Compositional Data Analysis (CoDA) for Microbiome Studies

Kipoong Kim

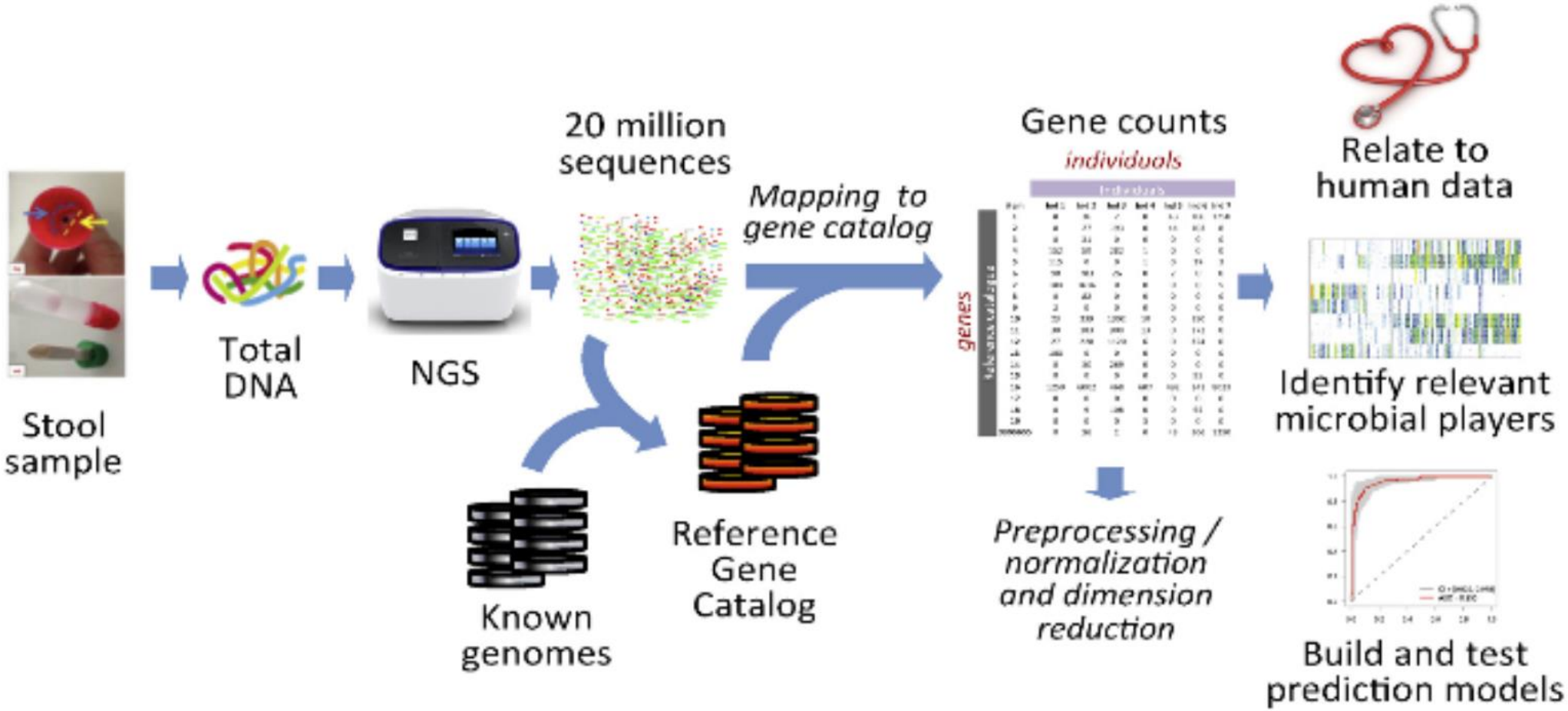
Department of Statistics, Changwon National University

January 2026

Table of Contents

NGS Technologies

- Samples → DNA extraction → PCR+Library prep. → Sequencing & Mapping → Microbiome count data



Library size

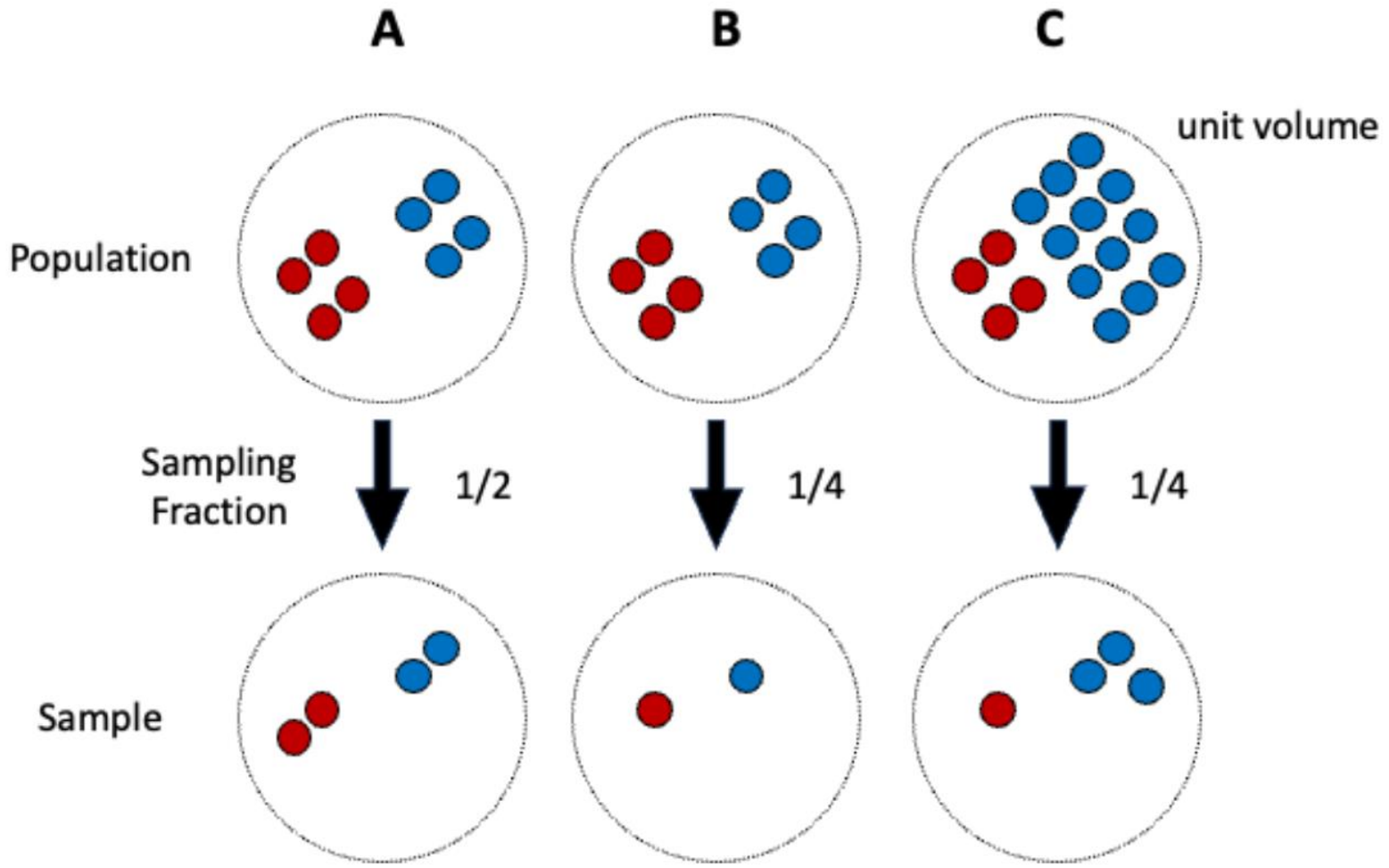


Figure: A vs B: different library size; A vs C: different sampling fraction

The Nature of Microbiome Data

- **Key Concepts:**
  - **Compositionality:** The total read count (Library Size) is arbitrary.
    - An increase in Taxon A mathematically forces a decrease in Taxon B (Spurious Correlation).
  - **Sparsity:** The data matrix is full of zeros (60-90%). Are they true absence or sampling failure?
- **Solution:** We move from the Simplex space to Euclidean space using **CLR (Centered Log-Ratio) Transformation**.

