Intro to Biodiversity Analysis

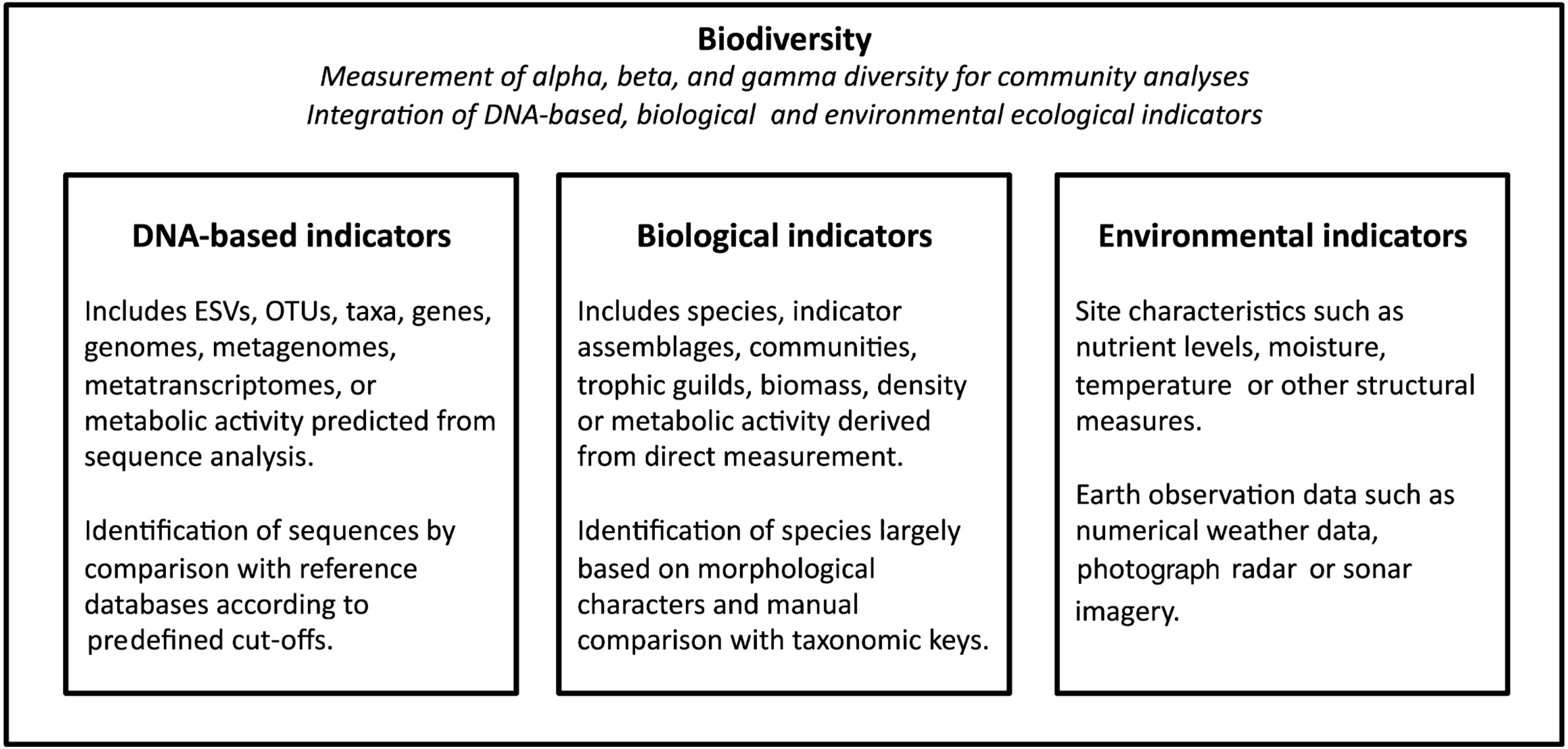
Teresita M. Porter

August 2022

A bit of theory

Biomonitoring

Repeated biodiversity measurements across time and space



Biodiversity

Measurement of alpha, beta, and gamma diversity for community analyses
Integration of DNA-based, biological and environmental ecological indicators

DNA-based indicators

Includes ESVs, OTUs, taxa, genes, genomes, metagenomes, metatranscriptomes, or metabolic activity predicted from sequence analysis.

Identification of sequences by comparison with reference databases according to predefined cut-offs.

Biological indicators

Includes species, indicator assemblages, communities, trophic guilds, biomass, density or metabolic activity derived from direct measurement.

Identification of species largely based on morphological characters and manual comparison with taxonomic keys.

Environmental indicators

Site characteristics such as nutrient levels, moisture, temperature or other structural measures.

Earth observation data such as numerical weather data, photograph radar or sonar imagery.

