Major Project Presentation

on

“Insilico Metagenomics Analysis(18s)using bash tools”

As

Junior Bioinformatics Associate

At

EdGene BioMed pvt.ltd., Gururgram, Haryana

Submitted by -

Yashika Sharma
Btech Biotechnology (2021-25)
2117000026

Submitted To-

Dr. Anuja Mishra
Assistant Professor
Dept. of Biotechnology
GLA University

Introduction

Metagenomics- Metagenomics is the study of genetic material recovered directly from environmental samples without the need for culturing organisms. It uses sequencing technologies to analyze the diversity, structure, and function of microbial communities. This approach helps identify novel microbes, genes, and metabolic pathways in diverse environments like soil, oceans, and the human gut_

What is 18s rRNA ?

- 18S rRNA is a component of the small eukaryotic ribosomal subunit.
- It contains both conserved and variable regions, allowing for the identification of eukaryotic organisms like protozoa, fungi and algae.

: Comparison of 16S and 18S rRNA sequencing, highlighting target organisms, primer regions, and applications.

16S RRNA	18S RRNA
The RNA component of the 30S subunit of a prokaryotic ribosome	The RNA component of the eukaryotic ribosomal small subunit
Occurs in the small subunit of the prokaryotic ribosome	Occurs in the small subunit of the eukaryotic ribosome
Forms 30S subunit	Forms 40S rRNA subunit
Binds to the Shine-Dalgarno sequence	Important in phylogenetic analysis

Sample Collection & Preparation

Sample Source & Type:

- Environmental samples were retrieved from the NCBI SRA (Sequence Read Archive).

Sample Nature:

- The dataset comprises single-end of paired-end FASTQ files, representing eukaryotic microbial communities from biological sources.

DNA Extraction & Amplification (Pre-sequencing):

- DNA was extracted from the original biological material (e.g., feces, water, or soil)
- The 18S rRNA gene was PCR-amplified using conserved primers to target eukaryotic sequences

Manifest file

File Edit View

sample-id	forward-absolute-filepath	reverse-absolute-filepath
SRR32838797	/mnt/y/training/work/SRR32838797_1.fastq	/mnt/y/training/work/SRR32838797_2.fastq
SRR32838801	/mnt/y/training/work/SRR32838801_1.fastq	/mnt/y/training/work/SRR32838801_2.fastq
SRR32838806	/mnt/y/training/work/SRR32838806_1.fastq	/mnt/y/training/work/SRR32838806_2.fastq
SRR32838811	/mnt/y/training/work/SRR32838811_1.fastq	/mnt/y/training/work/SRR32838811_2.fastq
SRR32838816	/mnt/y/training/work/SRR32838816_1.fastq	/mnt/y/training/work/SRR32838816_2.fastq
SRR32838821	/mnt/y/training/work/SRR32838821_1.fastq	/mnt/y/training/work/SRR32838821_2.fastq
SRR32838826	/mnt/y/training/work/SRR32838826_1.fastq	/mnt/y/training/work/SRR32838826_2.fastq
SRR32838831	/mnt/y/training/work/SRR32838831_1.fastq	/mnt/y/training/work/SRR32838831_2.fastq
SRR32838836	/mnt/y/training/work/SRR32838836_1.fastq	/mnt/y/training/work/SRR32838836_2.fastq
SRR32838841	/mnt/y/training/work/SRR32838841_1.fastq	/mnt/y/training/work/SRR32838841_2.fastq
SRR32838846	/mnt/y/training/work/SRR32838846_1.fastq	/mnt/y/training/work/SRR32838846_2.fastq
SRR32838850	/mnt/y/training/work/SRR32838850_1.fastq	/mnt/y/training/work/SRR32838850_2.fastq
SRR32838818	/mnt/y/training/work/SRR32838818_1.fastq	/mnt/y/training/work/SRR32838818_2.fastq
SRR32838833	/mnt/y/training/work/SRR32838833_1.fastq	/mnt/y/training/work/SRR32838833_2.fastq
SRR32838828	/mnt/y/training/work/SRR32838828_1.fastq	/mnt/y/training/work/SRR32838828_2.fastq

Introduction to the Galaxy Platform

- Galaxy is an open-source, web-based platform designed to make powerful bioinformatics tools accessible to everyone — even if you don't know programming or coding.
- It allows researchers to build and run complete biological data analysis workflows through a simple point-and-click interface.

