HOLOGENOMIX

Q-2024-015 REPORT

October 2024

TABLE OF CONTENTS

1. ORDER INFORMATION

Client Name	Timmy Paez Watson
Client Organization	TU Delft
Order Number	Q-2024-015
Application	3 metagenomics samples for bioinformatics analysis

2. INTRODUCTION

This report presents the results of a metagenomic analysis of 3 biological samples obtained from TU Delft reactors. The goal of the analysis was to assess the quality of the data derived from the sequencing of these samples and to generate statistical insights into the microbial communities present in the reactors as well as to classify generated bins to reference genomes provided by the client.

The quality of the sequencing data was assessed using a variety of metrics, including the number of reads per sample, and the average read length. The results of the quality assessment showed that the data was of high quality, with an average read length of 150 nucleotides.

A variety of statistical methods were used to analyze the microbial communities present in the samples. These methods included sample ordination plots, taxonomical analysis. **(i)** Sample ordination plots were used to visualize the relationships between the samples based on the metadata provided by the client. **(ii)** Taxonomical analysis was used to classify the microbial communities at the phylum and genus level. **(iii)** Bins generation. **(iv)** Bins classification based on reference genomes provided by the client, **(v)** phylogenetic tree based on *ppk1* gene from *Accumulibacter* bins.