Porter and Hajibabaei, 2018 Molec Ecol

What is biodiversity?

Typical biodiversity analyses include:

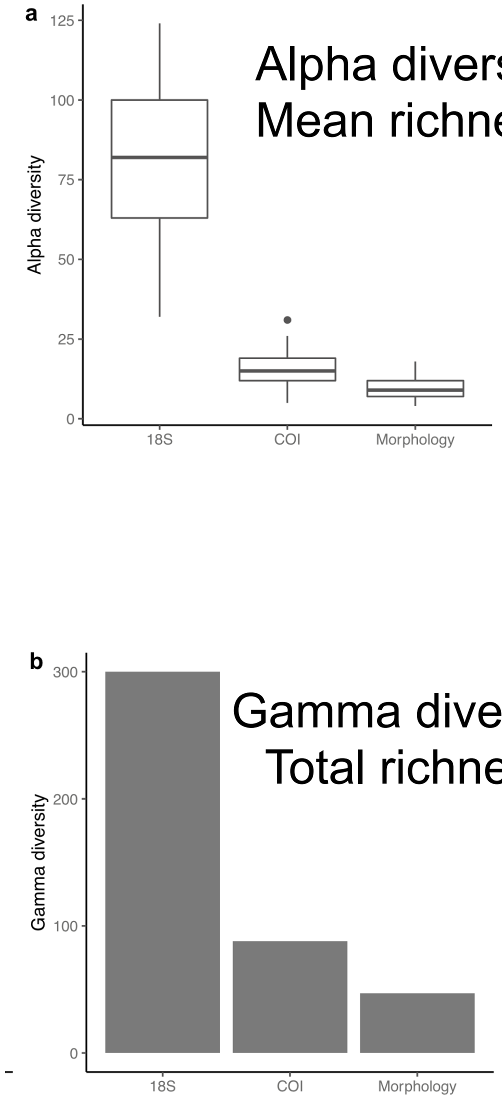
1. **Alpha diversity** (richness) - average number of unique species at local scale
   ex. average number of species per site/treatment/condition
   - gamma diversity is landscape scale diversity, ex. total number of unique species across all sites
   - visualized as box plots, bar plots

2. **Beta diversity** - ratio between regional and local diversity, comparison of diversity between pairs of communities, looks at how beta diversity changes over space/time/treatment/conditions
   ex. diversity indexes such as Bray Curtis/Sorensen/Jaccard dissimilarities/distances, Shannon (evenness), Simpson (richness+evenness), etc
   - visualized using unconstrained NMDS with fitted environmental parameters or constrained RDA (hypothesis-testing) where the PCA is constrained using a few selected env params
   - PERMANOVA (usually accompanies NMDS) can be used to see whether groupings explain a significant amount of variation in beta diversity or ANOVA to see whether variation explained by axes/constraints is significant (usually accompanies RDA)

3. **Community composition** - summarized to a particular taxonomic rank
   - visualized using stacked bar plots, heatmaps

