Grant agreement No. 764810

Science for Clean Energy

H2020-LCE-2017-RES-CCS-RIA
Competitive low-carbon energy

D4.9
Quantification of the biological activity of microbial samples from selected field sites

WP 4 – Samples and Fluid Characterisation

Due date of deliverable	Month 30 – February 2020
Actual submission date	11/02/ 2020
Start date of project	01/09/2017
Duration	36 months
Lead beneficiary	Université de Bretagne Occidentale (UBO)
Last editor	Mohamed Jebbar (UBO), Ashley Grosche (UBO)
Contributors	Mohamed Jebbar (UBO), Ashley Grosche (UBO)
Dissemination level	Public (PU)

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 764810

History of the changes

Version	Date	Released by	Comments
1.0	10-02-20	Ashley Grosche	Basic information, introduction, methodology, results, conclusions
1.1	11-02-20	Mohamed Jebbar	Proofreading
1.2	12-02-20	Ronny Pini	Proofreading and reviewing
1.3	12-02-20	Mohamed Jebbar	Proofreading
1.4	13-02-20	Camille Voirin	Formatting
1.5	18-02-20	Alberto Striolo	Minor edits
1.6	26-02-20	Ashley Grosche	Final edits

Table of Figures

Table of contents

Key word list

Deep-subsurface biosphere, microbial diversity and function, metagenomics, metagenomic libraries, gene screening, metagenome-assembled genomes (MAGs), next-generation sequencing (NGS), 16S metabarcoding analysis.

Definitions and acronyms

Acronyms	Definitions
MAGs	Metagenome-Assembled Genomes
NGS	Next-Generation Sequencing
Read	Sequenced fragment of DNA
CCS	Carbon Capture and Storage
16S Metabarcoding	A tool to identify bacteria that combines DNA-based identification (16S rRNA) and high-throughput DNA sequencing (metabarcoding) to analyze many samples in parallel
ASV	Amplicon Sequence Variant, an individual DNA sequence recovered from a high-throughput marker gene analysis following the removal of sequencing errors generated during PCR amplification and sequencing, ASVs can be resolved down to a single-nucleotide variant

1 Introduction

Within the S4CE consortium, Deliverable 4.9 provides the quantification of the biological activity of microbial samples from selected field sites. This workflow is associated with Work Package 4: Sample and Fluids Characterization.

1.1 General context

S4CE conducts research and development activities that are connected with several field sites across Europe. One of these sites, in Iceland, benefits from long-term operations intended to fixate carbon dioxide. This site is often referred to as CarbFix 1, which also refers to the project, supported in part by the European Commission, which was conducted there over several years. The activities discussed here were conducted suing samples from 2 field sites available to S4CE: CarbFix 1 and Nesjavellir. Both sites are in Iceland.

The carbon capture and storage technologies implemented during Carbfix 1 rely on the sequestration of magmatic gases released by the geothermal powerplant via injection and mineral precipitation into the basaltic subsurface. It has been demonstrated that the injection of gas-charged brine stimulates microbial growth in the subsurface aquifers proximate to the injection site, accompanied by a drop in microbial diversity (i.e., by stimulating a microbial bloom)[1].

The workflow outlined in Deliverable D4.3 was implemented herein with the goal of linking the change in microbial diversity associated with gas injection with a potential shift in microbial activity, and to further characterize the microbial component of fluids.

1.2 Deliverable objectives

The objectives of this deliverable are:

- To analyze and process the metagenome and diversity data obtained from environmental samples;

- To determine molecular markers relevant for biodiversity analyses (16S rDNA and functional genes), so as to identify the various metabolisms associated with gas injection at low temperature.

2 Methodological approach

Microbial samples collected from CarbFix 1 (30x samples) and Nesjavellir (2x) well water samples were collected as outlined in Deliverable 4.3.

2.1 16S rRNA Metabarcoding Analysis

Libraries were prepared at UBO following the VAMPS protocol for targeting the bacterial and archaeal V4-V5 region of the 16S rRNA gene (https://vamps.mbl.edu/resources/primers.php).

The 16S rRNA gene (gene encoding for the small subunit of ribosomal RNA) was amplified via polymerase chain reaction (PCR) with bacterial primers 518F and 926R and archaeal primers 517F and 958R. A combination of barcodes and indexes were used to multiplex the PCR to ensure a unique molecular signature for each sample, and to enable downstream partitioning (demultiplexing) of sequences by sample origin.

