Metagenome Analysis Report

Eurofins Project ID: **EF-DEMO**

Date of Processing: **24 March, 2025**

Pipeline: **Metagenome Analysis Pipeline**

Version: **v2.4.6**

This report is not a diagnostic / clinical report and is intended for Research Use Only!

Eurofins Genomics

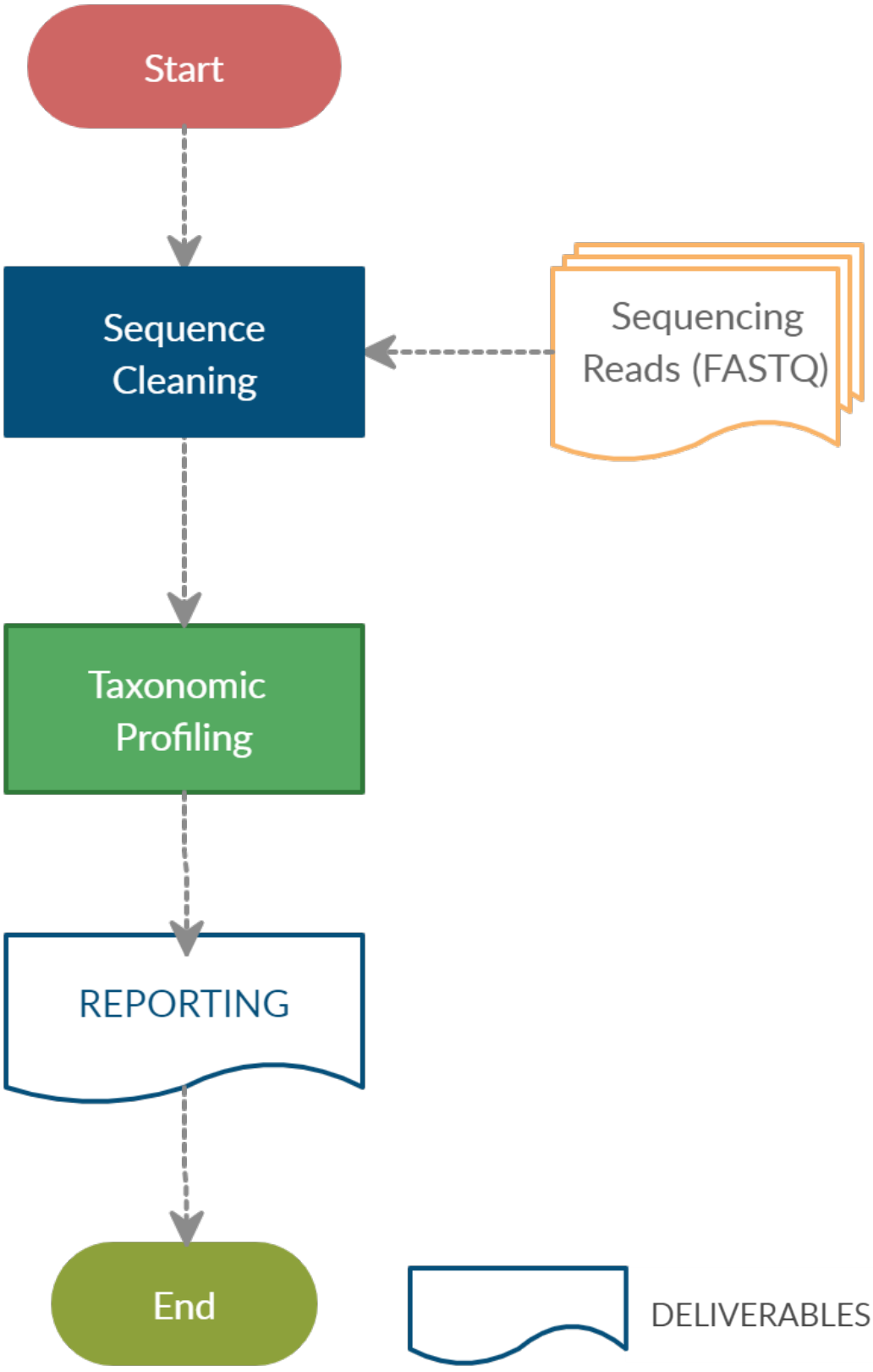
24 March, 2025

Contents

1 Metagenome Analysis

1.1 Analysis Workflow

Schematic diagram showing the main steps of the analysis method followed to perform the data analysis.


