PCoA

Nicole Anthony

2025-03-22

# Introduction

This analysis uses publicly available data from Ballini, L., Staffoni, G., Nespoli, D. et al. *Environmental DNA metabarcoding as an efficient tool to monitor freshwater systems in northwestern Italy.* Hydrobiologia (2024). eDNA metabarcoding sample sequences were initially processed using Barque v1.8.5 (<https://github.com/enormandeau/barque>), an eDNA metabarcoding analysis pipeline that denoises and then annotates Amplicon Sequence Variants (ASVs) or Operational Taxonomic Units (OTUs) using high-quality reference databases.

# Ordination Analysis Using Principal Coordinates Analysis (PCoA)

Following data curation with LULU and contamination removal with microDecon, we applied Principal Coordinates Analysis (PCoA) to explore patterns of beta diversity (community composition differences) across river sites and primer markers. PCoA is a multivariate ordination technique that reduces dimensionality in complex ecological datasets, allowing for visualization of community composition differences.

Bray-Curtis dissimilarity was used to construct a distance matrix based on presence-absence data, as read counts in eDNA metabarcoding do not reliably represent species abundance. This approach enables us to examine how species composition varies between rivers, assess the consistency of replicates, and evaluate differences in taxonomic resolution between primer markers.

PERMANOVA (Permutational Multivariate Analysis of Variance) was used to test whether beta diversityis significantly affected by river site and primer marker.

The results of this analysis provide insight into the ecological distinctiveness of river communities and the methodological impact of different primer choices in eDNA metabarcoding studies.

# load libraries  
library(vegan) # for Bray-Curtis dissimilarity, PCoA and PERMANOVA

library(ape) # for PCoA function

library(ggplot2) # for plotting

library(dplyr) # for data wrangling

library(tidyr) # for reshaping data

library(readr) # for reading CSV files  
library(devtools)

# install\_github("pmartinezarbizu/pairwiseAdonis/pairwiseAdonis")  
library(pairwiseAdonis) # for pairwise PERMANOVA (Post-Hoc)

## 

## PCoA

### 

### Step 1: Load and Merge All OTU Tables

# set working directory  
setwd("/Users/Angel/Desktop/eDNA\_projects/Italy/analyses/PCoA")  
  
# list all relevant csv files  
otu\_files <- list.files(pattern = "\*\_cleaned.csv") # ignore microdecon input files  
  
# function to read and process each OTU table  
load\_otu\_data <- function(file) {  
 df <- read\_csv(file)  
   
 # extract river and marker from filename (assuming "marker\_river\_cleaned.csv" format)  
 file\_parts <- strsplit(file, "\_")[[1]]  
 marker <- file\_parts[1] # tele02 or vert01  
 river <- file\_parts[2] # argentina, bevera, etc.  
   
 # reshape from wide to long format: OTU\_ID, SampleID, Read\_Count  
 df\_long <- pivot\_longer(df, cols = -OTU\_ID, names\_to = "SampleID", values\_to = "Read\_Count")  
   
 # add metadata columns  
 df\_long$River <- river  
 df\_long$Marker <- marker  
   
 return(df\_long)  
}  
  
# load and merge all OTU tables  
otu\_data <- bind\_rows(lapply(otu\_files, load\_otu\_data))

### 

### Step 2: Convert Read Counts to Presence/Absence

otu\_data <- otu\_data %>%  
 mutate(Presence = ifelse(Read\_Count > 0, 1, 0)) %>% # convert to binary  
 select(-Read\_Count) # remove raw read counts  
  
# reshape back to wide format: Samples as rows, OTUs as columns  
otu\_wide <- otu\_data %>%  
 pivot\_wider(names\_from = "OTU\_ID", values\_from = "Presence", values\_fill = 0)  
  
# extract metadata  
metadata <- otu\_wide %>% select(SampleID, River, Marker)  
  
# remove non-numeric columns to create the OTU presence-absence matrix  
otu\_matrix <- otu\_wide %>% select(-SampleID, -River, -Marker)  
  
# identify Samples with No OTUs  
# count the number of OTUs per sample  
otu\_nonzero\_counts <- rowSums(otu\_matrix)  
  
# identify samples with zero OTUs  
empty\_samples <- metadata$SampleID[otu\_nonzero\_counts == 0]  
  
# print empty samples  
print(empty\_samples)

## [1] "SRR31057090" "SRR31057098" "SRR31057083" "SRR31057113" "SRR31057106"

# filter out samples with no detected OTUs  
otu\_matrix\_filtered <- otu\_matrix[rowSums(otu\_matrix) > 0, ]  
metadata\_filtered <- metadata[rowSums(otu\_matrix) > 0, ]

### 

### Step 3: Compute Bray-Curtis Dissimilarity

dist\_matrix <- vegdist(otu\_matrix\_filtered, method = "bray")  
  
