Microbial community analysis
CP942

NIRAS

DNASense ApS

07-02-2022, Aalborg, Denmark

Contents

1 Project scope

This project concerns the analysis of microbial communities in WWTP sediment samples. DNASense received a total of 3 samples from Steffen L. Aggerholm at NIRAS and these were analysed using gene amplicon sequencing targeting the bacteria 16S rRNA gene region V3-4 in combination with taxonomic classification against the SILVA 132 database.

1.1 DNASense microbial community analysis

The general microbial community analysis workflow from raw sample to the final report is outlined in figure 1. First, incoming samples are registered in the laboratory and DNA is extracted from all organisms in the sample. Afterwards, the extracted microbial community DNA is processed and sequencing libraries prepared for DNA sequencing on state-of-the-art equipment in DNASense's laboratory. DNA from each microbe in the community contains specific taxonomic marker genes (also called fingerprint genes) that can be used for organism identification and abundance estimation. Examples of marker genes are the *16S ribosomal RNA (16S rRNA)* gene for bacteria and archaea, and the *internal transcribed spacer (ITS)* for fungi. DNA sequencing is used to count the number of marker gene copies from each microbe in a sample, and that count is in turn used as an estimate of the relative abundance of the microbe in the sample community. At least 10,000 marker genes are DNA sequenced from each sample to provide a high resolution of the community structure. The taxonomic marker genes can be matched with a reference database for identification of the microbes in the community.

Sample preparation and DNA sequencing for the taxonomic marker genes were conducted in agreement with the latest research standards. The raw sequencing data was processed using the research standard UPARSE workflow and data analysed through Rstudio using the ampvis2 package developed at Aalborg University. The abundances of the organisms - presented in the analysis - represent the count of each taxonomic marker gene in the sample. The abundances are influenced by DNA extraction, gene copy number and primer biases and does not necessarily represent the absolute *in situ* organism abundances.

