STØTTET AF
Promilleafgiftsfonden for landbrug

Dada2 pipeline

by BiomCare

10/12/2019

Customer	SEGES
Customer ID	DA00201-19
Project ID	Differences between bacetria and fungi in plowed and unplowed clay and sand samples
Sample Type	Soil

In this Data Processing Report, you will find information regarding DNA sequencing, data quality evaluation and filtering, and microbiome profiling.

All your data, illustrations, supplementary files and reports are now available for download in your private project folder on BiomCare's cloud. 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 metagenomic sequencing was performed, targeting 10.000 reads per sample (details on sequencing are available below). The performed amplicon sequencing resulted in a mean read count of 17758 reads across samples.

Evaluation of the quality of the raw data show a base quality score (phred score) above [28] across read lengths. No sample (single fastq file) has reads flagged as poor quality. The GC content is on average 56.54167 (min: 56 , max: 57). Reads across samples has a mean read length of 301bp.

The raw data does not contain adaptors/primers due to the use of customized primers for library preparsation, and therefore no removal of non-biological sequences was performed. Quality filtering was performed using DADA2 and removed on average 844.6667 (min: 255, max: 2949) reads per sample, while denoising of reads removed on average 2369.292 reads across samples. The read count following filtering and microbiome profiling was above the target threshold of 10.000 reads per sample for *** samples. The remaining samples (n=) had a low read-count following processing and we recommend removing these from further analysis. Inspection of the included quality control samples, as well as the mock community, indicated no contamination and a successful sequencing, filtering and microbiome profiling.

Table of contents

- Evaluation of raw data quality
- Quality filtering
    - Removal of adapter/primer sequences
    - Read quality and length filtering.
    - Calculate error rates
    - Sample Inference and Denoising
    - Remove chimeras
    - Read counts
- Software, settings, thresholds and reference data used

Data generation using Massive Parallel Sequencing

Sequencing is performed using polymerase chain reaction (PCR) for DNA amplification of the selected variable region(s) of the bacterial 16S rRNA gene. The product from PCR is normalized using the SequalPrep Normalization Kit, followed by sequencing on an Illumina MiSeq platform using v3 chemistry and multiplexing of up to 384 samples per run. This result in 2x300bp paired-end reads, that are demultiplexed while allowing for no mismatches in the index sequences.

Evaluation of raw data quality

BiomCare use FastQC and DADA2 to evaluate the quality of the raw sequencing data. BiomCare use the results from FastQC and DADA2 to evaluate the quality of the sequencing data, and to design appropriate quality filtering settings and steps. 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). The summary section above contains the key information form the data quality evaluation.

If you wish to further evaluate the results of the data quality assessment, please refer to the supportive documents in your BiomCare folder (location: 2_Quality_filtering_and_profiling/Illustrations) and the guide on how to read the illustrations found in the document Report Attachment 2.

Below you find a description of BiomCares data quality filtering. If data contain adaptors, adaptor removal is performed as part of the quality filtering.

Quality filtering

Quality filtering steps

- Removal of adapter/primer sequences
- Read quality and length filtering

Removal of adapter/primer sequences

Our sequencing facility uses custom primers wherever possible, leading to no sequence trimming of adapter/primer data needed for those samples. However, for those other samples, we always evaluate what measures are required to remove adaptor sequences from their next generation sequencing data. For some sequencing platforms, this is done during demultiplexing, however for

others it is necessary to remove the adaptor sequences during quality filtering. It is good practices to evaluate the presence of adaptor sequences and remove any remaining adaptor sequences before further processing. This step is performed using the software Cutadapt.

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. Reads that are shorter than a defined length threshold following quality filtering are removed. This step is performed using DADA2.

We do this for a number of reasons: a) to remove noise that may limited the information obtainable from the quality evaluation, b) to see if quality issues are easily handled with standard processing and c) to evaluate first-choice quality settings.

Calculate error rates

The errors made in the sequencing process generates a specific profile for each dataset. In order for 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**.