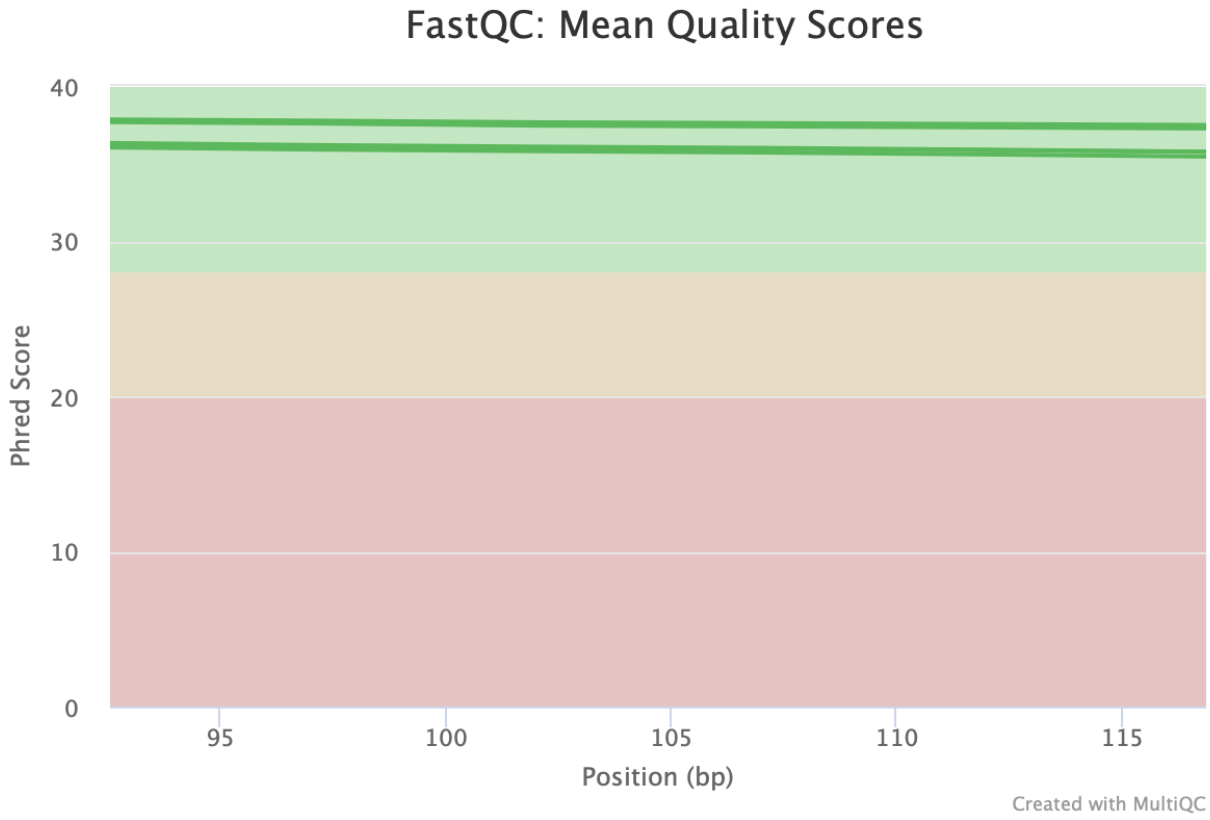
3. RAW DATA STATISTICS

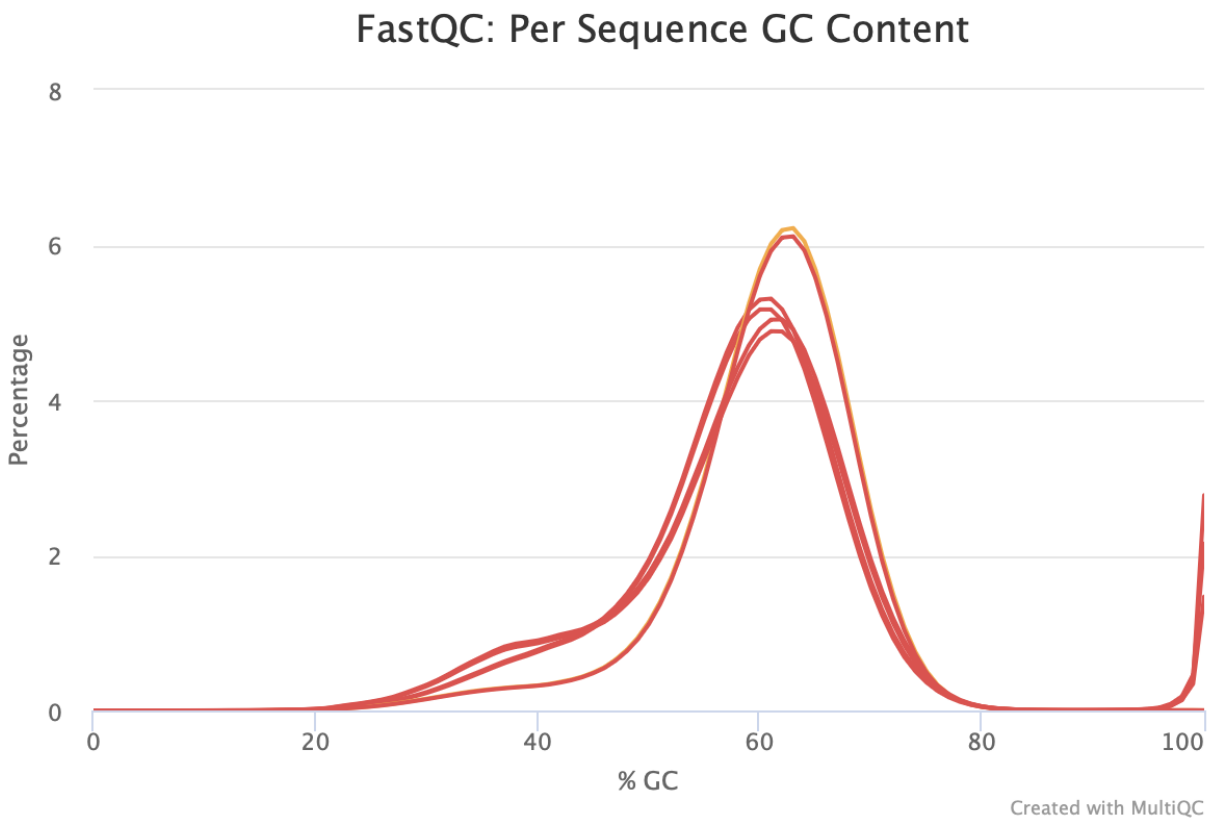
Sample ID	Total reads (Million)	GC(%)	Q30(%)	Average Read Length
P015-2p67n1	85.89	57.8	83.3	150 bp
P015-2p67n2	67.89	60.4	84.5	150 bp
P015-2p68n1	83.1	58.0	83.6	150 bp

- **Sample ID:** Sample name.
- **Total reads:** Total number of reads. This value refers to the sum of read1 and read2 for Illumina paired-end sequencing.
- **GC (%):** Ratio of GC content.
- **Q30(%):** Percentage of bases in a sequencing read with a Phred quality score greater than 20 and 30, respectively, which correspond to a base call accuracy of 99% and 99.9%.
- **Average Read Length: Paired**-end reads length in basepairs

4. READS QUALITY CONTROL

MultiQC generates per-base sequence quality scores, allowing you to visualize the quality of each base position across all reads. This information helps identify potential regions of low-quality bases, which might affect downstream analyses. The figures here displayed show the quality scores per sample after low-quality reads have been trimmed. Full reports are found in the attached folder (holo-23-15-multiqc.html).





5. DIVERSITY INDEXES

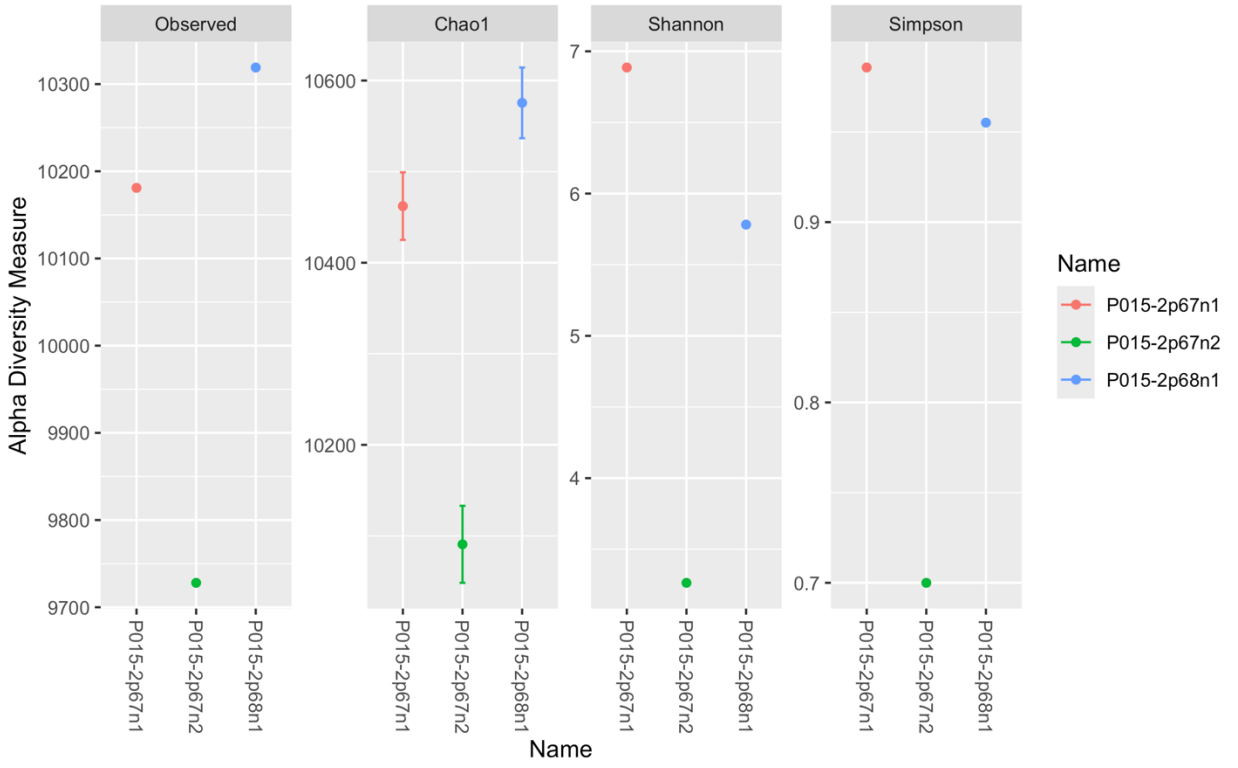
Diversity indexes are statistical measures used to quantify the variety and distribution of species or entities within a specific ecosystem, community, or dataset. They provide a way to assess the richness and evenness of species or elements present.