Zero replacement strategies

- Zero replacement
  - **Substitution:** Replace zeros with a small constant (e.g., pseudo-count)

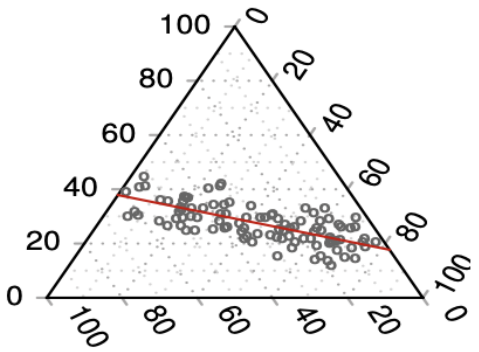
    Closure: Renormalize vector to maintain the sum constraint (compositional)

- Substitution:
  - Count-level data: 1/2
  - Compositional-level data: 0.5 x {minimum}

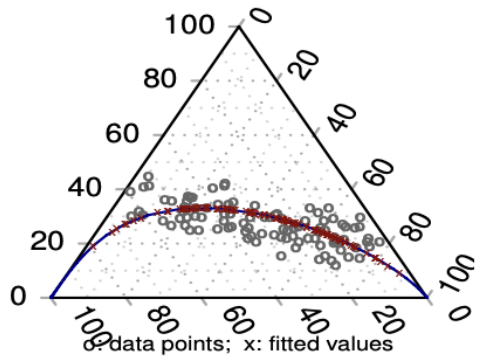
Zero replacement

Without zeros

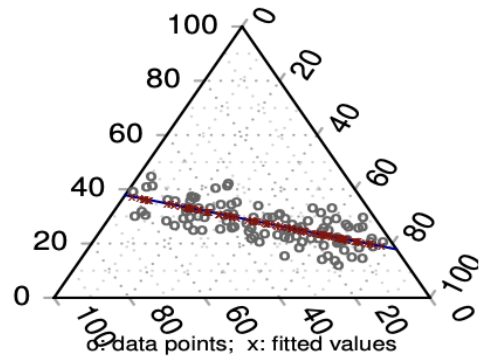
Data generation



lrPCA with $\delta_{(0)}$

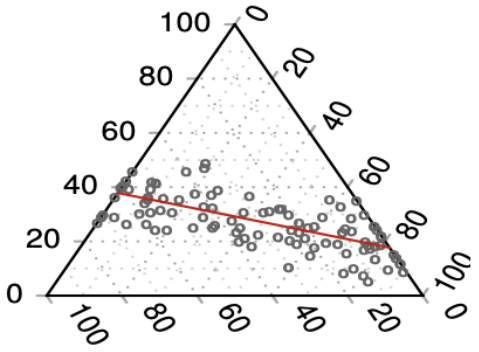


PCA

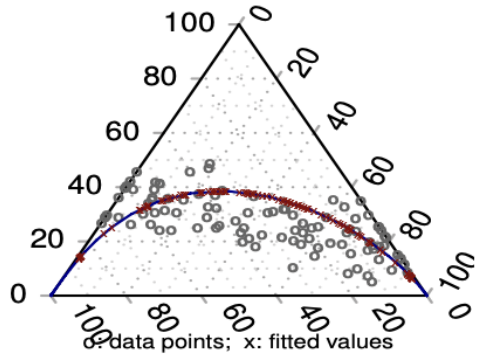


With 10% zeros

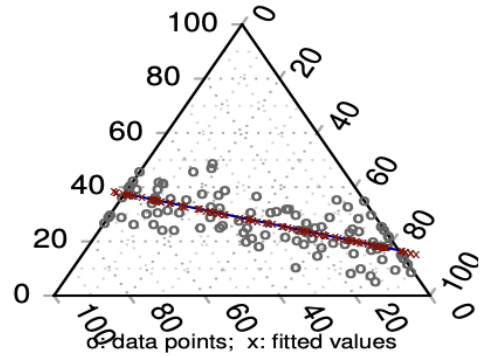
Data generation



lrPCA with $\delta_{(0)}$



PCA



$$\delta_{(0)} = \min\{x_{ij} \in \mathbf{X} : x_{ij} > 0\}$$

Analysis Pipeline: (1) Upstream Processing

- **QC & Trimming:** Quality filtering of raw FASTQ files, removing primers/adapters (FastQC, Trimmomatic).
- **Denoising & Clustering:**
  - **ASV (Amplicon Sequence Variant):** Uses error-correction models to resolve single-nucleotide differences (DADA2, Deblur).
  - **OTU (Operational Taxonomic Unit):** Clustering based on 97% similarity (UPARSE, QIIME1). *Legacy Standard.*
- **Taxonomic Assignment:** Mapping ASVs/OTUs against reference databases (Silva, Greengenes, GTDB).
- **Phylogenetic Tree Construction:** Required for calculating phylogenetic distances (e.g., UniFrac).

Analysis Pipeline: (2) Downstream Analysis

- **Data Import & Preprocessing (Phyloseq):**
  - Data Integration: Merging OTU Table, Taxonomy, Metadata, and Phylogenetic Tree.
  - Filtering**:** Removing low-depth samples, rare taxa (Sparsity handling), and contaminants.
- **Normalization & Transformation:**
  - *Legacy:* Rarefying (Subsampling) – Not recommended due to data loss.
  - *Modern:* CoDA (CLR, ILR), CSS, TMM, DESeq2 size factors.

Analysis Pipeline: (2) Downstream Analysis

❑ Alpha Diversity (Within-sample):

- ➢ Indices: Observed, Shannon, Simpson, Chao1, Faith's PD.
- ➢ Testing: Wilcoxon, T-test, ANOVA / GLM (for covariate adjustment).

❑ Beta Diversity (Between-sample):

- ➢ Distance Metrics: Bray-Curtis, Jaccard / UniFrac / Aitchison (CoDA).
- ➢ Ordination: PCoA (MDS), PCA (Biplot), NMDS, t-SNE, UMAP.
- ➢ Testing: PERMANOVA (adonis2), ANOSIM, Betadisper.

Analysis Pipeline: (2) Downstream Analysis

- **Differential Abundance Analysis (Biomarker Discovery):**
  - *Compositional:* ANCOM-BC2, ALDEx2.
  - *Count-based:* DESeq2, edgeR (Designed for RNA-seq; use with caution).
  - *Multivariable:* MaAsLin2.

- **Functional Prediction:** PICRUSt2, Tax4Fun (Predicting metabolic pathways from 16S data).

- **Network Analysis:** SPIEC-EASI, SparCC (Inferring microbial interactions).

Building the Phyloseq Object

- We start by merging three disparate files: Count Matrix, Metadata, and Taxonomy.
  - **ID Mapping:** We use a dictionary approach to standardize Sample IDs across files.
  - **Taxonomy Parsing:** We use tidyr::separate to split strings like "Bacteria; Bacteroidetes; …" into clean columns (Kingdom to Genus).
- Finally, we encapsulate everything into a single ps_gut object using phyloseq().

Quality Control (QC)

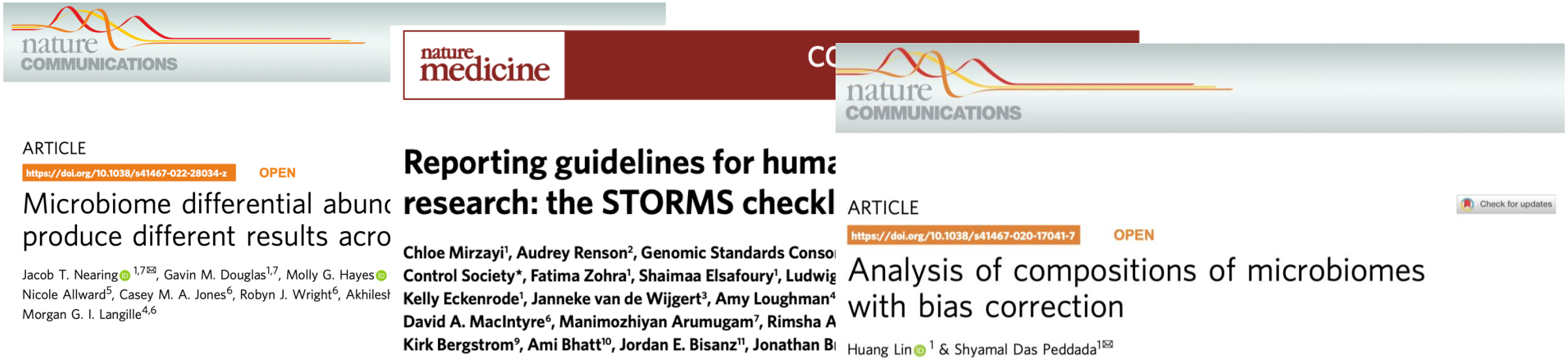
- Microbiome data is highly sparse (many zeros).

- **Filtering Criteria:**
  - **Read Depth:** Remove samples with < 1,000 reads.
  - **Prevalence:** Keep taxa appearing in at least 5% of samples.
  - **Contaminants:** Remove non-bacterial signals (Mitochondria, Chloroplasts).

- **Result:** The ps_clean object represents our high-quality dataset for analysis.