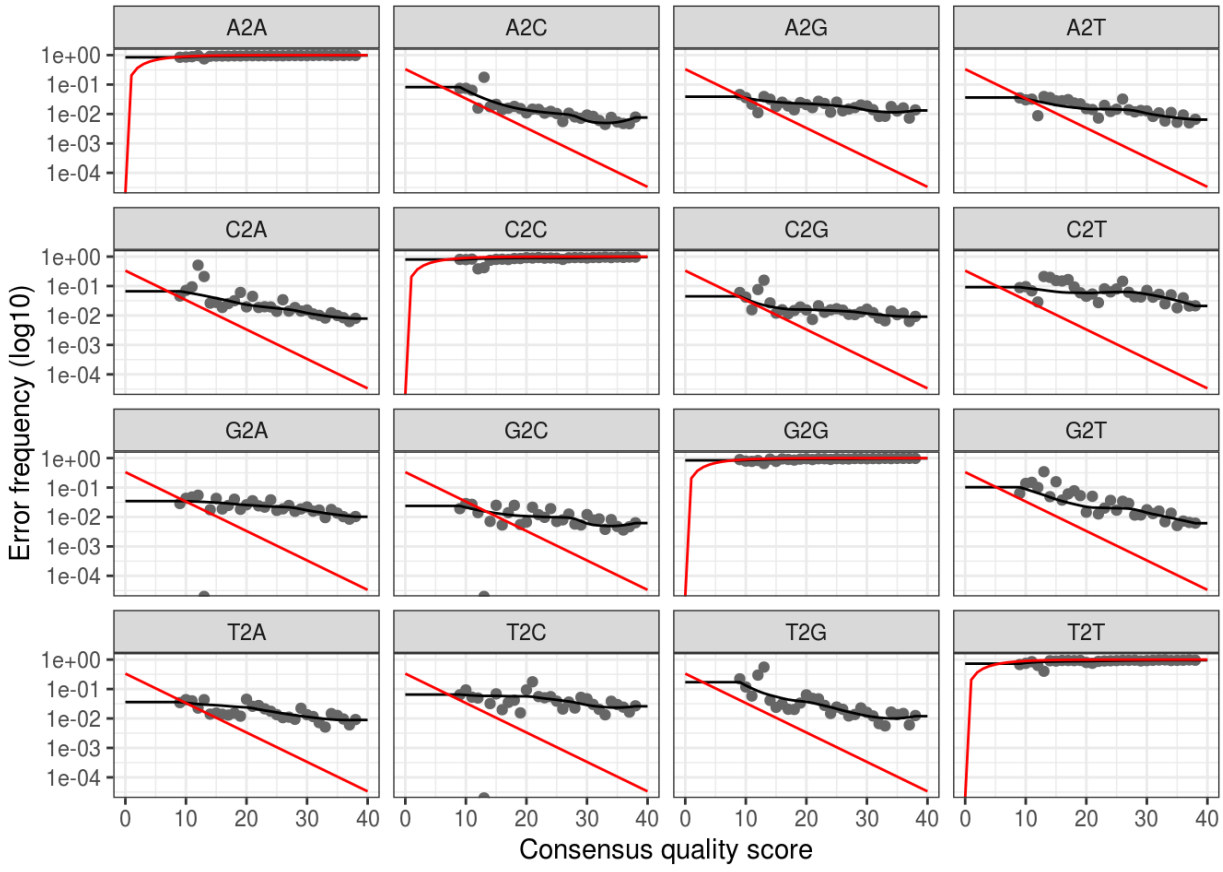

Figure 1: Error rates for all nucleotide transitions.

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, establishing what reads belong to 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.

Remove chimeras

DNA sequencing data contain 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 bimera. The identified chimeras are then removed.

Read counts

For each sample, reads may be removed at each of the quality filtering steps. Inspecting the number of reads filtered provide important information on the data quality. Table 1 below gives an easy overview of the data at each step through quality filtering. The table include an overview of the reads across samples in the raw and cleaned data (first and last row of the table).