Start
Sequence Cleaning
Sequencing Reads (FASTQ)
Taxonomic Profiling
REPORTING
End
DELIVERABLES

1.2 Sequence Quality Control

1.2.1 Quality Control of raw sequencing data

Raw sequencing data are preprocessed to generate clean data for downstream analysis.

Quality of raw sequencing data is checked and filtered to retain only high quality bases by performing adapter trimming, quality filtering and per-read quality pruning. Quality is interpreted as the probability of an incorrect base call or, equivalently, the base call accuracy. The quality score is logarithmically based, so a quality score of 10 reflects a base call accuracy of 90%, but a quality score of 20 reflects a base call accuracy of 99% and a quality score of 30 reflects a base call accuracy of 99.9%. These probability values are the results from the base calling algorithm and depend on how much signal was captured for the base incorporation.

Sequencing reads representing reads with quality score at least Q30 is above 90% is of very good quality. For a reasonably good sample source material, according to Illumina specifications, one could expect >75% reads with at least Q30 Phred quality.

Raw sequencing data is processed using fastp[2] software to remove poor quality bases (below Phred Quality 20) using the sliding window approach where in if the average quality of the bases drops below Q20, those bases are removed from the reads. After quality trimming, program checks for presence of any adapters in the reads and removes from the reads. Further, shorter reads which are <30bp length are also removed to retain only high quality sequencing reads for each sample in the analysis. In case of paired-end reads, both the sequencing reads which pass the QC criteria are considered for downstream analysis.

After QC processing, QC metrics such as Q30 reads and GC content can be used to assess the sequencing and sample quality across the samples.

1.2.2 Read Statistics

- Table 1: Sequence Quality Metrics overview. For each sample, the following QC metrics are provided:
  - Sample Name: name of the sample.
  - Total Raw Reads: the total number of raw sequencing reads generated for the sample.
  - Total HQ Reads: the total number of high quality reads after sequence cleaning and filtering.
  - HQ Bases (Q30): Percentage of high quality bases having at least phred quality 30.
  - GC Content: GC content in percentile of high quality sequencing reads.
  - Mean Read Length (bp): Average read length in bp of high quality sequencing reads.
  - HQ Reads %: High Quality Reads percentage.

ID	Sample Name	Total Raw Reads	Total HQ Reads	HQ Bases (Q30)	GC Content	Mean Read Length (bp)	HQ Reads %
1	MS_zymo_1	36.33 M	35.7 M	90.56%	48.05%	149	98.26%
2	MS_zymo_2	36.21 M	35.6 M	89.93%	47.78%	150	98.32%
3	MS_zymo_3	32.15 M	31.61 M	90.46%	47.68%	150	98.32%

1.3 Host Removal

The host removal is done using Kraken[7]. Kraken classifes reads by breaking each read into overlapping k-mers. Each k-mer is mapped to the lowest common ancestor (LCA) of the genomes containing that k-mer in a precomputed reference database . For each read, a classifcation tree is found by pruning the taxonomy and only retaining taxa (including ancestors) associated with k-mers in that read. Each node is weighted by the number of k-mers mapped to the node, and the path from root to leaf with the highest sum of weights is used to classify the read. KrakenUniq[1] computes the number of unique k-mers observed for each taxon, which allows to filter more false positives. The fastq files were filtered for non-host sequences using SeqKit[5] for further downstream analysis. The final host classifed, unclassifed and filter passed reads are reported in the table below.

- Table 2: Host removal profile metrics per sample:

ID	Sample Name	Total Reads	Classified Reads (Host)	Classified Reads (Host) %	Unclassified Reads (Host)	Unclassified Reads (Host) %
1	MS_zymo_1	35,696,570	56,644	0.16 %	35,639,926	99.84 %
2	MS_zymo_2	35,598,386	59,296	0.17 %	35,539,090	99.83 %
3	MS_zymo_3	31,610,568	53,894	0.17 %	31,556,674	99.83 %

- Table 3: Number of reads assigned to different host species per sample:

ID	Sample Name	Homo_sapiens
1	MS_zymo_1	56644 (0.16 %)
2	MS_zymo_2	59296 (0.17 %)
3	MS_zymo_3	53894 (0.17 %)

1.4 Taxonomic Profiling

Taxonomic profiling is done using MetaPhlAn[6]. MetaPhlAn (Metagenomic Phylogenetic Analysis) is a computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species level resolution. MetaPhlAn relies on unique clade-specific marker genes identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic). Unclassifed reads are than subjected to KrakenUniq[1]. Kraken[7] classifes reads by breaking each read into overlapping k-mers. Each k-mer is mapped to the lowest common ancestor (LCA) of the genomes containing that k-mer in a precomputed reference database. For each read, a classifcation tree is found by pruning the taxonomy and only retaining taxa (including ancestors) associated with k-mers in that read. Each node is weighted by the number of k-mers mapped to the node, and the path from root to leaf with the highest sum of weights is used to classify the read. KrakenUniq computes the number of unique k-mers observed for each taxon, which allows to filter more false positives. The final classifed, unclassifed and filter passed reads are reported in the table below.

