

Sequencing and Data Processing Report

Biomcare ApS

27/12/2024

Customer	Tove Pedersen
Customer ID	DA00206-23
Project	Processing and analysis of ITS region
Sample Type	Soil
Number of samples	14 samples
Type of data	Sequencing of ITS2 region

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 ITS2 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 302,967 reads across samples, with a high data quality (see section *Evaluation of raw data quality* below for more details).

Both raw data 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 9627 (min: 5,617, max: 1.488e+04) reads per sample, while denoising of reads removed on average 3603 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 Massive Parallel 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 μ L of sample, 480 μ L Sodium Phosphate Buffer and 120 μ L 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).

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 53% (min: 51% , 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 sequences (performed by sequencing facility)
- Removal of target primer sequences (performed by Biomcare using Cutadapt v2.8 with default settings and specification of primers)
- 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 set thresholds.

Microbiome profiling

After cleaning the data, we are ready to generate 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**.

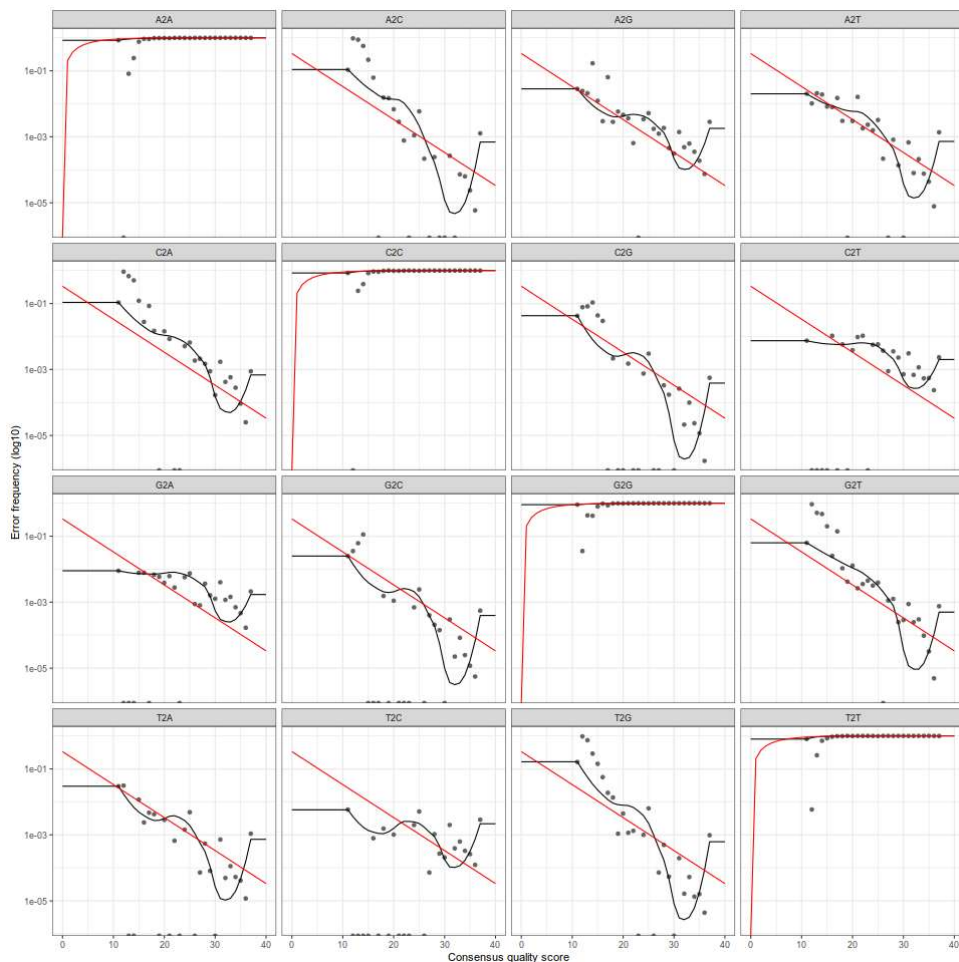


Figure 1: Error rates for all nucleotide transitions. The black dots are the observed error rates for each consensus quality score (x-axis). The black line shows the estimated error rates and the red line shows the expected error rates. The estimated error rates (black lines) should be modelling the 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 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 establishing what reads come 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. Chimera 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 chimera. 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 UNITE 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 provide 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	213706	244618	316804	302852	341916	408441
Filtered	5617	7619	9470	9627	10774	14884
Denoised	2564	3412	3684	3603	4059	4261
Merged	6418	9902	12592	13051	14373	27912
Nonchim	3578	5421	6827	7677	8124	16338
Clean	190183	217796	283370	268893	306553	360580

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.

