Quick Reference Guide: Diversity Analysis

Quick Start (Copy-Paste Ready)

```r
# 1. Install packages (run once)
install.packages(c("vegan", "iNEXT", "betapart", "dplyr", "tidyr"))

# 2. Load libraries
library(vegan)
library(iNEXT)
library(betapart)
library(dplyr)
library(tidyr)

# 3. Set working directory and load data
setwd("path/to/your/data")  # CHANGE THIS
master_taxa <- read.csv("masterTaxaGenus.csv")
benthics <- read.csv("stationBenthicsTESTSITE.csv")
station_info <- read.csv("stationInfoBenSampsTESTSITE.csv")

# 4. Run the main script
source("diversity_analysis.R")
```

Understanding the Three Diversity Metrics

Visual Concept

```
Stream System with 5 Sites (riffles)
=====================================

Site 1: Species A, B, C, D (4 species)
Site 2: Species A, B, E, F (4 species)
Site 3: Species C, D, G, H (4 species)
Site 4: Species E, F, I, J (4 species)
Site 5: Species A, C, E, G (4 species)

ALPHA (α): Average diversity within sites
= (4 + 4 + 4 + 4 + 4) / 5 = 4 species per site

GAMMA (γ): Total diversity across all sites
= 10 unique species total (A, B, C, D, E, F, G, H, I, J)

BETA (β): How much communities differ
= γ / α = 10 / 4 = 2.5
= "Species composition changes 2.5 times across sites"
```

The Three Beta Diversity Methods Compared

Method	Data Used	Range	Best For
Hill Numbers (β = γ/α)	Abundance	1 to ∞	Understanding how many distinct communities exist
Jaccard	Presence/Absence	0 to 1	Comparing which species are present
Bray-Curtis	Abundance	0 to 1	Detecting abundance-driven differences

Example Output Interpretation

Scenario 1: Low Beta Diversity (Homogeneous Community)

```
Year: 2015
gamma_q0 = 25 (total species)
mean_alpha_q0 = 20 (avg species per site)
beta_q0 = 1.25 (25/20)

mean_jaccard = 0.15 (low dissimilarity)
mean_braycurtis = 0.20 (low dissimilarity)
```

Interpretation: Sites are very similar! Most species are found at most sites. Low environmental heterogeneity or high dispersal among sites.

Scenario 2: High Beta Diversity (Heterogeneous Community)

```
Year: 2015
gamma_q0 = 50 (total species)
mean_alpha_q0 = 12 (avg species per site)
beta_q0 = 4.17 (50/12)

mean_jaccard = 0.75 (high dissimilarity)
mean_braycurtis = 0.80 (high dissimilarity)
```

Interpretation: Sites are very different! Species composition changes dramatically. Suggests strong environmental gradients, habitat heterogeneity, or dispersal barriers.

Scenario 3: Rare Species Dominate Pattern (q values differ)

```
Year: 2015
beta_q0 = 3.5  (many rare species drive turnover)
beta_q1 = 2.1  (intermediate)
beta_q2 = 1.4  (common species are similar across sites)
```

Interpretation: Many rare species are site-specific, but common/dominant species are shared across sites. This is typical in streams with good overall habitat quality but local microhabitat variation.

Formulas Reference

Hill Numbers

```
q = 0: Species Richness
     = Number of species

q = 1: Hill-Shannon
     = exp(-Σ(pi × ln(pi)))
     where pi = proportion of species i

q = 2: Hill-Simpson
     = 1 / Σ(pi²)
```

Beta Diversity

```
Hill approach:
β = γ / α

Where:
γ = total diversity across all sites
α = average diversity within sites
β = "effective number of communities"
```

Jaccard Index

```
Jaccard dissimilarity = (b + c) / (a + b + c)

Where:
a = species present at both sites
b = species only at site 1
c = species only at site 2
```

Bray-Curtis

```
Bray-Curtis = Σ|xij - xik| / Σ(xij + xik)

Where:
xij = abundance of species i at site j
xik = abundance of species i at site k
```

Common Patterns in Stream Systems

Headwaters → Downstream Gradient

```
Typical pattern:
- High beta diversity (β = 3-5)
- High Bray-Curtis dissimilarity (0.6-0.8)
- q=2 beta lower than q=0 (common species track river continuum)
```

Disturbed vs. Reference Sites

```
Disturbed systems:
- Lower alpha diversity
- Lower gamma diversity
- Sometimes higher beta (patchy degradation)

Reference systems:
- Higher alpha diversity
- Higher gamma diversity
- Moderate beta (natural heterogeneity)
```

Seasonal Variation

```
Spring vs. Fall:
- Often different species pools (gamma)
- May have similar beta patterns
- Life cycle phenology drives differences
```

Quick Diagnostics

Is my analysis working correctly?

