Biodiversity analysis in R

Steven Kembel

UQAM

steve.kembel@gmail.com

General background

In this workshop we are going to analyze a data set on the biodiversity of grassland plants in Alberta. This data set consists of data on the occurrence of grassland plants at several different sites in Alberta, along with information on their functional traits and phylogenetic relationships.

I described this data set in more detail in a recent paper: S.W. Kembel and J.F. Cahill, Jr. 2011. Independent evolution of leaf and root traits within and among temperate grassland plant communities. PLoS ONE 6(6): e19992. (doi:10.1371/journal.pone.0019992).

How to use these workshop materials

This workshop will walk through the process of loading and analyzing biodiversity data in R. If you want to work through the entire workshop you can follow along from the beginning. If you want to jump in to try an analysis at any point in the workshop, make sure you have loaded the picante package and the workspace image that contains all of the data files by running the following commands.

```
library(picante)
```

```
load("R_biodiversity_workspace.RData")  # or replace filename with
file.choose()
```

Getting biodiversity data into R

The first thing we need to do is import all the data we need into R.

We will want to make sure the different packages we are going to use are loaded. We will be using functions from the **ape**, **picante**, and **vegan** packages today. Since **picante** depends on the other two packages, loading it will load the other two as well.

```
library(picante)
```

```
## Loading required package: ape
## Loading required package: vegan
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.0-9
## Loading required package: nlme
```

To make it easier to load files, we can set our working directory to the folder containing the grassland data. The exact format of a filename will vary depending on your operating system. The format below works for Mac or Linux (although you'll need to change the location to wherever you put the files on your system). For Windows the file naming convention is different, a typical location would be something like "c://Documents//grassland_data".

```
setwd("/Users/steve/Dropbox/work/R_Workshop_files/biodivR/")
```

Also, remember that you could use the `file.choose()` function for each of the file-reading commands in this tutorial to interactively select files to load, rather than setting the working directory and writing out the filenames.

Community data

Ecological community data consist of observations of the (relative) abundance of species in different samples. In our case, the abundance measure is percent cover of different plant species in 20x20m quadrats in grasslands in different habitat types.

The format for community data is a data.frame with samples in the rows and species in the columns. Our data are already in this format so we can load them using the following command. Note that since we've set our working directory to the folder containing all the data files, we just have to type the filename.

```
# read community data use plot IDs as rownames (first column of
data) use
# species names as colnames (default read.csv is header=TRUE)
replace
# filename with file.choose() to open interactive window
comm <- read.csv("grassland_data/grassland.community.csv", header =
TRUE, row.names = 1)
```

By reading the data in this way, we have set the species names as the column names, and the sample names as the row names. This is important to note - we didn't load these labels in as data - they are the *names* of the rows/columns. Later this will make it easier for us to link different data sets. Let's check to make sure our rows and columns have reasonable-looking names.

```
class(comm)
```

```
## [1] "data.frame"
```

```
# get the dimension of the community object (rows x columns)
dim(comm)
```

```
## [1] 27 76
```

```
rownames(comm)
```

```
##  [1] "mix-O-1"  "mix-O-2"  "mix-O-3"  "mix-O-4"  "mix-O-5"
"mix-O-6"
##  [7] "mix-O-7"  "fes-K-8"  "fes-K-9"  "fes-K-10" "fes-K-11"
"fes-K-12"
## [13] "fes-K-13" "fes-K-14" "fes-K-15" "fes-K-16" "fes-K-17"
"mix-H-18"
## [19] "mix-H-19" "mix-H-20" "mix-H-21" "mix-H-22" "mix-H-23"
"mix-H-24"
## [25] "mix-H-25" "mix-H-26" "mix-H-27"
```

```
head(colnames(comm))
```

```
## [1] "Antennaria_parvifolia"
## [2] "Artemisia_cana"
## [3] "Artemisia_frigida"
## [4] "Symphyotrichum_ericoides_var._ericoides"
## [5] "Bouteloua_gracilis"
## [6] "Carex_filifolia"
```

```
# take a peek at the data (just the first five rows/columns)
comm[1:5, 1:5]
```

```
##         Antennaria_parvifolia Artemisia_cana Artemisia_frigida
## mix-O-1                    10             10                50
## mix-O-2                     0             10                50
## mix-O-3                    20             20                30
## mix-O-4                     0              0                 0
## mix-O-5                     0             10                 0
##         Symphyotrichum_ericoides_var._ericoides
Bouteloua_gracilis
## mix-O-1                                      10
70
## mix-O-2                                      10
90
## mix-O-3                                      10
60
## mix-O-4                                       0
90
## mix-O-5                                       0
100
```