The reaction mixture for PCR was as follows:

Component	20 µL rxn	Final conc.
H2O	add to 20 µL	
5x Phusion HF Buffer	4 µL	1X
10 mM dNTPs	0.4 µL	200 µM
Forward primer	X	0.15 µM
Reverse primer	X	0.15 µM
Template DNA		
Phusion High-Fidelity DNA polymerase	0.2 µL	0.02 U/µL

Samples were amplified using the following thermocycling program:

Cycle step	Temp.	Time	Cycles
Initial Denaturation	98ºC	2 min	1
1. Denaturation	98ºC	30s	30
2. Annealing	57ºC	45s	30
3. Extension	72ºC	1min	30
Final Extenstion	72ºC	2 min	1
Hold	4ºC	Hold	Hold

Purification of the PCR product was performed using AMPure XP (Agencourt) magnetic beads following manufacturer's recommendations with a ratio of 5:9 of sample: bead solution (https://www.beckman.fr/reagents/genomic/cleanup-and-size-selection/pcr). DNA quantification, size distribution, and quality assessment were carried out using the Agilent Bioanalyzer 2100 High Sensitivity DNA kit (Agilent 5067-4626). PCR products from all of the samples were pooled in equimolar amounts and shipped for Illumina MiSeq sequencing at Josephine Bay Paul Center in Woods Hole, Massachusetts, USA.

Sequencing reads (sequenced fragments of DNA) were processed using the DADA2 pipeline (v 1.12). The workflow (https://benjjneb.github.io/dada2/tutorial.html) was followed with adjustments made to improve read recovery:

```
filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(240,160),
              maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
              compress=TRUE, multithread=TRUE)
```

truncLen=0
trimLeft=c(23,18) for bacterial samples and trimLeft=c(22,18) for archaeal samples
maxEE=c(2,6)

Taxonomy was assigned using the Silva database (v 132). The ASV table and taxonomy table generated via the DADA2 pipeline were merged with a metadata table to create a phyloseq object[2]. The phyloseq object was subsetted to remove unwanted taxa (*e.g.*, mitochondria, chloroplasts) and analyzed using the R package decontam (https://benjjneb.github.io/decontam/vignettes/decontam_intro.html) to remove potential contaminants via the "frequency" method. Samples with < 10k reads were pruned and alpha diversity metrices were calculated. Normalized data were subject to NMDS (non-metric multidimensional scaling) analysis using Bray-Curtis distance and plots were colored categorically. The top 100 ASVs were subsetted, and taxonomic diversity was plotted according to sample origin.

2.2 Metagenomic Analysis

Library preparation and sequencing were done as outlined in Deliverable 4.3.

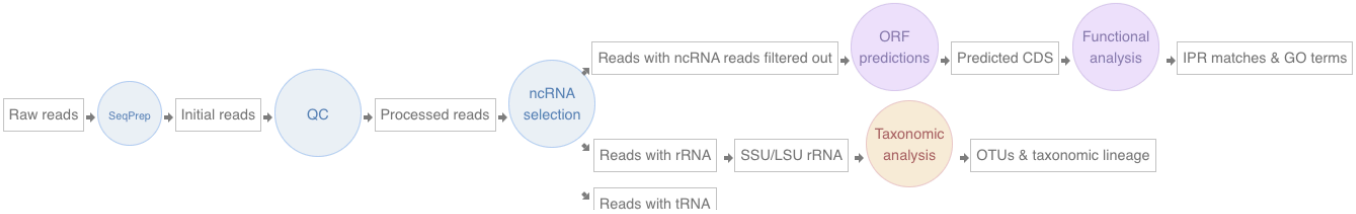

Figure 1: MGnify pipeline, version 4.1, for metagenomic analysis, modified from MGnify the website

Read-based Analysis

For read-based analysis, raw reads were submitted to MGnify (https://www.ebi.ac.uk/metagenomics/pipelines/2.0), a platform for the assembly, analysis and archiving of microbiome data, which is shown schematically in **Fig. 1**.

Assembly-based analysis

For assembly-based analysis, reads were quality filtered using Minoche[3], and processed using the Anvi'o snakemake workflow for metagenomics (http://merenlab.org/2018/07/09/anvio-snakemake-workflows/#metagenomics-workflow, **Fig. 2**) using the "all-against-all" mode with samples grouped by well. Assembly was done using MEGAHIT[4] (sensitive mode) and automatic binning was performed using CONCOCT[5] using n/2 estimated genomes.