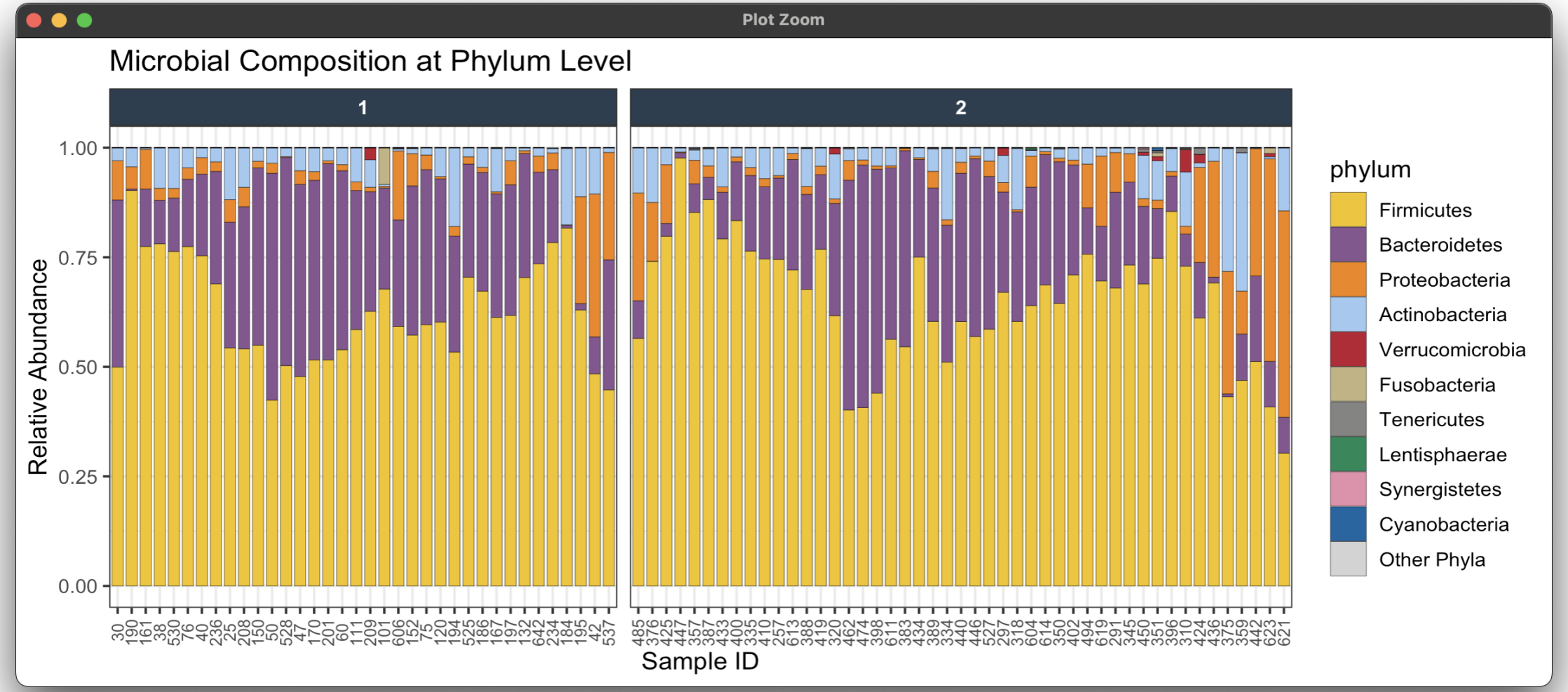
Quality Control (QC) - Rationale

- **Prevalence Filtering**: Concordance between analysis tools (e.g., ALDEx2, ANCOM-BC) is highest and false positives are reduced when taxa found in less than 10% of all samples are removed.
- **Read Count Filtering**: A minimum read count per sample (e.g., 1,000 or 5,000) must be specified, and the number of excluded samples must be reported.
- **No Rarefying**: In differential abundance analysis (DAA), rarefaction compromises statistical power.
- **Bias Correction**: Differences in total sequencing depth (library size) among samples are due to differences in the 'sampling fraction,' and this must be mathematically estimated and corrected (Bias-corrected LFC).



Diversity Analysis: Microbial Composition (Phylum Level)

- We visualize the Top 10 Phyla and group the rest as "Other" for a clean presentation.



Diversity Analysis: Alpha Diversity

- Key Components:
  - How many taxa are there? (Richness)
  - How strictly are they balanced? (Evenness)

- Why it matters:
  - Low alpha diversity is often associated with disease states (e.g., IBD, obesity, antibiotic usage).

Diversity Analysis: Alpha Diversity

- **Richness Estimators (Focus on Count)**
  - **# of Observed Features:** The raw count of unique taxa found.
    - (A, B, C) = (3, 2, 1) >> **3**
  - **Chao1 / ACE:** Statistical estimators that account for **unseen/rare species** to predict the *true* richness of the population.
- **Diversity Indices (Richness + Evenness)**
  - **Shannon Index:** It is sensitive to **rare species**. (Most commonly used).
    - (A, B, C) = (0.5, 0.33, 0.17) >> **- (0.5 x log 0.5 + 0.33 x log 0.33 + 0.17 x log 0.17)**
  - **Simpson Index:** Represents the probability that two randomly selected individuals belong to different species. It gives more weight to **dominant species**.
    - Prob. (A-A) >> 0.5 x 0.5; Prob. (B-B) >> 0.33 x 0.33; Prob. (C-C) >> 0.17 x 0.17
    - >> 1 – 0.5 x 0.5 – 0.33 x 0.33 – 0.17 x 0.17
- **Phylogenetic Diversity (Evolutionary Distance)**
  - **Faith's PD:** Sums the branch lengths on the phylogenetic tree. It captures the **evolutionary diversity** of the community.

Diversity Analysis: Alpha Diversity

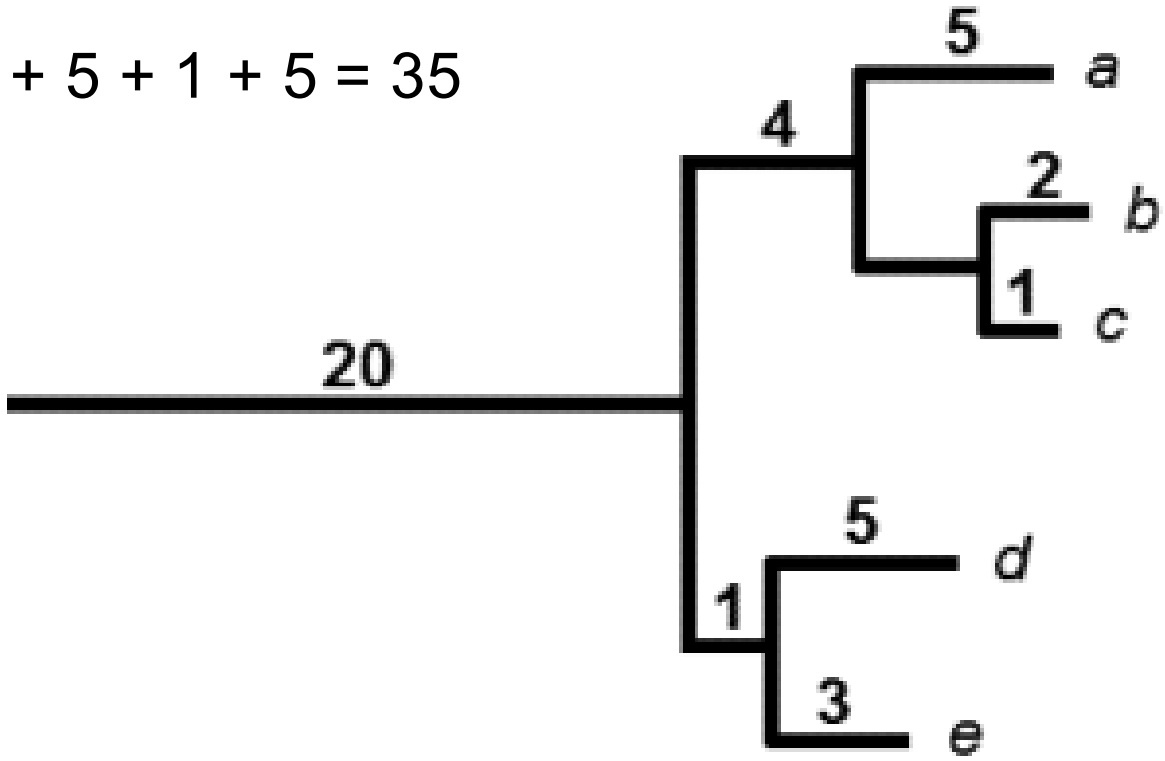
- **Richness Estimators (Focus on Count)**
  - **# of Observed Features:** The raw count of unique taxa found.
    - (A, B, C) = (3, 2, 1) >> **3**
  - **Chao1 / ACE:** Statistical estimators that account for **unseen/rare species** to predict the *true* richness of the population.
- **Diversity Indices (Richness + Evenness)**
  - **Shannon Index:** It is sensitive to **rare species**. (Most commonly used).
    - (A, B, C) = (0.5, 0.33, 0.17) >> **- (0.5 x log 0.5 + 0.33 x log 0.33 + 0.17 x log 0.17)**
  - **Simpson Index:** Represents the probability that two randomly selected individuals belong to different species. It gives more weight to **dominant species**.
    - Prob. (A-A) >> 0.5 x 0.5; Prob. (B-B) >> 0.33 x 0.33; Prob. (C-C) >> 0.17 x 0.17
    - >> 1 – 0.5 x 0.5 – 0.33 x 0.33 – 0.17 x 0.17
- **Phylogenetic Diversity (Evolutionary Distance)**
  - **Faith's PD:** Sums the branch lengths on the phylogenetic tree. It captures the **evolutionary diversity** of the community.