- Table 4: Taxonomic profling metrics per sample:

ID	Sample Name	Reads	Classified Reads	Classified Reads %	Unclassified Reads	Unclassified Reads %
1	MS_zymo_1	35,639,926	30,304,221	85.03 %	5,335,705	14.97 %
2	MS_zymo_2	35,539,090	30,145,848	84.82 %	5,393,242	15.18 %
3	MS_zymo_3	31,556,674	26,812,721	84.97 %	4,743,953	15.03 %

- Table 5: Number of reads assigned to different kingdoms per sample,

- Ambigious: Reads which can not be assigned to one specific kingdom.
- Eukaryota: Parasitic and non-parasitic Protozoa.

ID	Sample Name	Archaea	Bacteria	Eukaryota	Fungi	Viruses	Ambiguous
1	MS_zymo_1	66 (0.00 %)	28,860,121 (95.23 %)	8,353 (0.03 %)	1,115,040 (3.68 %)	236,263 (0.78 %)	84,378 (0.28 %)
2	MS_zymo_2	62 (0.00 %)	28,700,611 (95.21 %)	8,429 (0.03 %)	1,119,298 (3.71 %)	233,394 (0.77 %)	84,054 (0.28 %)
3	MS_zymo_3	52 (0.00 %)	25,538,545 (95.25 %)	7,360 (0.03 %)	986,208 (3.68 %)	206,576 (0.77 %)	73,980 (0.28 %)

1.4.1 Taxa Abundance

Read counts of input samples observed at various taxa levels (Phylum, Genus, and Species) are collected and normalized by using the rarefy function implemented in the vegan bioconductor package[3] to compare species richness from all samples in the analysis run. Rarefied read counts enable better comparisons of OTU profiles between samples with different sample sizes. Abundance measured by the percentage of OTU assigned reads from various taxonomic levels is determined and are used to generate heatmaps and bar plots at Phylum, Genus and Species levels. Heatmap and bar plots representing the taxonomic abundance at species level are provided below.