Feature	Description
Web-Based Access	No installation needed; accessible via browser.
Reproducible	Every analysis is stored as **history**, making it easy to track and re-run
Tool Integration	Includes hundreds of built-in tools (e.g., FastQC, Cutadapt, Kraken2)
Workflow Automation	You can build and save complete **automated workflows**
Secure	Your data is private and stored in your account

Key Features of Galaxy:

Ideal for amplicon-based metagenomic analysis
Easily supports 18S rRNA workflows
Drag-and-drop interface suits both beginners and experts
No need to install tools like FastQC, DADA2, Kraken — all are preloaded!

WORKFLOW

SRA (Data Download from NCBI)

FastQC(Quality Check)

Cutadapt (Adapter & Quality Trimming)

DADA2 / VSearch (Denoising or Clustering)

SILVA

Diversity Analysis (Alpha & Beta Diversity)

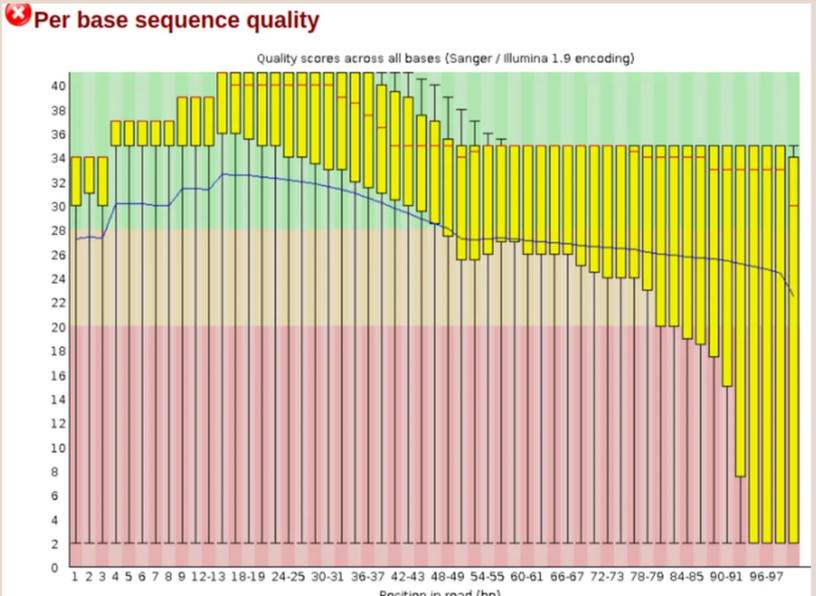
Quality Control (FastQC & MultiQC)

Purpose of Quality Control

Before analyzing sequencing data, it's essential to ensure that the raw reads are of high quality. Poor quality data can lead to incorrect taxonomic assignments and false diversity estimations.

Key Quality Metrics Checked:

- Per-base sequence quality: Ensures most reads have high Phred scores (≥ Q30)
- GC content: Checks for abnormal GC% which might suggest contamination
- Adapter Content: Identifies leftover sequencing adapters
- Per-sequence quality scores: Looks for low-quality reads



Trimming

Cutadapt is a widely used tool that removes unwanted sequences from high-throughput sequencing data, such as:

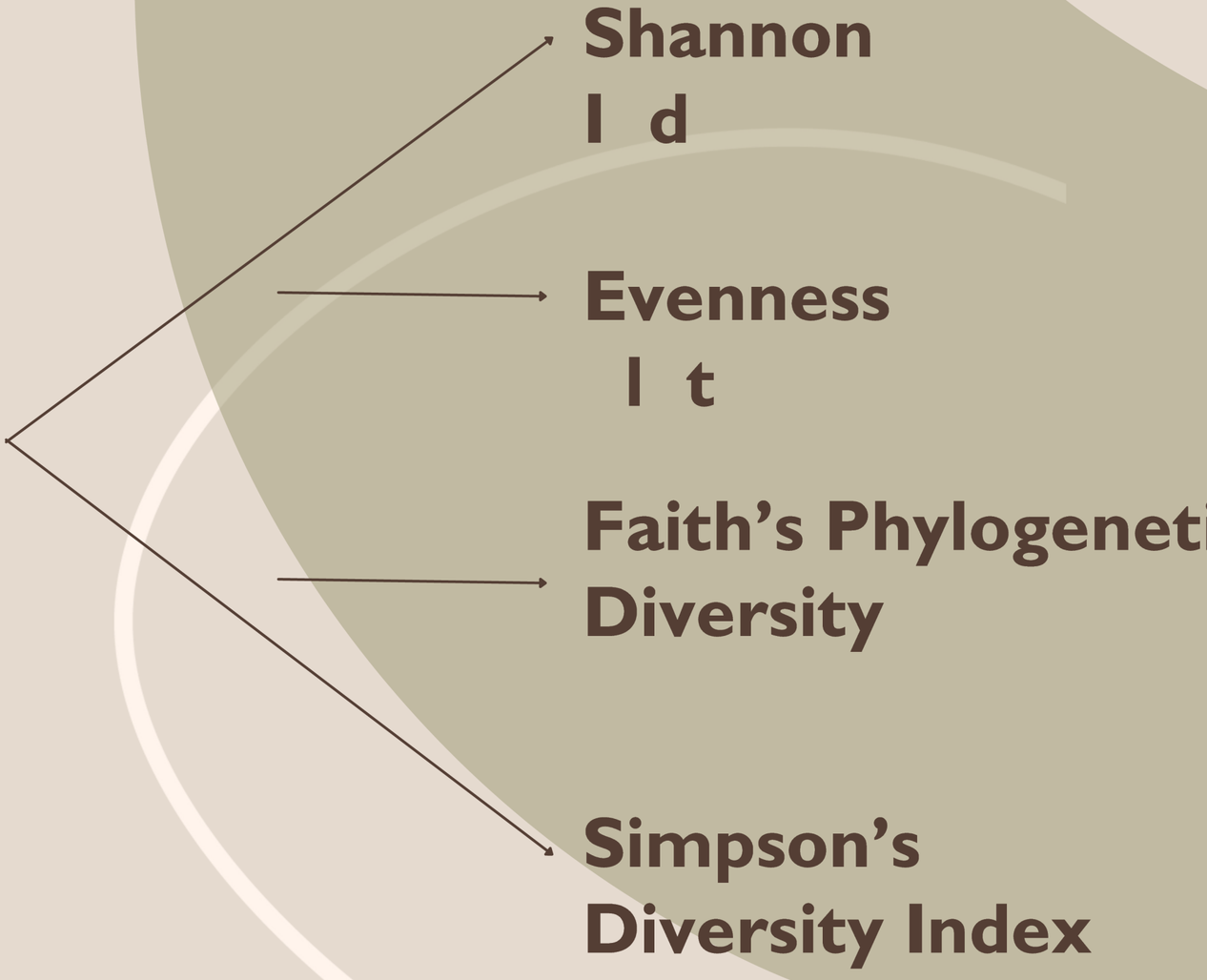
- Adapters
- Primers
- Low-quality bases (from 3' and 5' ends)

This cleaning step is crucial before clustering or taxonomy, as it reduces errors, bias, and improves data accuracy.

Task	**Output**
Adapter Removal	Eliminates non-biological sequences
Quality Trimming	Removes low-Q score bases (usually Q<20)
Length Filtering	Discards too-short reads after trimming
Output File	Cleaned .fastq.gz files ready for clustering

Diversity Analysis

Alpha - Diversity



Shannon Diversity index

It considers both **species richness** and **species evenness.**

Formula -
$H' = -\sum(pi \cdot lnpi)$

Interpretation:

- **Higher Shannon Index**: Greater diversity (more species and/or more even distribution).
- **Lower Shannon Index**: Lower diversity (fewer species or dominance by a few species).
- If $H' = 0$H' = 0$H' = 0$, the sample has only one species.