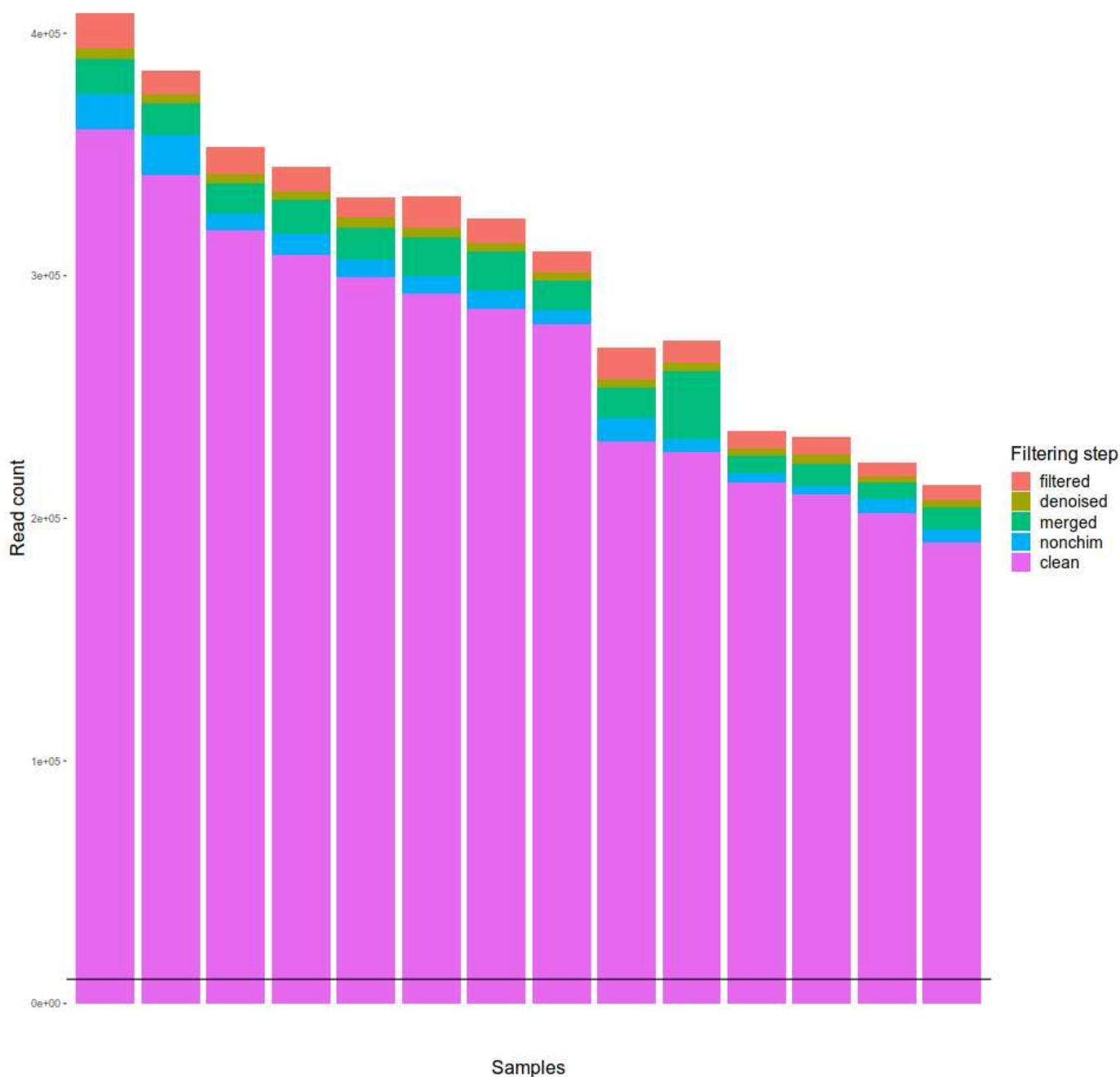


Figure 2: Summary of reads across samples in the project. Each bar shows reads for one sample. The clean reads are the number of reads (merged if paired-end data) that remain after processing the raw data and calling the microbiome profiles. Filtered is the number of reads removed during read filtering (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
Truncation length	F:0, R:0
Left trimming length	F:0, R:5
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	UNITE
Licence or reference for reference data	Abarenkov, Kessy; Zirk, Allan; Piirmann, Timo; Pöhönen, Raivo; Ivanov, Filipp; Nilsson, R. Henrik; Kõljalg, Urmas (2021): UNITE general FASTA release for Fungi 2. Version 10.05.2021. UNITE Community. https://doi.org/10.15156/BIO/1280089 (https://doi.org/10.15156/BIO/1280089)

Table 2: Reference and parameter values.

Software and package versions

Package	Version	Package	Version
OS	Ubuntu 20.04.4 LTS	rstudioapi	0.16.0
R	4.3.3	abind	1.4-5
bitops	1.0-7	yaml	2.3.9
deldir	2.0-4	vegan	2.6-6.1

Package	Version	Package	Version
permute	0.9-7	codetools	0.2-20
