Sequencing and Data Processing Report

Biomcare ApS

28/12/2024

Customer	Tove Pedersen
Customer ID	DA00206-23
Project	Processing and analysis of 16S rRNA gene data
Sample Type	Soil
Number of samples	14 samples
Type of data	Sequencing of 16S rRNA gene

In this *Sequencing and Data Processing Report* you will find information regarding data generation (DNA extraction, library preparation and sequencing), data quality evaluation and filtering as well as microbiome profiling.

All your data, illustrations, supplementary files and reports are now available for download in your private project folder on Biomcare's server. Please refer to the supplementary document *How to navigate your Biomcare folder* for details on how to find specific files and for instructions on how to interpret illustrations that are not introduced in this report.

Summary

Paired-end 16S rRNA gene amplicon sequencing was performed, targeting 30,000 reads per sample (details on sequencing are available below). The performed amplicon sequencing resulted in a mean read count of 122,936 reads across samples, with a high data quality (see section *Evaluation of raw data quality* below for more details).

The input data does not contain adapters/primers/barcodes as these were removed during splitting of the data from the sequencing platform (performed by the sequencing facility using FLASH). Both raw and adapters/primers/barcodes cleaned data is available in the "1_Raw_sequencing_data" folder. Quality filtering was performed using DADA2 and removed on average 6750 (min: 4,051, max: 9,573) reads per sample, while denoising of reads removed on average 8862 reads across samples. The read count following filtering was below the set threshold of 10000 reads per sample for 0 samples. As a result 0 samples were excluded from further analysis.

Data generation using Next Generation Sequencing

DNA extraction was performed by DNAsense, DK, and sequencing was performed by BMKgene, DE, on behalf of Biomcare.

The samples were centrifuged down for 15 min to remove stabilizer and dissolved in 1000 uL buffer.

Sampels were added 10 uL spike-in.

DNA was extracted from the soil samples using the FastDNA SPIN Kit for Soil, followed by DNA quality evaluation using a combination of Nanodrop, Qubit and Gel electrophoresis methods. A total of samples (of the 14 samples in total) passed DNA quality evaluation and were passed on to library preparation and sequencing with no remarks.

DNA extraction of samples was done using a slightly modified version of the standard protocol for FastDNA Spin kit for Soil (MP Biomedicals, USA) with the following exceptions: 500 µL of sample, 480 µL Sodium Phosphate Buffer and 120 µL MT Buffer were added to a Lysing Matrix E tube. Bead beating was performed at 6 m/s for 4x40s [3]. Gel electrophoresis using Tapestation 2200 and Genomic DNA screentapes (Agilent, USA) was used to validate product size and purity of a subset of DNA extracts. DNA concentration was measured using Qubit dsDNA HS/BR Assay kit (Thermo Fisher Scientific, USA).

The primer set was used to amplify the target region of 16S rRNA gene from the genomic DNA extracted from each sample. Both the forward and reverse 16S primers were tailed with sample-specific Illumina index sequences to allow for deep sequencing. The PCR was performed in a total reaction volume of 10 µl: DNA template 5-50 ng, *Vn F (10µM) 0.3 µl,* Vn R (10µM) 0.3 µl, KOD FX Neo Buffer 5 µl, dNTP (2 mM each) 2 µl, KOD FX Neo 0.2 µl, ddH2O up to 10 µl. Vn F and Vn R are selected according to the amplification area. After with initial denaturation at 95 °C for 5 min, followed by 25 cycles of denaturation at 95 °C for 30 s, annealing at 50 °C for 30 s, and extension at 72 °C for 40 s, and a final step at 72 °C for 7 min. The total of PCR amplicons is purified with Agencourt AMPure XP Beads (Beckman Coulter,Indianapolis, IN) and quantified using the Qubit dsDNA HS Assay Kit and Qubit 4.0 Fluorometer (Invitrogen, Thermo Fisher Scientific, Oregon, USA). After the individual quantification step, amplicons were pooled in equal amounts. For the constructed library, use Illumina NovaSeq 6000 (Illumina, Santiago CA, USA) for sequencing.