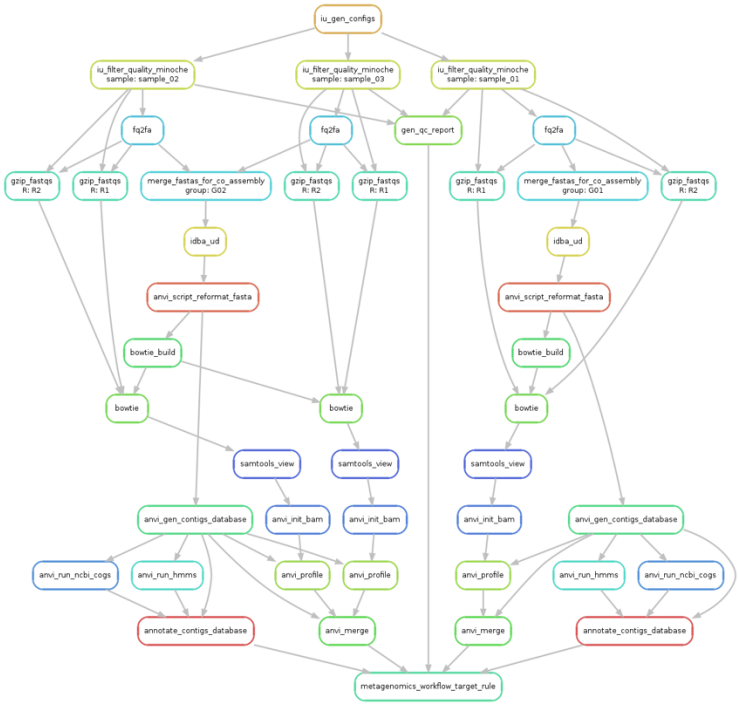

Figure 2: Anvi'o Snakemake workflow for metagenomics.

3 Summary of activities and research findings

16S Analysis: Alpha Diversity

A diversity index is a mathematical measure of species diversity in a community. Shannon's index accounts for abundance and evenness of the species present. This is given by the equation:

$$H = \Sigma[(pi) \times ln(pi)]$$

where *pi* = proportion of total sample represented by species *i*, obtained by dividing the number of individuals of species *i* by total number of samples. In this study, alpha diversity has been estimated using the Shannon's index. Species richness is given by the number of species, *S.* Accordingly, the maximum diversity possible is given by

$$Hmax = \ln(S)$$

while the evenness is obtained as:

$$E = H/Hmax$$

A drop in alpha diversity can represent a change in community structure where certain microbial members become more abundant while others diminish in relative proportion. This can result from microbial dynamics and/or environmental disturbance. The samples from Carbfix1 show a natural fluctuation in alpha diversity (**Fig. 3**) typical of a microbial community, with slightly higher diversity seen in pre-injection samples (excluding HK12, a shallow well that does not experience gas injection).

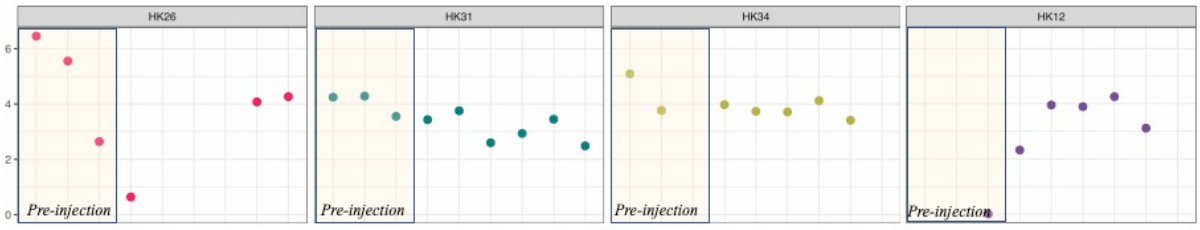

Figure 3: Alpha diversity measured by the Shannon Index

16S Analysis: Beta diversity and Taxonomy

Non-metric multidimensional scaling (NMDS) analysis is a method for visualizing the level of similarity between samples in a dataset. Multidimensional scaling creates an ordination plot using a similarity metric (Bray-Curtis), using dimension reduction to plot multi-dimensional data in two-dimensional space, with more similar samples clustering together. Plots were colored by well origin and well depth. The results from this analysis show that samples appeared to cluster by sample origin (**Fig. 4a**) as well as depth (**Fig. 4b**). Investigation of the taxonomic diversity reveals distinct communities of microorganisms in each of the wells which may be driving this clustering pattern (**Fig. 5**).