Diversity Analysis: Alpha Diversity

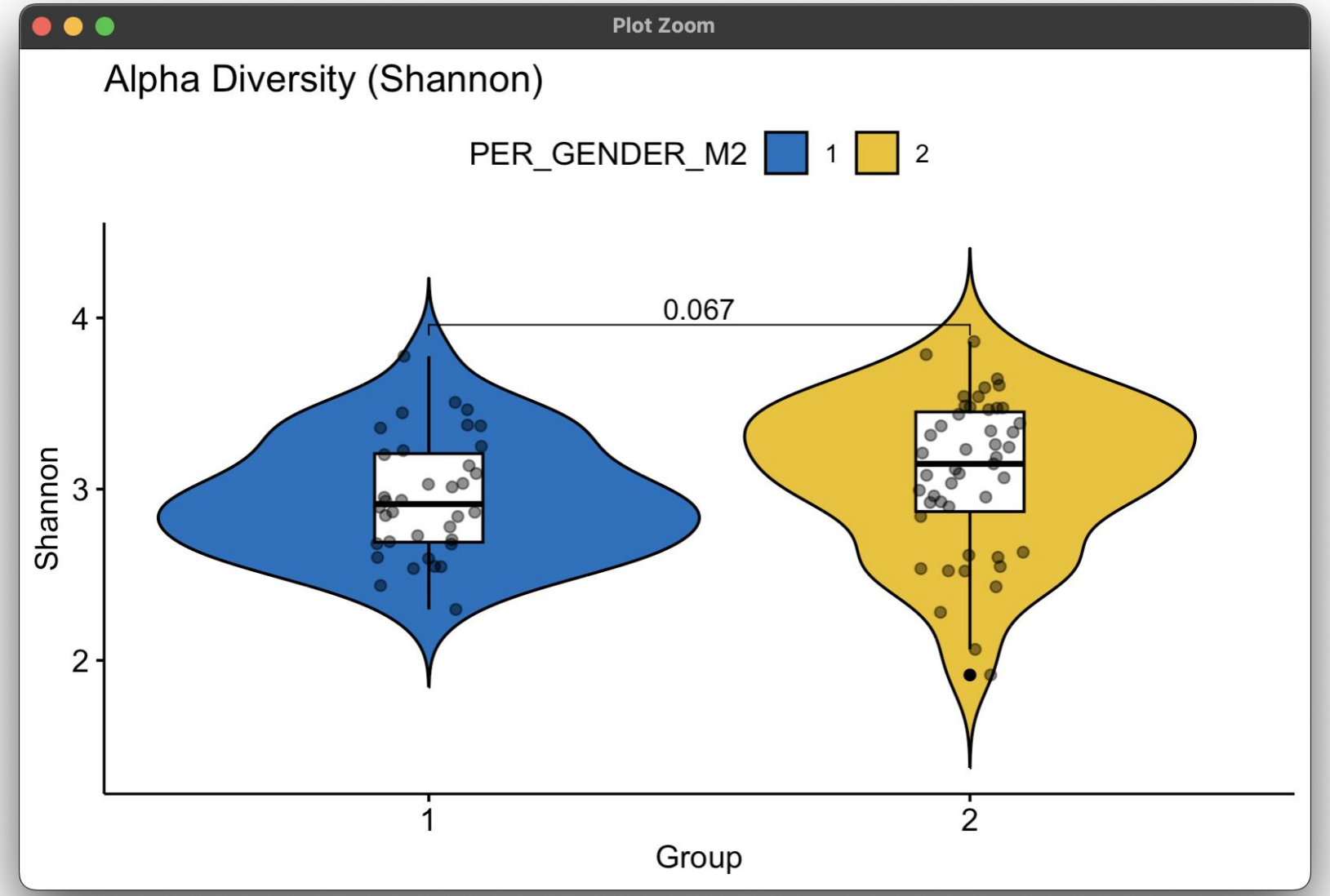
- **Richness Estimators (Focus on Count)**
  - **# of Observed Features:** The raw count of unique taxa found.
    - (A, B, C) = (3, 2, 1) >> **3**
  - **Chao1 / ACE:** Statistical estimators that account for **unseen/rare species** to predict the *true* richness of the population.
- **Diversity Indices (Richness + Evenness)**
  - **Shannon Index:** It is sensitive to **rare species**. (Most commonly used).
    - (A, B, C) = (0.5, 0.33, 0.17) >> **- (0.5 x log 0.5 + 0.33 x log 0.33 + 0.17 x log 0.17)**
  - **Simpson Index:** Represents the probability that two randomly selected individuals belong to different species. It gives more weight to **dominant species**.
    - Prob. (A-A) >> 0.5 x 0.5; Prob. (B-B) >> 0.33 x 0.33; Prob. (C-C) >> 0.17 x 0.17
    - >> 1 – 0.5 x 0.5 – 0.33 x 0.33 – 0.17 x 0.17
- **Phylogenetic Diversity (Evolutionary Distance)**
  - **Faith's PD:** Sums the branch lengths on the phylogenetic tree. It captures the **evolutionary diversity** of the community.

Diversity Analysis: Alpha Diversity

- The number of species within our samples = 20 (a, b, c, d, e)
- (b, c) >> 20 + 4 + 2 + 1 = 27
- (a, d) >> 20 + 4 + 5 + 1 + 5 = 35



Diversity Analysis: Alpha Diversity



Diversity Analysis: Beta Diversity (Indices)

❑ Distance between two samples

- Sample X: (A, B) = (10, 0) vs Sample Y: (A, B) = (2, 8)

➢ (1) Count-level

- Jaccard index: $1 - \frac{|A \cap B|}{|A \cup B|}$ = 1 – 1/2 = 0.5
- Bray-Curtis index: $\frac{\sum |A_i - B_i|}{\sum (A_i + B_i)}$ = (8 + 8) / (12 + 8) = 0.8

➢ (2) Compositional-level

- Aitchison Distance: $\ln\left(\frac{\text{관측값}}{\text{기하평균}}\right)$ = Euclidean distance on CLR data

Diversity Analysis: Beta Diversity (Indices)

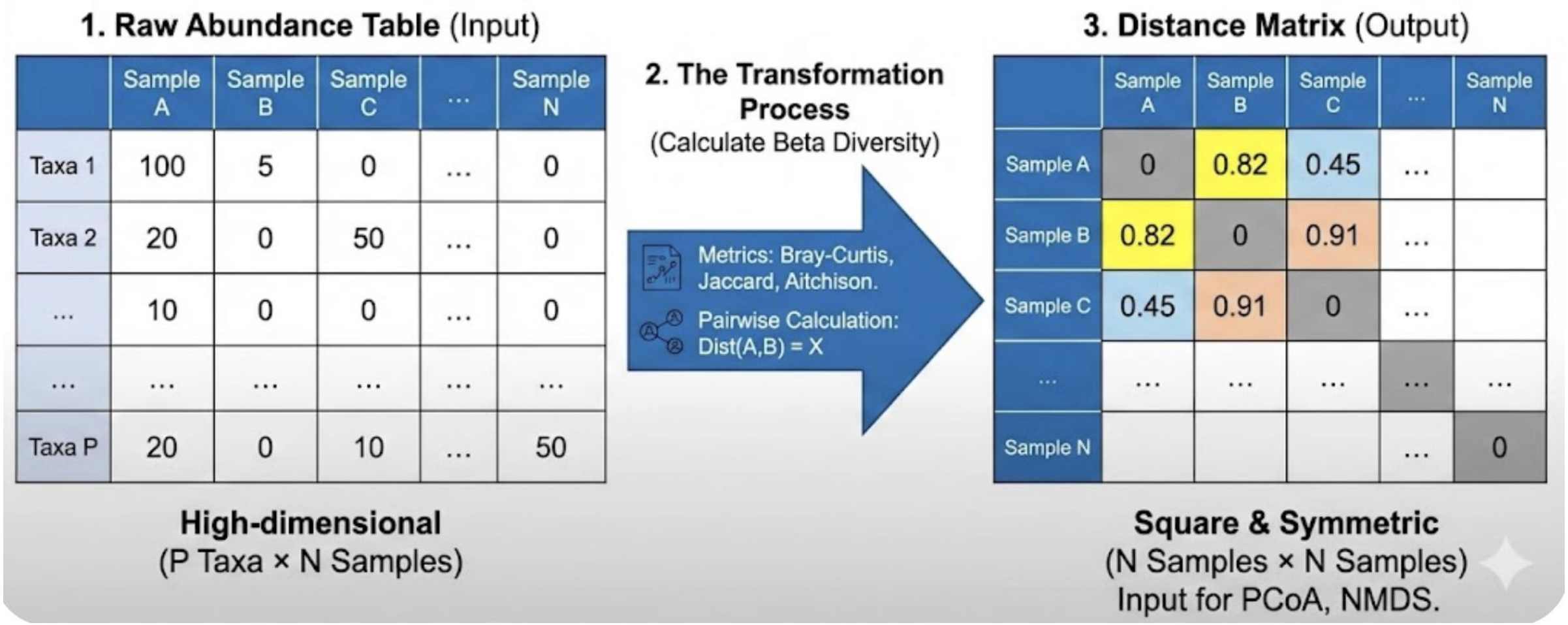
Limitation of count-level Bray-Curtis index

- In microbiome data, the total number of reads (sequencing depth) is determined arbitrarily by the machine.
  - Sample A: The machine worked well and detected a total of 10,000 reads.
  - Sample B: A machine error resulted in only 100 reads detected.

- Even if the actual proportions of microbes are identical, they are dramatically misinterpreted as "hugely different samples" when viewed at the count level.
  - Therefore, forced correction using rarefaction* is often performed.