Alpha-diversity

It is a measure of biodiversity that assesses the species richness and diversity within a particular local or specific habitat, ecosystem, or community. It focuses on the diversity of species within a single location or sample without considering the differences between multiple locations or samples. Alpha-diversity metrics aim to answer questions like "How many different species are present in this area?" or "How evenly are these species distributed within the habitat?" Common measures of alpha-diversity include

species richness (the total number of different species) and various diversity indices such as the Shannon Index, Chao1 and Simpson's Diversity Index.

Visualized by sample



Beta-diversity

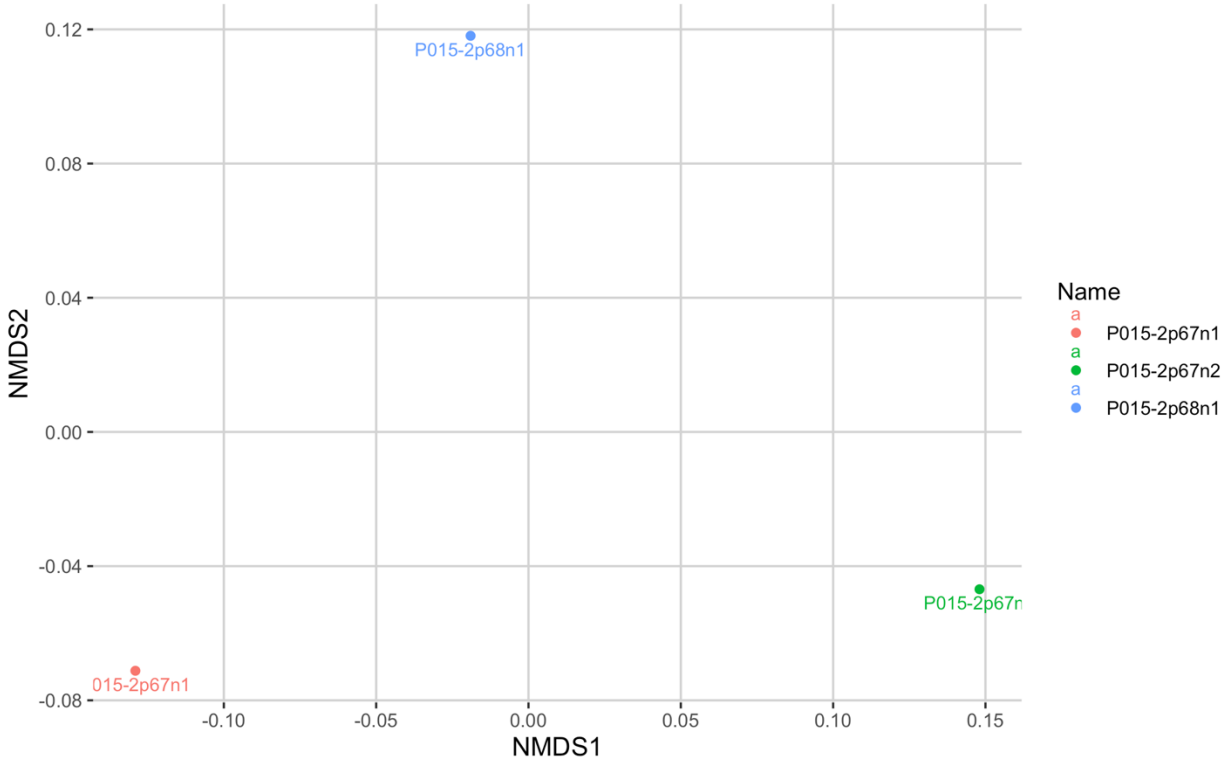
Beta-diversity is a measure of biodiversity that assesses the differences in species composition and diversity between multiple habitats, ecosystems, or communities. It quantifies the turnover or change in species composition as one moves from one location or sample to another. In essence, beta-diversity provides insights into the variation in species identities and abundances between different sites or environments.

There are various ways to calculate beta-diversity, but the fundamental idea is to compare the species composition and relative abundances across different sampling units (such as different plots in a forest or different geographic locations). Beta-diversity can help scientists understand how different habitats or regions differ in terms of biodiversity and can provide valuable information for conservation efforts,

habitat management, and ecological research. Common methods for calculating beta-diversity include Bray-Curtis Dissimilarity.