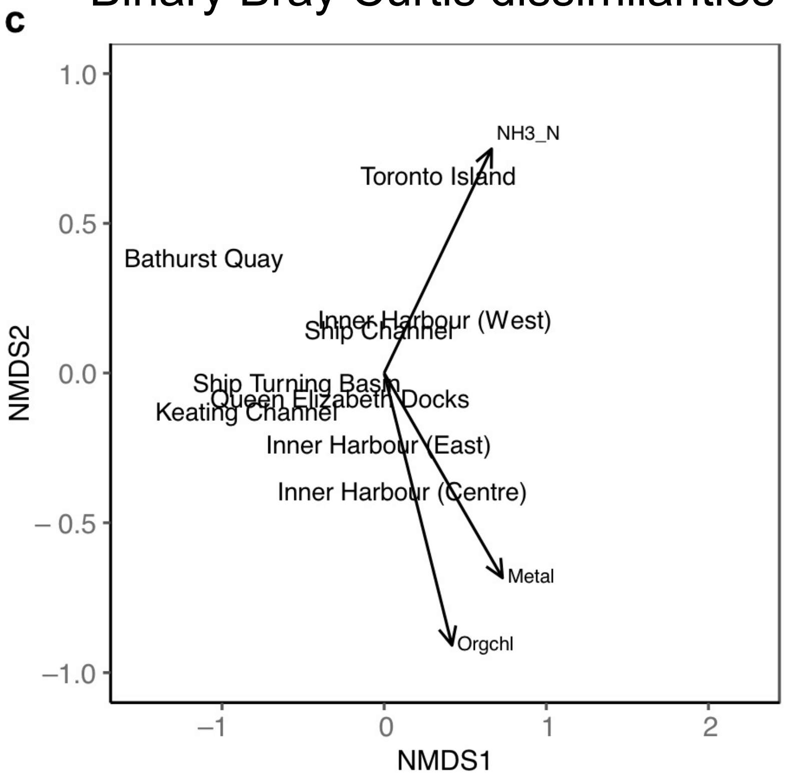
Alpha diversity
Mean richness

Gamma diversity
Total richness

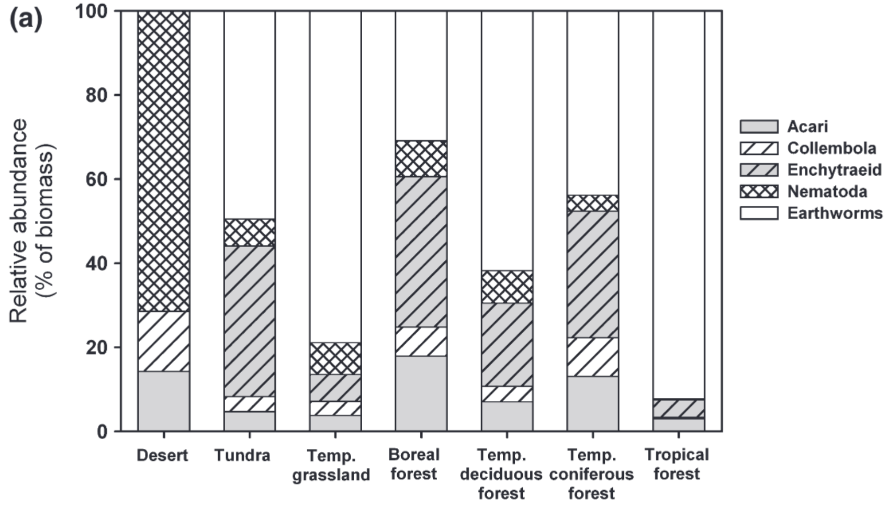


Robinson et al., 2022 Sci Rep

NMDS
Binary Bray Curtis dissimilarities



Community composition



Fierer et al., 2009 Ecol Letters

Related descriptive concepts:

Hill’s numbers / Expected richness, Zeta diversity (new)
Phylogenetic diversity / Unifrac distances
Indicator analysis
Network analysis / trophic analysis
Nestedness analysis

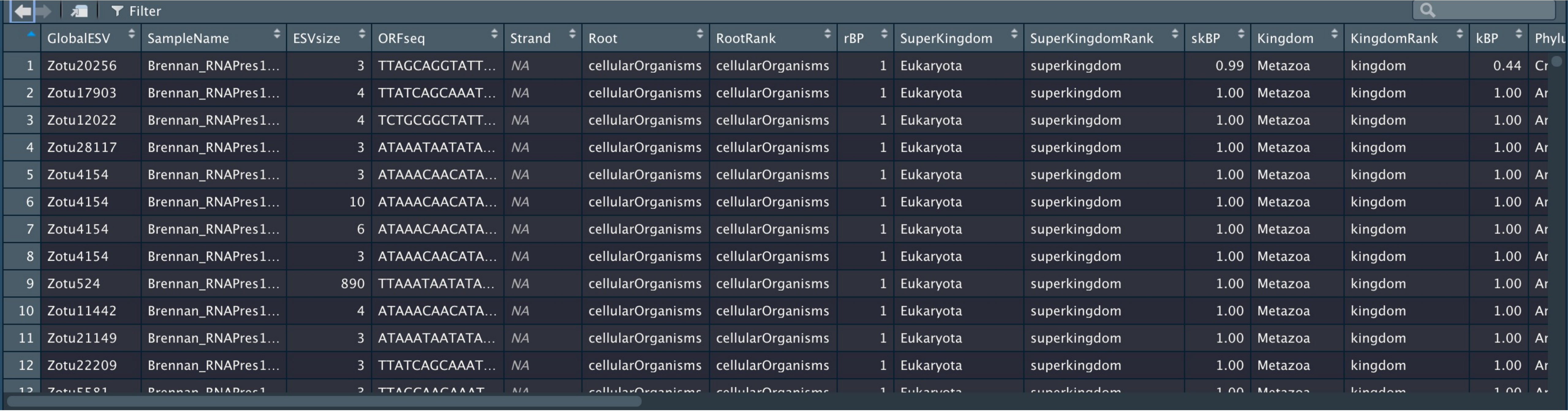
Related predictive methods:

Regression/Hierarchical partitioning/GLMs
Simper

Preparation for exploratory analysis

Prep work for exploratory analysis:

raw sequence data + primer sequences -> MetaWorks bioinformatic processing -> results.csv