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

	Min.	1st Q.	Median	Mean	3rd Q.	Max.
Input	9332	14737.00	16164.0	17758.00	21026.00	28724
Filtered	255	349.20	498.0	844.70	817.00	2949
Denoised	1637	1914.00	2286.0	2369.00	2676.00	3261
Nonchim	0	43.75	116.5	178.12	284.25	487
Clean	6953	11310.00	12286.0	14366.00	17491.00	23809

A supportive file is provided for visual inspection of read counts in each sample at each filtering step (see file: Read_counts_per_sample.pdf). Please see Figure 2 below for an introduction to the barplots found in the supportive pdf.

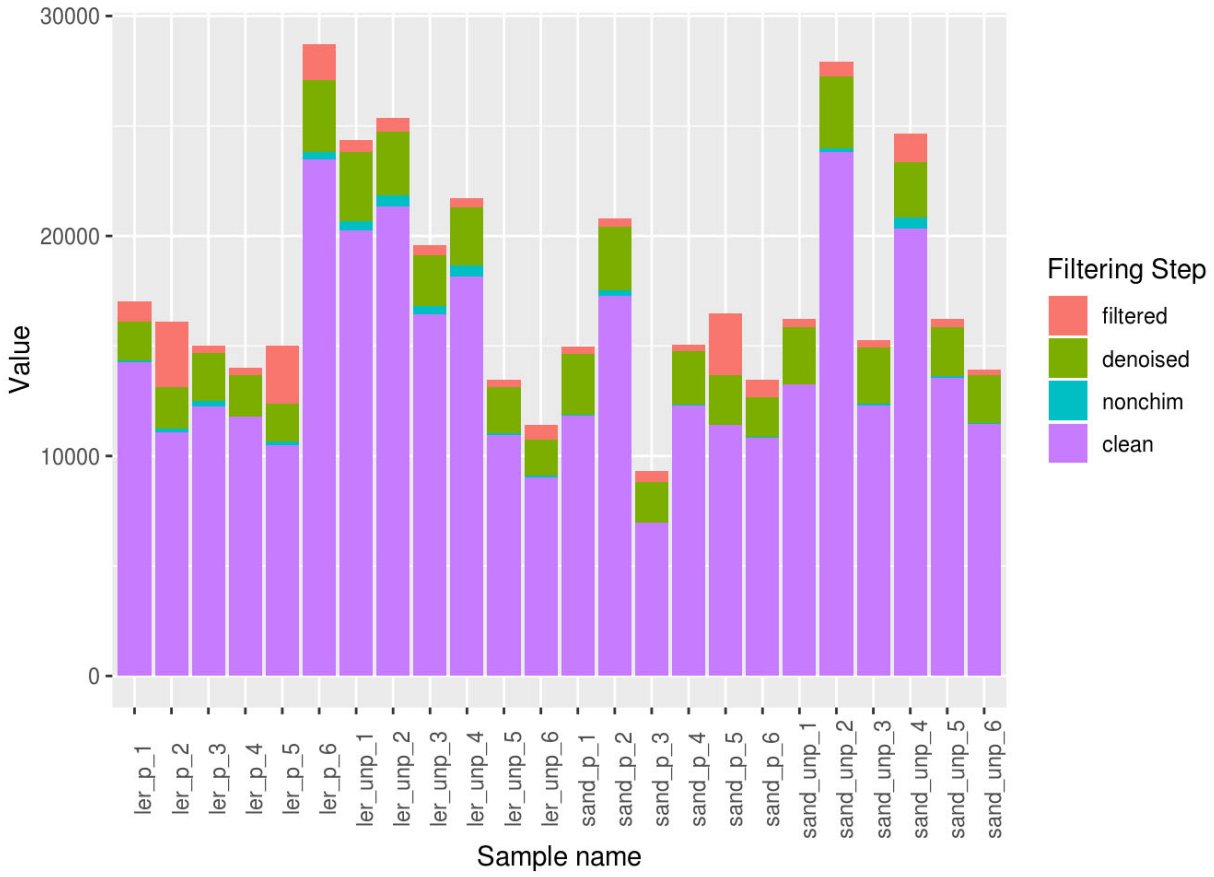

Figure 2: Summary of reads across samples in the project. Each bar show reads for each sample. The “cleaned” reads are the number of reads (merged if paired-end data) that remain after processing the raw data and calling the microbiome profiles. Filtered is the number of reads removed during read filtering (removed due to low read quality or short length). Denoised is the number of reads removed due to duplication.