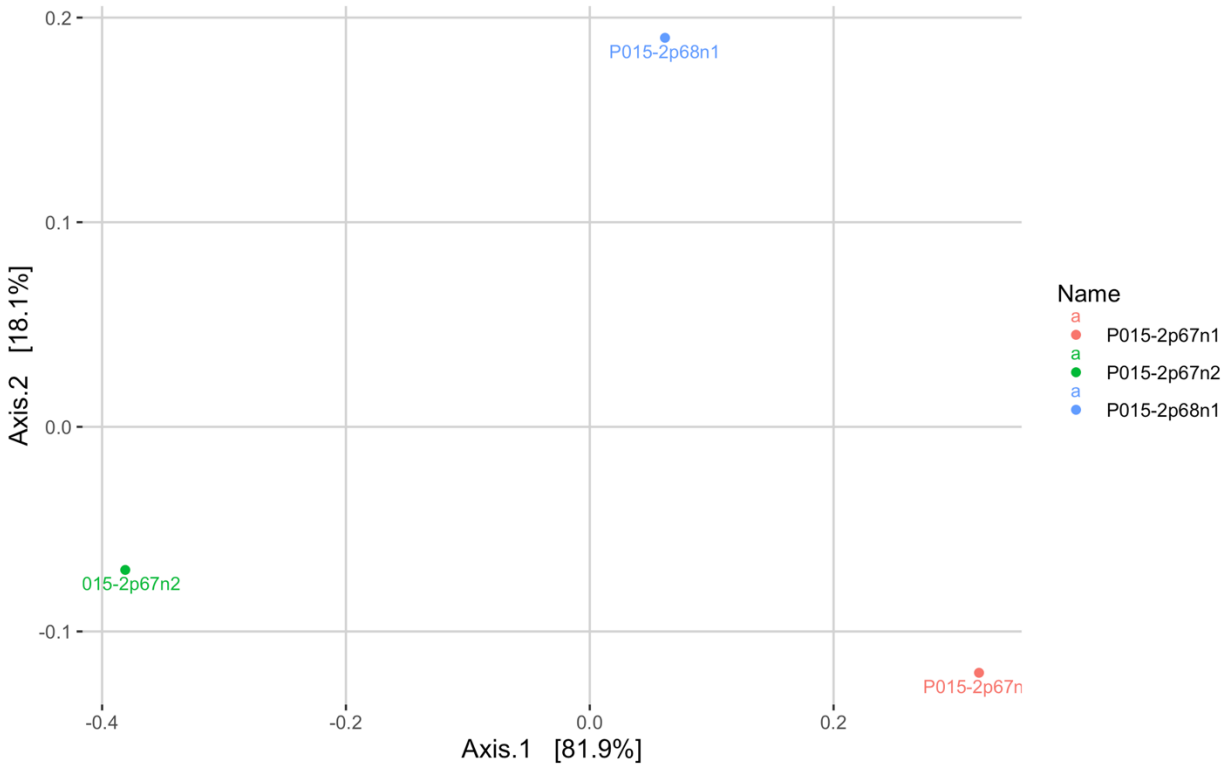
NMDS – Non-Metric Multidimensional Scaling

It is a statistical technique commonly used in ecology to assess beta-diversity, which measures differences in species composition between multiple sampling sites or habitats. NMDS is a dimensionality reduction method that transforms complex species composition data into a lower-dimensional space while preserving the dissimilarity or distance relationships between the sites.



PCoA – Principal Coordinates Analysis

It is a statistical method used in ecology to visualize and analyze the patterns of similarity or dissimilarity in multidimensional data. PCoA is particularly useful for exploring and visualizing the structure of datasets that involve distances or dissimilarities between objects of samples.

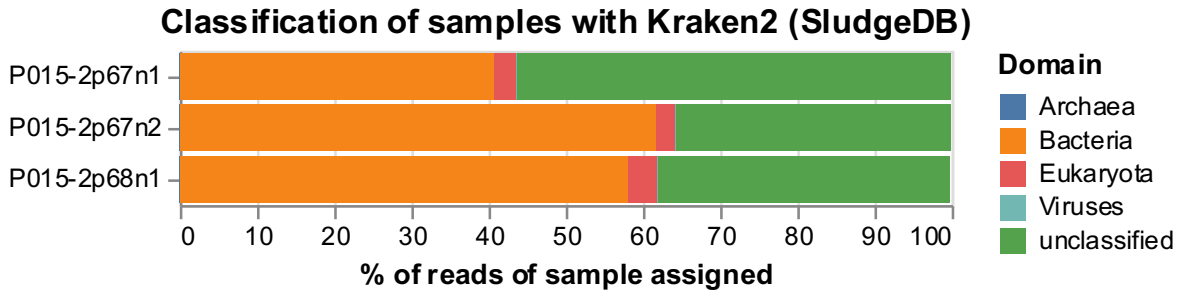

0.2
0.1
0.0
-0.1
Axis.2 [18.1%]
P015-2p68n1
015-2p67n2
P015-2p67n
-0.4
-0.2
0.0
0.2
Axis.1 [81.9%]
Name
a
P015-2p67n1
a
P015-2p67n2
a
P015-2p68n1