* Random Subsampling

Diversity Analysis: Beta Diversity (Distance Matrix)



	Sample A	Sample B	Sample C	…	Sample N
Taxa 1	100	5	0	…	0
Taxa 2	20	0	50	…	0
…	10	0	0	…	0
…	…	…	…	…	…
Taxa P	20	0	10	…	50

	Sample A	Sample B	Sample C	…	Sample N
Sample A	0	0.82	0.45	…	
Sample B	0.82	0	0.91	…	
Sample C	0.45	0.91	0	…	
…	…	…	…	…	…
Sample N				…	0

Diversity Analysis: Beta Diversity (PCoA)

- PCoA: visualizes relationships between samples by plotting their similarities/dissimilarities in a lower-dimensional space, often a 2D graph

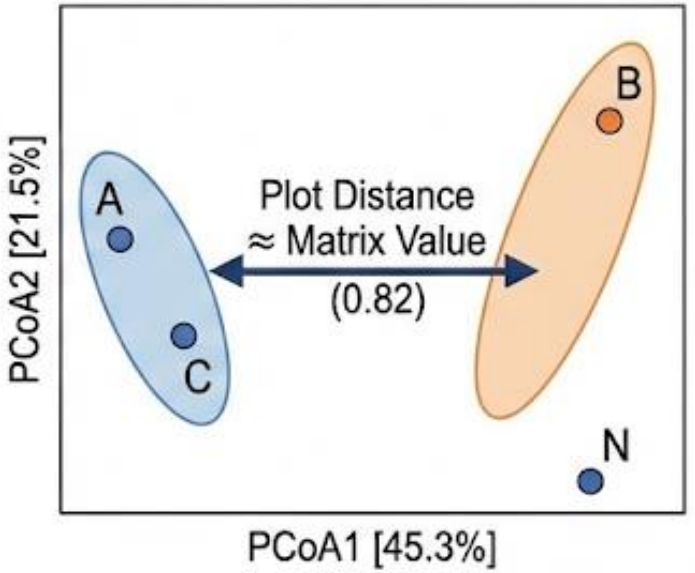
1. Input: Distance Matrix ($N \times N$)

Sample IDs	A	B	C	...	N
A	0	0.82	0.45	...	0.92
B	0.82	0	0.91	...	0.67
C	0.45	0.82	0	...	0.85
⋮	⋮	⋮	⋮	⋱	⋮
N	0.91	0.45	0.82	...	0.91

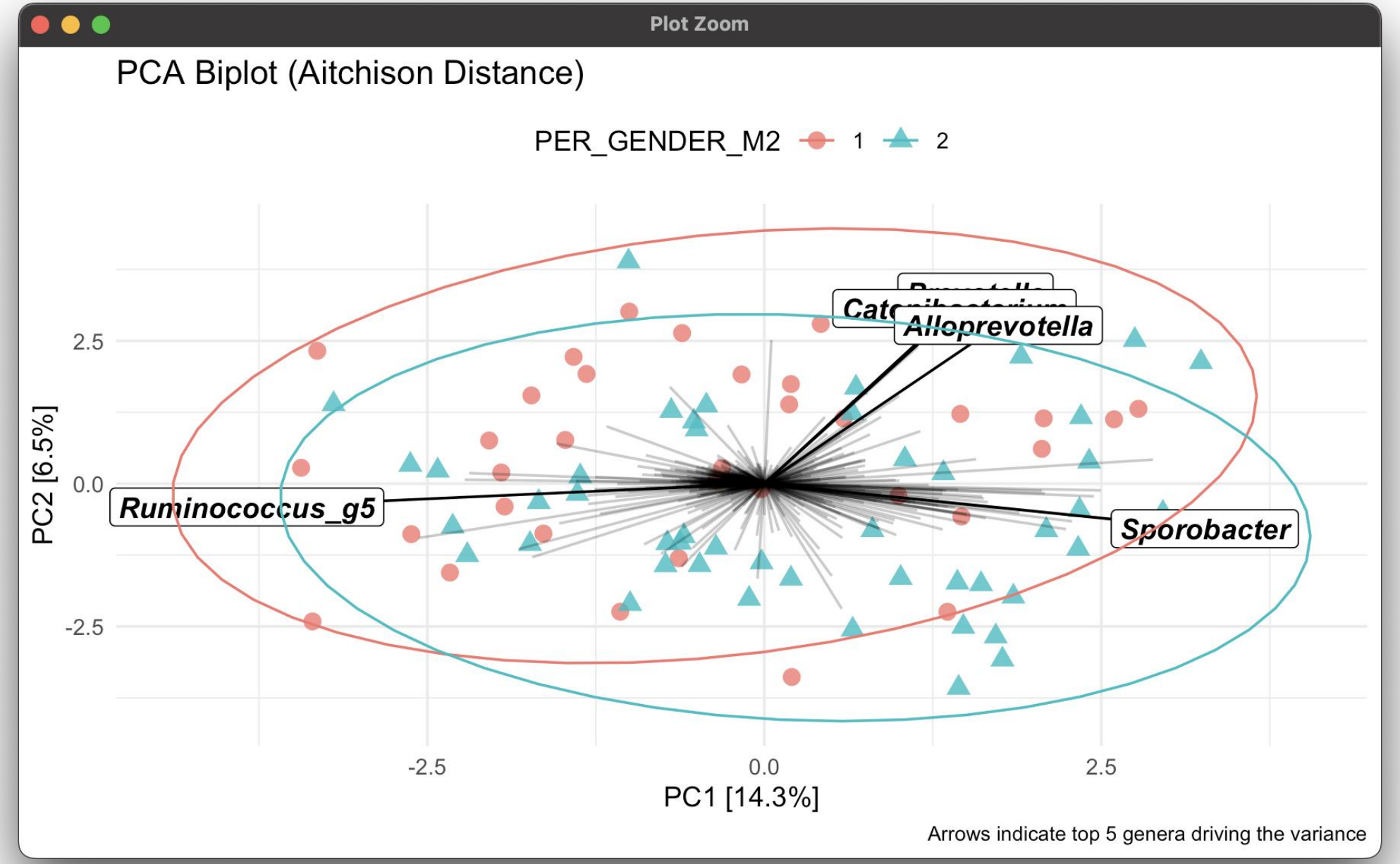
2. The PCoA Algorithm (Eigen-decomposition)

Find new axes that maximize variance explained

3. Output: 2D PCoA Plot



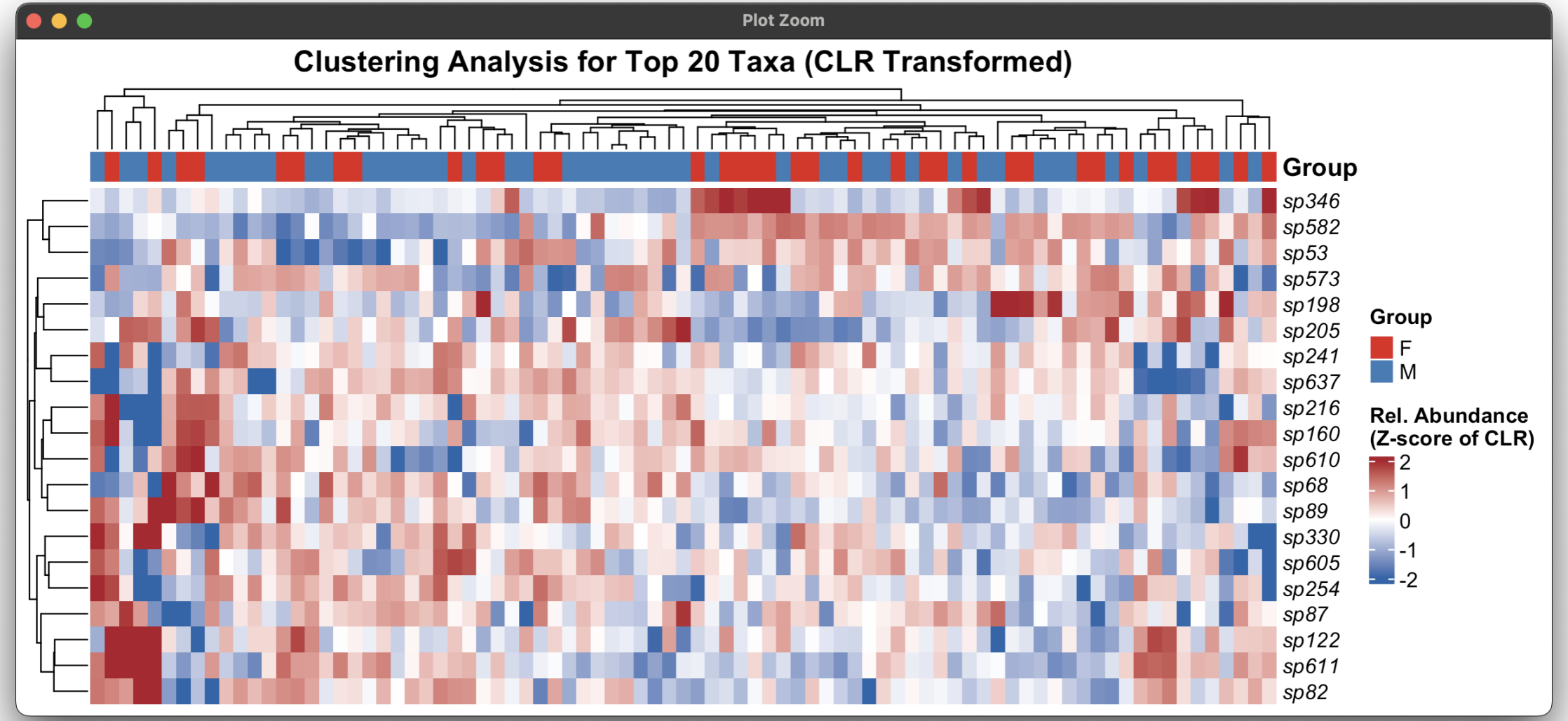
Interpretation
- Close points = Similar Communities
- Distant points = Different Communities
- Clusters = Group Differences


