Data Analysis Report: Microbiome Profiling

Project / Study: EF-DEMO

Date: May 24, 2021

Table of Contents

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 FASTQ Read Statistics

The processing of sequencing reads according to primer sequences has been performed with in-house scripts. Only read pairs where the expected forward primer as well as the expected reverse primer were found have been kept for further analysis. For the identification of primer sequences no mismatches were allowed. The following table provides various statistics describing the sorted reads.

No	Sample	Read Pairs	Yield (Kbp)	%Q30	Mean Q
1	Sample1.AV3V4a	148 641	83 287	82.82	33.97
2	Sample2.AV3V4a	171 535	95 353	81.73	33.70
3	Sample3.V3V4a	110 850	63 037	79.60	33.21
	Total/Average	**431 026**	**241 677**	**81.38**	**33.63**

Table 2: FASTQ processing results.

Remarks:

- All reads are passed filter, i.e. reads have passed the default Illumina filter procedure (chastity filter).
- ”Yield (Kbp)”: number of bases called in kilobases.
- ”%Q30”: represents the percentage of bases with a quality score of at least 30 (inferred base call accuracy of 99.9%). The Q-score is a prediction of the probability of a wrong base call.

3 Read Merging

Paired-end reads were merged using the software FLASH (2.2.00, Magoc and Salzberg, 2011). Briefly, the FLASH algorithm considers all possible overlaps at or above a minimum length between the reads in a pair and chooses the overlap that results in the lowest proportion of mismatched bases in the overlapped region. FLASH computes a consensus sequence in the overlapped region by selecting at each overlapped position the base with the higher quality value. If both bases have an identical quality value, one is selected randomly. Pairs were merged with a minimum overlap size of 10bp to reduce falsepositive merges.

No	Sample	Total Pairs	Percent combined	Mean of lengths
1	Sample1.AV3V4a	148641	99.23%	391
2	Sample2.AV3V4a	171535	99.04%	383
3	Sample3.V3V4a	110850	97.35%	413

Table 3: Results of read merging.

Citation:
Magoc T and Salzberg S (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27 (21), 2957-63

Eurofins Genomics' products, services and applications reach the best quality and safety levels. They are carried out under strict QM and QA systems and comply with the following standards:

ISO 17025	Accredited analytical excellence	GLP	The gold standard to conduct non-clinical safety studies
ISO 13485	Oligonucleotides according to medical devices standard	GCP	Pharmacogenomic services for clinical studies
cGMP	Products and testing according to pharma and biotech requirements		

Eurofins Genomics Europe Sequencing GmbH • Jakob-Stadler-Platz 7 • 78467 Constance • Germany