✓ **Check 1**: Is gamma ≥ alpha for all years? - If not, check your grouping and calculations

✓ **Check 2**: Is beta ≥ 1 for Hill numbers? - Beta cannot be less than 1 mathematically

✓ **Check 3**: Are Jaccard and Bray-Curtis between 0 and 1? - If not, check for negative abundances or data errors

✓ **Check 4**: Do sample sizes make sense? - Check n_comparisons in output tables - For N sites, expect N(N-1)/2 pairwise comparisons

Red Flags in Your Data

⚠ **Warning signs**: - Beta = 1 exactly (suggests only one site or duplicate sites) - Gamma = Alpha (suggests only one site analyzed) - All dissimilarities = 0 (all sites identical - check data) - All dissimilarities = 1 (no shared species - unusual!)

File Output Guide

alpha_diversity_results.csv

```
Columns:
- SiteID: Unique sample identifier
- q0_richness: Number of species (Hill q=0)
- q1_shannon: Shannon diversity (Hill q=1)
- q2_simpson: Simpson diversity (Hill q=2)
- total_abundance: Total individuals counted
- StationID: Station identifier
- BenSampID: Sample identifier
```

```
- Year: Collection year
- Season: Collection season

Use for:
- Plotting diversity over time
- Comparing sites
- Identifying high/low diversity samples
```

diversity_summary_by_year.csv

```
Columns:
- Year: Collection year
- gamma_q0, gamma_q1, gamma_q2: Regional diversity
- mean_alpha_q0, mean_alpha_q1, mean_alpha_q2: Avg site diversity
- beta_q0, beta_q1, beta_q2: Beta diversity (Hill numbers)
- mean_jaccard: Average Jaccard dissimilarity
- mean_braycurtis: Average Bray-Curtis dissimilarity

Use for:
- Annual trends in diversity
- Comparing beta diversity metrics
- Summary statistics for reports
```

Visualization Ideas

Plot Alpha Diversity Over Time

```r
library(ggplot2)

ggplot(alpha_diversity, aes(x = Year, y = q0_richness)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Species Richness Over Time",
       y = "Alpha Diversity (q=0)",
       x = "Year")
```

Plot Beta Diversity Comparison

```r
beta_long <- diversity_summary %>%
  select(Year, beta_q0, beta_q1, beta_q2) %>%
  pivot_longer(cols = starts_with("beta"),
               names_to = "metric",
               values_to = "beta_diversity")

ggplot(beta_long, aes(x = Year, y = beta_diversity, color = metric)) +
  geom_line() +
  geom_point() +
  labs(title = "Beta Diversity by q Value",
       y = "Beta Diversity",
       x = "Year")
```

Compare Dissimilarity Indices

```
dissim_long <- diversity_summary %>%
  select(Year, mean_jaccard, mean_braycurtis) %>%
  pivot_longer(cols = c(mean_jaccard, mean_braycurtis),
               names_to = "index",
               values_to = "dissimilarity")

ggplot(dissim_long, aes(x = Year, y = dissimilarity, color = index)) +
  geom_line() +
  geom_point() +
  labs(title = "Community Dissimilarity Over Time",
       y = "Mean Dissimilarity",
       x = "Year")
```

Statistical Testing Ideas

Test for temporal trends

```
# Linear model for alpha diversity
model_alpha <- lm(q0_richness ~ Year, data = alpha_diversity)
summary(model_alpha)

# Linear model for beta diversity
model_beta <- lm(beta_q0 ~ Year, data = diversity_summary)
summary(model_beta)
```

Compare seasons

```
# t-test comparing Spring vs Fall
spring_data <- alpha_diversity %>% filter(Season == "Spring")
fall_data <- alpha_diversity %>% filter(Season == "Fall")

t.test(spring_data$q0_richness, fall_data$q0_richness)
```

PERMANOVA for community composition

```
# Test if years differ in community composition
adonis2(species_matrix ~ Year,
        data = site_metadata,
        method = "bray")
```

Troubleshooting Checklist

Problem: Script crashes when loading data - [ ] Are all three CSV files in working directory? - [ ] Check file names match exactly (case-sensitive) - [ ] Try: `list.files()` to see what files R sees

Problem: No results for certain years - [ ] Check date format in station_info file - [ ] Verify: `unique(site_metadata$Year)` - [ ] Look for NA values: `sum(is.na(station_info$Year))`

Problem: Alpha diversity values seem wrong - [ ] Check Individuals column for negative values - [ ] Verify species names in FinalID column - [ ] Look for duplicate taxa: `table(benthics$FinalID)`

Problem: Beta diversity > 10 (seems too high) - [ ] Check if you have very few individuals per site - [ ] Verify gamma calculation is working - [ ] May indicate poor sampling or data quality issues

Citation Template

If using these methods in a publication:

```
We calculated alpha, beta, and gamma diversity using Hill numbers (q = 0, 1, 2) following Chao et al. (2014). Beta diversity was calculated as β = γ/α (Jost 2007). We also calculated pairwise dissimilarity using Jaccard (presence/absence) and Bray-Curtis (abundance-based) indices (Anderson et al. 2011). All analyses were performed in R v4.x using the vegan (Dixon 2003), iNEXT (Hsieh et al. 2016), and betapart (Baselga & Orme 2012) packages.
```