QIIME 2

qiime2view

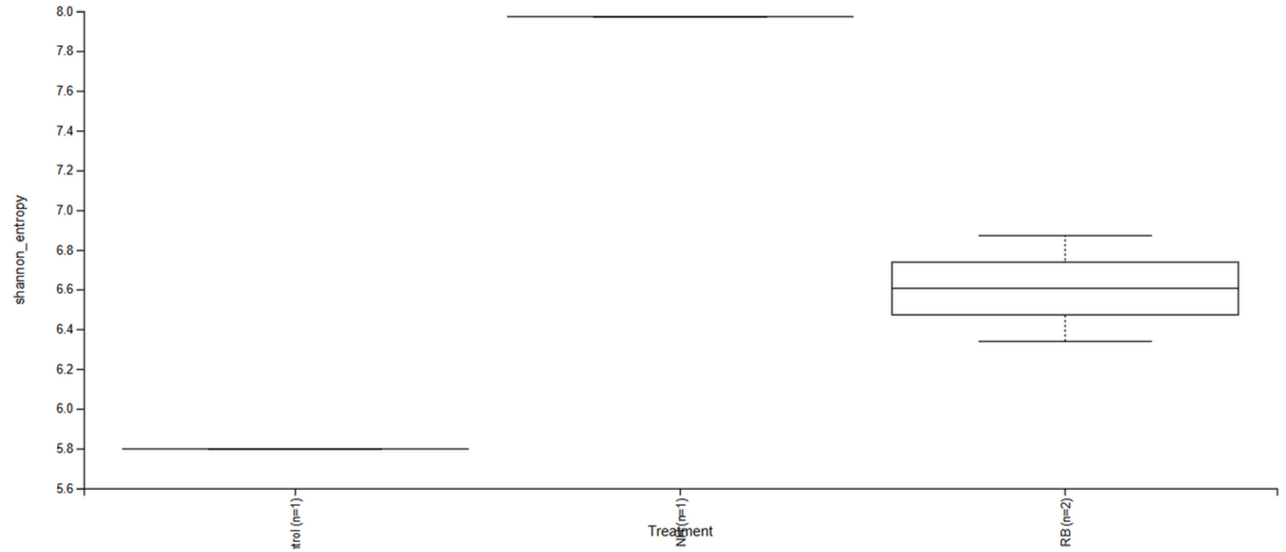
File: **shannon_group.qzv** ×

Alpha Diversity Boxplots

Column

Download SVG

Treatment



Download raw data as TSV

Kruskal-Wallis (all groups)

	Result
H	2.6999999999999993
p-value	0.2592402606458915

Kruskal-Wallis (pairwise)

Download CSV

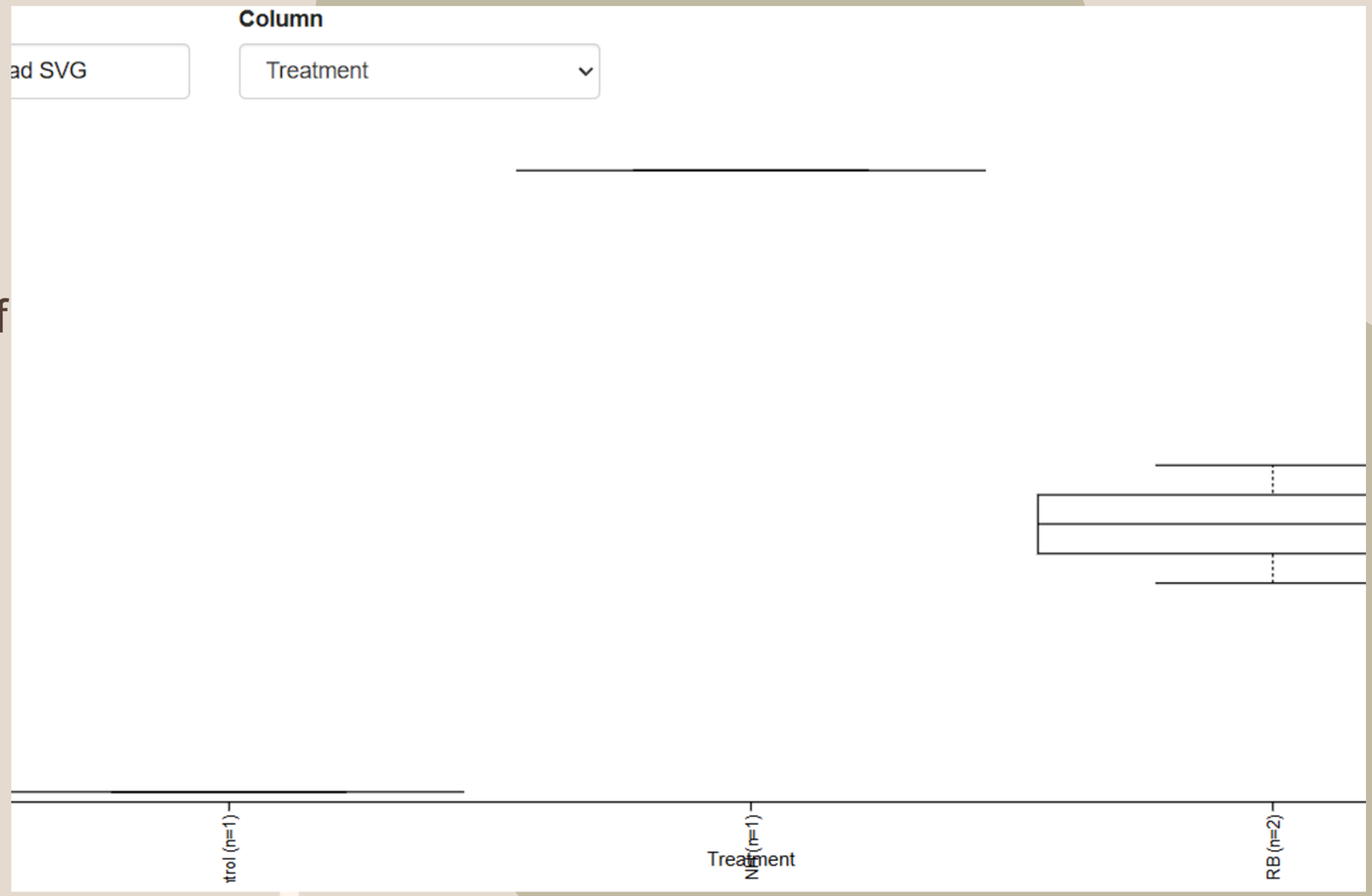
		H	p-value	q-value
Group 1	**Group 2**			
Control (n=1)	**NH (n=1)**	1.0	0.317311	0.317311
	RB (n=2)	1.5	0.220671	0.317311
NH (n=1)	**RB (n=2)**	1.5	0.220671	0.317311