6. TAXONOMIC ANALYSIS

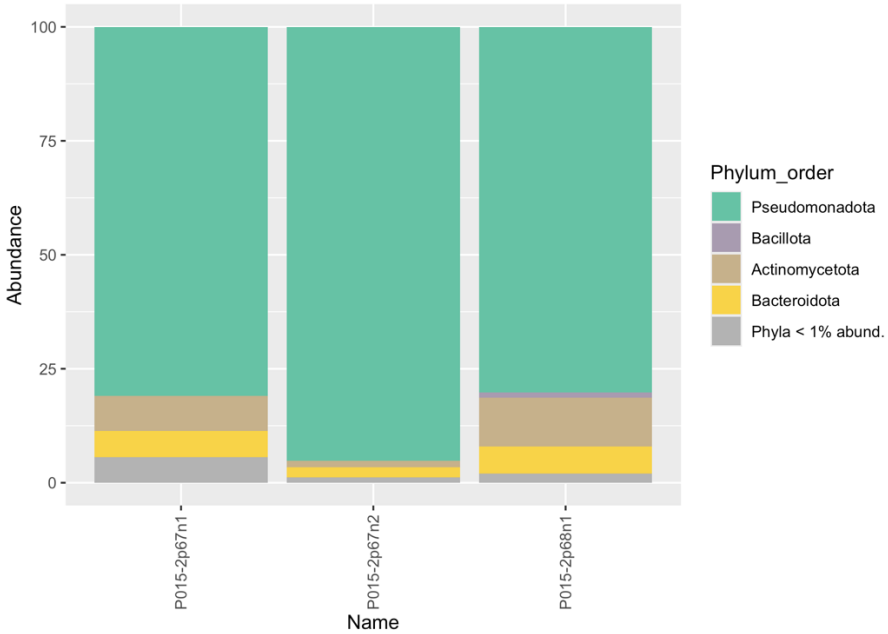
6.1. Kraken results from raw-reads

Taxonomic analysis is a systematic approach to classifying and categorizing living organisms based on their shared characteristics and evolutionary relationships. It involves identifying, naming, and organizing species into hierarchical groups, such as genera, families, orders, and more, to understand the diversity of life on Earth. Taxonomic analysis provides a foundation for biological research, conservation efforts, and our overall understanding of the natural world. We display here Phylum (plural: Phyla) level and Genus (plural: Genera) level from raw-reads.