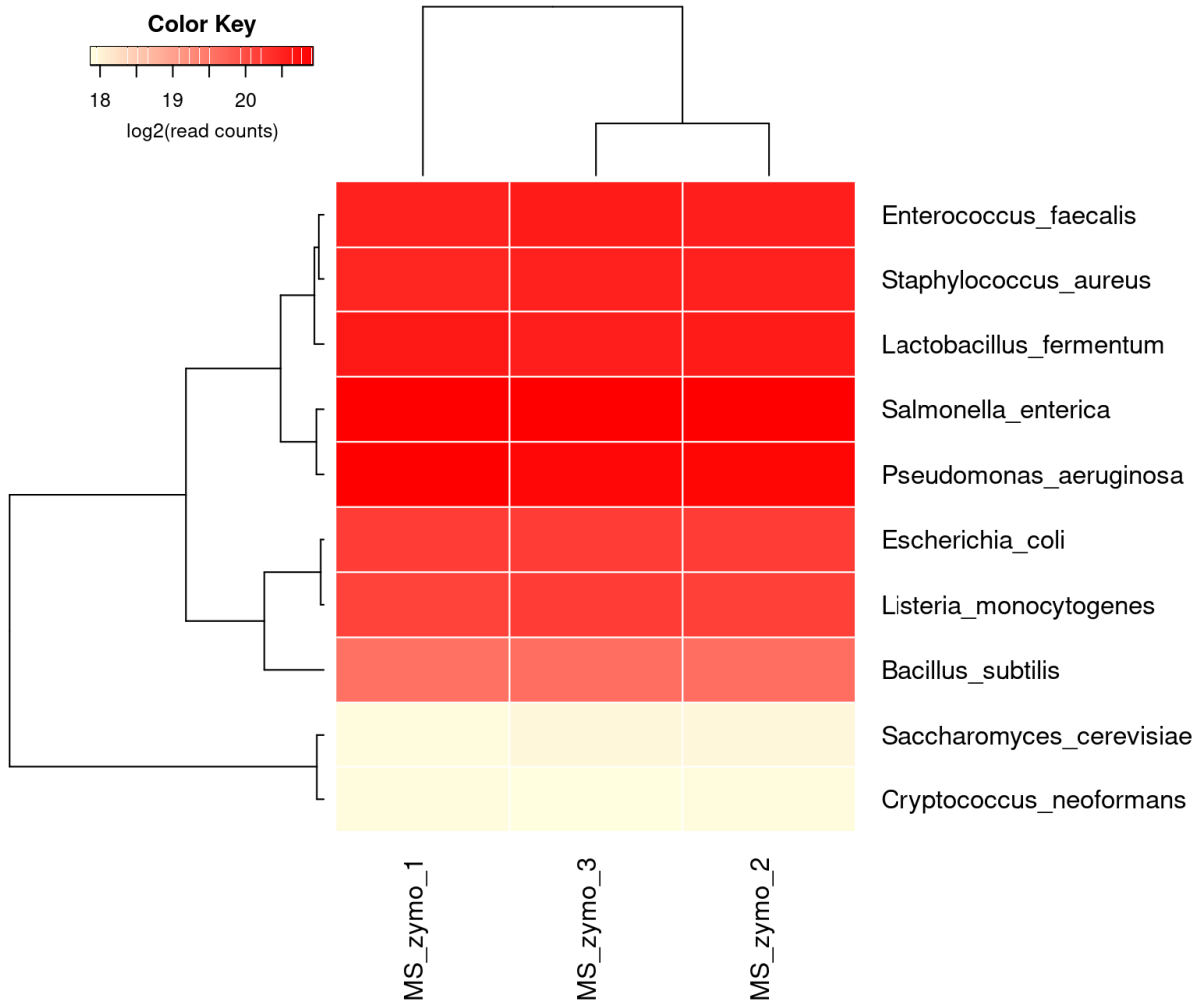


Figure 1: Heat map(s) showing the taxonomic abundance and their relation across the samples. Dendrograms determined by computing hierarchical clustering from the abundance levels shows the relationship between the species (left) and the samples (top). The abundance levels (number of reads associated with each taxa) are logarithmically transformed to base 2 for clarity. Taxa-level: Species

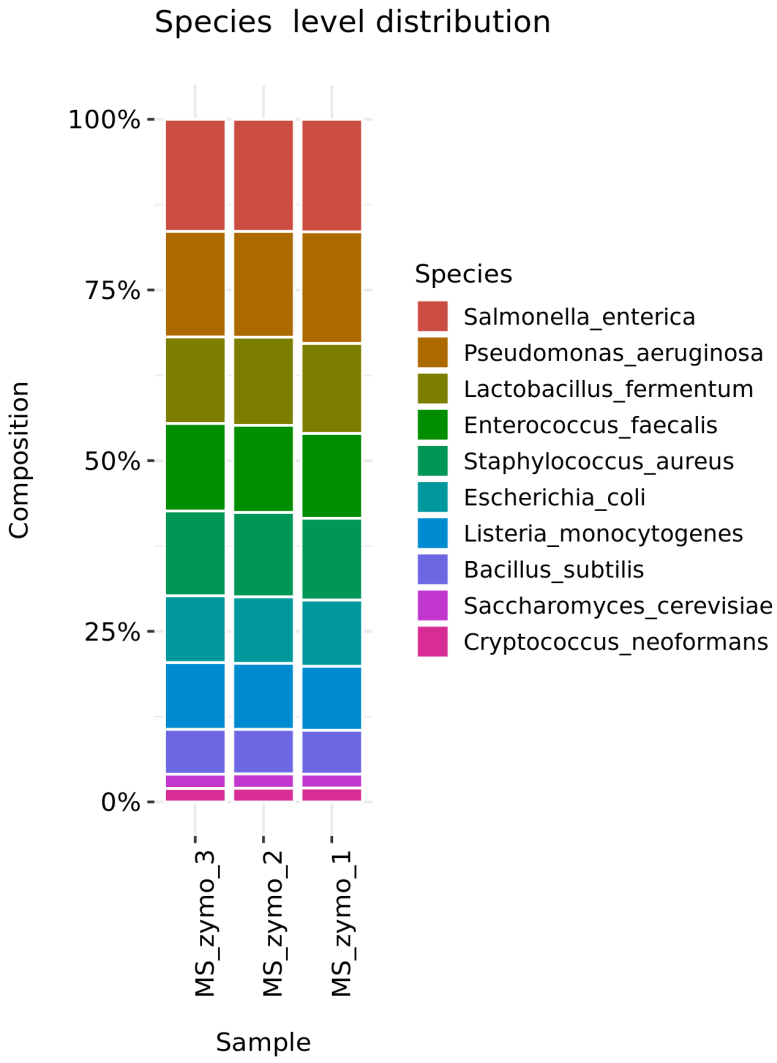


Figure 2: Bar plot(s) showing the taxonomic abundance across the samples. Taxa-level: Species