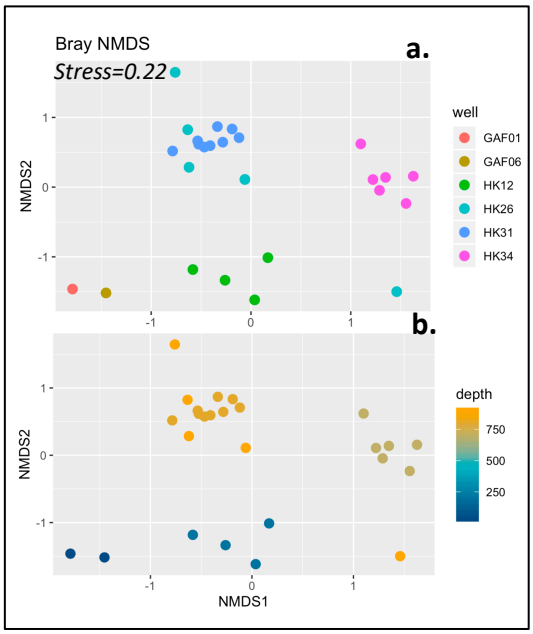

Figure 4: NMDS analysis using Bray-Curtis distance

Analysis of the top 100 ASVs at CarbFix1 revealed several taxa that could be ecologically relevant to subsurface operations. *Candidatus Desulforudis* are thought to survive on H_2 produced from radiolysis of water and sulfate derived from oxidation of pyrite by radiolytically produced O_2 and H_2O_2, and has the metabolic capabilities for CO_2 and N_2 fixation. This species was originally found as the predominate microbe 2.8 km beneath Earth's surface at Mponeng Gold Mine near Johannesburg, South Africa[6]. *Candidatus Tenderia electrophaga* is thought to be an electroautotroph, utilizing energy directly from a conductive surface (such as stainless steel) as a metabolic electron donor while fixing carbon dioxide into biomass[7]. *Desulfotomaculum*, *Hydrogenophaga,* and *Paracoccus* are capable of oxidizing hydrogen gas, with some species from these genera also capable of utilizing carbon dioxide. Several of these taxonomic groups have isolated cultivates with the capability to oxidize reduced sulfur and iron species (*Sulfuricella*, *Thiobacillus*).

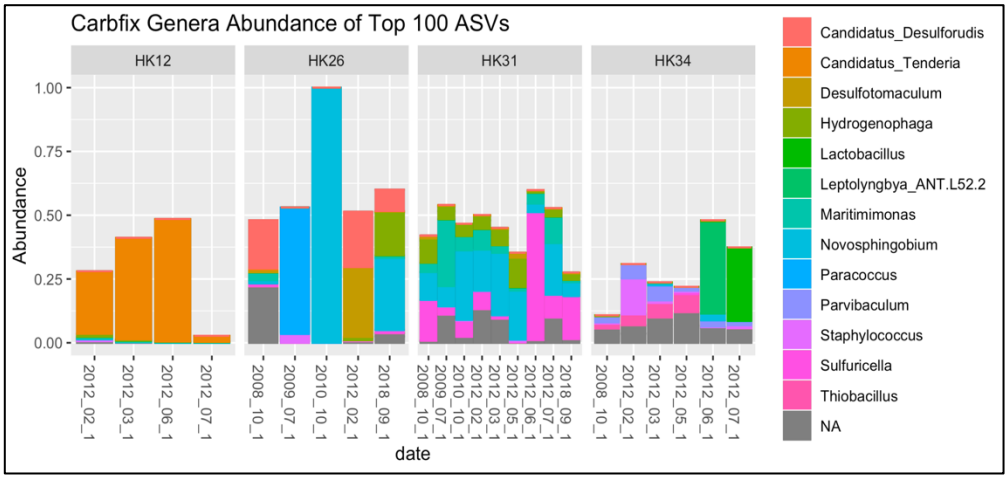

Figure 5: Taxonomic diversity of the top 100 ASVs classified at the genus level

Metagenomics: Read-based Analysis

Read-based metagenomic analysis is used to classify single reads from Whole Genome Sequencing (WGS) with regard to taxonomy and function, without primer biases introduced by metabarcoding analysis. This type of analysis can answer questions related to the taxonomical composition of a sample or related to the presence or absence of organisms, genes, or metabolic pathways. The output from the MGnify pipeline includes taxonomic analysis of large subunit rRNA and small subunit rRNA to asses taxonomic diversity that can include groups not detected by metabarcoding analysis (**Fig. 6a**). More than 44% of the bacterial diversity is dominated by the Proteobacteria phylum, followed by uncultivated members of the divison *Candidatus Kaiserbacteria* (16.29%) and then unassigned bacteria (11.73%).