Evaluation of raw data quality

Biomcare uses the two software solutions *FastQC* and *DADA2* to evaluate the quality of the raw data generated using Next Generation Sequencing. If results from the quality evaluation indicate an issue with data generation, we bring this back to our sequencing facility (if Biomcare has generated the data) or to you (if you have provided us with raw sequencing data).

Evaluation of the quality of the raw data shows a high average base quality score (phred score) across read lengths as seen in the illustrations *QualityProfiles_F* and *QualityProfiles_R* (note F indicates forward reads and R indicate reverse reads). No sample (single fastq file) has reads flagged as poor quality. The GC content is on average 55% (min: 54% , max: 56%). Reads across samples have a mean read length of 250 bp.

If you wish to further evaluate the results from the data quality assessment, please refer to the supportive documents in your Biomcare folder and the guide on how to locate and interpret the relevant files (*How to navigate your Biomcare folder*).

Based on the evaluations, Biomcare selects appropriate quality filtering settings and steps. Below you will find a description of the performed data quality filtering.

Quality filtering

Quality filtering steps

- Removal of adapter/primer sequences (performed by the sequencing facility)
- Read quality and length filtering

Read quality and length filtering

When bases are called based on the data obtained from the sequencing platform, each base is annotated with a quality score. The quality score of the called bases (often in the form of phred scores) is accessed and used to remove both low-quality reads and low-quality bases at the ends of reads. We also remove a set number of bases from the left end of the reads as these generally have low base quality, remove reads with any uncalled bases and set a max on the expected error rate. Reads that are shorter than a defined length threshold following quality filtering are removed. Finally, reads are discarded if they match against the phiX genome. These steps are performed using DADA2. Please see **Table 2** for the applied quality filtering thresholds.

Microbiome profiling

After cleaning the data, we are ready to start generating the microbiome profiles. To do this we use the software DADA2, which is the most used solution in the field due to its high resolution, stemming from its very efficient denoising algorithm. The denoising steps aim to correct introduced sequencing errors and thereby facilitate higher precision in the identification of microbial variation. The microbiome profiling process includes four main steps:

- Calculation of error rates
- Sample inference and denoising
- Removal of chimeras
- Assigning taxonomy

Calculation of error rates

The errors made in the sequencing process generate a specific profile for each dataset. In order to denoise the sequences, an error model needs to be built. This is done using machine learning to estimate the error rates for each possible nucleotide transition for this specific dataset. This step is performed using DADA2 and the error rates can be seen below in **Figure 1**.