1.4.2 Species Diversity

A diversity index is a quantitative measure that reflects how many different types (such as species) are in a dataset, and simultaneously takes into account how evenly the basic entities (such as individuals) are distributed among those types. The value of a diversity index increases both when the number of species increases and when all species are present at nearly the same level. For a given number of species, the value of a diversity index is maximized when all species are equally abundant. The following diversity indices are computed using vegan[3] package in R. Simpson refers to Simpson diversity index and has values ranging from 0 to 1. Values near 1 are simple environments and smaller values are diverse environments. InvSimpson refers to inverse Simpson diversity and has values >0. A larger value means greater diversity. Shannon refers to Shannon diversity index and has values >0. A higher value means greater diversity. Alpha refers to Fischer's model of predicting species richness by computing alpha diversity and has values >0. A larger value means greater diversity. Evenness refers to the distribution of individuals across species and is determined by Pielou's measure of species evenness. The index tends to 0 as the evenness decreases in simple environments (species-poor communities). SpeciesNo refers to the absolute number of species found in each sample.

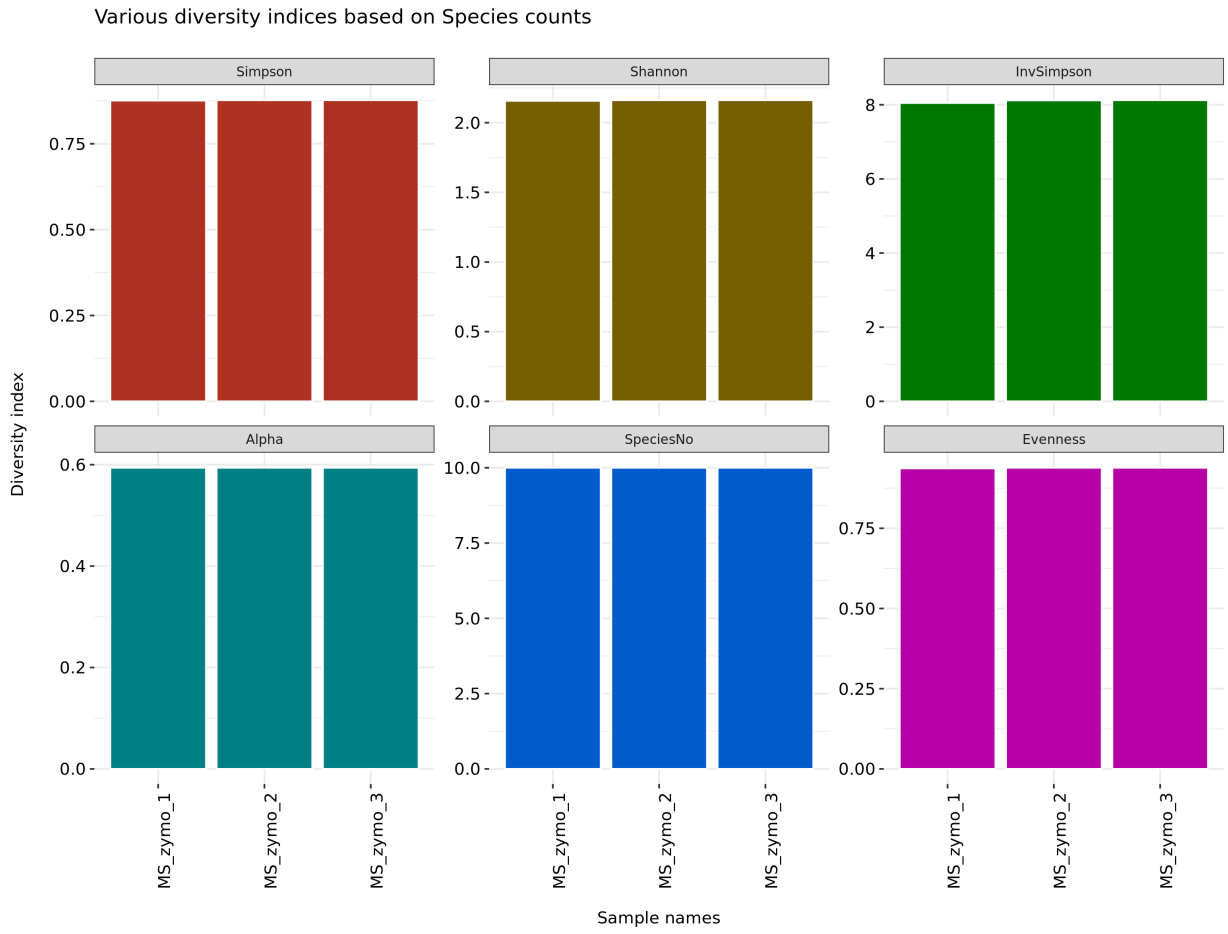


Figure 3: Various diversity indices computed based on the species counts found in each sample

1.4.3 Rarefaction Curves

Rarefaction allows the calculation of species richness for a given number of individual samples, based on the construction of rarefaction curves. This curve is a plot of the total number of distinct species found as a function of the number of sequences sampled. Sampling curves generally rise very quickly at first and then level off towards an asymptote as fewer new species are found in each sample. These rarefaction curves are calculated from the table of species abundance. The curves represent the average number of different species found for subsamples of the complete dataset.

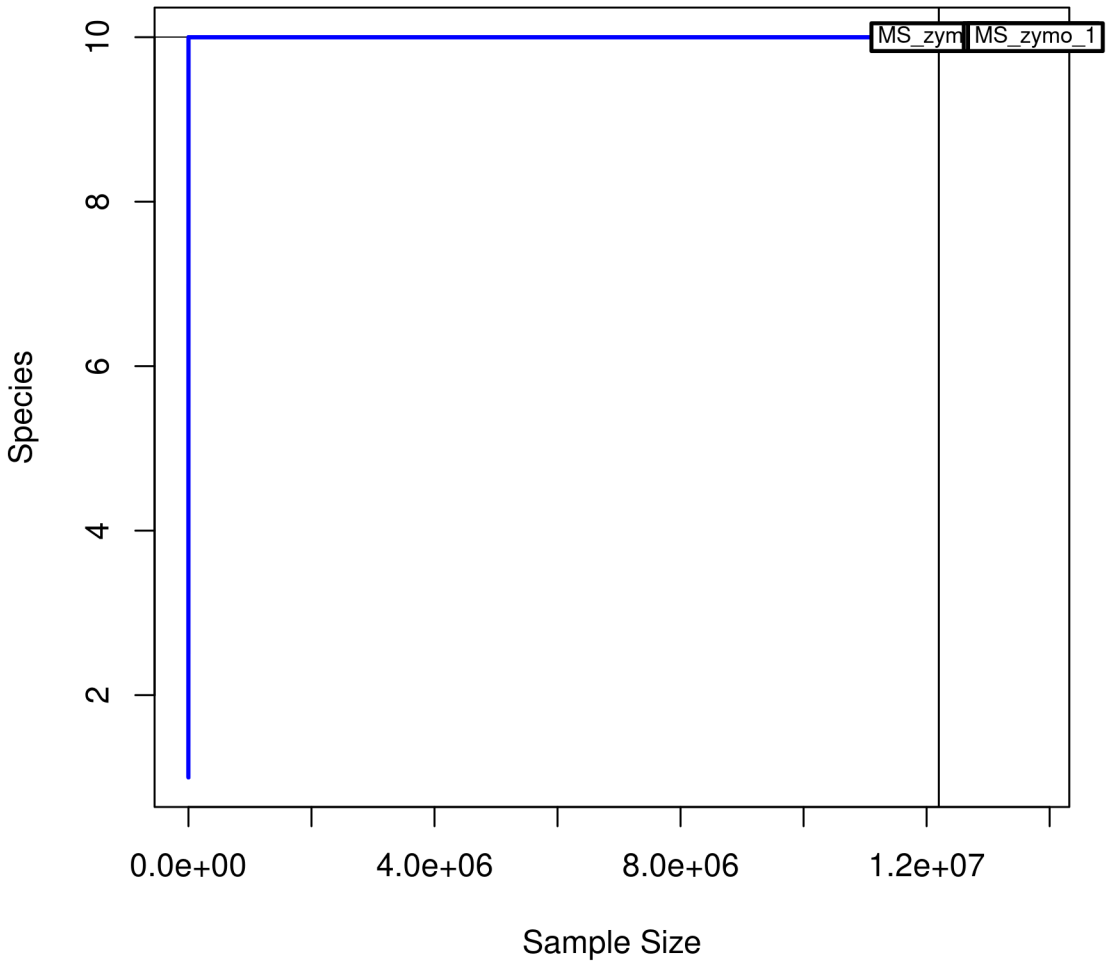


Figure 4: Rarefaction curve of annotated species richness

1.5 Resistome Profiling

Profiling the collective antimicrobial resistance (AMR) within a metagenome is referred as resistome, which facilitates greater understanding of AMR gene diversity and dynamics in metagenomic environments.