MGnify also categorizes the functional content of each metagenome, using matches to the InterPro database (**Fig. 6b**), which provides functional analysis of proteins using predictive models, and classifies them using Gene Ontology terms (biological classes for molecular functions, cellular locations, and processes gene products may carry out, **Fig. 6c**). At this stage of analysis, it is challenging to retrieve specific functional metabolisms for the environment being studied.

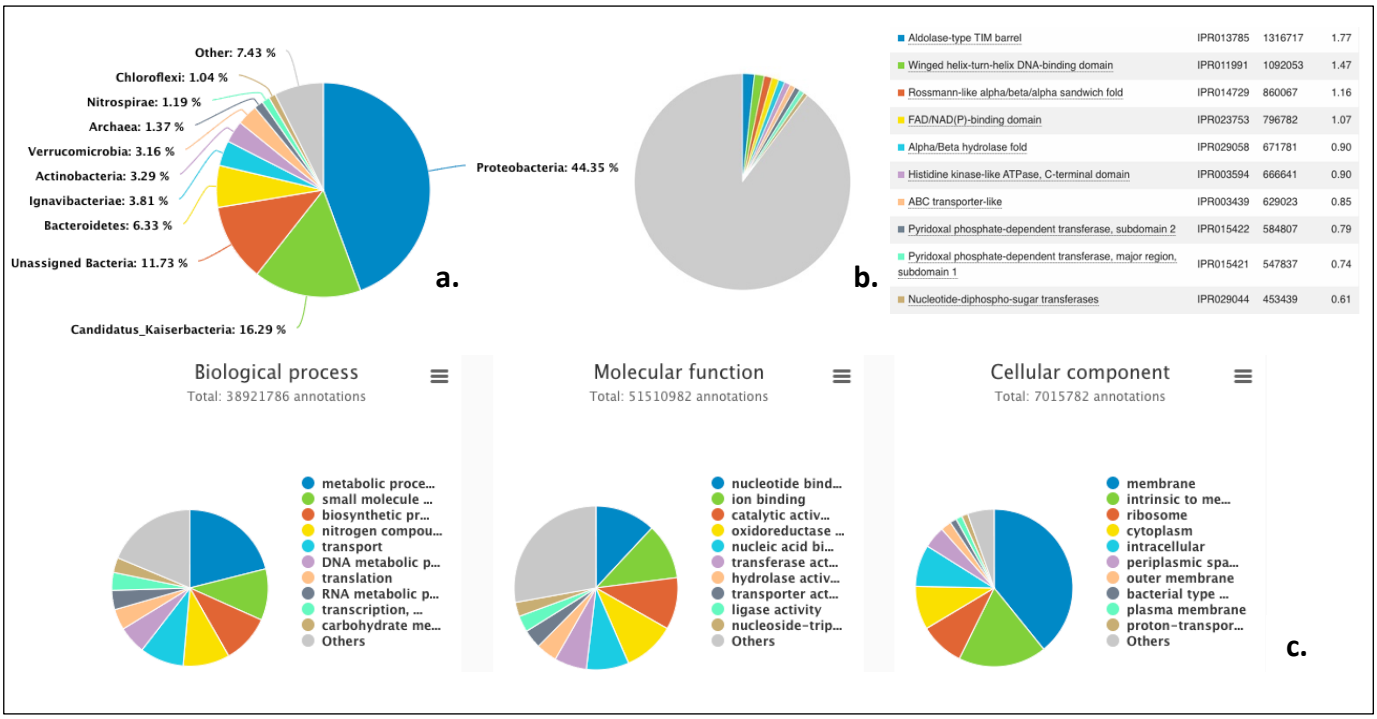

Figure 6: Figure 6. Read-based analysis of HK31 (March 2012) with a.) taxonomy using the SS and LS rRNA, b.) Interpro match summary, and c.) the summary of GO terms derived from the Interpro matches