	GlobalESV	SampleName	ESVsize	ORFseq	Strand	Root	RootRank	rBP	SuperKingdom	SuperKingdomRank	skBP	Kingdom	KingdomRank	kBP	Phylu
1	Zotu20256	Brennan_RNAPres1...	3	TTAGCAGGTATT...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	0.99	Metazoa	kingdom	0.44	Cr
2	Zotu17903	Brennan_RNAPres1...	4	TTATCAGCAAAT...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
3	Zotu12022	Brennan_RNAPres1...	4	TCTGCGGCTATT...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
4	Zotu28117	Brennan_RNAPres1...	3	ATAAATAATATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
5	Zotu4154	Brennan_RNAPres1...	3	ATAAACAACATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
6	Zotu4154	Brennan_RNAPres1...	10	ATAAACAACATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
7	Zotu4154	Brennan_RNAPres1...	6	ATAAACAACATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
8	Zotu4154	Brennan_RNAPres1...	3	ATAAACAACATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
9	Zotu524	Brennan_RNAPres1...	890	TTAAATAATATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
10	Zotu11442	Brennan_RNAPres1...	4	ATAAACAACATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
11	Zotu21149	Brennan_RNAPres1...	3	ATAAATAATATA...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
12	Zotu22209	Brennan_RNAPres1...	3	TTATCAGCAAAT...	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar
13	Zotu5581	Brennan_RNAPres1	3	TTAGCAACAAAT	NA	cellularOrganisms	cellularOrganisms	1	Eukaryota	superkingdom	1.00	Metazoa	kingdom	1.00	Ar

Prep work for exploratory analysis:

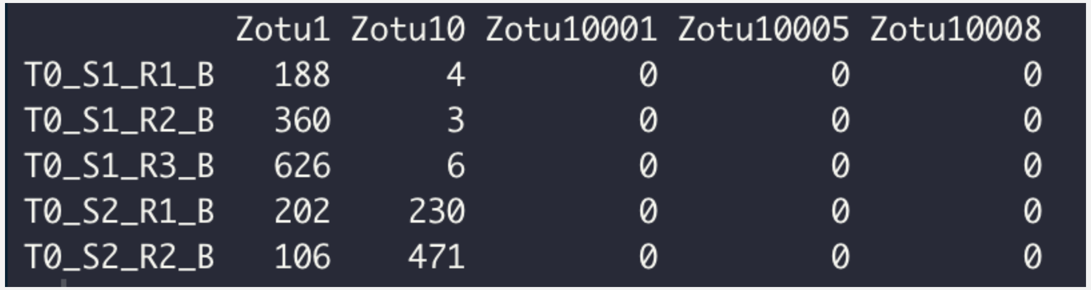
1. Extract ESV table from results.csv

```
#read in MetaWorks results
COI <- read.csv("results.csv", header=TRUE, stringsAsFactors=FALSE)

# see the top of the COI object (results file)
head(COI)[1:5,1:5]

# Create pivot table using reshape2 library
ESV.table <- reshape2::dcast(COI, SampleName ~ GlobalESV, value.var =
"ESVsize", fun.aggregate = sum)

# see the top of the table
head(ESV.table)[1:5,1:5]
```

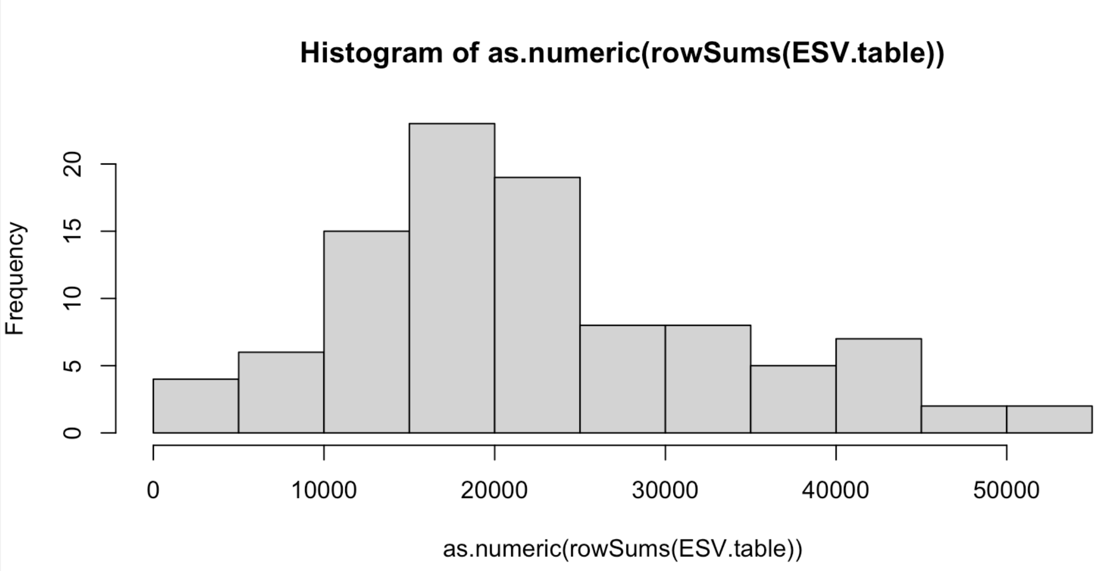


```
           Zotu1 Zotu10 Zotu10001 Zotu10005 Zotu10008
T0_S1_R1_B   188      4         0         0         0
T0_S1_R2_B   360      3         0         0         0
T0_S1_R3_B   626      6         0         0         0
T0_S2_R1_B   202    230         0         0         0
T0_S2_R2_B   106    471         0         0         0
```