Antimicrobial resistant genes (ARGs) from the metagenomic samples are screened using Graphing Resistance Out Of meTagenomes (GROOT[4]) software, which combines a variation graph representation of gene sets with a locality-sensitive hashing forest indexing scheme to allow for fast classification of metagenomic sequence reads using similarity-search queries. Subsequent hierarchical local alignment of classified reads against graph traversals enables accurate reconstruction of full-length gene sequences using a scoring scheme.

Reference ARG database contains >6000 well curated ARGs sourced from the public repositories.

1.5.1 Resistome Profile Results

Antimicrobial resistant genes and the associated antibiotic classes detected in each metagenomic sample are summarized in the following table.

- Table 6: ARG profiling metrics per sample:

ANTIBIOTIC	ANTIBIOTIC GENES	MS_zymo_1	MS_zymo_2	MS_zymo_3
acriflavine	nan	Ye	Ye	Yes
beta-lactam	blaZ, blaPDC, ampC, blaI, bla, BlaZ, mecA	Yes	Yes	Yes
fosfomycin	fosX	Ye	Ye	Yes
multidrug	mexR, mexC, mdfA, ebrA, emrD, emrY, mdtK, mexG, emrB, acrF, mdtO, yjiO, emrK, mdtH, mdtG, mdtA, mdtN, oprN, acrD	Yes	Yes	Yes
oxacillin	blaOXA	Ye	Ye	Yes
penicillin	PbP4B	Yes	Yes	Yes
quinolone	QnrB	Ye	Ye	Yes

1.5.2 Supplementary Information : Detected Antiobiotics Appendix

- Table 7: Resistome profiling appendix:

ANTIBIOTIC	Description
acriflavine	Acriflavine is a topical antiseptic. It has the form of an orange or brown powder. It may be harmful in the eyes or if inhaled. Acriflavine is also used as treatment for external fungal infections of aquarium fish.
beta-lactam	ACT beta-lactamases, also known as AmpC beta-lactamases, are cephalosporinases that cannot be inhibited by clavulanate. These enzymes are encoded by genes located on the chromosome and can be induced by the presence of beta-lactam antibiotics.
fosfomycin	Fosfomycin (also known as phosphomycin and phosphonomycin) is a broad-spectrum produced by certain Streptomyces species. It is effective on gram positive and negative bacteria as it targets the cell wall, an essential feature shared by both bacteria. Its specific target is MurA (MurZ in E.coli), which attaches phosphoenolpyruvate (PEP) to UDP-N-acetylglucosamine, a step of commitment to cell wall synthesis. In the active site of MurA, the active cysteine molecule is alkylated which stops the catalytic reaction.
multidrug	Antibiotics reistant to a broad range of drugs
oxacillin	Oxacillin is a penicillinase-resistant beta-lactam. It is similar to methicillin, and has replaced methicillin in clinical use. Oxacillin, especially in combination with other s, is effective against many penicillinase-producing strains of Staphylococcus aureus and Staphylococcus epidermidis.
penicillin	Penicillin (sometimes abbreviated PCN) is a beta-lactam used in the treatment of bacterial infections caused by susceptible, usually Gram-positive, organisms. It works by inhibiting the synthesis of the peptidoglycan layer of bacterial cell walls. The peptidoglycan layer is important for cell wall structural integrity, especially in Gram-positive organisms.
quinolone	A quinolone antibiotic is a member of a large group of broad-spectrum bacteriocidals that share a bicyclic core structure related to the substance 4-quinolone. They are used in human and veterinary medicine to treat bacterial infections, as well as in animal husbandry

1.6 Relevant Programs

- Table 8: The programs/softwares used in this pipeline.

Tool	Version	Description
fastp	0.20.0	Fastp is a tool designed to provide fast all-in-one preprocessing for FastQ files.
GROOT	1.1.2	Indexed variation graphs for efficient and accurate resistome profiling
KrakenUniq	0.5.8	KrakenUniq: confident and fast metagenomics classification using unique k-mer counts
KronaTools	2.7.1	Interactive metagenomic visualization in a Web browser
MetaPhlAn	3.0.7	MetaPhlAn for enhanced metagenomic taxonomic profiling
R	4.1.3	R is a programming language and environment for statistical computing.
SeqKit	0.12.0	SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation
vegan	2.6.4	The functions in the vegan package contain tools for diversity analysis, ordination methods and tools for the analysis of dissimilarities

1.7 Sequence Data Used

- Table 9: The samples used in this pipeline.