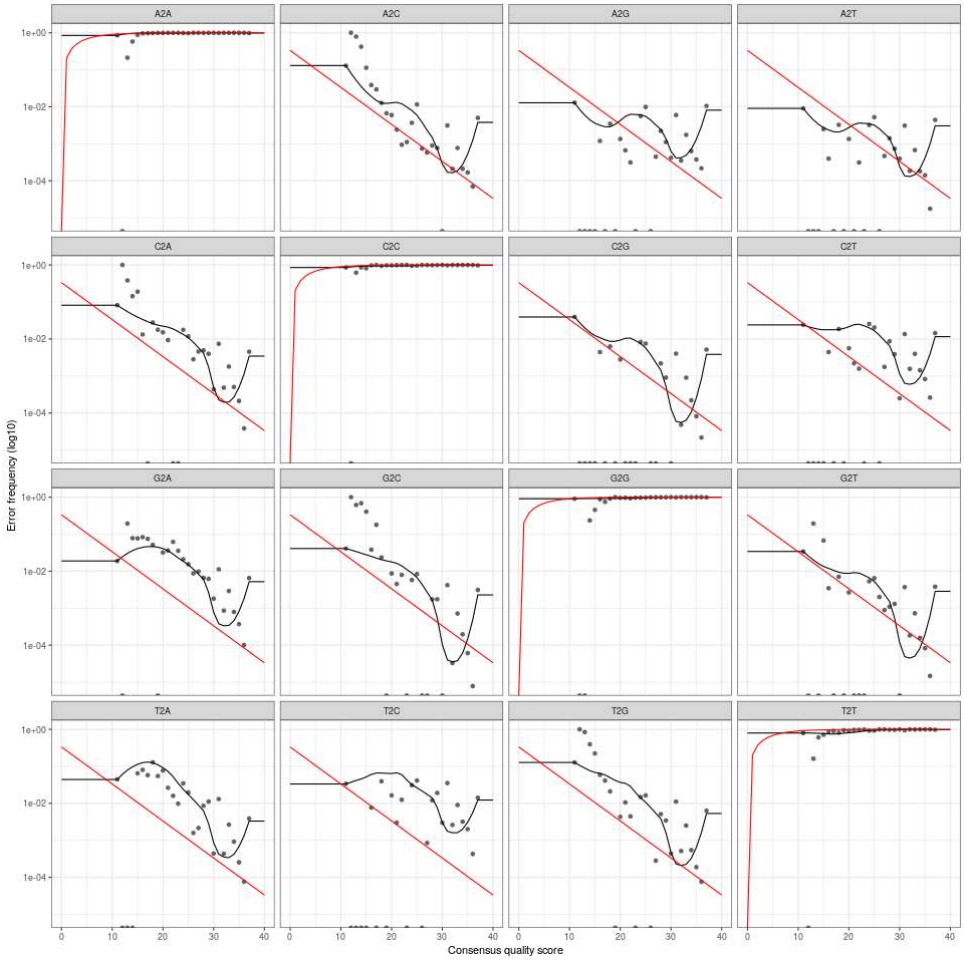

Figure 1: Error rates for all nucleotide transitions. The black dots are the observed error rates for each consensus quality score (x-axis). The black line shows the estimated error rates and the red line shows the expected error rates. The estimated error rates (black lines) should be modeling the observed error rates (black dots). This is generally the case in this example. Also, all error rates are decreasing with increasing quality score. We can thus proceed with the data processing.

Sample inference and denoising

DADA2 uses the error rates calculated in the previous step in order to correct for any sequencing errors present. This denoising step allows for sequence inference, thereby identifying which reads that came from the same sequence. DADA2 performs these steps as one step in its signature Divisive Amplicon Denoising Algorithm, leaving only those sequences which are likely to be truly present in the samples. Using these sequences, the abundance table of amplicon sequence variants (ASV) is constructed.

Removal of chimeras

DNA sequencing data contains a low number of sequences called chimeras. Chimeric sequences are artifacts that are formed when two or more biological sequences are incorrectly joined. Because these sequences do not represent a true microorganism, they must be removed before further data processing. Chimeras are identified by first locating the two parent sequences of the chimeras using the Needleman-Wunsch global algorithm. Finding two parent sequences which overlap exactly the child sequence from the left and right indicates the finding of a chimeric sequence. The identified chimeras are then removed. The resulting ASV table is saved and can be found in your project folder (location: 3_Microbiome_profiles/Microbiome_profile_tables, name: ASV.rds)

Assigning taxonomy

Sequences in the ASV table are annotated using the naive Bayesian classifier method available in DADA2, with reference data from the SILVA database. The taxonomy table is saved and can be found in your project folder (location: 3_Microbiome_profiles/Microbiome_profile_tables, name: taxa.rds)

Read counts through data generation, processing and microbiome profiling

For each sample, reads may be removed at each of the quality filtering steps and during microbiome profiling. Inspecting the number of reads filtered provides important information on the data quality and the validity of the performed processing. **Table 1** below provides an easy overview of the data at each step through data processing. The first row (Input) presents the raw sequencing data while the last row presents the clean microbiome profiles.

	Min.	1st Q.	Median	Mean	3rd Q.	Max.
Input	83068	98241	124481	122739	141448	182397
Filtered	4051	5481	6474	6750	8216	9573
Denoised	3841	6928	8652	8862	10981	14444
Merged	2680	3503	3899	4448	5789	6698
Nonchim	23771	29501	36378	37885	39460	78027
Clean	45369	53637	64200	64794	76850	85065

Table 1: Summary statistics for each step of data processing. The first and last row provide summary information on read counts in the raw sequencing data (input) and the cleaned data following the full microbiome profiling process (clean), respectively. The four rows in-between (filtered, denoised, merged and nonchim) are the summary information for the number of reads removed at each of these steps across all samples.