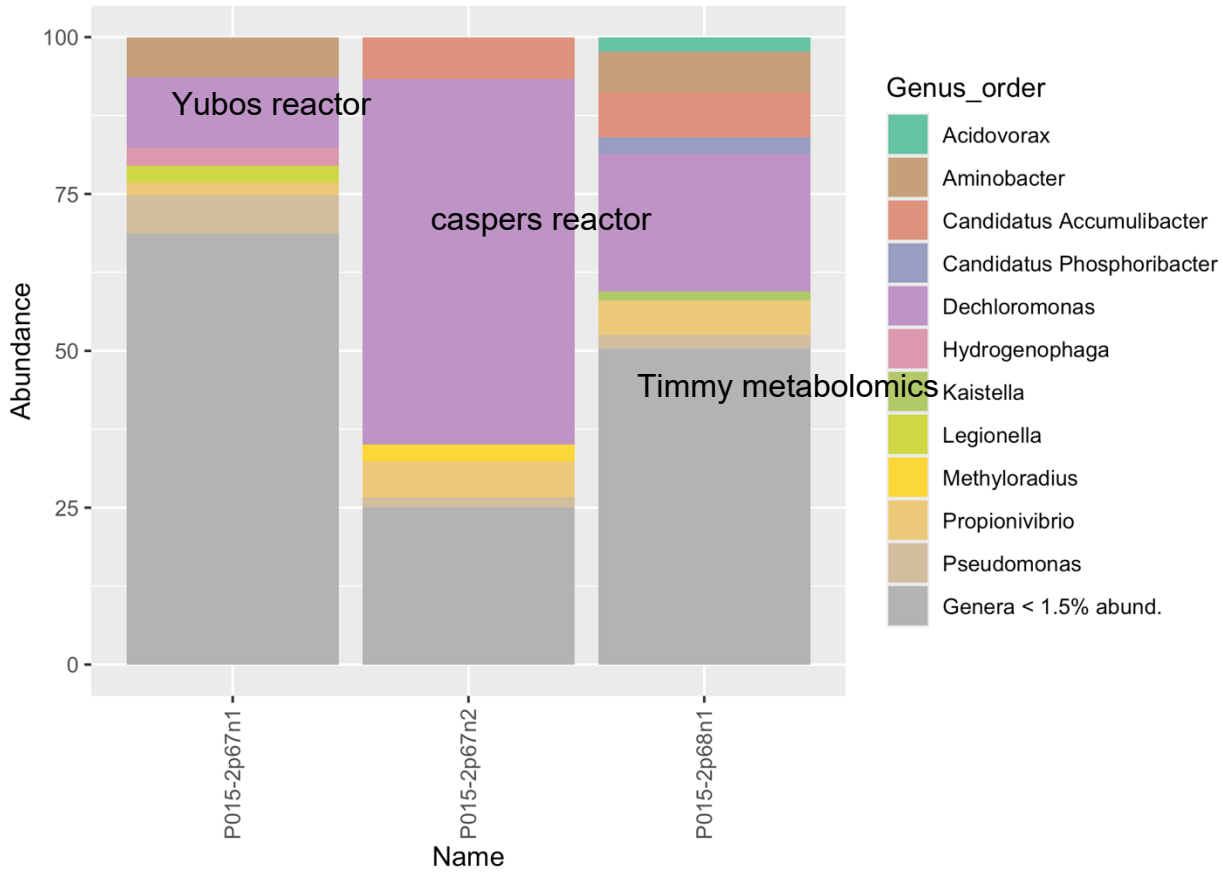
Kingdom



Phylum



Genera

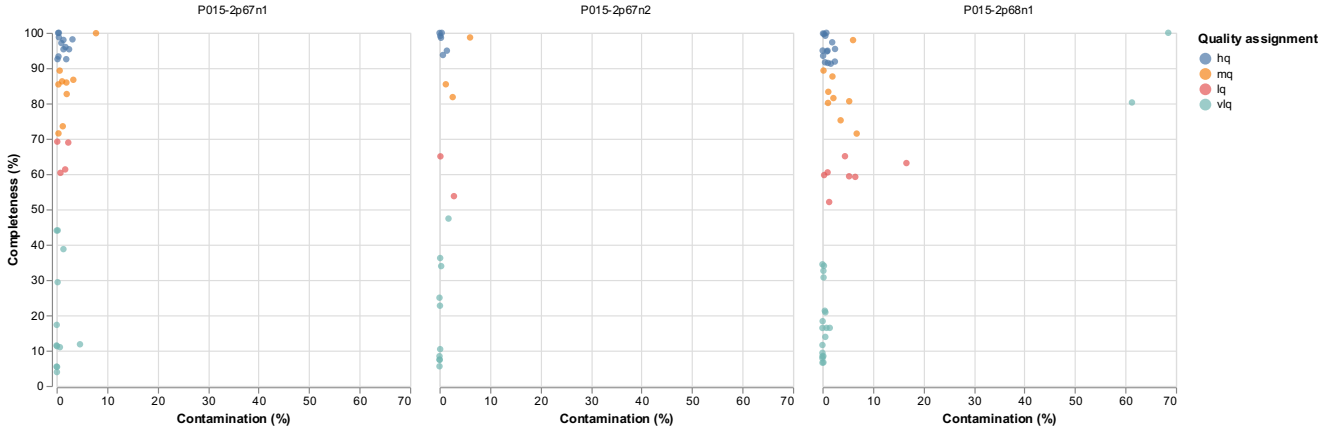

Yubos reactor
caspers reactor
Timmy metabolomics
Abundance
100
75
50
25
0
P015-2p67n1
P015-2p67n2
P015-2p68n1
Name
Genus_order
Acidovorax
Aminobacter
Candidatus Accumulibacter
Candidatus Phosphoribacter
Dechloromonas
Hydrogenophaga
Kaistella
Legionella
Methyloradius
Propionivibrio
Pseudomonas
Genera < 1.5% abund.

6.2. Taxonomic classification of generated bins

This process assigns bins, created from your metagenome data.

6.2.1. Bins quality

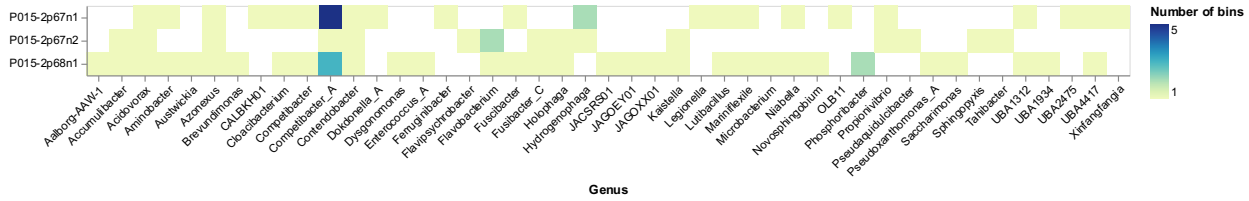
CheckM assesses the completeness and contamination of your bins. Completeness tells you how much of the original genome is captured in the bin, while contamination indicates the presence of genetic material from other organisms. Good quality bins have high completeness (>90%) and low contamination (<5%). For the taxonomical classification using GTDB and the assignment to the reference genomes provided by client, we use bins only with low contamination while keeping bins that are incomplete (<90%). Figures and tables are available in the attached results folder.



Note: HQ (High Quality) = 90%+ Com, 5%- Con. MQ (Medium Quality)= 70%+ Com, 10%- Con. LQ (Low Quality) = 50%+ Com, 30%- Con. VLQ (Very Low Quality) =remaining. Com=Completeness. Con=Contamination

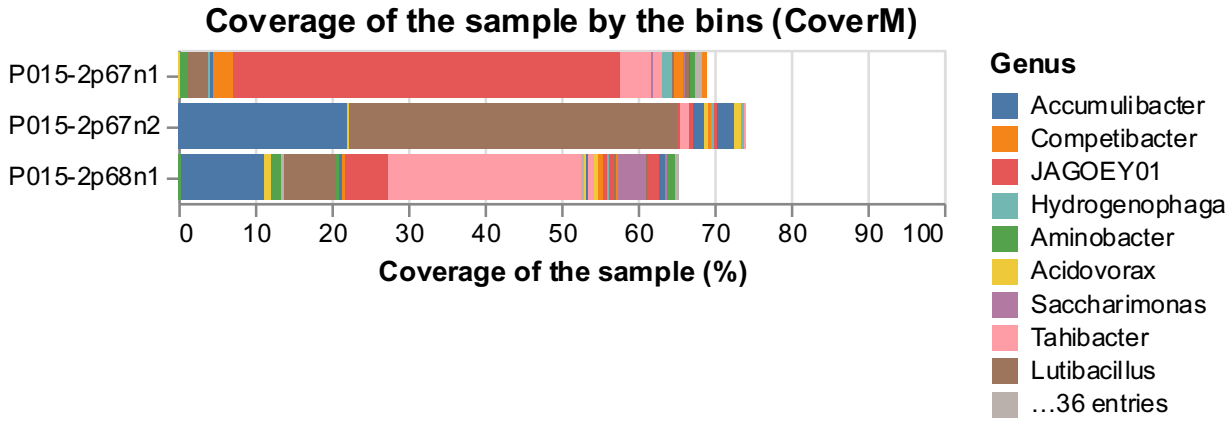
6.2.2. GTDB classification

GTDB (Genome Taxonomy Database) is a resource that proposes a new way to classify bacteria and archaea based on their entire genomes, rather than just a few genes. This approach aims to be more accurate and reflect evolutionary relationships. GTDB-tk is the software tool that works with GTDB database. Results are displayed in the attached folder in file *gtdbk.bac120.summary.tsv*.



6.2.3. Bins coverage with trimmed raw-reads

This refers to how well the trimmed reads (filtered, high-quality reads from your sequencing data) map back to the generated bins. High coverage means a good portion of the reads originates from the organisms represented by those bins.



7. PHYLOGENETIC TREE ANALYSIS

The microbial composition and functional potential of the environmental samples were analyzed through a series of bioinformatic approaches. First, GTDB-tk was used to assign taxonomic classifications to the genomic bins (section 6.2.), allowing for the identification of their phylogenetic affiliations. This step provided detailed insights into the diversity of organisms present, including the identification of key taxa relevant to polyphosphate accumulation. Two bins were annotated as *Accumulibacter* from 2 out of the 3 samples provided (P015-2p68n1 and P015-2p67n2).

