Sequencing and Data Processing Report

Biomcare ApS

28/12/2024

Customer	Tove Mariegaard Pedersen
Customer ID	DA00206-23
Project	Regenerativt landbrug til videreudvikling af den økologiske planteproduktion
Sample Type	Soil
Number of samples	14 samples
Type of data	Shotgun Metagenomic Sequencing

In this report you will find information regarding data generation (DNA extraction, library preparation and sequencing), data quality evaluation, data 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 navigate your folder. The document will tell you how you find specific files.

Summary

DNA extraction, library preparation and Shotgun Metagenomic Sequencing was performed (details on sequencing are available below). The performed sequencing resulted in a mean read count of 29,820,358 reads across samples, with a high data quality (see section "Evaluation of raw data quality" below for more details).

Data quality was evaluated using FastQC, and raw data was processed through quality filtering, removal of adapters, as well as removal of host (if applicable) and phiX contamination. Together, quality filtering removed on average 47,718 reads per sample (min: 38,899 , max: 62,927).

Taxonomic profiling and/or functional profiling was performed as requested. You find information on the used software in the "Microbiome Profiling" section below, and a detailed evaluation of the microbiome profiles in "Report 2: Microbiome Profiling Report".

Data generation using Shotgun Metagenomic Sequencing

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

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 14 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 genomic DNA was fragmented using a enzyme-based fragmentation with FEA Enzyme Mix and for library construction the VAHTS universal Plus DNA Library Prep Kit for Illumina V2 was used. For the constructed library, use Illumina NovaSeq X (Illumina, Santiago CA, USA) was used for sequencing.

Evaluation of raw data quality

Biomcare uses a suite of different QC software to evaluate the quality of the raw data generated using next generation sequencing. These software solutions include FastQC and cutadapt, which run in the wrapper Trim Galore. 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 in your QC folder. No sample file (single fastq file) was flagged as poor quality. The GC content is on average 61% (min: 60% , max: 62%). Reads across samples have a mean read length of 150 bp (min: 150bp , max: 150 bp).

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

Based on the evaluations of data quality, we have selected appropriate quality filtering settings and performed the steps described below.

Read quality and length filtering

When bases are called from 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. In addition to trimming on low quality reads and bases, and removal of short reads, adapters are removed.

Both read trimming and adapter removal is performed with the default settings for Trim Galore. Key settings used are; paired-end mode, quality Phred score cutoff: 20, Maximum trimming error rate: 0.1, minimum read length: 20bp.

Reads that map to the PhiX genome or known sequencing artifacts are removed using the software BBduk. If applicable, reads that map to a host reference genome are removed using the software BMtagger. If this step has been performed on the data, it will appear in **Table 1**.

In your project folder you will find two FastQC reports per file in the project. The “pre-QC-report” reports the quality of the raw data and the “post-QC-report” reports the data quality following the data quality processing steps described in this section.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Raw_reads	27,032,473	29,126,959	29,946,453	29,820,358	30,739,736	31,719,152
Trim_Galore	38,898	44,773	46,247	47,711	49,402	62,920
PhiX_and_artifacts	0	10	16	15	22	28
Total_removed	38,899	44,698	46,254	47,718	49,558	62,927
Clean_reads	26,993,574	29,077,401	29,898,358	29,772,640	30,693,767	31,675,314

Table 1: Summary statistics for each step of the quality control. The first and last row provide summary information on number of paired-end reads in the raw sequencing data (Raw_reads) and the cleaned data (Clean_reads), respectively. The row “Total_removed” lists summary statistics for the total number of reads removed across all filtering steps, and the remaining rows provide summary information on the number of reads removed at the specific steps of the quality filtering.

Microbiome profiling

In order to characterize the microbiome, we performed both a taxonomic profiling and a functional profiling. Please note that we have chosen to perform taxonomic profiling using a pipeline that combines the two key software, Kraken 2 and Bracken. This pipeline was selected for this project as it includes the most comprehensive reference database available for shotgun metagenomic data, including organisms from multiple kingdoms such as bacteria and fungi. The functional profiling was performed using the software tool HUMAnN3, as described below.