Plot Zoom
PCA Biplot (Aitchison Distance)
PER_GENDER_M2
1
2
Alloprevotella
Ruminococcus_g5
Sporobacter
2.5
0.0
-2.5
PC2 [6.5%]
-2.5
0.0
2.5
PC1 [14.3%]
Arrows indicate top 5 genera driving the variance

Biplot Interpretation

- A Biplot simultaneously visualizes both the **Samples (Observations)** and the **Variables (Taxa)** in a single low-dimensional space.
  - **Points:** Samples; **Arrows (Vectors):** Variables (Taxa).
- *How to Interpret*
  - Points closer together have similar microbial compositions.
  - **Longer arrows** indicate taxa that strongly drive the variation between samples (high contribution to the Principal Components).
  - **Angle between Arrows (Correlation)** indicates Positive correlation (Taxa increase/decrease together).
  - If a sample point lies in **the direction of a taxon arrow**, that sample has a **higher-than-average abundance** of that taxon. If it lies in **the opposite direction**, it has **lower abundance**.

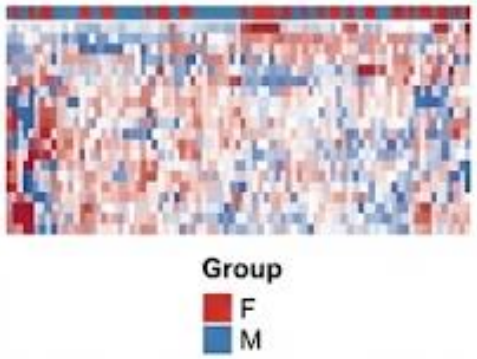
Heatmap Visualization



Next Steps: Differential Abundance Analysis

❑ Identifying significant taxa between groups (F vs M) using CoDA

1. Heatmap Cluster Insights

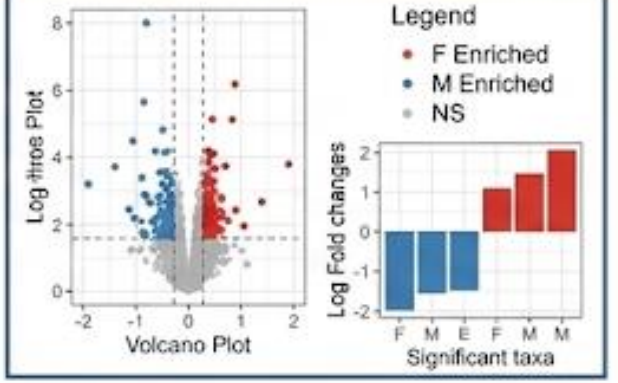


Hypothesis Generation

2. CoDA-based Statistical Testing

ANCOM-BC, ALDEx2, or Maaslin2. Accounting for Compositionality.

3. Result Visualization



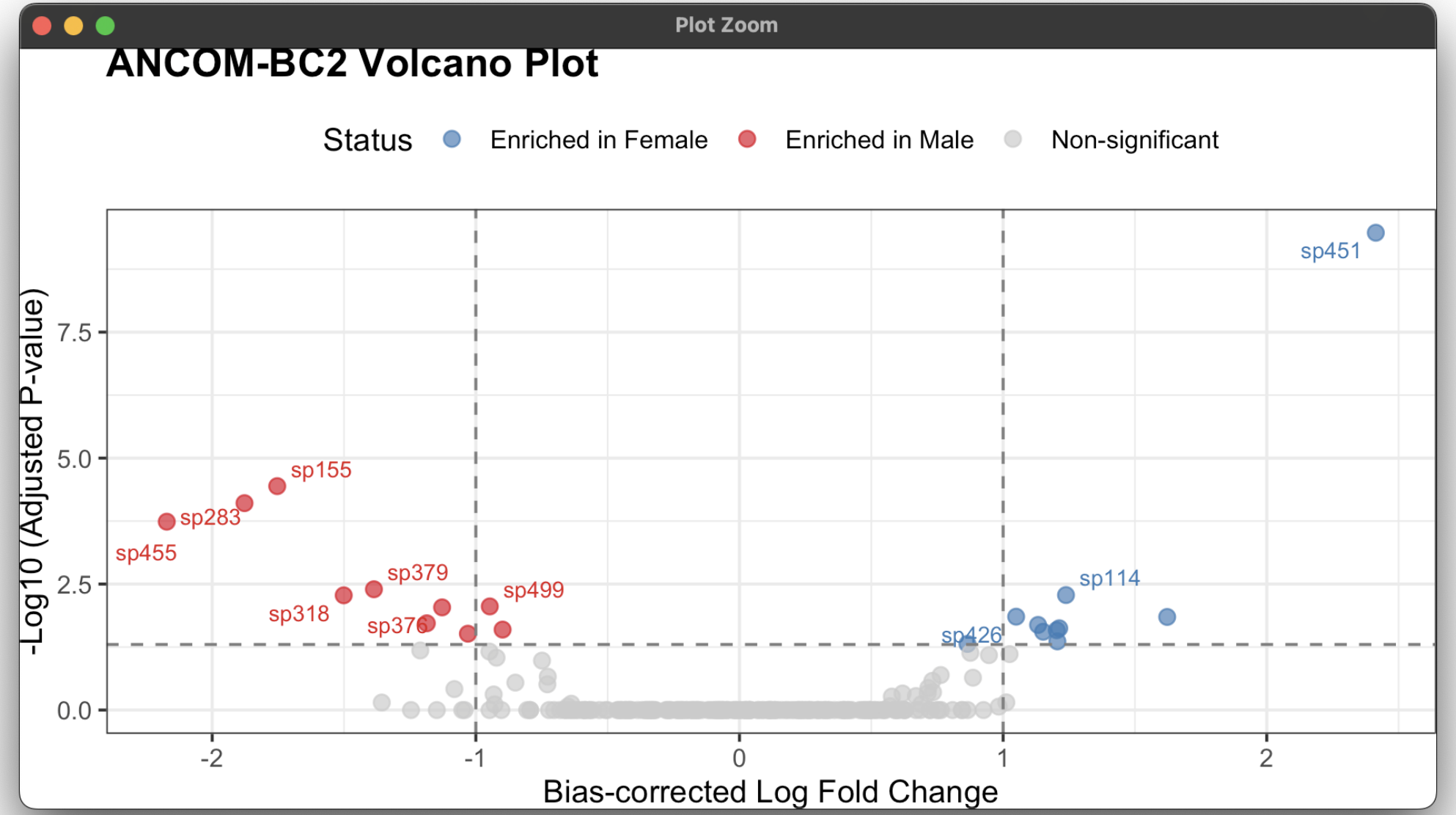
- Expected Outcomes
  - Significantly differentially abundant taxa ($q < 0.05$)
  - Effect sizes (Log Fold Change)
  - Volcano plots and bar charts of top features.

ANCOM-BC2 (Analysis of Compositions of Microbiomes)