Sample	Bin	Species	Completeness	Contamination	Closest reference	ANI to closest reference	AF to closest reference
P015-2p68n1	bin.13	Accumulibacter sp000585075	99.66	0.31	GCA_000585075.1	98.99	0.88
P015-2p67n2	bin.10	Accumulibacter sp000585075	98.68	6.05	GCA_000585075.1	98.92	0.839

Following this, bakta was applied to annotate the bins identified as *Accumulibacter*, a genus known for its role in enhanced biological phosphorus removal. This annotation process facilitated the identification of relevant genes and metabolic pathways associated with this genus.

To further investigate the potential for polyphosphate accumulation, HMMer was used to search for the *ppk1* gene, which encodes polyphosphate kinase, a key enzyme in the polyphosphate metabolism pathway. Predicted genes across the bins were screened, with specific attention given to those from *Accumulibacter*.

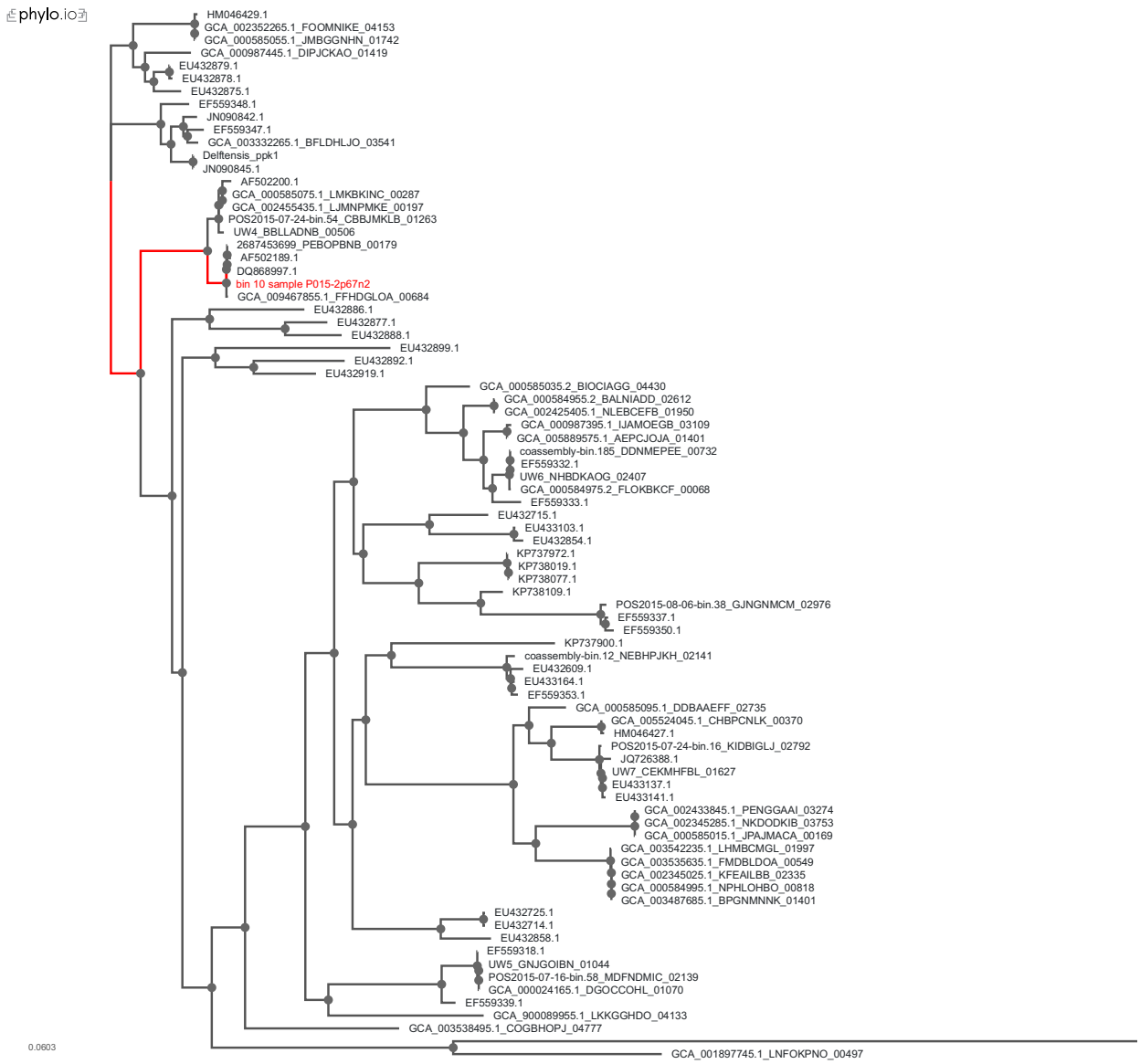
BIN 10

```
#                                                            --- full sequence ---- --- best 1 domain ---- --- domain number estimation ----
# target name        accession  query name           accession    E-value  score  bias   E-value  score  bias   exp reg clu  ov env dom rep inc description of target
#------------------- ---------- -------------------- ---------- --------- ------ ----- --------- ------ -----   --- --- ---  -- --- --- --- --- ---------------------
OLFJAH_12200         -          ppk1-refs            -          1.1e-157  520.5   0.1  1.6e-157  520.0   0.1   1.2   1   0   0   1   1   1   1 polyphosphate kinase 1
```

BIN 13

```
sequence ---- --- best 1 domain ---- --- domain number estimation ----
# target name        accession  query name           accession    E-value  score  bias   E-value  score  bias   exp reg clu  ov env dom rep inc description of target
#------------------- ---------- -------------------- ---------- --------- ------ ----- --------- ------ -----   --- --- ---  -- --- --- --- --- ---------------------
JLFGNB_03800         -          ppk1-refs            -          1.1e-157  520.5   0.1  1.5e-157  520.0   0.1   1.2   1   0   0   1   1   1   1 polyphosphate kinase 1
```