After data processing, some samples are likely to have a very low read count. This will in most instances be caused by low DNA concentration or quality of the sample that was sequenced. **Figure 2** below shows a stacked barplot of all samples in the project and gives a visual overview of data quality and read depth.

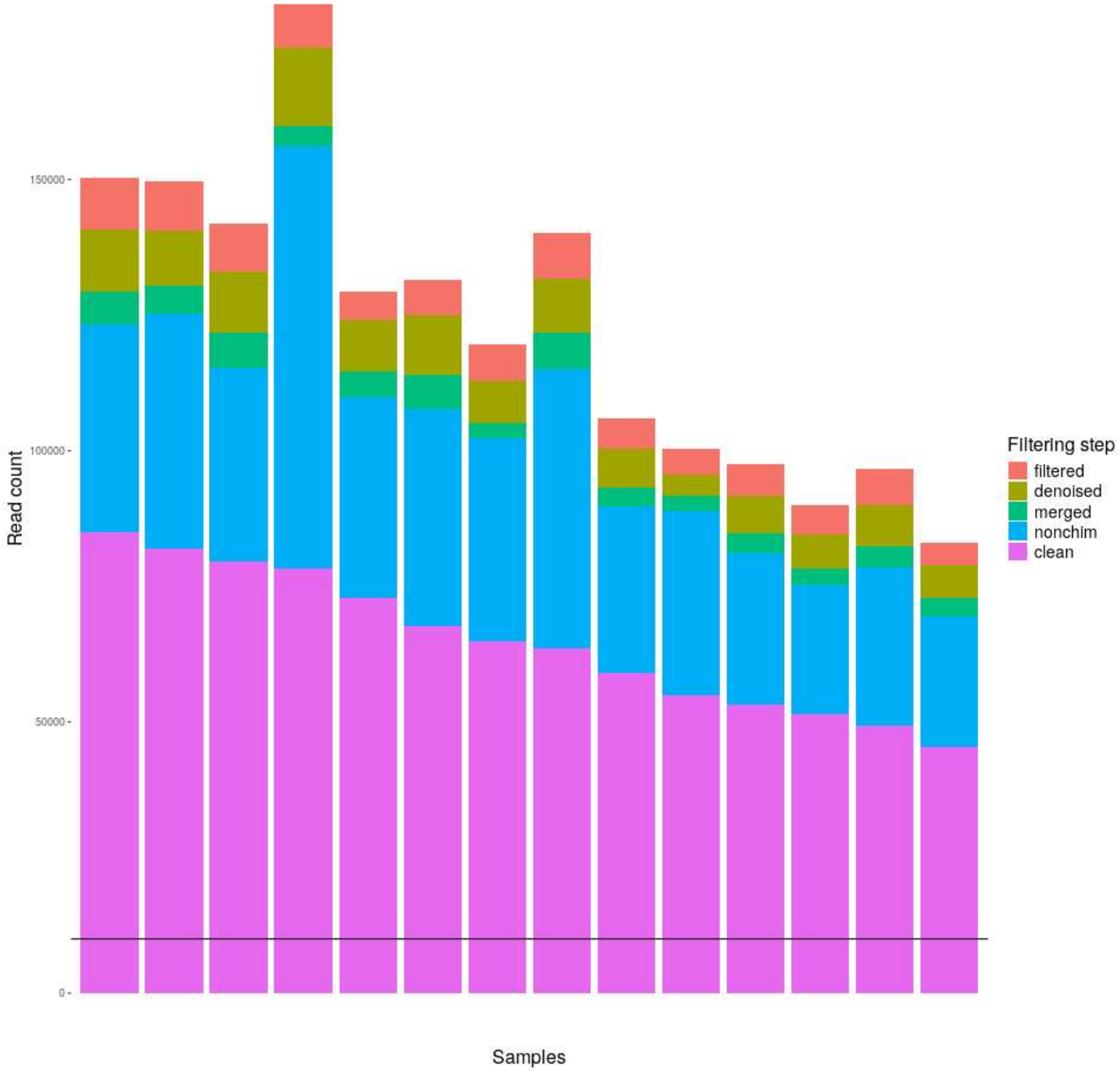

Figure 2: Summary of reads across samples in the project. Each bar shows reads for one sample. The clean reads are the number of reads (merged if paired-end data) that remain after processing the raw data and calling the microbiome profiles. Filtered is the number of reads removed during read filtering (removed due to low read quality or short length). Denoised is the number of reads removed due to duplication. Merged is the number of reads removed because the read-pair was missing. Nochim is the reads removed because they were found to be chimeras. The horizontal black line indicates the 10,000 reads threshold.

Software, settings, thresholds and reference data used

Reference databases and parameters

Parameters	Settings
Truncation quality	5

Parameters	Settings
Truncation length	F:0, R:0
Left triming length	F:0, R:0
Right trimming length	F:5, R:5
Max. length	Inf
Min. length	20
Max. #N	0
Min. Q	0
Max. Expected Errors	F:4, R:4
Remove PhiX	TRUE
Low complexity filter	0
Reference Data	SILVA
Licence or reference for reference data	https://creativecommons.org/licenses/by/4.0/ (https://creativecommons.org/licenses/by/4.0/)

Table 2: Reference and parameter values.

