(http://allgenetics.eu)

Project ID MetaFeren

Offer 202300709

Organisation / Institution Ayuntamiento de Gijón

Authorised client (PI) Dr Zuzana Ferencova / zferencova@gijon.es (mailto:zferencova@gijon.es)

Report ID MetaFeren_01_2023.10.24

Report date 24/10/2023

Project specialist Dr Fátima Sánchez-Barreiro / fatima@allgenetics.eu (mailto:fatima@allgenetics.eu)

Analysis of the prokaryotic community associated to soil samples using DNA metabarcoding

Sample reception

A total number of 30 soil samples were received at AllGenetics on 02/08/2023.

DNA isolation and quantification

We isolated the DNA from each soil sample using the DNeasy PowerSoil Pro DNA isolation kit (Qiagen), strictly following the manufacturer's instructions. The DNA was eluted in a final volume of 50 µL.

For sample 23, however, the above-mentioned protocol did not yield a DNA extract of enough concentration. Therefore, we isolated the DNA of sample 23 using the FastDNA Spin Kit for Soil (MPBio). First, the sample was homogenised in a TissueLyser (Qiagen) at maximum speed for 3 minutes. Then, the DNA isolation was carried out strictly following the manufacturer's instructions, and the DNA was eluted in a final volume of 70 µL.

We included an extraction blank (Bex) in each round of DNA extraction, and treated them as regular samples to check for contamination.

We quantified the DNA concentration in each extract using the Qubit High Sensitivity dsDNA Assay (Thermo Fisher Scientific).

Sample IDs and DNA quantification values are reported in Table 1.

DNA metabarcoding library preparation and sequencing

For prokaryotic library preparation, a fragment of the 16S rRNA gene of around 300 bp (including the primer sequences) was amplified using the following primers:

Forward - 515F-Y (5' GTGYCAGCMGCCGCGGTAA 3') (Parada et al., 2016).

Reverse - 806RB (5' GGACTACNVGGGTWTCTAAT 3') (Apprill et al., 2015).

These primers also included the Illumina sequencing primer sequences attached to their 5' ends.

In the first amplification step, PCRs were carried out in a final volume of 12.5 µL, containing 1.25 µL of template DNA (1:2 diluted in PCR round 3; not diluted in PCR round 11), 0.5 µM

of the primers, 3.13 µL of Supreme NZYTaq 2x Green Master Mix (NZYTech), and ultrapure water up to 12.5 µL. The reaction mixture was incubated as follows: an initial denaturation at 95 ºC for 5 min, followed by 25 cycles of 95 ºC for 30 s, 47 ºC for 45 s, 72ºC for 45 s, and a final extension step at 72 ºC for 7 min.

The oligonucleotide indices that are required for multiplexing different libraries in the same sequencing pool were attached in a second amplification step with identical conditions but only 5 cycles and 60 ºC as the annealing temperature. For a schematic overview of the library preparation process, please see Figure 1 in Vierna et al. (2017).

A negative control that contained no DNA (BPCR) was included in every PCR round to check for contamination during library preparation.

We verified the library size by running the libraries on 2 % agarose gels stained with GreenSafe (NZYTech) and imaging them under UV light. Then, we purified the libraries using the Mag-Bind RXNPure Plus magnetic beads (Omega Bio-tek), following the instructions provided by the manufacturer.

Finished libraries were pooled in equimolar amounts according to the results of a Qubit dsDNA HS Assay (Thermo Fisher Scientific) quantification.

The pool was sequenced in a fraction of a NovaSeq PE250 flow cell (Illumina) aiming for a total output of 2 gigabases.

Quality control of sequencing data

Illumina paired-end raw data for each library consists of forward (R1) and reverse (R2) reads stored in separate files, which also include the reads' quality scores. We removed any potential traces of adapter dimers using Cutadapt v3.5 (Martin, 2011).

The FASTQ files after the removal of adapter dimers can be accessed by following the link below:

Raw data (http://services.allgenetics.eu/MetaFeren/2STP-KASY/16S/raw_data_clean.tgz)

(md5: 100cc05fb239282d75db56756acd813b)

We assessed the quality of the FASTQ files with the software FastQC (Andrews, 2010), and summarised the output using MultiQC (Ewels et al., 2016). The quality control reports of R1 and R2 reads can be accessed by clicking on the links below:

QC report R1 (http://services.allgenetics.eu/MetaFeren/2STP-KASY/16S/R1_multiqc.html)

QC report R2 (http://services.allgenetics.eu/MetaFeren/2STP-KASY/16S/R2_multiqc.html)

Processing of sequencing data and inference of ASVs

We first used Cutadapt v3.5 to trim off non-biological DNA sequences (primers, indices, and sequencing adapters) that might appear at the end of some reads due to potential length variability of the amplified marker.