Taxonomic profiling using Kraken 2 and Bracken

After cleaning the data, we are ready to start generating the taxonomic profiles. For your project, we have chosen to use the two software, Kraken 2 and Bracken. Kraken 2 is a taxonomic classification software that uses exact k-mer matching to map each sequence in the data to a known reference with high accuracy. This approach allows for the annotation of sequences belonging to an organism found in the reference data.

We are using an extensive reference database to allow for annotation of as many sequences as possible. We used the Kraken 2 standard database which includes the NCBI taxonomic information, as well as the complete genomes in RefSeq for the bacterial, archaeal, and viral domains, along with the human genome and a collection of known vectors (UniVec_Core). Please see the overview table below of reference databases for details. To the standard database we added RefSeq complete fungal genomes and RefSeq complete protozoan genomes, together named “StandardPlusPF” database.

When faced with ambiguous read assignment at a certain taxonomic level, Kraken 2 will assign that read to a higher taxonomic level (e.g. if a read matches equally well with two different species, the read will be assigned to the latest shared node in the phylogeny, the Last Common Ancestor (LCA)). While this is a sensible approach for annotation, it does not allow for proper quantification of organisms at a specific taxonomic level. In order to obtain accurate abundance data, we use Bracken. Bracken takes the data from Kraken 2, and for each read assigned at a higher taxonomic level, estimates the most probable species assignment, thereby allowing us to estimate the actual microbiome profile at each taxonomic level. Together, this makes for highly accurate microbiome profiling.

Computing bacterial-to-fungal abundance ratio

In order to estimate the bacterial-to-fungal abundance ratio of each sample, we use Kraken 2 and Bracken together with the Silva138 Small Subunit (SSU) database as the fungus kingdom is better represented in the SSU database compared to databases of complete reference genomes (e.g. RefSeq). The bacterial-to-fungal abundance ratios were calculated based on the number of SSU reads assigned to either bacteria or fungi.

We emphasize that this estimate can be very different from estimates obtained using other methods such as cell counting and biomass. As the bacterial-to-fungal abundance ratios reported here are based on sequencing of the DNA, we will obtain DNA from both living and dead cells. Furthermore, a number of sequences in the dataset will originate from spores, which are generally not detected in non-DNA based methods. Lastly, as the generated taxonomic profiles only detect known organisms with sequences available in the reference database, the obtained values will reflect the completeness of the reference database. Thus, when comparing bacterial-to-fungal abundance ratios estimated from sequencing of the DNA with ratios estimated using other methods, it is important to keep in mind that the values are likely to reflect different aspects of the ratio. Especially for estimates that consider biomass it is important to consider the differences in the cell size between fungi and bacteria, which will result in very different values compared to estimates based on DNA ratios.

Functional profiling using HUMAnN3

To perform the functional profiling, we used the software HUMAnN3 in which the reads are mapped to the comprehensive UniRef50 protein database. This database has been created by clustering the protein sequences found in the UniProt Knowledgebase (UniProtKB) at 50% identity and selecting representative sequences for each cluster. Thus, UniRef50 is a non-redundant protein database. To map the reads to UniRef50, HUMAnN3 uses the software tool DIAMOND, which performs a “translated search” (i.e. the DNA sequences are translated to protein sequences and subsequently, mapped to the protein database to annotate the proteins). For each protein with a known function, the best match (protein) in the database is reported by DIAMOND.

