eurofins | Genomics

Data Analysis Report: Microbiome Profiling

Project / Study: EF-Demo

Date: August 23, 2022

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
V3V5	ca. 600 bp	not expected

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	215 277	100.0%
Remaining sequences after preprocessing and quality filtering	215 263	100.0%
Remaining sequences after chimera detection and filtering	215 228	100.0%
Total number of sequences assigned to OTUs	155 307	72.1%
Total number of sequences assigned to taxa	155 307	72.1%
Copy-number corrected total count	39 136	-
Total number of OTUs	130	100.0%
Number of OTUs assigend to taxa	130	100.0%

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. A copy-number correction was performed for bacterial species only, see Angly FE et al. (2014). To do so, the number of reads assigned to a species was divided by the known or assumed copy-number of marker genes/regions. The resulting corrected total count may be significantly lower than the (raw) total number of assigned reads.

Sample	**1)**	**2)**	**3)**	**4)**	**5)**	**6)**
Zymo100.pool.purified.3.V3V4a	71 750	100.0%	72.7%	72.7%	13 281	427
Zymo100.pool.purified.4.V3V4a	71 890	100.0%	72.9%	72.9%	13 323	427
Zymo100.pool.purified.5.V3V4a	71 637	100.0%	70.8%	70.8%	12 532	427

Table 3: **1)** Input sequences. **2)** Sequences after preprocessing and chimera removal. **3)** Sequences assigned to OTUs. **4)** Sequences assigned to taxa. **5)** Count after lineage-specific copy-number correction. **6)** 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 corrected 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.
A copy-number correction was perfor for bacterial species only, see Angly FE et al. (2014). If the listed normalized fraction and raw fraction are identical, either no copy-number correction factor was available in the database or the factor is exactly one.

Sample Name (copy-number corrected read counts)

Taxonomic Level	Taxonomic Unit	Normalized Fraction	Raw Fraction
Zymo100.pool.purified.3.V3V4a (13 281 reads)			
g	**Salmonella** (5 OTUs with 99-100% identity in 422-424bp to: 2 unclassified Salmonella strains, Salmonella enterica)	**23.1%**	12.2%
g	**Listeria** (20 OTUs with 99-100% identity in 286-427bp to: Listeria innocua, Listeria ivanovii, Listeria monocytogenes, Listeria seeligeri, Listeria welshimeri)	**19.7%**	17.7%
f	**Enterobacteriaceae** (2 OTUs with 100% identity in 422-424bp to: Escherichia coli, Escherichia fergusonii, Escherichia marmotae, Salmonella sp. S13, Shigella boydii, Shigella flexneri, Shigella sonnei, Shigella sp.)	**16.7%**	8.8%
g	**Staphylococcus** (11 OTUs with 99-100% identity in 425-427bp to: Staphylococcus argenteus, Staphylococcus aureus)	**9.7%**	12.7%
g	**Lactobacillus** (7 OTUs with 99-100% identity in 425-427bp to: Lactobacillus fermentum, Lactobacillus oris, Lactobacillus sp.)	**7.7%**	11.1%
f	**Bacillaceae** (8 OTUs with 99-100% identity in 286-428bp to: Alkalihalobacillus halodurans, Bacillus atrophaeus, Bacillus halotolerans, Bacillus mojavensis, Bacillus sp. (in, Bacillus subtilis, Bacillus tequilensis)	**6.1%**	14.0%
c	**Bacilli** (8 OTUs with 99-100% identity in 425-427bp to: Enterococcus faecalis, Enterococcus faecium, Enterococcus sp., Staphylococcus aureus)	**5.3%**	11.1%
g	**Pseudomonas** (3 OTUs with 99-100% identity in 422-424bp to: Pseudomonas aeruginosa, Pseudomonas fluorescens, Pseudomonas knackmussii, Pseudomonas sp.)	**3.7%**	4.2%
s	**Staphylococcus aureus** (9 OTUs with 100% identity in 425-427bp to: Staphylococcus aureus)	**3.2%**	4.2%
s	**Escherichia coli** (1 OTU with 100% identity in 424bp to: Escherichia coli)	**3.0%**	1.6%
g	**Bacillus** (1 OTU with 100% identity in 428bp to: Bacillus intestinalis, Bacillus subtilis)	**0.9%**	1.8%
s	**Enterococcus faecium** (1 OTU with 100% identity in 427bp to: Enterococcus faecium)	**0.7%**	0.5%
g	**Escherichia** (1 OTU with 100% identity in 424bp to: Escherichia coli, Escherichia sp. UIWRF0665)	**0.2%**	0.1%
	Other	**0.0%**	0.0%
	Unclassified (0 reads)		
	Filtered (0 reads)		
Zymo100.pool.purified.4.V3V4a (13 323 reads)			
g	**Salmonella** (8 OTUs with 99-100% identity in 422-424bp to: 2 unclassified Salmonella strains, Salmonella enterica)	**23.1%**	12.2%
g	**Listeria** (19 OTUs with 99-100% identity in 286-427bp to: Listeria innocua, Listeria ivanovii, Listeria monocytogenes, Listeria seeligeri, Listeria welshimeri)	**19.3%**	17.4%
f	**Enterobacteriaceae** (3 OTUs with 100% identity in 280-424bp to: Escherichia coli, Escherichia fergusonii, Escherichia marmotae, Escherichia sp., Salmonella sp. S13, Shigella boydii, Shigella flexneri, Shigella sonnei, Shigella sp.)	**16.8%**	8.9%
g	**Staphylococcus** (11 OTUs with 99-100% identity in 425-427bp to: Staphylococcus argenteus, Staphylococcus aureus)	**9.5%**	12.4%