Then, the amplicon reads were processed using QIIME 2 (release 2023.7) (Bolyen et al., 2019). Specifically, we used the tool DADA2 (Callahan et al., 2016), implemented in QIIME 2, to: remove the PCR primers, filter the reads according to their quality, denoise and infer Amplicon Sequence Variants (ASVs), merge the forward and reverse reads, and remove chimaeric sequences.

The first step in the DADA2 pipeline consists in trimming and filtering the data to remove the amplification primers and avoid low-quality bases. In this case, after checking the read quality profiles, reads were truncated at position 175 for forward reads, and at position 131 for reverse reads.

Then, error rates were learned from the dataset to denoise, using the parametric error model implemented in DADA2.

Before the inference of sequence variants, dereplication of the dataset was carried out, i.e. the combination of all identical reads into unique reads to reduce computational effort. Then, these dereplicated forward and reverse reads were used to infer ASVs with the *core sample inference algorithm* (Callahan et al., 2016).

Subsequently, corresponding R1 and R2 reads were merged into pairs with a minimum overlap of 12 identical base pairs.

The DADA2 pipeline includes a final step to reduce the impact of artefacts in the dataset. These artefacts, such as chimaeras, are produced during PCR and sequencing, and could lead to an overestimation of the number of ASVs if not removed.

The number of sequences per sample that passed each of the DADA2 processing steps is shown in Table 1.

The resulting output of the DADA2 pipeline is a table containing the count of reads of every observed ASV in each sample, as well as a table listing all ASVs and their corresponding representative sequences. We used the latter table to conduct the taxonomic assignment of the ASVs (see Taxonomic assignment below).

Table 1

CSV Excel **Search:**

Sample ID	File ID	Remarks	DNA isolation protocol	Extract DNA concentration (ng/µL)
1	1-16S	01/06/2023 N4A La Coría	PowerSoil Pro	>100
2	2-16S	02/06/2023 N2B Cabueñes	PowerSoil Pro	>100
3	3-16S	05/06/23	PowerSoil Pro	76.6
4	4-16S	05/06/23	PowerSoil Pro	>100
5	5-16S	05/06/23	PowerSoil Pro	>100

6	6-16S	07/06/23	PowerSoil Pro	>100
7	7-16S	07/06/23	PowerSoil Pro	>100
8	8-16S	07/06/23	PowerSoil Pro	>100
9	9-16S	08/06/23	PowerSoil Pro	>100
10	10-16S	08/06/23	PowerSoil Pro	>100
11	11-16S	08/06/23	PowerSoil Pro	95.2
12	12-16S	12/06/23	PowerSoil Pro	>100
13	13-16S	12/06/23	PowerSoil Pro	78
14	14-16S	12/06/23	PowerSoil Pro	>100

Table 1. Sample IDs, sample information, Qubit quantification values in DNA extracts and libraries, identifier of PCR round, number of raw reads, and number of sequences retained after each processing step. * Number of raw reads from paired-end sequencing, i.e., count of either R1 or R2 reads.

Taxonomic assignment

We conducted the taxonomic assignment of each ASV using a pre-trained classifier of the SILVA reference database (Quast et al. (2013); release 138.1 August 2020). To compare the representative sequences of the ASVs to the reference database, and compute the taxonomic assignment, we employed the algorithm *sklearn*, which is implemented in QIIME 2 (Bokulich et al., 2018) as the feature-classifier approach *classify-sklearn*. The resulting table lists the number of sequences of each ASV found in each sample, and their corresponding taxonomic information (Table 2). Subsequently, based on the results of this table, we applied several different filters.

Singletons, i.e. ASVs containing only one member sequence in the whole data set, were excluded.

In DNA metabarcoding studies, it has been observed that a low percentage of the reads of a given library might be erroneously assigned to another library. This phenomenon, referred to as mistagging (also tag jumping, index hopping, or index jumping) is the result of the misassignment of the indices during library preparation, sequencing, and/or demultiplexing steps (Bartram et al., 2016; Esling et al., 2015; Guardiola et al., 2016; Illumina, 2017). In order to correct for this bias, ASVs occurring at a frequency below 0.01 % in each sample were removed.

Additionally, non-prokaryotic ASVs such as eukaryotic sequences of plastid (Moore et al., 2019) and mitochondrial (Emelyanov, 2001) origin were removed from the filtered ASV table.

The resolution of taxonomic assignments depends largely on the completeness of the reference databases available; some taxa might result unidentified even when working

with recently updated, taxon-specific databases. Therefore, unidentified sequences, and those assigned only at domain level ('Bacteria') were removed from the filtered ASV table.