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. To support an effective evaluation of the samples with a low read count, stacked barplots for these samples are shown in a separate PDF file (file_name). Samples were selected based on a read count below 5000 after data processing, as this is the number of reads that will be available for further analysis.

Software, settings, thresholds and reference data used

Table 2: Reference and parameter values

Parameters	Values
Truncation quality	5
Truncation length	F:250, R:200
Left triming length	F:5, R:5

Parameters	Values
Right trimming length	0
Max. length	Inf
Min. length	20
Max. #N	0
Min. Q	0
Max. Expected Errors	F:2, R:2
Remove PhiX	TRUE
Low complexity filter	0
Reference Data	RDP training set 16

Table 3: Software and package versions

	version
OS	Ubuntu 16.04.6 LTS
R	3.6.1
colorspace	1.4-1
hwriter	1.3.2
ellipsis	0.3.0
rprojroot	1.3-2
XVector	0.26.0
GenomicRanges	1.38.0
fs	1.3.1
rstudioapi	0.10
farver	2.0.1
remotes	2.1.0
lubridate	1.7.4
xml2	1.2.2

	version
codetools	0.2-16
splines	3.6.1
knitr	1.26
pkgload	1.0.2
polyclip	1.10-0
zeallot	0.1.0
ade4	1.7-13
Rsamtools	2.2.1
broom	0.5.2
cluster	2.1.0
dbplyr	1.4.2
compiler	3.6.1
httr	1.4.1
backports	1.1.5
assertthat	0.2.1
Matrix	1.2-17
lazyeval	0.2.2
cli	1.1.0
tweenr	1.0.1
prettyunits	1.0.2
htmltools	0.4.0
tools	3.6.1
igraph	1.2.4.2
gtable	0.3.0
glue	1.3.1

	version
GenomeInfoDbData	1.2.2
reshape2	1.4.3
ShortRead	1.44.0
Biobase	2.46.0
cellranger	1.1.0
vctrs	0.2.0
Biostrings	2.54.0
multtest	2.42.0
ape	5.3
nlme	3.1-140
iterators	1.0.12
xfun	0.11
ps	1.3.0
testthat	2.3.1
rvest	0.3.5
lifecycle	0.1.0
devtools	2.2.1
zlibbioc	1.32.0
MASS	7.3-51.4
scales	1.1.0
hms	0.5.2
parallel	3.6.1
SummarizedExperiment	1.16.0
biomformat	1.14.0
rhdf5	2.30.1

	version
RColorBrewer	1.1-2
curl	4.3
yaml	2.2.0
memoise	1.1.0
latticeExtra	0.6-28
stringi	1.4.3
highr	0.8
desc	1.2.0
S4Vectors	0.24.1
foreach	1.4.7
permute	0.9-5
BiocGenerics	0.32.0
pkgbuild	1.0.6
zip	2.0.4
BiocParallel	1.20.0
GenomeInfoDb	1.22.0
rlang	0.4.2
pkgconfig	2.0.3
matrixStats	0.55.0
bitops	1.0-6
evaluate	0.14
lattice	p.20-38
Rhdf5lib	1.8.0
labeling	0.3
GenomicAlignments	1.22.1

	version
processx	3.4.1
tidyselect	0.2.5
plyr	1.8.4
magrittr	1.5
R6	2.4.1
IRanges	2.20.1
generics	0.0.2
DelayedArray	0.12.0
DBI	1.0.0
pillar	1.4.2
haven	2.2.0
withr	2.1.2
mgcv	1.8-28
survival	2.44-1.1
RCurl	1.95-4.12
modelr	0.1.5
crayon	1.3.4
rmarkdown	1.18
usethis	1.5.1
readxl	1.3.1
callr	3.3.2
vegan	2.5-6
webshot	0.5.2
reprex	0.3.0
digest	0.6.23

	version
RcppParallel	4.4.4
stats4	3.6.1
munsell	0.5.0
viridisLite	0.3.0
sessioninfo	1.1.1