Evenness Plot

- **Evenness** refers to how similar the abundances of different species are within a community.
- **Richness** is the total number of species present.
- **Diversity Indices** quantify both richness and evenness.

Interpreting an Evenness Plot:

- A higher evenness value (closer to 1) means the species are more evenly distributed.
- A lower evenness value (closer to 0) suggests dominance by a few species, with others being rare.
- The plot may display evenness across different samples or conditions, helping to compare microbial community structures.



Simpson's Diversity Index

Simpson's Diversity Index (D) is a metric used to measure **alpha diversity** by considering both:

1. **Species Richness**
2. **Species**

Formula

$$D = \sum \left(\frac{n_i(n_i - 1)}{N(N - 1)} \right)$$

Faith's Phylogenetic Diversity

- Faith's PD measures **alpha diversity** by summing the total **branch lengths** in a phylogenetic tree spanning all observed taxa in a sample.
- Unlike species richness, it considers **evolutionary relationships**, making it more informative for microbial ecology studies.
- **Higher PD** indicates a more evolutionarily diverse community, while **lower PD** suggests closely related species dominate.

Beta - Diversity

Taxonomic-Based Beta Diversity Metrics

Phylogenetic-Based Beta Diversity Metrics

Taxonomic-Based Beta Diversity Metrics

Bray-Curtis Dissimilarity

- Measures differences based on **species abundance**.

- Formula: $$BC = 1 - \frac{2C}{S_1 + S_2}$$

Jaccard Index

- Measures differences based on **presence/absence of species**.

- Formula: $$J = \frac{A}{A + B + C}$$

Phylogenetic-Based Beta Diversity Metrics

UniFrac (Unweighted & Weighted)

- Measures differences using **phylogenetic distances**.
- **Unweighted UniFrac** – Only considers **presence/absence** of taxa
- **Weighted UniFrac** – Accounts for **abundance** of taxa.
- **More robust for microbial ecology studies.**

Beta Diversity Visualization & Analysis

(A) Principal Coordinates Analysis (PCoA)