The identified *ppk1* genes, along with sequences from clones and an outgroup, were aligned using MUSCLE to assess their evolutionary relationships. This alignment provided a basis for phylogenetic analysis. A maximum-likelihood phylogenetic tree was then constructed using RAxML-NG to examine the evolutionary placement of the *ppk1* genes across the identified taxa, shedding light on the distribution and potential functional diversity of this key gene in the microbial community. Client suggested to add information only from **P015-2p67n2 (bin10).**



8. FILES ATTACHED

Attached is a list of files located within the **data** folder. Within this **data** folder, you will discover the results organized into folders named after each respective **sample**. Within each of these **sample** folders, you will find the following:

- **Raw_data** files. It contains the sequenced reads obtained from the samples provided. Provided in *sample.fastq.gz* format.
- **Trimmed_data** files. It contains the sequenced reads obtained from the samples provided with low-quality reads trimmed. Provided in *sample.fastp.fastq.gz* format.
- **FastQC** files. It contains the MultiQC and fastq report on data quality both from raw and trimmed reads. Provided in fastp*_fastqc.html* format.
- **Fastp** files. It contains the fastp results in *fastp.html* format.
- **Kraken_results** files. It contains the taxonomical classification tables obtained from Kraken 2.0 from the samples provided. Provided as *sample.trimmed.kraken.report*.
- **Sample.metaspades** folder. It contains the assemblies (contigs) files from each of the samples provided in *contigs.fasta* format as well as *sample.contigs.[sludgedb,standard].report* files with the kraken outputs.
- **Sample_metabat** folder. It contains the bins generated in each of the sample (sample.metabat folder).
- **Sample_bins** folder. It contains the bins generated after filtering high quality bins in each of the sample (sample.metabat folder).
- **Sample.checkm** folder. It contains stats about the quality of the bins (sample.checkm folder).
- **Sample.checkm2** folder. It contains stats about the quality of the bins (sample.checkm2 folder).
- **CoverM** files. It contains the raw-reads coverage results from reference genomes (provided by client) *sample.coverm.references.tsv* and from bins generated in *sample.coverm.tsv.*
- **gtdbtk** folder. It contains the output of the tool. Inside the **classify** subfolder, the assignments of the classification are found in file *gtdbtk.bac120.summary.tsv.*
- **Figures** folder. It contains the figures displayed in this report. Briefly, the **Diversity** folder contains the diversity and multidimensional reduced figures in *.tiff* (quality ready for publication*)* format. The **Taxonomy** folder contains the phylum and genera figures in *.tiff* (quality ready for publication*)* format. It contains the figures for the analysis of the bins per sample

(sample.bin_quality.tiff, sample.coverage-bins.tiff, sample.coverage-references.tiff, sample.heatmap-dist.tiff) including their associated tables in .csv format.

9. MATERIALS & METHODS

Quality control of sequenced reads

The quality of the sequenced raw reads was assessed by FastQC (version 0.11.7) with default parameters (Andrews, 2010) and visualized with MultiQC (version 1.19). Low-quality paired-end reads were trimmed and filtered by Fastp version 0.23.4 on the paired-end mode (Chen, 2023).

Microbiome profiling

Taxonomic classification of raw reads was performed to profile the microbiome from each sample using the standard Kraken2 (version 2) database (uses all complete bacterial, archaeal, and viral genomes in NCBI Refseq database) complemented with a curated wastewater database (sludgeDB) with default parameters (Wood et al., 2019). The taxonomic classification outcomes from Kraken2.0 were converted into abundance tables using the biom file converter tool to explore metagenomics classification datasets. Figures were generated with the R package “phyloseq” (McMurdie and Holmes, 2013).

Microbiome diversity

To identify patterns in the microbiome, different types of ordination and statistical numerical methods were performed using the R software. The input data was the output of the taxonomic classification from Kraken2.0.

Assembly of sequence reads

Clean reads were assembled into contigs using MetaSPAdes (version 3.15.5) with default parameters (Nurk et al., 2017).

Binning of DNA contigs

Contigs resulting from the sequencing were binned with MetaBAT version 2.2.15 (Kang et al., 2019) to reconstruct metagenome-assembled genomes (MAGs or bins) on default parameters. CheckM (version 1.2.2) (Parks et al., 2015) determined bin completeness and contamination using the "lineage_wf" workflow.

Read coverage

The relative abundance of the bins with contamination <5% in each sample was determined with CoverM (version 0.7.0, https://github.com/wwood/CoverM) with default parameters.

Phylogenetic tree generation

Followed method from https://github.com/elizabethmcd/ppk1_Database/:

Bins were classified with GTDBtk version 2.4.0 (Chaumeil et al., 2022) and GTDB release 220 as a standard method. In bins identified as *Accumulibacter*, hmmsearch (Johnson et al., 2010) was used with ppk1.hmm, taking the best hit as *ppk1* gene (same result as using bakta and searching for *ppk1*). The *ppk1* genes found were combined with those in the existing database and aligned with MUSCLE v5.1.linux64 (Edgar, 2004) . RaxML-NG v1.2.2 (Kozlov et al., 2019) was then used to create the phylogenetic tree.

10. REFERENCES

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.

Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2022. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics 38, 5315–5316. https://doi.org/10.1093/bioinformatics/btac672

Edgar, R.C., 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5. https://doi.org/10.1186/1471-2105-5-113

Johnson, L.S., Eddy, S.R., Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z., 2019. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019, 1–13. https://doi.org/10.7717/peerj.7359

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., 2019. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35, 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

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

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. MetaSPAdes: A new versatile metagenomic assembler. Genome Research 27, 824–834. https://doi.org/10.1101/gr.213959.116

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25, 1043–1055. https://doi.org/10.1101/gr.186072.114

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 1–13. https://doi.org/10.1101/762302