Prep work for exploratory analysis:

2. Remove under-sequenced samples
   ex. Remove samples with less than 10,000 reads

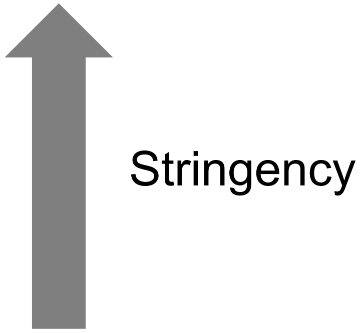
```
# visualize read depth distribution
hist(as.numeric(rowSums(ESV.table)))
```



```
# remove samples with < 10,000 reads
ESV.table2 <- ESV.table[!rowSums(ESV.table) < 10000,]
```

Prep work for exploratory analysis:

3. Remove rare ESVs
   - ex. filter ESV table to remove ESVs that represent less than 0.0001 or 0.01% of total reads
     - Set a cutoff to compensate for the the rate of expected index-hopping/tag-switching (Schnell et al. 2015, 2.5-2.7%; Elbrecht & Leese papers 0.01%)
     - Remove ESVs with < 8 reads (default setting in USEARCH)
     - Remove ESVs with < 3 reads because these rare clusters tend to contain poor-quality artefactual sequences (Tedersoo et al., et al., 2010 New Phytologist; Zhan et al., 2014 PLoS ONE)



Stringency

```
# remove ESVs that represent less than 0.0001 or 0.01% of all reads
cutoff <- sum(colSums(ESV.table2)) * 0.0001
ESV.table3 <- ESV.table2
ESV.table3[colSums(ESV.table2) < cutoff] <- NULL
```

Prep work for exploratory analysis:

4. (Optional) Remove infrequent ESVs (especially important for network analysis but can be done on large datasets to reduce the dataset size)

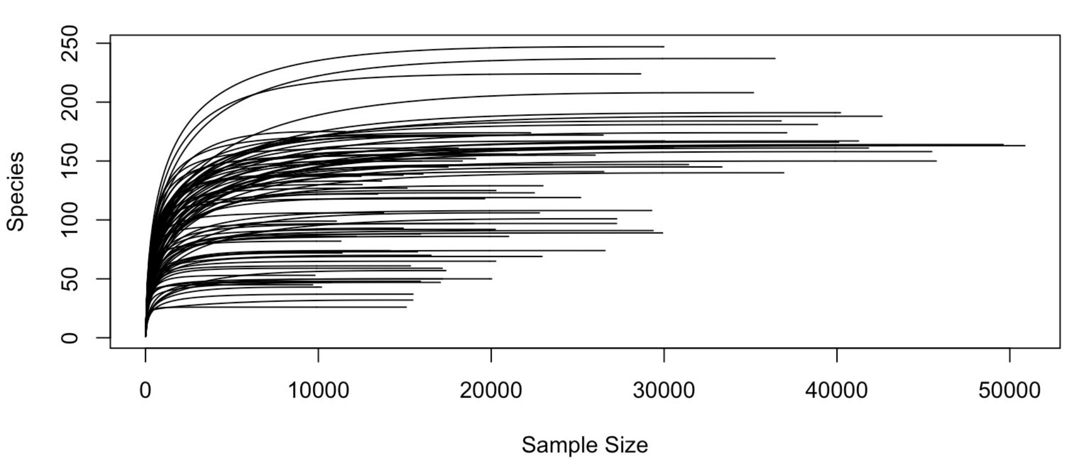
    - ex. remove ESVs if they are not present in at least 1/treatments of the samples

Prep work for exploratory analysis:

5. Plot rarefaction curves to assess that sequencing effort was sufficient
   - Do curves reach a plateau?
     - yes - may not need to rarefy, just normalize for multivariate analyses by converting read counts to proportions (reads in ESV / total reads in sample) or Hellinger transform
     - no - rarefy to lowest number of reads per sample (old school) or 15th percentile (my preference)