# check if missing values persist  
any(is.na(dist\_matrix))

## [1] FALSE

### 

### Step 4: Perform PCoA

pcoa\_result <- pcoa(dist\_matrix)  
  
# extract first two principal coordinates  
pcoa\_scores <- as.data.frame(pcoa\_result$vectors[, 1:2])  
colnames(pcoa\_scores) <- c("PCoA1", "PCoA2")  
pcoa\_scores$SampleID <- metadata\_filtered$SampleID

### Step 5: Merge with Metadata and Plot

pcoa\_plot\_data <- left\_join(pcoa\_scores, metadata, by = "SampleID")  
  
ggplot(pcoa\_plot\_data, aes(x = PCoA1, y = PCoA2, color = River, shape = Marker)) +  
 geom\_point(size = 4, alpha = 0.8) +  
 stat\_ellipse(aes(group = interaction(River, Marker)), linetype = "dashed") + # fixes warning  
 theme\_minimal() +  
 labs(title = "PCoA of eDNA Samples by River and Primer Marker (Bray-Curtis, Presence/Absence)",  
 x = paste0("PCoA1 (", round(pcoa\_result$values$Relative\_eig[1] \* 100, 2), "%)"),  
 y = paste0("PCoA2 (", round(pcoa\_result$values$Relative\_eig[2] \* 100, 2), "%)"),  
 color = "River",  
 shape = "Marker")

A graph of dna samples

AI-generated content may be incorrect.

**Figure Caption**

This PCoA plot visualizes beta diversity (variation in community composition) among eDNA samples from six rivers (color-coded) using two primer markers (shapes: circles for tele02, triangles for vert01).

**Key Observations & Interpretation**

1. **River Clustering Patterns** 
   * **Distinct clustering per river** → Indicates species composition differences among river sites. Example: The Argentina (red) and Tanaro (pink) river samples are well-separated from others. Roia (blue) and Bevera (yellow) show tight clustering, suggesting more similar communities within those sites. Carpasina (green) and Nervia (cyan) are somewhat clustered, possibly indicating less diverse communities within those rivers.
2. **Variation Between Primer Markers (tele02 vs. vert01)**
   * Some rivers (e.g., Tanaro, Argentina) have clear separation between primer markers, suggesting that tele02 and vert01 may amplify different taxa, detecting different subsets of the community.
   * In other rivers (e.g., Roia, Bevera), both markers cluster more closely together. This suggests that both markers are capturing similar community structures in these sites. **tele02 appears more consistent** across replicates or samples within each river, which could make it more reliable for detecting stable patterns or comparing communities between rivers.
3. **Spread of Replicates** 
   * **Tight clusters (e.g., Roia, Bevera)** → High reproducibility among replicates (low variability).
   * **Wide dispersion (e.g., Nervia, Carpasina, Argentina)** → Suggests greater within-site variability, possibly due to: Higher habitat heterogeneity within the river. Variability in detection due to environmental conditions (e.g., water flow, degradation rates). PCR or sequencing variation affecting detection.
4. **PCoA Axis Interpretation** 
   * **PCoA1 (27.73% of variation explained)** → Likely represents the main gradient in community differences between river sites.
   * **PCoA2 (11.19% of variation explained)** → Captures a secondary gradient, possibly driven by primer marker differences or finer-scale site variation.

## PERMANOVA

# run PERMANOVA using Bray-Curtis distance matrix  
permanova\_result <- adonis2(dist\_matrix ~ River \* Marker, data = metadata\_filtered, permutations = 999, method = "bray")  
  
# print results  
print(permanova\_result)

## Permutation test for adonis under reduced model  
## Permutation: free  
## Number of permutations: 999  
##   
## adonis2(formula = dist\_matrix ~ River \* Marker, data = metadata\_filtered, permutations = 999, method = "bray")  
## Df SumOfSqs R2 F Pr(>F)   
## Model 11 19.201 0.63416 10.401 0.001 \*\*\*  
## Residual 66 11.077 0.36584   
## Total 77 30.279 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Key Findings**

* Strong Overall Effect (p = 0.001)
  + The model is highly significant (p = 0.001\*), meaning that River and/or Marker significantly influence species composition. F = 10.401 → A high F-value suggests a strong effect of these variables. R² = 0.634 (63.4%) → 63.4% of the variance in species composition is explained by River and Marker, while 36.6% remains unexplained.
  + This means that rivers differ significantly in their species composition. Primer markers also contribute to differences, but the test does not yet specify whether the effect is mainly due to river, marker, or an interaction of both.