rlang	1.1.4	hwriter	1.3.2.1
magrittr	2.0.3	lattice	0.22-6
ade4	1.7-22	plyr	1.8.9
compiler	4.3.3	withr	3.0.0
mgcv	1.9-1	evaluate	0.24.0
systemfonts	1.1.0	survival	3.7-0
png	0.1-8	RcppParallel	5.1.8
vctrs	0.6.5	zip	2.3.1
reshape2	1.4.4	xml2	1.3.6
pkgconfig	2.0.3	pillar	1.9.0
crayon	1.5.3	foreach	1.5.2
fastmap	1.2.0	generics	0.1.3
utf8	1.2.4	RCurl	1.98-1.16
rmarkdown	2.27	hms	1.1.3
tzdb	0.4.0	munsell	0.5.1
xfun	0.46	scales	1.3.0
zlibbioc	1.48.2	glue	1.7.0
cachem	1.1.0	tools	4.3.3
biomformat	1.30.0	interp	1.1-6
highr	0.11	rhdf5	2.46.1
rhdf5filters	1.14.1	ape	5.8
DelayedArray	0.28.0	latticeExtra	0.6-30
Rhdf5lib	1.24.2	colorspace	2.1-0
jpeg	0.1-10	nlme	3.1-165
parallel	4.3.3	GenomeInfoDbData	1.2.11
cluster	2.1.6	cli	3.6.3
R6	2.5.1	fansi	1.0.6
bslib	0.7.0	viridisLite	0.4.2
stringi	1.8.4	S4Arrays	1.2.1
RColorBrewer	1.1-3	svglite	2.1.3
jquerylib	0.1.4	gtable	0.3.5

Package	Version	Package	Version
iterators	1.0.14	sass	0.4.9
timechange	0.3.0	digest	0.6.36
Matrix	1.6-5	SparseArray	1.2.4
splines	4.3.3	htmltools	0.5.8.1
igraph	2.0.3	multtest	2.58.0
tidyselect	1.2.1	lifecycle	1.0.4

Table 3: Software and package versions.