g	**Lactobacillus** (4 OTUs with 99-100% identity in 425-427bp to: Lactobacillus fermentum, Lactobacillus oris, Lactobacillus sp.)	**7.8%**	11.2%
f	**Bacillaceae** (11 OTUs with 99-100% identity in 286-428bp to: Alkalihalobacillus halodurans, Bacillus atrophaeus, Bacillus halotolerans, Bacillus mojavensis, Bacillus sp. (in, Bacillus subtilis, Bacillus tequilensis)	**6.6%**	15.0%
c	**Bacilli** (10 OTUs with 99-100% identity in 424-427bp to: Enterococcus faecalis, Enterococcus faecium, Enterococcus sp., Staphylococcus aureus)	**4.9%**	10.2%
g	**Pseudomonas** (4 OTUs with 99-100% identity in 283-424bp to: Pseudomonas aeruginosa, Pseudomonas fluorescens, Pseudomonas knackmussii, Pseudomonas sp.)	**3.9%**	4.4%
s	**Escherichia coli** (2 OTUs with 100% identity in 424bp to: Escherichia coli)	**3.1%**	1.6%
s	**Staphylococcus aureus** (10 OTUs with 100% identity in 425-427bp to: Staphylococcus aureus)	**3.0%**	3.9%
g	**Bacillus** (1 OTU with 100% identity in 428bp to: Bacillus intestinalis, Bacillus subtilis)	**0.9%**	2.0%
s	**Enterococcus faecium** (1 OTU with 100% identity in 427bp to: Enterococcus faecium)	**0.8%**	0.5%
g	**Escherichia** (1 OTU with 100% identity in 424bp to: Escherichia coli, Escherichia sp. UIWRF0665)	**0.3%**	0.1%
	Other	**0.0%**	0.0%
	Unclassified (0 reads)		
	Filtered (0 reads)		
Zymo100.pool.purified.5.V3V4a (12 532 reads)			
g	**Salmonella** (7 OTUs with 99-100% identity in 422-424bp to: 2 unclassified Salmonella strains, Salmonella enterica)	**19.7%**	10.1%
g	**Listeria** (14 OTUs with 99-100% identity in 425-427bp to: Listeria innocua, Listeria ivanovii, Listeria monocytogenes, Listeria seeligeri, Listeria welshimeri)	**19.0%**	16.6%
f	**Enterobacteriaceae** (3 OTUs with 100% identity in 280-424bp to: Escherichia coli, Escherichia fergusonii, Escherichia marmotae, Escherichia sp., Salmonella sp. S13, Shigella boydii, Shigella flexneri, Shigella sonnei, Shigella sp.)	**17.1%**	8.8%
g	**Staphylococcus** (9 OTUs with 99-100% identity in 425-427bp to: Staphylococcus argenteus, Staphylococcus aureus)	**10.8%**	13.7%
g	**Lactobacillus** (7 OTUs with 99-100% identity in 425-427bp to: Lactobacillus fermentum, Lactobacillus oris, Lactobacillus sp.)	**7.6%**	10.6%
f	**Bacillaceae** (16 OTUs with 99-100% identity in 286-428bp to: Alkalihalobacillus halodurans, Bacillus atrophaeus, Bacillus halotolerans, Bacillus mojavensis, Bacillus sp. (in, Bacillus subtilis, Bacillus tequilensis)	**7.5%**	16.7%
c	**Bacilli** (8 OTUs with 99-100% identity in 425-427bp to: Enterococcus faecalis, Enterococcus faecium, Enterococcus sp., Staphylococcus aureus)	**5.4%**	10.9%
g	**Pseudomonas** (3 OTUs with 99-100% identity in 422-424bp to: Pseudomonas aeruginosa, Pseudomonas fluorescens, Pseudomonas knackmussii, Pseudomonas sp.)	**3.9%**	4.4%
s	**Escherichia coli** (1 OTU with 100% identity in 424bp to: Escherichia coli)	**2.9%**	1.5%
s	**Staphylococcus aureus** (5 OTUs with 100% identity in 427bp to: Staphylococcus aureus)	**2.6%**	3.4%
o	**Enterobacterales** (1 OTU with 100% identity in 283bp to: Enterobacter cloacae, Enterobacter kobei, Enterobacter ludwigii, Enterobacter sp., Pantoea agglomerans, Salmonella enterica)	**1.3%**	0.6%
g	**Bacillus** (1 OTU with 100% identity in 428bp to: Bacillus intestinalis, Bacillus subtilis)	**1.0%**	2.0%
s	**Enterococcus faecium** (1 OTU with 100% identity in 427bp to: Enterococcus faecium)	**0.8%**	0.5%
g	**Escherichia** (2 OTUs with 100% identity in 424bp to: 2 unclassified Escherichia strains, Escherichia coli)	**0.5%**	0.3%
	Other	**0.1%**	0.1%
	Unclassified (0 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/). Abundances of bacterial taxonomic units were normalized using lineage-specific copy numbers of the relevant marker genes to improve estimates (Angly FE, 2014).

OTU-picking strategy: Minimum entropy decomposition
Reference database: /mnt/nsa3/projects/active/bioit_development/ebe_transfer/mdxMicrobiomeProfiling/ncbi_nt/n
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.
- **Copy number correction:** Angly FE et al. (2014) CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome 2:11.

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 abundancy-corrected fraction of reads assigned to the taxonomic unit.
- 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.