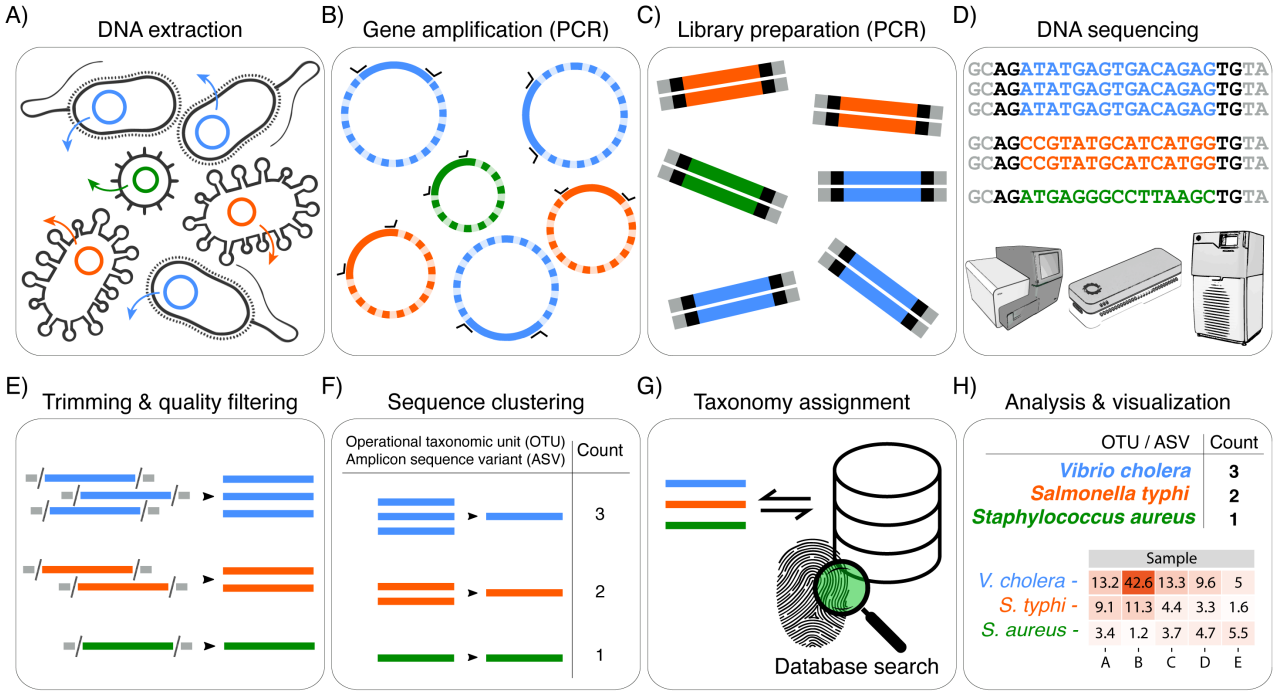

Figure 1: ***Overview of the general workflow from sample to microbial community profile****. **A)** Total community DNA is extracted and **B)** DNA amplicons are prepared using PCR with specific primers (black) targeting e.g. the 16S rRNA region V1-V9 for bacterial taxonomic marker genes. **C)** A second PCR amplification adds sequencing adapters. **D)** Resulting amplicon libraries are DNA sequenced and then basecalled. **E)** Basecalled sequences are adaptor trimmed and quality filtered. **F)** DNA reads are partitioned into clusters (either OTUs or ASVs) providing the read abundance of each cluster. **G)** Read sequence taxonomy is assigned by searching against a reference database. **H)** OTU/ASV tables correlating taxonomy and abundance are generated for further results analysis e.g. using Ampvis.*

2 Results

2.1 Data availability

The project data is available from the dropbox folder /CP942 with the password **bridge73celebs**. The file ***shiny/rawData.zip*** contains the raw unprocessed sequencing data, and the folder ***/figures*** includes all project figures in .pdf and .png format for general use.

The file ***shiny/otutable.txt*** contains the different OTUs that were identified in all samples, their abundances and taxonomic assignment. Each OTU is identified by a name e.g. *OTU_1* and the corresponding DNA sequence of the specific OTU can be found in the file ***shiny/OTUs.fa***. Data can also be explored, filtered and visualized in heatmaps and ordination plots, such as PCA, in the DNASense app using the link https://dnasense.shinyapps.io/dnasense/ with the username CP942 and the same password as above. Furthermore, summary data tables are provided in the file ***summary_tables.xlsx*** including a sequencing overview table, output of OTU count tables including their corresponding DNA sequence, and OTU count tables rarefied to 51991 reads; the latter for direct comparison between samples or sample groups.

2.2 DNA extraction, library preparation and sequencing

DNA extraction and sequencing library preparation was successful for 3 / 3 sample analyses (100 %), and yielded between 51991 and 58042 DNA reads after QC and bioinformatic processing (table 1). Failed samples were those yielding significantly less quality filtered DNA reads (filtReads) than 10,000; here filtReads < 8,000. Low-read samples were disregarded in all subsequent analyses. For this project no samples failed and all were therefore included in analyses. Table 1 provides an overview of the sequencing outcome, and much more extensive data details are given in the sequencing overview table in summary_tables.xlsx.

Table 1: ***Sequencing statistics***. ***seqID*** *is the name of the sequencing data file ID,* ***sampleName*** *is the customer provided sample label,* ***sampleGroup*** *is the sample's experiment group,* ***rawReads*** *is the total number of DNA reads from sequencing,* ***filtReads*** *is the number of reads after quality filtering,* ***OTUs*** *is the total number of OTUs observed in the sample, and* ***Shannon*** *is a commonly used alpha-diversity index for comparing sample diversities and essentially quantifies the number of typical OTUs observed in a given sample.*

seqID	sampleName	rawReads	filtReads	OTUs	Shannon
MQ211210-229	sediment-1	97815	54268	457	4.56
MQ211210-230	sediment-2	104673	58042	457	4.65
MQ211210-231	sediment-3	99137	51991	464	4.65

Sample sequencing rarefaction curves for evaluation of the obtained sequencing depth relative to sample complexity, here represented as unique OTUs (figure 2). Shown below are curves for the 3 samples from the bV34-A data set.