Subsequently, HUMAnN3 performs quantification of the DIAMOND output to obtain counts of each protein family in each of the samples. In order to increase the specificity, a threshold of 50% is applied for the sequence identity between the protein in the sample and the protein in the database (i.e. the percentage of identical amino acid residues aligned against each other must be equal to or greater than 50%). After the abundance tables of UniRef50 protein families and MetaCyc pathways were generated using HUMAnN3, the identified protein families were regrouped to 2 different classification systems: Gene Ontology (GO) and KEGG Orthogroup (KO). Finally, after regrouping, the protein family files were renamed to human readable output, which are then ready for further processing and analysis.

Software, settings, thresholds and reference data used

For documentation on **non-default software settings** see /0_Reports/0_Software_documentation/non_default_settings.txt

Software versions can be found in: /0_Reports/0_Software_documentation/.

Software tools used for running **QC** are found in: /0_Reports/0_Software_documentation/metawrap.yml

MetaWRAP version (QC module) is found in: /0_Reports/0_Software_documentation/metawrap.txt

Software tools used for **Taxonomic profiling** are found in: /0_Reports/0_Software_documentation/kraken2.yml

Software tools used for **Functional profiling** are found in:
/0_Reports/0_Software_documentation/humann3.yml

Database versions

Database	Software	Version
Silva SSU	Kraken2+Bracken	Silva138
Standard RefSeq Plus PF (StandardPlusPF)	Kraken2+Bracken	5/17/2021
UniRef50	HUMAnN (Diamond)	v201901b
Utility mapping	HUMAnN	v201901b

Table 2: Reference databases and version

Package versions

Package	Version	Package	Version
OS	Ubuntu 20.04.4 LTS	**rstudioapi**	0.16.0
R	4.3.3	**reshape2**	1.4.4
ade4	1.7-22	**tzdb**	0.4.0
tidyselect	1.2.1	**ape**	5.8
viridisLite	0.4.2	**cachem**	1.1.0
Biostrings	2.70.3	**rhdf5**	2.46.1
bitops	1.0-7	**splines**	4.3.3
fastmap	1.2.0	**zlibbioc**	1.48.2
RCurl	1.98-1.16	**parallel**	4.3.3
digest	0.6.36	**XVector**	0.42.0
timechange	0.3.0	**vctrs**	0.6.5
lifecycle	1.0.4	**Matrix**	1.6-5
cluster	2.1.6	**jsonlite**	1.8.8
survival	3.7-0	**IRanges**	2.36.0
magrittr	2.0.3	**hms**	1.1.3
compiler	4.3.3	**S4Vectors**	0.40.2
rlang	1.1.4	**systemfonts**	1.1.0
sass	0.4.9	**foreach**	1.5.2
tools	4.3.3	**jquerylib**	0.1.4
igraph	2.0.3	**glue**	1.7.0

Package	Version	Package	Version
utf8	1.2.4	**codetools**	0.2-20
yaml	2.3.9	**gtable**	0.3.5
knitr	1.48	**GenomeInfoDb**	1.38.8
xml2	1.3.6	**munsell**	0.5.1
plyr	1.8.9	**pillar**	1.9.0
withr	3.0.0	**htmltools**	0.5.8.1
BiocGenerics	0.48.1	**rhdf5filters**	1.14.1
grid	4.3.3	**GenomeInfoDbData**	1.2.11
stats4	4.3.3	**R6**	2.5.1
fansi	1.0.6	**kableExtra**	1.4.0
multtest	2.58.0	**evaluate**	0.24.0
biomformat	1.30.0	**Biobase**	2.62.0
colorspace	2.1-0	**lattice**	0.22-6
Rhdf5lib	1.24.2	**highr**	0.11
scales	1.3.0	**bslib**	0.7.0
iterators	1.0.14	**Rcpp**	1.0.13
MASS	7.3-60.0.1	**svglite**	2.1.3
cli	3.6.3	**permute**	0.9-7
vegan	2.6-6.1	**nlme**	3.1-165
rmarkdown	2.27	**mgcv**	1.9-1
crayon	1.5.3	**xfun**	0.46
generics	0.1.3	**pkgconfig**	2.0.3

Table 3: Package versions.