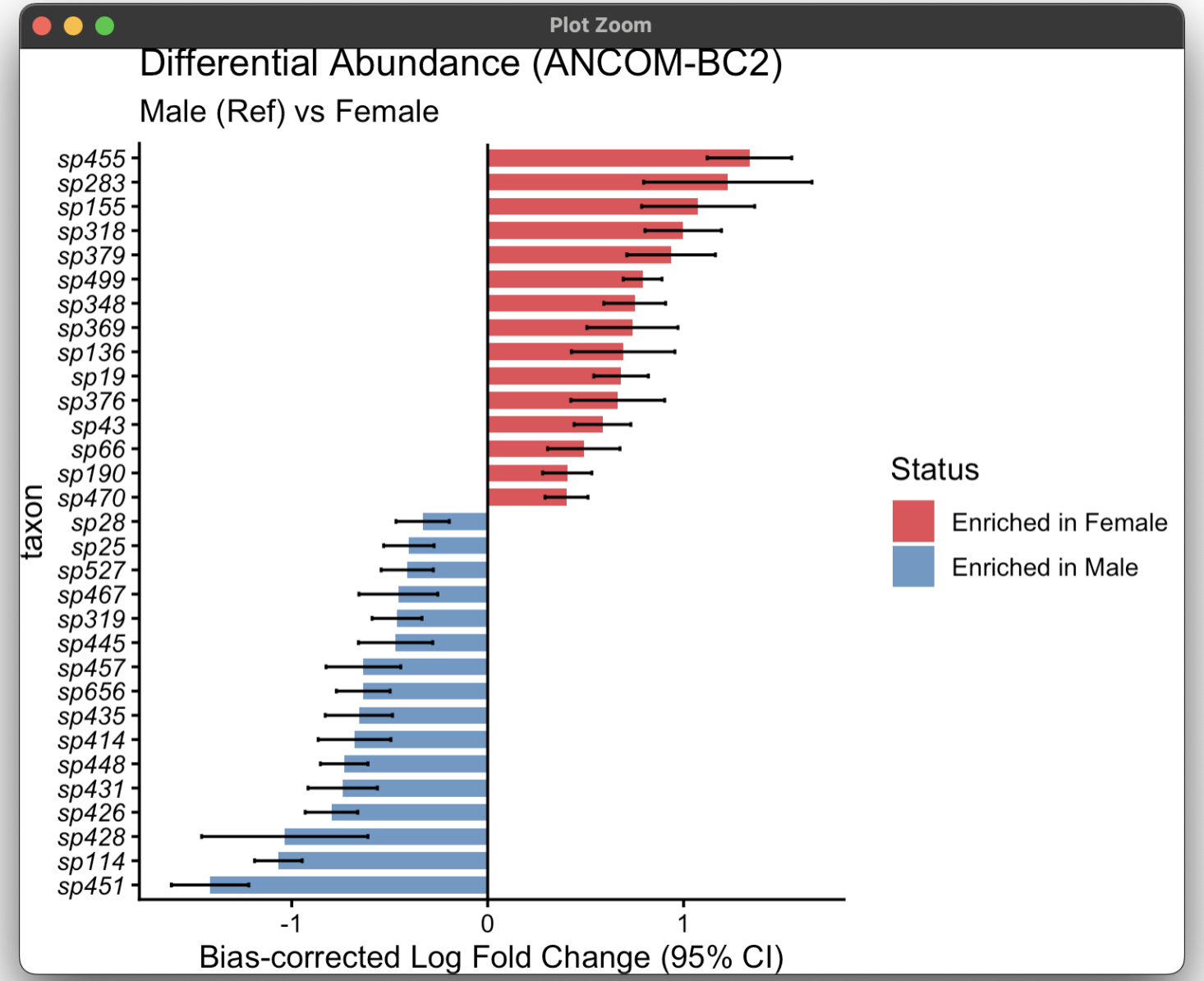
- Traditional tests (t-test, Wilcoxon) fail to account for Sampling Bias (different sequencing depths).

- ANCOM-BC2 estimates the sampling fraction and corrects the bias. It also handles Structural Zeros (taxa genuinely absent).

ANCOM-BC2: Volcano plot



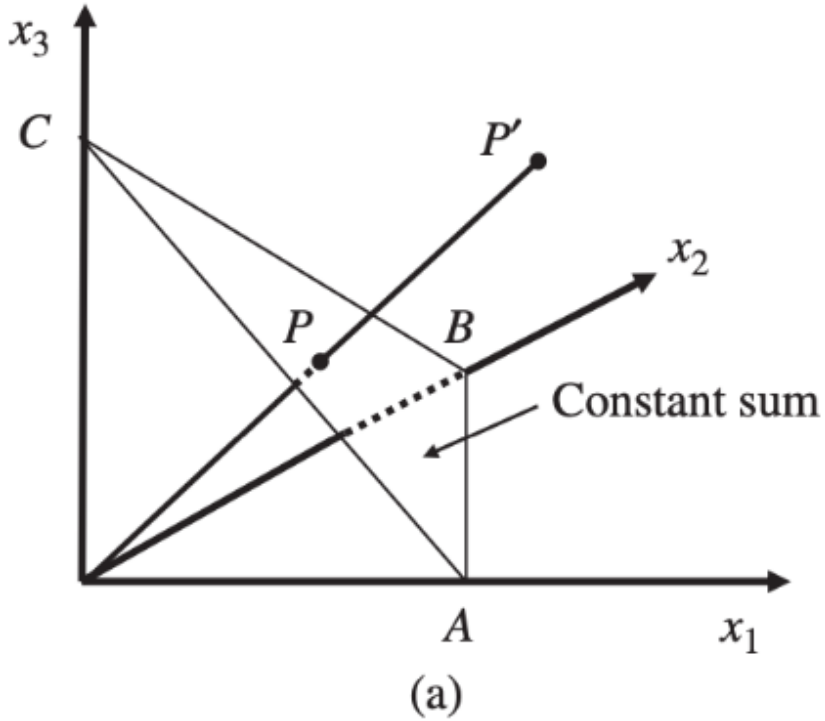
The Waterfall Plot



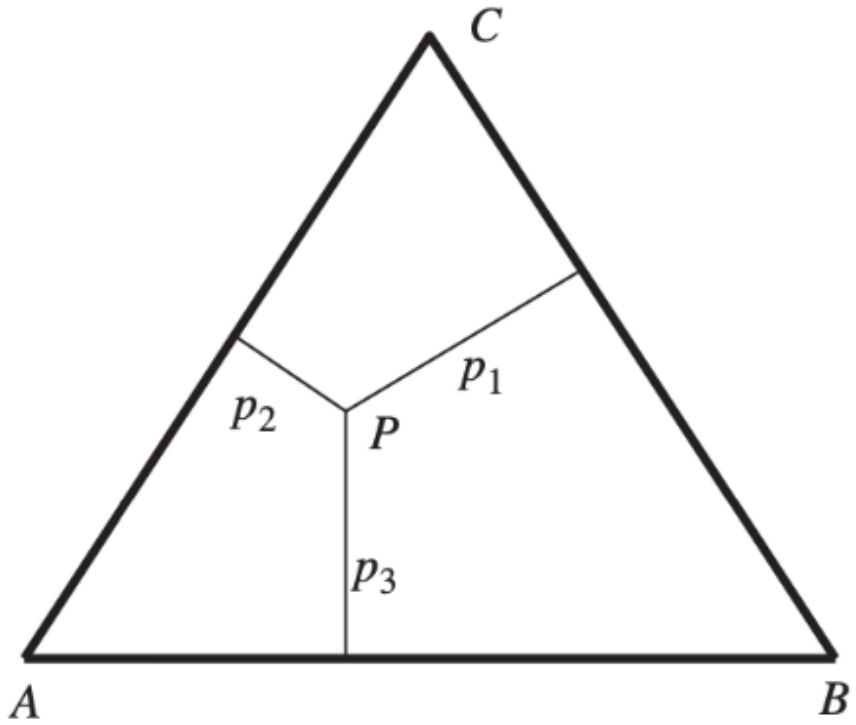
Regression on the Simplex

- Standard regression assumes independence, but compositional components are constrained & dependent (sum to 1).
- Standard regression fails to work.
- Our function performs regression in the CLR space and back-transforms the results to the Simplex for interpretation.

Simplex

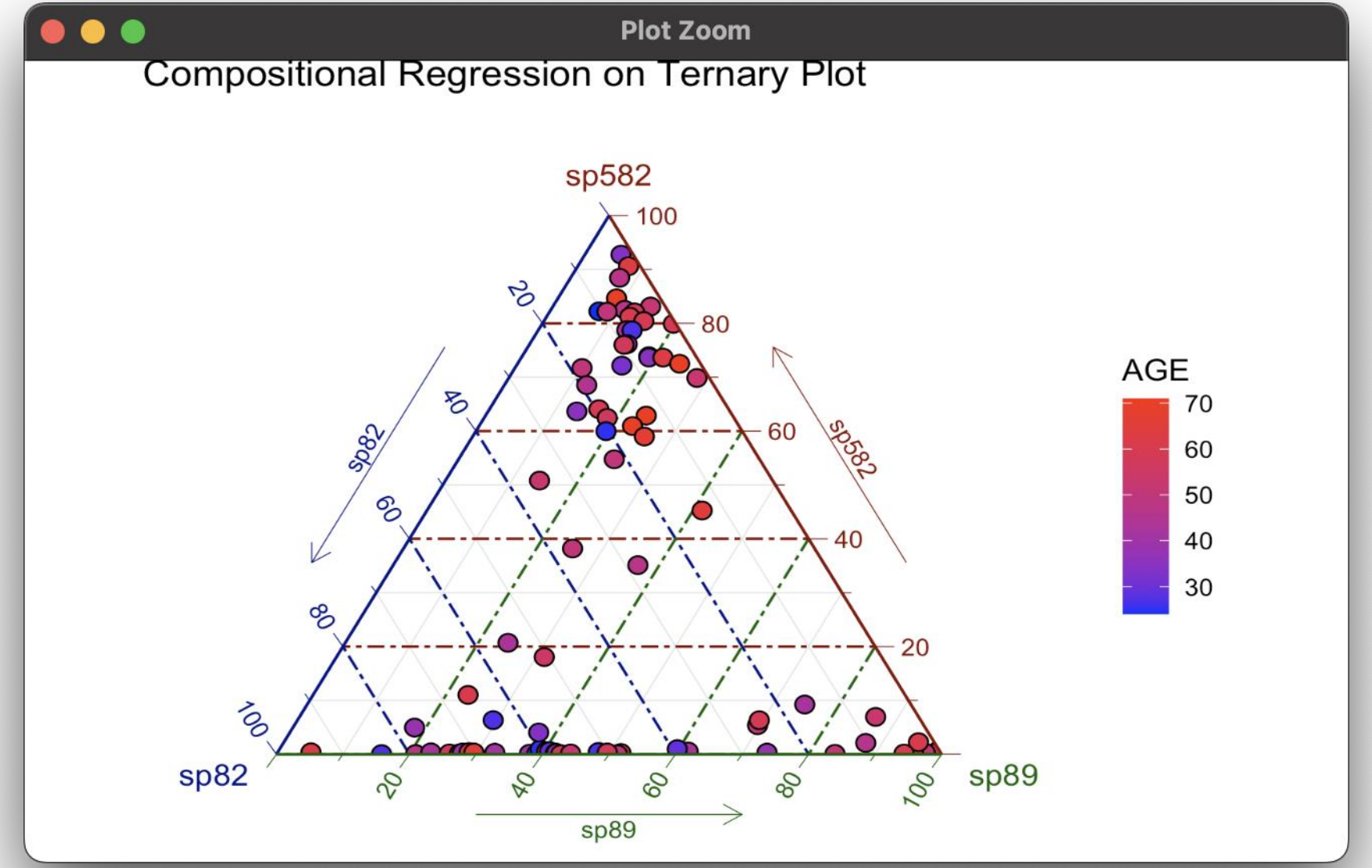


(a)