       - randomly subsample “x” number of reads per sample (ex, 1,000 reads per sample; old school, don’t do this)

```
# visualize curves
rc <- rarecurve(ESV.table3, step=100, label=FALSE)
```



Prep work for exploratory analysis:

6. Filter taxonomic assignments by the appropriate bootstrap support cutoffs to ensure a certain level of accuracy (depends on marker type, amplicon length, taxonomic rank)

```
#read in MetaWorks results
COI <- read.csv("results.csv", header=TRUE, stringsAsFactors=FALSE)

# filter COI taxonomic assignments for 95% correct species, 99% correct genus & up
# See https://github.com/terrimporter/CO1Classifier for cutoffs
COI$Species <- ifelse(COI$sBP >= 0.7, COI$Species, "")
COI$Genus <- ifelse(COI$gBP >= 0.3, COI$Genus, "")
COI$Family <- ifelse(COI$fBP >= 0.2, COI$Family, "")
```

Summary

Exploratory analysis:

1. Richness -> box plots
2. Beta diversity -> NMDS, fitted env vars
3. Community composition -> stacked bar charts or heatmaps summarized to ex. phyla (bacteria), class (fungi), order (arthropods)

Next steps:

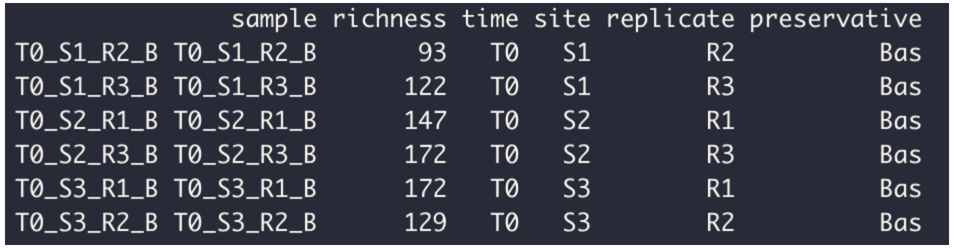
Address the specific hypotheses/objectives for your project using whatever descriptive/predictive analyses are most appropriate

Sample results

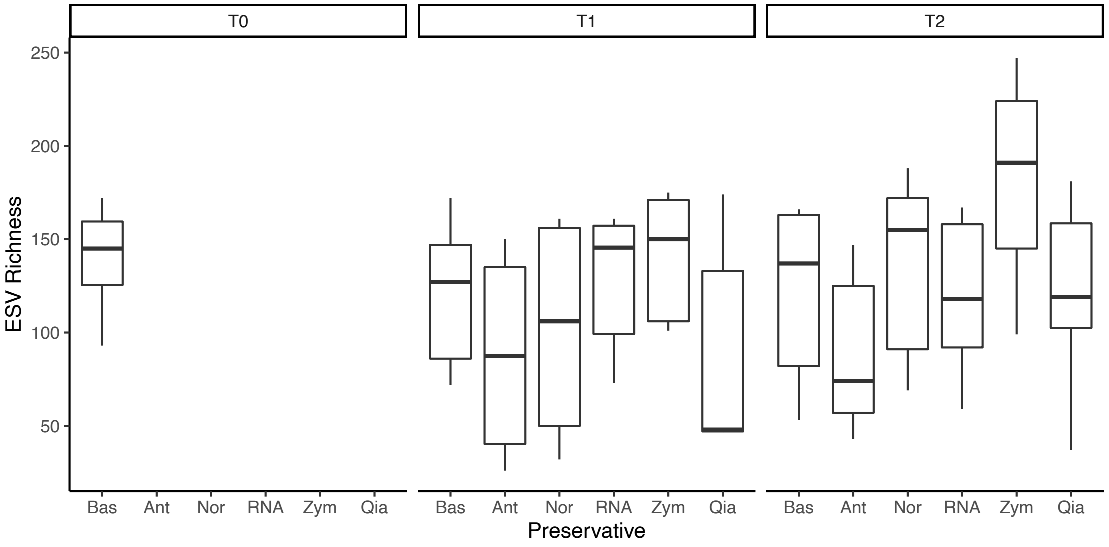
Richness

```
# ESV.table3 under-sequenced samples removed, rare ESVs removed
t <- read.csv("ESVtable.csv", header=TRUE, row.names=1, stringsAsFactors = FALSE)

# ESV Richness ----
r <- data.frame(sample=rownames(t), richness=specnumber(t))
```