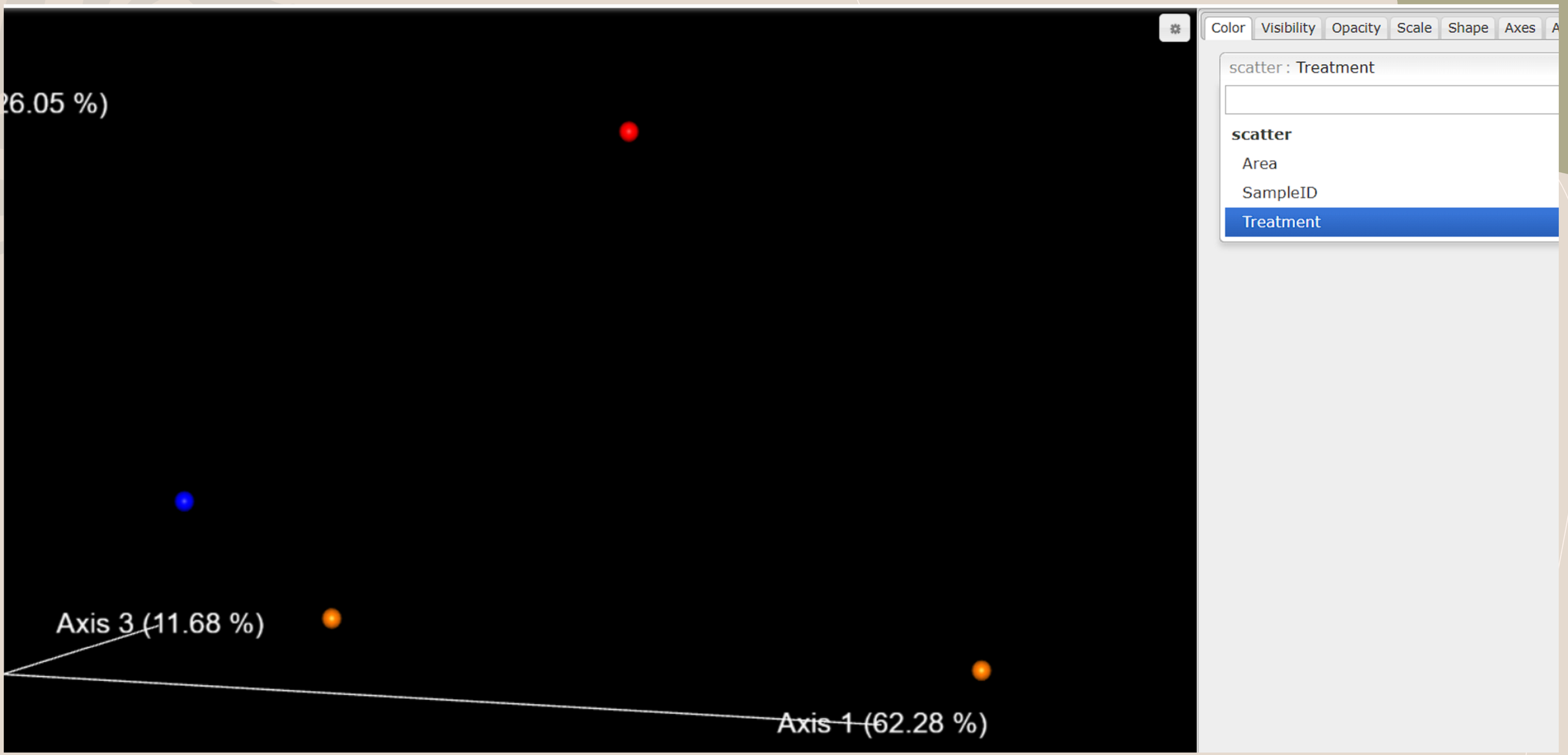
A dimensionality reduction technique to **visualize beta diversity distances** in 2D/3D plots.
Clusters indicate similarity; spread shows variation.
Commonly used with **Bray-Curtis & UniFrac**.

(B) Non-Metric Multidimensional Scaling (NMDS)

Similar to PCoA but **uses rank-based distances** for better visualization.
Preferred when **data are highly non-linear**.

(C) PERMANOVA (Permutational Multivariate Analysis of Variance)

Statistical test to check if groups differ significantly in beta diversity.
Used with distance matrices (e.g., Bray-Curtis, UniFrac).
Reported as **p-values & R^2 values** (higher R^2 means stronger group separation).


26.05 %)
Axis 3 (11.68 %)
Axis 1 (62.28 %)
Color
Visibility
Opacity
Scale
Shape
Axes
scatter : Treatment
scatter
Area
SampleID
Treatment

Applications of metagenomics

- Studied gut, skin, and oral microbiomes to understand health and disease.
- Detects pathogens and antibiotic resistance genes without culture.
- Analyze microbial communities for pollution control and climate studies.
- Enhances crop growth, soil fertility, and disease resistance using microbes
- Discovers enzymes, biofuels, and biomolecules for sustainable industries.
- Explores microbial evolution, symbiosis, and horizontal gene transfer.

Limitations of 18S rRNA-Based Metagenomics Analysis

- 18S rRNA gene sequences are highly conserved in eukaryotes.
- This often restricts classification to genus level only, making species-level identification difficult.
- Closely related species (e.g., fungal strains) may appear indistinguishable.

- Taxonomic assignment tools like Kraken2 rely on reference databases such as SILVA.
- Incomplete or outdated databases lead to incorrect or “unclassified” results.
- New or rare species might not be in the database at all.

THANK YOU