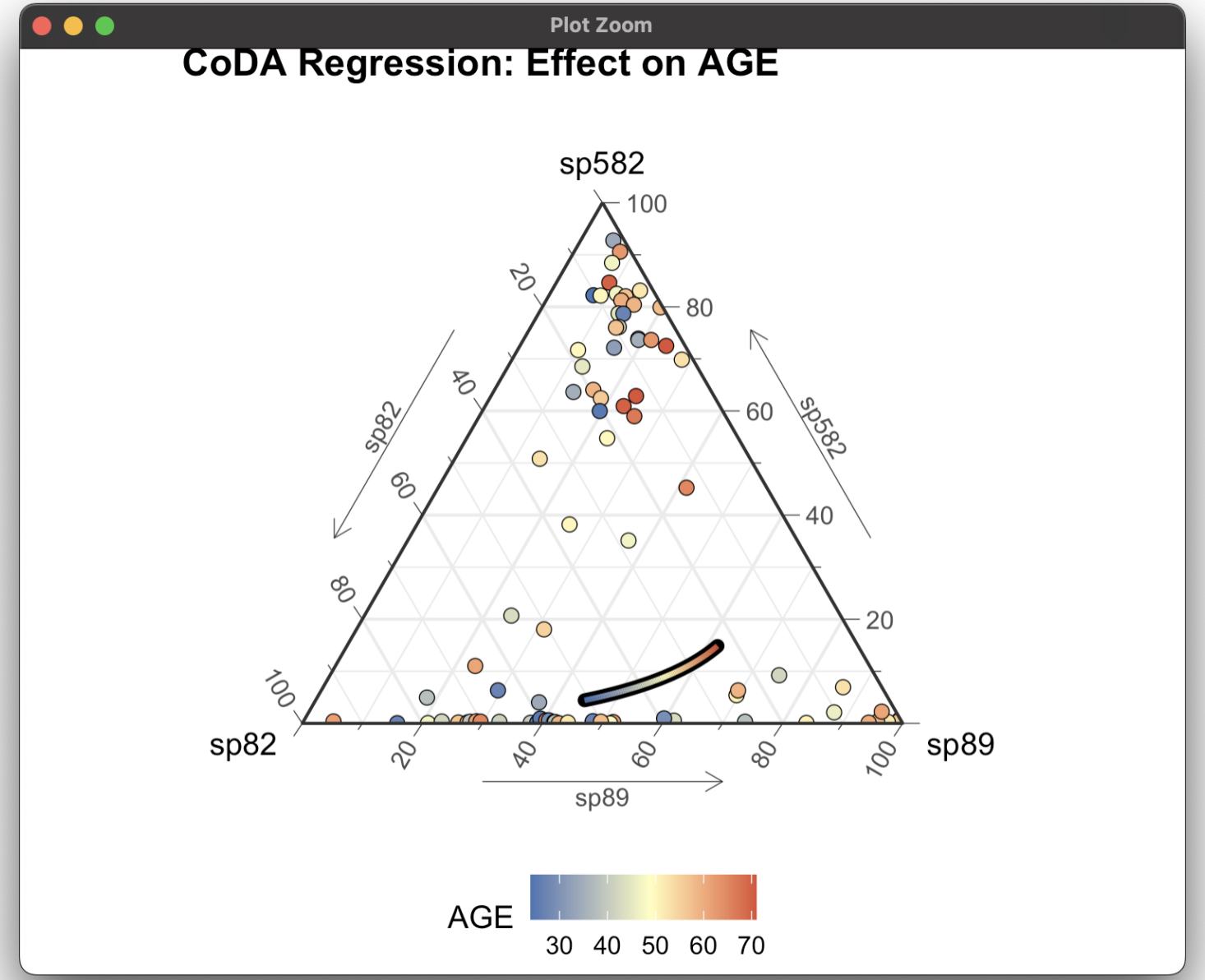


(b)

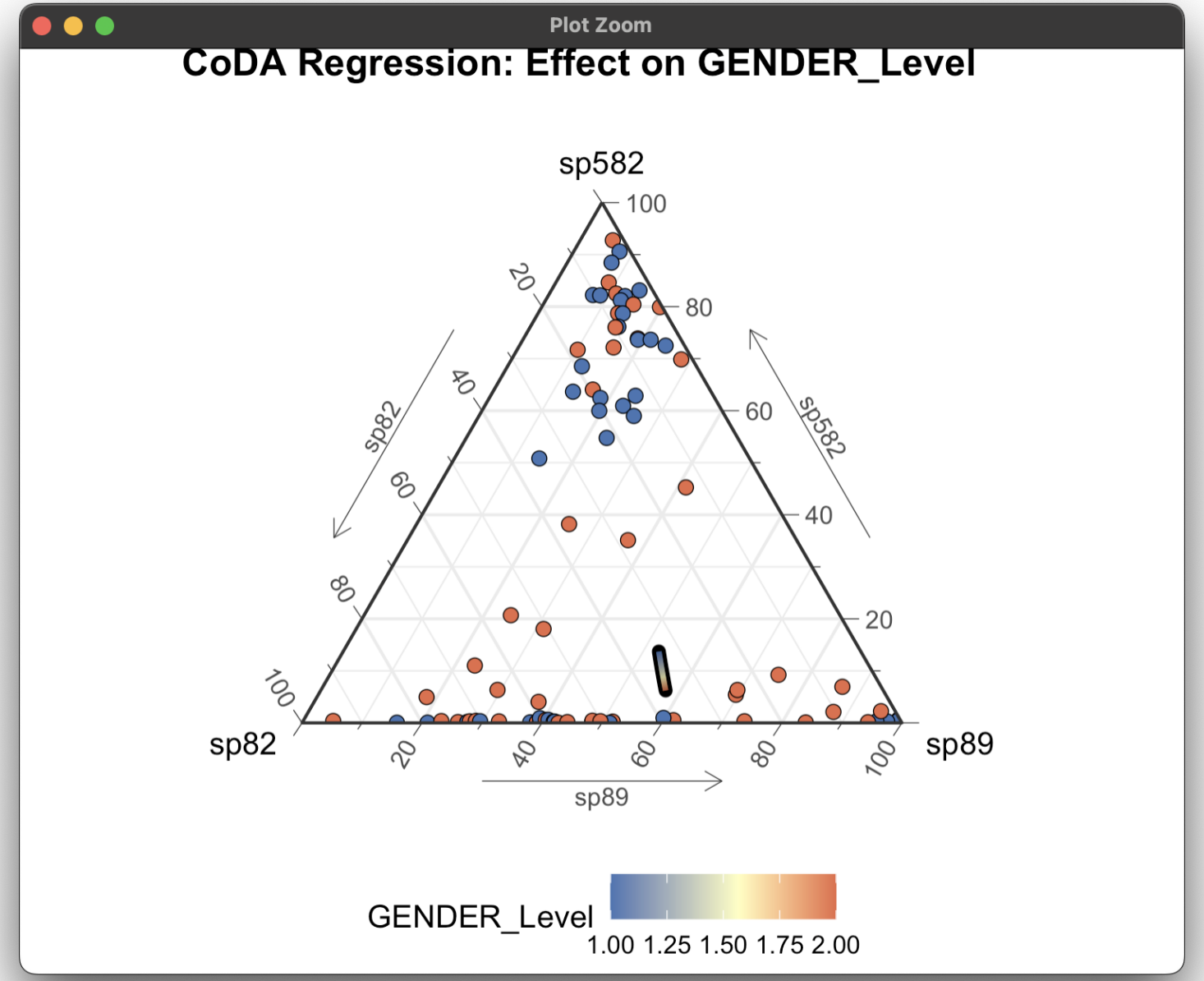
Data points on the Simplex (colored by Age)



Visualizing Continuous Covariates (Age)



Visualizing Categorical Covariates (Gender)



Conclusion

- Recap:
  - QC: Clean your data rigorously.
  - Transform: Use CLR for multivariate analysis (PCA, Heatmap).
  - Model: Use CoDA regression for continuous variables.
  - Test: Use ANCOM-BC2 for differential abundance.

- Take-home Message: "Respect the compositional nature of your data to avoid false discoveries."

Thank you for your attention !