Metagenomics: Assembly-based Analysis

Assembly-based workflows aim to combine the reads from one or more samples (assembly), subset these contigs into metagenome-assembled genomes (binning), analyze metabolic profiles contained in the MAGs to determine taxonomic assignment and calculate the abundance and distribution of the MAG across samples. The estimated number of MAGs present in the samples is generated after the Anvi'o Snakemake workflow by determining if each MAG contains a complete list of single-copy core genes specific and essential to either bacteria, archaea, or eukarya (**Fig. 7**). In order to ensure each MAG represents DNA from a single source or organism, manual refinement to account for genome fragmentation (the breaking apart of a MAG across multiple bins) or inflation (the inclusion of multiple MAGs in a single bin) is necessary before analysis of functional profiles, abundance, and distribution. Regardless of the sampling site, no eukaryotic MAGs were reconstructed. It can be stated that bacteria dominate at all sites and represent 89% of the genomes (1,524 MAGs) assembled from metagenomes.

Domain	Nesjavellir (2)	HK12 (7)	HK26 (7)	HK31 (9)	HK34 (7)	Total (32)
Archaea (Archaea_76)	112	3	40	12	0	167
Bacteria (Bacteria_71)	663	113	257	295	29	1,357
Eukarya (Protista_83)	0	0	0	0	0	0

***Figure 7**: Distribution of estimated 1,524 MAGs before manual refinement from Carbfix1 and Nesjavellir (GAF01 and GAF06)*

4 Conclusions and future steps

The objectives of this deliverable were to analyze and process the metagenome and diversity data obtained from environmental samples and determine molecular markers relevant for biodiversity analyses (16S rDNA and functional genes) to identify the various metabolisms associated with gas injection at low temperature. For the 16S Metabarcoding analysis, investigation of the taxonomic diversity is complete, while the significance of beta diversity needs to be statistically verified to determine chemical parameters driving differences in well communities. The process of identifying biomarkers for different well chemistries is currently being finished. The metabarcoding analysis shows unique taxonomic profiles for each of the monitoring wells, most likely being driven by physiochemical conditions including depth and temperature. Across monitoring wells, many of the dominant taxa present are known to metabolize inorganics such as hydrogen gas, carbon dioxide, and reduced iron and sulfur compounds, which is not surprising given the chemical constituents present in the aquifers.

Metagenomic analyses are still ongoing. For the read-based analysis, current work includes identifying patterns in functional changes as they relate to well conditions.
For the assembly-based analysis, the Snakemake workflow to generate bins is finished, resulting in subsets of data (bins), some of which contain MAGs. Manual refinement is needed to generate high-quality bins will low contamination, representing metagenome-assembled genomes (MAGs). The high quality MAGs that have been reconstructed so far are almost entirely derived from bacteria, which predominate this system, and with the functional diversity reflecting the taxonomic diversity found using metabarcoding analysis. The completion of this process will enable the identification of the various organisms and their respective metabolisms in response to geologic conditions not possible with metabarcoding alone.

5 Bibliographical references

[1] Trias, R., Ménez, B., le Campion, P. *et al.* High reactivity of deep biota under anthropogenic CO_2 injection into basalt. *Nat Commun* **8,** 1063 (2017). https://doi.org/10.1038/s41467-017-01288-8

[2] McMurdie, Paul J., and Susan Holmes. "phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data." *PloS one* 8.4 (2013).

[3]Eren AM, Vineis JH, Morrison HG, Sogin ML (2013). A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. PLOS ONE 8(6).

[4]Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674-1676.

[5]Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... & Quince, C. (2013). CONCOCT: clustering contigs on coverage and composition. *arXiv preprint arXiv:1312.4038*.

[6]Chivian, D., Brodie, E. L., Alm, E. J., Culley, D. E., Dehal, P. S., DeSantis, T. Z., ... & Moser, D. P. (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. *Science*, *322*(5899), 275-278.

[7]Eddie, B. J., Wang, Z., Malanoski, A. P., Hall, R. J., Oh, S. D., Heiner, C., ... & Strycharz-Glaven, S. M. (2016). 'Candidatus Tenderia electrophaga', an uncultivated electroautotroph from a biocathode enrichment. *International journal of systematic and evolutionary microbiology*, *66*(6), 2178-2185.