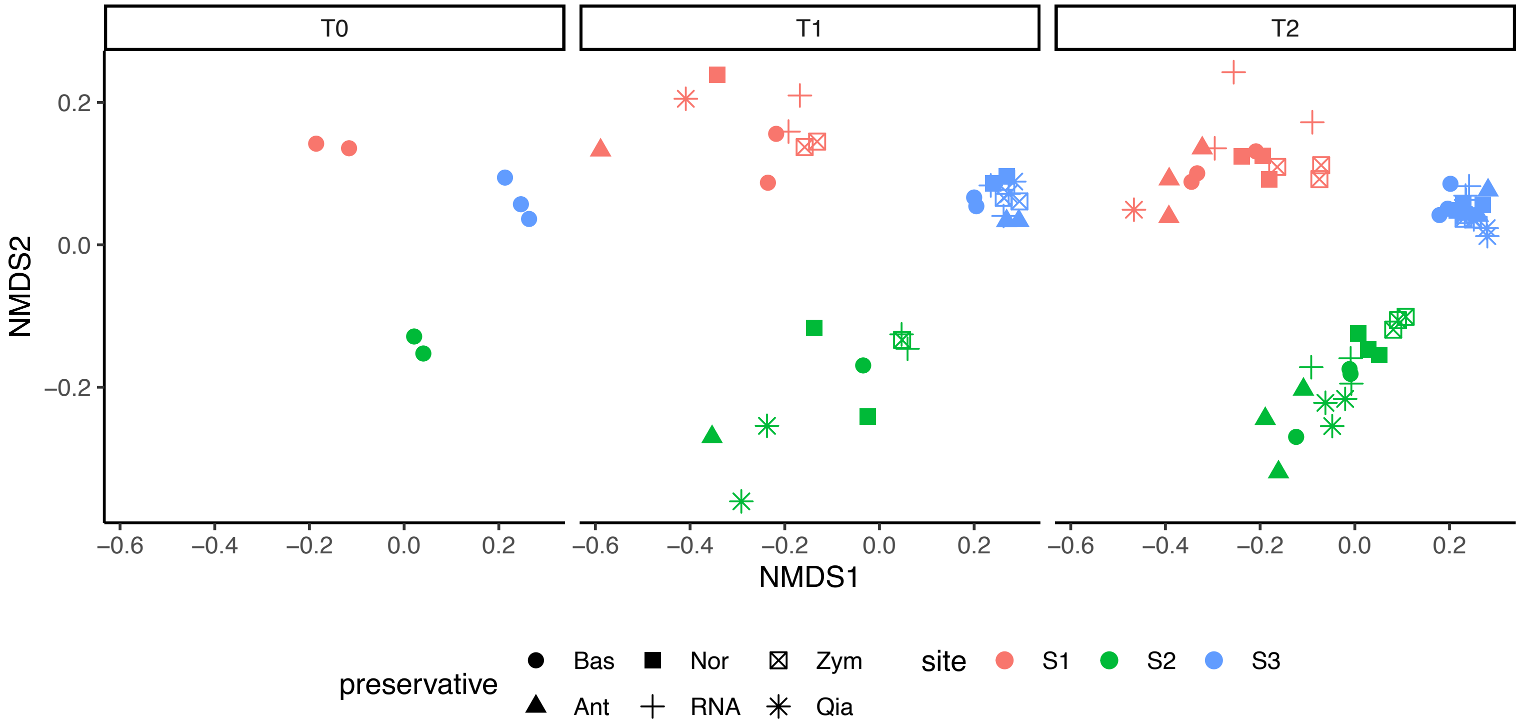
```
               sample richness time site replicate preservative
T0_S1_R2_B T0_S1_R2_B       93   T0   S1        R2          Bas
T0_S1_R3_B T0_S1_R3_B      122   T0   S1        R3          Bas
T0_S2_R1_B T0_S2_R1_B      147   T0   S2        R1          Bas
T0_S2_R3_B T0_S2_R3_B      172   T0   S2        R3          Bas
T0_S3_R1_B T0_S3_R1_B      172   T0   S3        R1          Bas
T0_S3_R2_B T0_S3_R2_B      129   T0   S3        R2          Bas
```



Beta diversity

```
# binary Bray Curtis = Sorensen dissimilarity (presence-absence)
m <- vegdist(t, method="bray", binary=TRUE)

# Do 3 dimensional NMDS
nmds3 <- metaMDS(m, k=3, trymax=100)
```



Site explains 50% (p = 0.001) and preservative explains 8.5% (p = 0.04) of the variation in beta diversity due to both variation within and among groups.