As previously described in Salter et al. (2014), the detection of a band in agarose gel assessment of PCR negative controls is recurrent in microbiome analyses. We took strict precautions to avoid environmental contamination, i.e. a laminar flow hood and filter tips were used at all times, and all surfaces were periodically wiped with bleach. Even with these precautions, however, negative controls contained some sequences that were taxonomically assigned. Therefore, ASVs present in the negative controls that were assigned to prokaryotic taxa, and were more abundant in the negative controls than in the samples were discarded in the filtered ASV table.

The final filtered ASV table (Table 3) was converted into a Biological Observation Matrix file (.biom) that was directly imported into R v4.2.2 (R Core Team, 2022) using the package *phyloseq* v1.44.0 (McMurdie and Holmes, 2013) to plot the results of the analyses.

Table 2 Table 3

Table before ASV filtering (http://services.allgenetics.eu/MetaFeren/2STP-KASY/16S/Table2.tgz)

Alpha rarefaction curves

The alpha rarefaction plots show the number of ASVs obtained with a rarefied number of sequences in each sample. These plots were generated using the ASV tables before and after filtering (see Rarefaction plots). The vertical axis displays the number of ASVs observed, and the horizontal axis shows the subsampling depth. When the rarefaction curves tend towards saturation, i.e. they reach a plateau, the sequencing depth is considered to be sufficient to retrieve the existing diversity of the taxa of interest.

Rarefaction plots

Rarefaction curves before ASV filtering (http://services.allgenetics.eu/MetaFeren/2STP-KASY/16S/Rarefaction/RarefPlot_before_filtering.html)

Rarefaction curves after ASV filtering (http://services.allgenetics.eu/MetaFeren/2STP-KASY/16S/Rarefaction/RarefPlot_after_filtering.html)

Taxonomy summary plots

In order to easily visualise the breakdown of taxonomic composition, we generated stacked bar plots at each taxonomic level showing the relative abundance of each ASV in each sample (see Figure 1 for an example). In addition, all the stacked bar plots were exported as a compressed QIIME visualisation file (".qzv"), which can be opened and viewed at the web interface QIIME 2 View (https://view.qiime2.org). To do so, please, open QIIME 2 View in Google Chrome (version 49 or later) or Mozilla Firefox (version 47 or later) web browser, and drag and drop the ".qzv" file (see QIIME2_MetaFeren_16S_barplots.qzv at the Stacked bar plots link) directly from your computer.

We also created a zoomable pie chart for each sample with the Krona package (Ondov et al., 2011) (see Krona pie charts).

Finally, we extracted the representative sequences of the ASVs before (from Table 2) and after the ASV filtering process (from Table 3) (see Representative sequences).

In DNA metabarcoding studies, ASV relative abundance is defined as the number of reads assigned to that ASV in a given sample divided by the total number of reads of that sample. Please note, however, that the PCR may cause biases due to differences in primer specificity. These biases might result in taxa with low representation in the DNA extract becoming more abundant in the final results, which impedes the correct inference of species abundance in the original sample. For instance, if *Species A* is represented by 35 % of the sequences in SAMPLE 1, and *Species B* is represented by 50 % of the sequences in the same sample, we cannot reliably conclude that there was more *Species B* DNA in the original sample.

That being said, it is expected that, within the same study, the PCR bias affects all samples equally. Therefore, it is possible to compare how the abundance of a given taxon varies across different samples with a similar composition. For example, if *Species A* is represented by 35 % of the sequences in SAMPLE 1 and by 10 % in SAMPLE 2, we can conclude that there was less *Species A* DNA in SAMPLE 2.

For more information on the inference of abundances, please refer to Schloss et al. (2011), Geisen et al. (2015), Thomas et al. (2016), and Matesanz et al. (2019).

Figure 1	Stacked bar plots	Krona pie charts	Representative sequences

Stacked bar plots for all taxonomic levels (http://services.allgenetics.eu/MetaFeren/2STP-KASY/16S/Stacked_barplots.tgz)

Data download deadline

Please, kindly download all your files by **23/12/2023**. Thank you very much.

How to cite our services

If you would like to cite our services in your scientific articles, we suggest you do it as follows: “DNA metabarcoding analyses were carried out by AllGenetics & Biology SL (www.allgenetics.eu)”.

