AM, 2013 and 2015). Minimum Entropy Decomposition (MED) provides a computationally efficient means to partition marker gene datasets into OTUs (Operational Taxonomic Units). Each OTU represents a distinct cluster with significant sequence divergence to any other cluster. By employing Shannon entropy, MED uses only the information-rich nucleotide positions across reads and iteratively partitions large datasets while omitting stochastic variation. The MED procedure outperformes classical, identity based clustering algorithms. Sequences can be partitioned based on relevant single nucleotide differences without being susceptible to random sequencing errors. **This allows a decomposition of sequence data sets with a single nucleotide resultion.** Furthermore, the MED procedure identifies and filters random "noise" in the dataset, i.e. sequences with a very low abundance (less than $\approx$ 0.02% of the average sample size).
To assign taxonomic information to each OTU, DC-MEGABLAST alignments of cluster representative sequences to the sequence database were performed. A most specific taxonomic assignment for each OTU was then transferred from the set of best-matching reference sequences (lowest common taxonomic unit of all best hits). Hereby, a sequence identity of 70% accross at least 80% of the representative sequence was a minimal requirement for considering reference sequences.
Further processing of OTUs and taxonomic assignments was performed using the QIIME software package (version 1.9.1, http://qiime.org/).

OTU-picking strategy: Minimum entropy decomposition
Reference database: /mnt/nsa3/projects/active/bioit_development/ebe_transfer/mdxMicrobiomeProfiling/ncbi_nt/nt
02-03_well_classified_only/nt.filtered.fa (Release 2020-02-03)

References:

- **OTU picking:** Eren AM et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16s rRNA gene data. Methods Ecol Evol (4), 1111-1119.
  Eren AM et al. (2015) Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME Journal advance online publication, doi: 10.1038/ismej.2014.195.
- **Taxonomic assignment:** Altschul SF et al. (1990) Basic local alignment search tool. J Mol Biol 215(3), 403-410.
- **QIIME:** Caporaso JG et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5), 335-336.
- **Chimera detection:**
  Rognes T et al. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584 https://doi.org/10.7717/peerj.2584.
  Edgar RC et al. (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27(16), 2194-2200.

2.3 Output Files and Descriptions

The *MicrobiomeProfiling* directory contains the result files. All relevant files are described below. Some of these descriptions were excerpted from the official QIIME tutorials (http://qiime.org/tutorials/index.html).

01_Taxonomy_shortlist.txt: One of the **main results** of the microbiome analysis. This file can be used to get a quick overview of the microbiome. It contains a summarized list of indentified taxonomic units for each sample. The first two columns are the sample name and the total number of reads that where assigned to OTUs in this sample. The following columns list all taxonomic units with at least 0.1% of reads assigned to them. The individual columns state:

- The number of reads assigned to the taxonomic unit.
- The number of different OTUs that where classified as this taxonomic unit.
- The taxonomic level of the taxonomic unit. One of k...kingdom, p...phylum, c...class, o...order, f...family, g...genus, s...species.
- The fraction of reads assigned to the taxonomic unit.
- The identity and length of the best BLAST hit(s) to the database and a list of species that match with these alignment scores (not for all analysis types).

All taxonomic units with less than 0.1% of reads are collapsed in the category "Other". If the representative sequence of an OTU had no significant database match, no taxonomic unit could be assigned. The total number of reads of these unclassified OTUs is stated as category "Unclassified".
Depending on the type of analysis, some taxonomic units might be removed as they to not match the expected clade, e.g. eukaryotes in a bacterial microbiome analysis. The number of removed reads is stated as category "Filtered". If this category is not listed, no filtering was performed.
Please consider the provided identity and length of the best BLAST hits. The stated taxonomic unit was derived as lowest common ancestor of the best hits, but in case of a low sequence identity, it might be more appropriate to assign a higher taxonomic level than that of the lowest common ancestor.

02_Taxonomy_table.txt: One of the **main results** of the microbiome analysis. There is one line for each taxonomic unit and one column for each sample. The entries of the matrix are the estimated abundances of the respective taxonomic unit/sample combination. The file can be imported into Excel for further processing (sorting, calculations, diagrams).

03_OTU_representative_sequences.fasta: One of the **main results** of the microbiome analysis. Contains all read sequences of OTU representatives in FASTA format. The FASTA header contains the OTU identifier, the read identifier of the representative, the number of reads in the corresponding OTU, and the taxonomic classification. Representatives without taxonomic assignment are marked as "Unassigned", "Unclassified" or as "NOHIT", depending on the OTU picking method. Please note that representative sequences are not sample specific, i.e. a representative read subsumes similar reads of all samples. Thus, the given number of reads is the total number of reads of all samples that were assigned to the corresponding OTU.
Please note that OTUs only subsume sequences with identical lengths. Thus, OTU representatives may be prefixes of other OTU representatives. This occurs if assembled read pairs and (unassembled) single reads are processed together.

04_OTU_table.biom: One of the **main results** of the microbiome analysis. A file in BIOM format (http://biom-format.org/). This file is used as input by many QIIME scripts and is useful for downstream processing. OTUs of all samples are contained in this file.

05_OTU_table.txt: There is one line for each OTU and one column for each sample. The entries of the matrix are the estimated abundances of the respective OTU/sample combinations. The last column contains the taxonomic assignment of the OTU. OTUs without taxonomic assignment are marked

as "Unassigned", "Unclassified", or "NOHIT", depending on the OTU picking method. Please see file `02_Taxonomy_table.txt` for the abundances per taxonomic unit and sample. The file can be imported into Excel for further processing (sorting, calculations, diagrams).

06_OTU_table_summary.txt: Contains a summary describing `05_OTU_table.txt`.

07_OTU_table_per_sample_statistics.txt: Contains statistics for each sample in `05_OTU_table.txt`.

08_Processed_reads.fasta.gz: Contains all read sequences in FASTA format that went into the OTU-picking process. Reads that were identified as chimeric are not contained in this file. Processed-read identifiers consist of the sample name and a sequential number, followed by the raw-read identifier and the length of the read. Reads of all samples are contained in this file.

09_OTU_read_assignment.txt: A mapping of OTU identifier to read identifier, i.e. each line represents one OTU, the first column contains the OTU identifier, all other columns contain the identifier of reads that are part of the OTU. OTUs/Reads of all samples are contained in this file.

10_Taxonomy_plots: This directory contains files `area_charts.html` and `bar_charts.html`. These files can be opened with any web browser. The data of `02_Taxonomy_table.txt` (as relative abundances) will be displayed as either area or bar chart plots. There are several plots, each for a different level of taxonomy: from phylum to species. Hereby, higher level plots give a more coarse-grained view on the data than lower level plots. Mouseover the plots to see which taxa are contributing to the percentage shown, and a click on the hyperlinks in the legend starts a web-search using the most specific taxonomic unit. Charts, legends, and tables can be exported by clicking on the respective hyperlinks.