Software and package versions

Package	Version	Package	Version
OS	Ubuntu 20.04.4 LTS	**abind**	1.4-5
R	4.3.3	**yaml**	2.3.9
bitops	1.0-7	**vegan**	2.6-6.1
deldir	2.0-4	**codetools**	0.2-20
permute	0.9-7	**hwriter**	1.3.2.1
rlang	1.1.4	**lattice**	0.22-6
magrittr	2.0.3	**plyr**	1.8.9
ade4	1.7-22	**withr**	3.0.0
compiler	4.3.3	**evaluate**	0.24.0
mgcv	1.9-1	**survival**	3.7-0
systemfonts	1.1.0	**RcppParallel**	5.1.8
png	0.1-8	**zip**	2.3.1
vctrs	0.6.5	**xml2**	1.3.6

Package	Version	Package	Version
reshape2	1.4.4	**pillar**	1.9.0
pkgconfig	2.0.3	**foreach**	1.5.2
crayon	1.5.3	**generics**	0.1.3
fastmap	1.2.0	**vroom**	1.6.5
utf8	1.2.4	**RCurl**	1.98-1.16
rmarkdown	2.27	**hms**	1.1.3
tzdb	0.4.0	**munsell**	0.5.1
bit	4.0.5	**scales**	1.3.0
xfun	0.46	**glue**	1.7.0
zlibbioc	1.48.2	**tools**	4.3.3
cachem	1.1.0	**interp**	1.1-6
biomformat	1.30.0	**rhdf5**	2.46.1
highr	0.11	**ape**	5.8
rhdf5filters	1.14.1	**latticeExtra**	0.6-30
DelayedArray	0.28.0	**colorspace**	2.1-0
Rhdf5lib	1.24.2	**nlme**	3.1-165
jpeg	0.1-10	**GenomeInfoDbData**	1.2.11
parallel	4.3.3	**cli**	3.6.3
cluster	2.1.6	**fansi**	1.0.6
R6	2.5.1	**viridisLite**	0.4.2
bslib	0.7.0	**S4Arrays**	1.2.1
stringi	1.8.4	**svglite**	2.1.3
RColorBrewer	1.1-3	**gtable**	0.3.5
jquerylib	0.1.4	**sass**	0.4.9
iterators	1.0.14	**digest**	0.6.36
timechange	0.3.0	**SparseArray**	1.2.4
Matrix	1.6-5	**htmltools**	0.5.8.1
splines	4.3.3	**multtest**	2.58.0
igraph	2.0.3	**lifecycle**	1.0.4
tidyselect	1.2.1	**bit64**	4.0.5
rstudioapi	0.16.0	**MASS**	7.3-60.0.1

Table 3: Software and package versions.