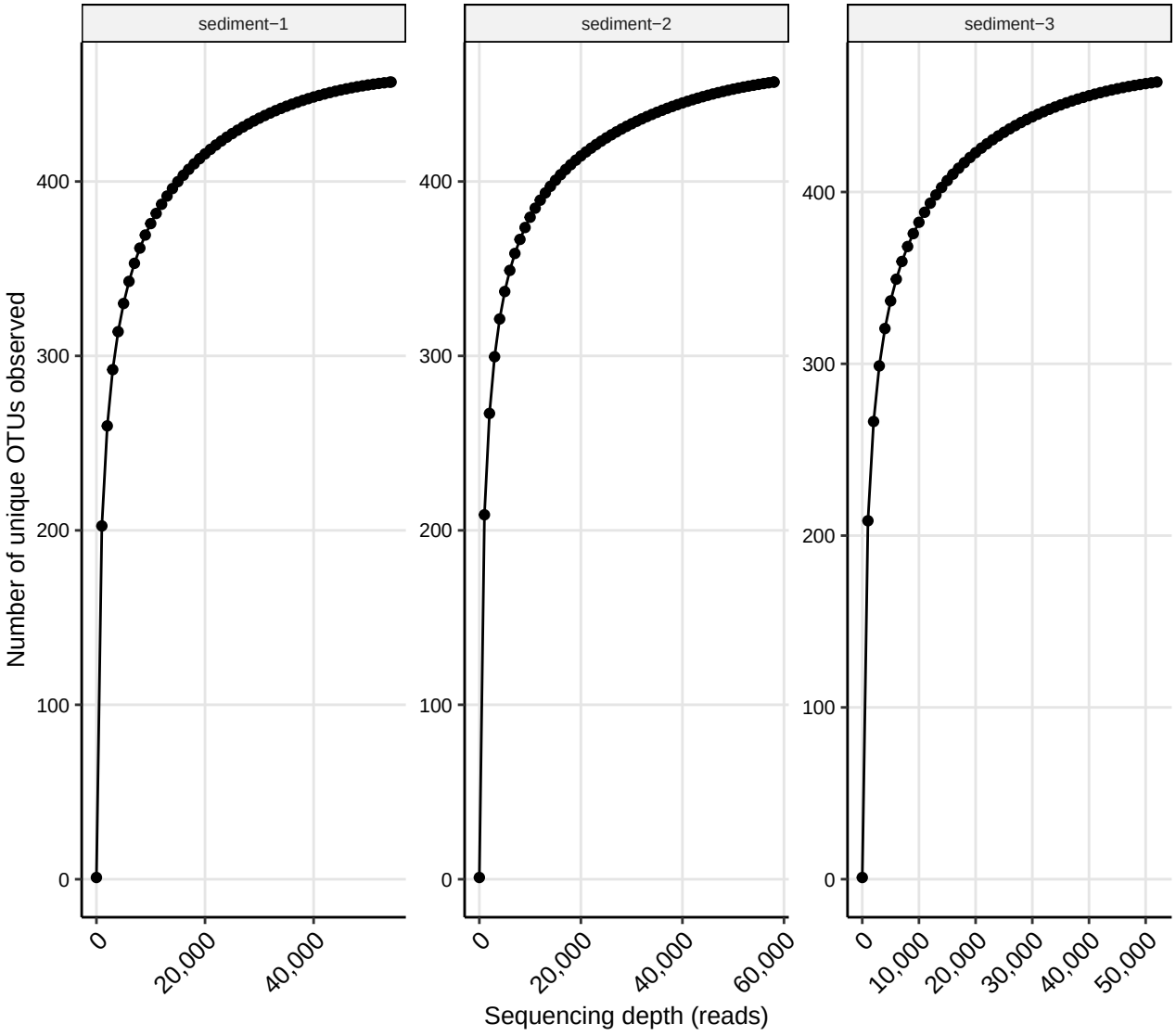

Figure 2: ***Rarefaction curves for sequencing of samples***. *The x-axis represents the number of sequences sampled whereas the y-axis depicts the number of observed OTUs.*

2.3 Microbial community composition

Figure 3 gives an overview of the 25 most abundant genera across all samples.

Figure 3: ***Heatmap of the 25 most abundant genera.*** *The most abundant genera in all samples arranged by sampleName. Where available the OTU's phylum classification is provided along with genus, and if no genus level classification could be obtained, the lowest assigned taxonomic classification is given. Values are shown as normalised fraction of total sequences (%).*

3 Materials and methods

The project data analysis and reporting was done using DNASense's custom bioinformatic workflow (version *MCA_DS220120*).

3.1 Sample DNA extraction

3.1.1 FastDNA SPIN Kit for Soil

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 (Albertsen et al., 2015). 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).