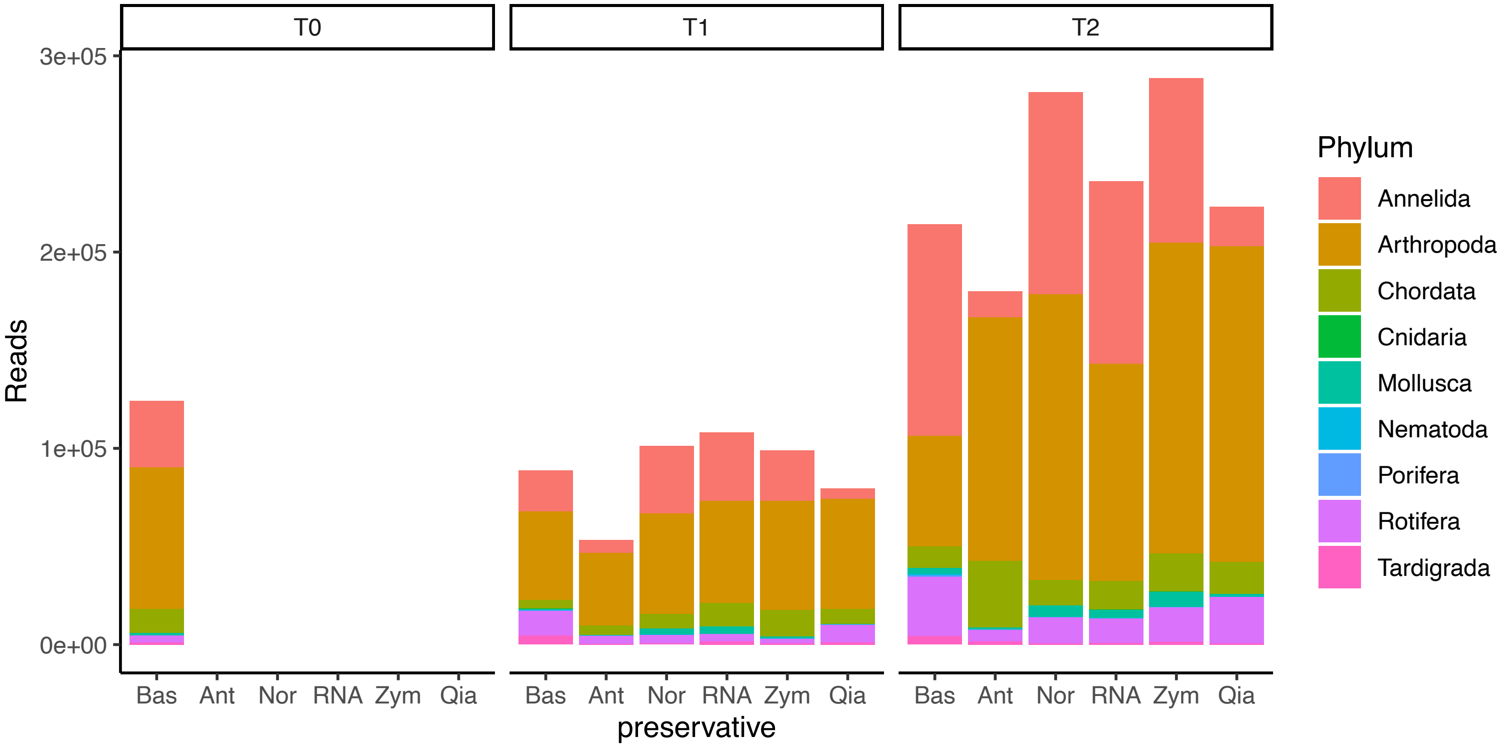
Time does not explain a significant amount of variation in beta diversity.

Stress = 0.07, Linear R2 = 0.976

Community composition

```
# read in ESV table and fix up formatting
# read in taxonomy and fix up formatting
# merge ESV and taxonomy

# summarize community composition at the phylum rank
gg <- data.frame(tab2 %>% group_by(sample,Phylum) %>% dplyr::summarize(sum(reads)))
```



Resources

For users new to RStudio & biodiversity analysis

STREAM data workshop presentation by Wendy Monk
https://youtu.be/aQGKXHrxaiw?t=4908

Sample scripts

Available from https://github.com/terrimporter/IntroBiodiversityAnalysis2022

Big biodiversity papers

Fierer, N., Strickland, M. S., Liptzin, D., Bradford, M. A., & Cleveland, C. C. (2009). Global patterns in belowground communities. *Ecology Letters*, *12*(11), 1238–1249. doi: 10.1111/j.1461-0248.2009.01360.x

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, *405*(6783), 212–219. doi: 10.1038/35012221

Tedersoo, L., Bahram, M., Polme, S., **Koljalg**, U., Yorou, N. S., Wijesundera, R., … **Abarenkov**, K. (2014). Global diversity and geography of soil fungi. *Science*, *346*(6213), 1256688–1256688. doi: 10.1126/science.1256688

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., **Knight**, R., **Gilbert**,J.A., Zhao, H. (2017). A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature*. doi: 10.1038/nature24621