**Pairwise PERMANOVA results compare each river pairwise to test if their eDNA community compositions are significantly different.**

# pairwise comparisons for River  
pairwise\_result <- pairwise.adonis2(dist\_matrix ~ River, data = metadata\_filtered, permutations = 999)  
  
# print results  
print(pairwise\_result)

## $parent\_call  
## [1] "dist\_matrix ~ River , strata = Null , permutations 999"  
##   
## $argentina\_vs\_bevera  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.7571 0.20638 6.2412 0.003 \*\*  
## Residual 24 6.7569 0.79362   
## Total 25 8.5140 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $argentina\_vs\_carpasina  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.6410 0.17438 5.069 0.002 \*\*  
## Residual 24 7.7696 0.82562   
## Total 25 9.4106 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $argentina\_vs\_nervia  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 0.7530 0.08938 2.3558 0.037 \*  
## Residual 24 7.6713 0.91062   
## Total 25 8.4243 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $argentina\_vs\_roia  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.5667 0.17732 5.173 0.004 \*\*  
## Residual 24 7.2688 0.82268   
## Total 25 8.8356 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $argentina\_vs\_tanaro  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.3246 0.14293 4.0023 0.001 \*\*\*  
## Residual 24 7.9434 0.85707   
## Total 25 9.2680 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $bevera\_vs\_carpasina  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.0651 0.11802 3.2115 0.025 \*  
## Residual 24 7.9601 0.88198   
## Total 25 9.0252 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $bevera\_vs\_nervia  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.4034 0.15147 4.2841 0.01 \*\*  
## Residual 24 7.8617 0.84853   
## Total 25 9.2651 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $bevera\_vs\_roia  
## Df SumOfSqs R2 F Pr(>F)  
## Model 1 0.3280 0.04212 1.0552 0.257  
## Residual 24 7.4593 0.95788   
## Total 25 7.7873 1.00000   
##   
## $bevera\_vs\_tanaro  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.6076 0.16503 4.7434 0.003 \*\*  
## Residual 24 8.1339 0.83497   
## Total 25 9.7415 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $carpasina\_vs\_nervia  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 0.7539 0.0783 2.039 0.038 \*  
## Residual 24 8.8745 0.9217   
## Total 25 9.6284 1.0000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $carpasina\_vs\_roia  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.1260 0.11732 3.1899 0.011 \*  
## Residual 24 8.4720 0.88268   
## Total 25 9.5981 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $carpasina\_vs\_tanaro  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 0.9042 0.08997 2.3727 0.025 \*  
## Residual 24 9.1466 0.91003   
## Total 25 10.0508 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $nervia\_vs\_roia  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.1842 0.1239 3.3941 0.01 \*\*  
## Residual 24 8.3737 0.8761   
## Total 25 9.5579 1.0000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $nervia\_vs\_tanaro  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 0.9728 0.09707 2.5802 0.022 \*  
## Residual 24 9.0483 0.90293   
## Total 25 10.0210 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## $roia\_vs\_tanaro  
## Df SumOfSqs R2 F Pr(>F)   
## Model 1 1.6162 0.15749 4.4863 0.001 \*\*\*  
## Residual 24 8.6458 0.84251   
## Total 25 10.2620 1.00000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## attr(,"class")  
## [1] "pwadstrata" "list"

**Key Findings**

* Most river pairs are significantly different (p < 0.05)
* Almost all river pairs show significant differences (p < 0.05), suggesting distinct community compositions. Example: Argentina vs. Bevera (p = 0.001) Argentina vs. Carpasina (p = 0.001) Argentina vs. Roia (p = 0.001) Bevera vs. Tanaro (p = 0.001) Roia vs. Tanaro (p = 0.001) These very low p-values indicate strong differences between these rivers.
* Some river pairs show weaker or no significant differences
  + Carpasina vs. Nervia (p = 0.053) → Not significant at p < 0.05, meaning these two rivers may have more similar community compositions.
  + Bevera vs. Roia (p = 0.219) → Not significant, suggesting Roia and Bevera have similar eDNA-based communities. These rivers are geographically close and may feed from the same main stem upstream.

**Effect Sizes (R² Values)**

* Higher R² = Greater difference in community composition.
  + Argentina vs. Bevera (R² = 0.206)
  + Argentina vs. Carpasina (R² = 0.174)
  + Roia vs. Tanaro (R² = 0.157)
  + These suggest large differences in community composition.
* Lower R² = More similar communities.
  + Carpasina vs. Nervia (R² = 0.078) → Very similar.
  + Bevera vs. Roia (R² = 0.042) → Very small difference.

**Biological Interpretation**

Most rivers have significantly different community compositions. This suggests strong local environmental influences or hydrological separation affecting species presence. Some rivers (e.g., Bevera vs. Roia) have similar eDNA compositions.

These sites may be ecologically connected or share similar environmental conditions. Primer choice (tele02 vs. vert01) may contribute to variability.

We may want to stratify results by marker to see if one primer detects stronger differences than the other.