References

Andrews, S., 2010. FastQC: A quality control tool for high throughput sequence data (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Apprill, A., McNally, S., Parsons, R., Weber, L., 2015. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquatic Microbial Ecology 75, 129–137. https://doi.org/10.3354/ame01753 (https://doi.org/10.3354/ame01753)

Bartram, J., Mountjoy, E., Brooks, T., Hancock, J., Williamson, H., Wright, G., Moppett, J., Goulden, N., Hubank, M., 2016. Accurate sample assignment in a multiplexed, ultrasensitive, high-throughput sequencing assay for minimal residual disease. The Journal of Molecular Diagnostics 18, 494–506. https://doi.org/10.1016/j.jmoldx.2016.02.008 (https://doi.org/10.1016/j.jmoldx.2016.02.008)

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A., Caporaso, J.G., 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 22019s q2-feature-classifier plugin. Microbiome 6, 90. https://doi.org/10.1186/s40168-018-0470-z (https://doi.org/10.1186/s40168-018-0470-z)

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., others, 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9 (https://doi.org/10.1038/s41587-019-0209-9)

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581. https://doi.org/10.1038/nmeth.3869 (https://doi.org/10.1038/nmeth.3869)

Emelyanov, V.V., 2001. Evolutionary relationship of *Rickettsiae* and mitochondria. FEBS Letters 501, 11–18. https://doi.org/10.1016/S0014-5793(01)02618-7 (https://doi.org/10.1016/S0014-5793(01)02618-7)

Esling, P., Lejzerowicz, F., Pawlowski, J., 2015. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. Nucleic Acids Research 43, 2513–2524. https://doi.org/10.1093/nar/gkv107 (https://doi.org/10.1093/nar/gkv107)

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354 (https://doi.org/10.1093/bioinformatics/btw354)

Geisen, S., Laros, I., Vizcaíno, A., Bonkowski, M., De Groot, G., 2015. Not all are free-living: High-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoa. Molecular Ecology 24, 4556–4569. https://doi.org/10.1111/mec.13238 (https://doi.org/10.1111/mec.13238)

Guardiola, M., Wangensteen, O.S., Taberlet, P., Coissac, E., Uriz, M.J., Turon, X., 2016. Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. PeerJ 4, e2807. https://doi.org/10.7717/peerj.2807 (https://doi.org/10.7717/peerj.2807)

Illumina, I., 2017. Effects of index misassignment on multiplexing and downstream analysis. https://www.illumina.com (https://www.illumina.com).

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200 (https://doi.org/10.14806/ej.17.1.200)

Matesanz, S., Pescador, D.S., Pías, B., Sánchez, A.M., Chacón-Labella, J., Illuminati, A., Cruz, M. de la, López-Angulo, J., Marí-Mena, N., Vizcaíno, A., others, 2019. Estimating belowground plant abundance with DNA metabarcoding. Molecular Ecology Resources 19,

1265–1277. https://doi.org/10.1111/1755-0998.13049 (https://doi.org/10.1111/1755-0998.13049)

McMurdie, P.J., Holmes, S., 2013. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PloS One 8. https://doi.org/10.1371/journal.pone.0061217 (https://doi.org/10.1371/journal.pone.0061217)

Moore, K.R., Magnabosco, C., Momper, L.M., Gold, D.A., Bosak, T., Fournier, G.P., 2019. An expanded ribosomal phylogeny of Cyanobacteria supports a deep placement of plastids. Frontiers in Microbiology 10, 1612. https://doi.org/10.3389/fmicb.2019.01612 (https://doi.org/10.3389/fmicb.2019.01612)

Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12, 385. https://doi.org/10.1186/1471-2105-12-385 (https://doi.org/10.1186/1471-2105-12-385)

Parada, A.E., Needham, D.M., Fuhrman, J.A., 2016. Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environmental Microbiology 18, 1403–1414. https://doi.org/10.1111/1462-2920.13023 (https://doi.org/10.1111/1462-2920.13023)

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41, D590–D596. https://doi.org/10.1093/nar/gks1219 (https://doi.org/10.1093/nar/gks1219)

R Core Team, 2022. R: A language and environment for statistical computing (http://www.r-project.org) (http://www.R-project.org).

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W., 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology 12, 87. https://doi.org/10.1186/s12915-014-0087-z (https://doi.org/10.1186/s12915-014-0087-z)

Schloss, P.D., Gevers, D., Westcott, S.L., 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PloS One 6. https://doi.org/10.1371/journal.pone.0027310 (https://doi.org/10.1371/journal.pone.0027310)

Thomas, A.C., Deagle, B.E., Eveson, J.P., Harsch, C.H., Trites, A.W., 2016. Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. Molecular Ecology Resources 16, 714–726. https://doi.org/10.1111/1755-0998.12490 (https://doi.org/10.1111/1755-0998.12490)

Vierna, J, Doña, J., Vizcaíno, A., Serrano, D., Roger, J., 2017. PCR cycles above routine numbers do not compromise high-throughput DNA barcoding results. Genome 60, 868–873. https://doi.org/10.1139/gen-2017-0081 (https://doi.org/10.1139/gen-2017-0081)