No	Sample	File Name
1	MS_zymo_1	IC-2490_MS_zymo_1_lib607411_8151_3_1.fastq.gz
		IC-2490_MS_zymo_1_lib607411_8151_3_2.fastq.gz
2	MS_zymo_2	IC-2490_MS_zymo_2_lib607412_8151_3_1.fastq.gz
		IC-2490_MS_zymo_2_lib607412_8151_3_2.fastq.gz
3	MS_zymo_3	IC-2490_MS_zymo_3_lib607413_8151_3_1.fastq.gz
		IC-2490_MS_zymo_3_lib607413_8151_3_2.fastq.gz

1.8 Filter Settings

- Table 10: Filters used in postprocessing of taxonomic profiling results.

Filter	Value
Top OTUs to include in plots	20.00
Minimum read count proportion	0.01

1.9 Deliverables

- Table 11: List of delivered files, format and recommended programs to access the data.

File	Format	Program To Open File
<PROJECT-ID>.Metagenome_Analysis_Report.html	html	Web Browser
<PROJECT-ID>.kingdom_assignment.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.taxonomic_profile_metrics.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.krona_plot.html	html	Web Browser
<PROJECT-ID>.Genus.barplot.png	png	Image Viewer
<PROJECT-ID>.Species.barplot.png	png	Image Viewer
<PROJECT-ID>.Phylum.barplot.png	png	Image Viewer
<PROJECT-ID>.Genus.diversity_indices.png	png	Image Viewer
<PROJECT-ID>.Species.diversity_indices.png	png	Image Viewer
<PROJECT-ID>.Species.rarefaction_heatmap.log2scale.png	png	Image Viewer
<PROJECT-ID>.Genus.rarefaction_heatmap.log2scale.png	png	Image Viewer
<PROJECT-ID>.Phylum.rarefaction_heatmap.log2scale.png	png	Image Viewer
<PROJECT-ID>.Phylum.composition.proportion.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Species.composition.proportion.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Genus.composition.proportion.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Species.composition.reads.normalized.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Phylum.composition.reads.normalized.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Genus.composition.reads.normalized.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Phylum.composition.reads.raw.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Genus.composition.reads.raw.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Species.composition.reads.raw.tsv	tsv	Spreadsheet Editor
<SAMPLE-ID>.taxonomic_profile.tsv	tsv	Spreadsheet Editor
<SAMPLE-ID>.krona.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.host_profile_metrics.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.host_species_assignment.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>_<SAMPLE-ID>_<LIB-ID>_non_host_reads_1.fastq.gz	fastq.gz	tar

(continued)

File	Format	Program To Open File
<PROJECT-ID>_<SAMPLE-ID>_<LIB-ID>_non_host_reads_2.fastq.gz	fastq.gz	tar
<PROJECT-ID>.ARG_profile_metrics.tsv	tsv	Spreadsheet Editor
<PROJECT-ID>.Detected_Antibiotics_Supplementary_File.tsv	tsv	Spreadsheet Editor

Note: All the deliverables have been compressed and are available as tar.gz file with the file name EF-DEMO.Metagenome_Analysis_Results.tar.gz

1.10 References

[1] Florian P Breitwieser, DN Baker, and Steven L Salzberg. 2018. KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. *Genome biology* 19, 1 (2018), 1–10.

[2] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, 17 (September 2018), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

[3] Jari Oksanen, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, RB Ohara, Gavin L Simpson, Peter Solymos, M Henry H Stevens, Helene Wagner, and others. 2013. Package vegan. *Community ecology package, version* 2, 9 (2013), 1–295.

[4] Will PM Rowe and Martyn D Winn. 2018. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics* 34, 21 (2018), 3601–3608.

[5] Wei Shen, Shuai Le, Yan Li, and Fuquan Hu. 2016. SeqKit: A cross-platform and ultrafast toolkit for fasta/q file manipulation. *PloS one* 11, 10 (2016), e0163962.

[6] Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. MetaPhlAn for enhanced metagenomic taxonomic profiling. *Nature methods* 12, 10 (2015), 902–903.

[7] Derrick E Wood and Steven L Salzberg. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15, 3 (2014), 1–12.