3.2 Sequencing library preparation

Amplicon libraries for the bacteria 16S rRNA gene region V3-4 were prepared by a custom protocol based on an Illumina protocol (Illumina, 2015). Up to 10 ng of extracted DNA was used as template for PCR amplification of the bacteria 16S rRNA gene region V3-4 amplicons. Each PCR reaction (25 μL) contained (12.5 μL) PCRBIO Ultra mix and 400 nM of each forward and reverse tailed primer mix. PCR was done with the following program: Initial denaturation at 95 °C for 2 min, 30 cycles of amplification (95 °C for 15 s, 55 °C for 15 s, 72 °C for 50 s) and a final elongation at 72 °C for 5 min. Duplicate PCR reactions were performed for each sample and the duplicates were pooled after PCR. The forward and reverse, tailed primers were designed according to (Illumina, 2015) and contain primers targeting the bacteria 16S rRNA gene region V3-4: [341F] CCTACGGGNGGCWGCAG and [805R] GACTACHVGGGTATCTAATCC (Herlemann et al., 2011). The primer tails enable attachment of Illumina Nextera adaptors necessary for sequencing in a subsequent PCR. The resulting amplicon libraries were purified using the standard protocol for CleanNGS SPRI beads (CleanNA, NL) with a bead to sample ratio of 4:5. DNA was eluted in 25 μL of nuclease free water (Qiagen, Germany). DNA concentration was measured using Qubit dsDNA HS Assay kit (Thermo Fisher Scientific, USA). Gel electrophoresis using Tapestation 2200 and D1000/High sensitivity D1000 screentapes (Agilent, USA) was used to validate product size and purity of a subset of sequencing libraries.

Sequencing libraries were prepared from the purified amplicon libraries using a second PCR. Each PCR reaction (25 μL) contained PCRBIO HiFi buffer (1x), PCRBIO HiFi Polymerase (1 U/reaction) (PCRBiosystems, UK), adaptor mix (400 nM of each forward and reverse) and up to 10 ng of amplicon library template. PCR was done with the following program: Initial denaturation at 95 °C for 2 min, 8 cycles of amplification (95 °C for 20 s, 55 °C for 30 s, 72 °C for 60 s) and a final elongation at 72 °C for 5 min. The resulting sequencing libraries were purified using the standard protocol for CleanNGS SPRI beads with a bead to sample ratio of 4:5. DNA was eluted in 25 μL of nuclease free water. DNA concentration was measured using Qubit dsDNA HS Assay kit. Gel electrophoresis using Tapestation 2200 and D1000/High sensitivity D1000 screentapes was used to validate product size and purity of a subset of sequencing libraries.

3.3 DNA sequencing

The purified sequencing libraries were pooled in equimolar concentrations and diluted to 2 nM. The samples were paired-end sequenced (2x300 bp) on a MiSeq (Illumina, USA) using a MiSeq Reagent kit v3 (Illumina, USA) following the standard guidelines for preparing and loading samples on the MiSeq. > 10 % PhiX control library was spiked in to overcome low complexity issues often observed with amplicon samples.

3.4 Bioinformatic processing

Forward and reverse reads were trimmed for quality using Trimmomatic v. 0.32 (Bolger et al., 2014) with the settings SLIDINGWINDOW:5:3 and MINLEN: 275 . The trimmed forward and reverse reads were merged using

FLASH v. 1.2.7 (Magoč and Salzberg, 2011) with the settings -m 10 -M 250. The trimmed reads were dereplicated and formatted for use in the UPARSE workflow (Edgar, 2013). The dereplicated reads were clustered, using the usearch v. 7.0.1090 -cluster_otus command with default settings. OTU abundances were estimated using the usearch v. 7.0.1090 -usearch_global command with -id 0.97 -maxaccepts 0 -maxrejects 0. Taxonomy was assigned using the uclust classifier as implemented in the assign_taxonomy.py script in QIIME (Caporaso et al., 2010) and the SILVA database, release 132 (Quast et al., 2013). All bioinformatic processing was done via RStudio IDE (1.4.1106) running R version 4.1.1 (2021-08-10) and using the R packages: ampvis (2.7.10) (Albertsen et al., 2015), tidyverse (1.3.1), seqinr (4.2.8), ShortRead (1.50.0) and iNEXT (2.0.20) (Chao et al., 2014; Hsieh et al., 2016).

References

Albertsen, M., Karst, S.M., Ziegler, A.S., Kirkegaard, R.H., and Nielsen, P.H. (2015) Back to Basics – The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. *PLOS ONE* **10**: e0132783.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., et al. (2010) QIIME allows analysis of high-throughput community sequencing data. **7**: 335–336.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., and Ellison, A.M. (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84**: 45–67.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods* **10**: 996–8.

Herlemann, D.P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J.J., and Andersson, A.F. (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME journal* **5**: 1571–9.

Hsieh, T.C., Ma, K.H., and Chao, A. (2016) iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol Evol* **7**: 1451–1456.

Illumina, I. (2015) 16S Metagenomic Sequencing Library Preparation, Part # 15044223 Rev. B.

Magoč, T. and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics (Oxford, England)* **27**: 2957–63.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucl. Acids Res.* **41**: D590–D596.