Each cell contains the percent cover of a species in a sample. Many multivariate methods are sensitive to the total abundance in a sample, so we should probably convert these absolute abundance estimates to a relative abundance estimate. We can do this with a function from the **vegan** package.

```
# check total abundance in each sample
apply(comm, 1, sum)
```

```
##  mix-O-1  mix-O-2  mix-O-3  mix-O-4  mix-O-5  mix-O-6  mix-O-7
fes-K-8
##      640      630      710      350      400      650      560
960
##  fes-K-9 fes-K-10 fes-K-11 fes-K-12 fes-K-13 fes-K-14 fes-K-15
fes-K-16
##      960      980      830      980      980      830      640
1080
## fes-K-17 mix-H-18 mix-H-19 mix-H-20 mix-H-21 mix-H-22 mix-H-23
mix-H-24
##      710      440      590      540      340      420      400
600
## mix-H-25 mix-H-26 mix-H-27
##      540      590      420
```

```
# Turn percent cover to relative abundance by dividing each value
by sample
# total abundance
comm <- decostand(comm, method = "total")
# check total abundance in each sample
apply(comm, 1, sum)
```

```
##  mix-O-1  mix-O-2  mix-O-3  mix-O-4  mix-O-5  mix-O-6  mix-O-7
fes-K-8
##        1        1        1        1        1        1        1
1
##  fes-K-9 fes-K-10 fes-K-11 fes-K-12 fes-K-13 fes-K-14 fes-K-15
fes-K-16
##        1        1        1        1        1        1        1
1
## fes-K-17 mix-H-18 mix-H-19 mix-H-20 mix-H-21 mix-H-22 mix-H-23
mix-H-24
##        1        1        1        1        1        1        1
1
## mix-H-25 mix-H-26 mix-H-27
##        1        1        1
```

```
# look at the transformed data
comm[1:5, 1:5]
```

```
##         Antennaria_parvifolia Artemisia_cana Artemisia_frigida
## mix-O-1               0.01562        0.01562            0.07812
## mix-O-2               0.00000        0.01587            0.07937
## mix-O-3               0.02817        0.02817            0.04225
## mix-O-4               0.00000        0.00000            0.00000
## mix-O-5               0.00000        0.02500            0.00000
##         Symphyotrichum_ericoides_var._ericoides
Bouteloua_gracilis
## mix-O-1                                 0.01562
0.10938
## mix-O-2                                 0.01587
0.14286
## mix-O-3                                 0.01408
0.08451
## mix-O-4                                 0.00000
0.25714
## mix-O-5                                 0.00000
0.25000
```

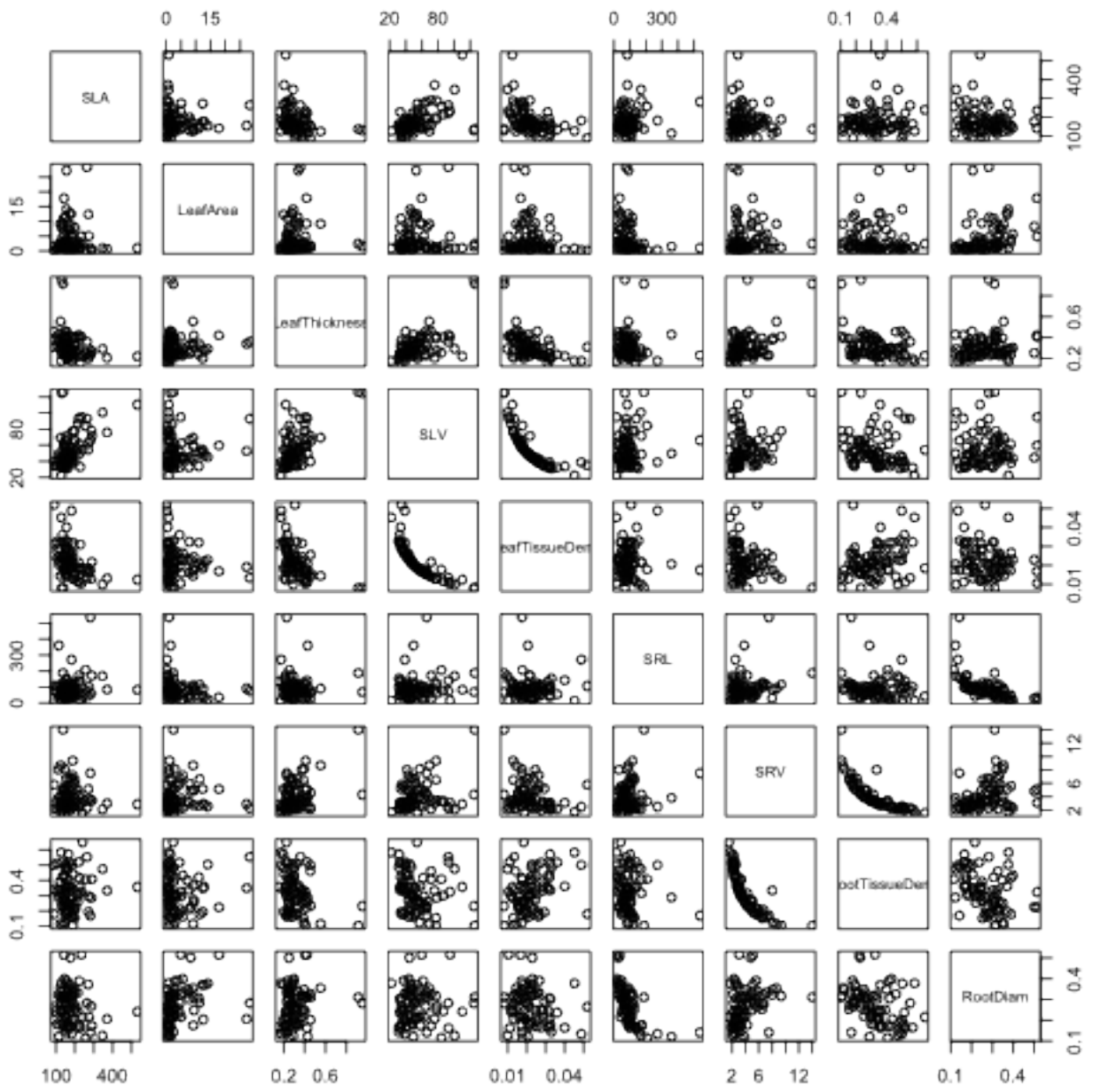
Trait data

We also have information on the leaf and root traits of each species. We can load these data in the same way as the community data, but now we will have species in the rows and traits in the columns.

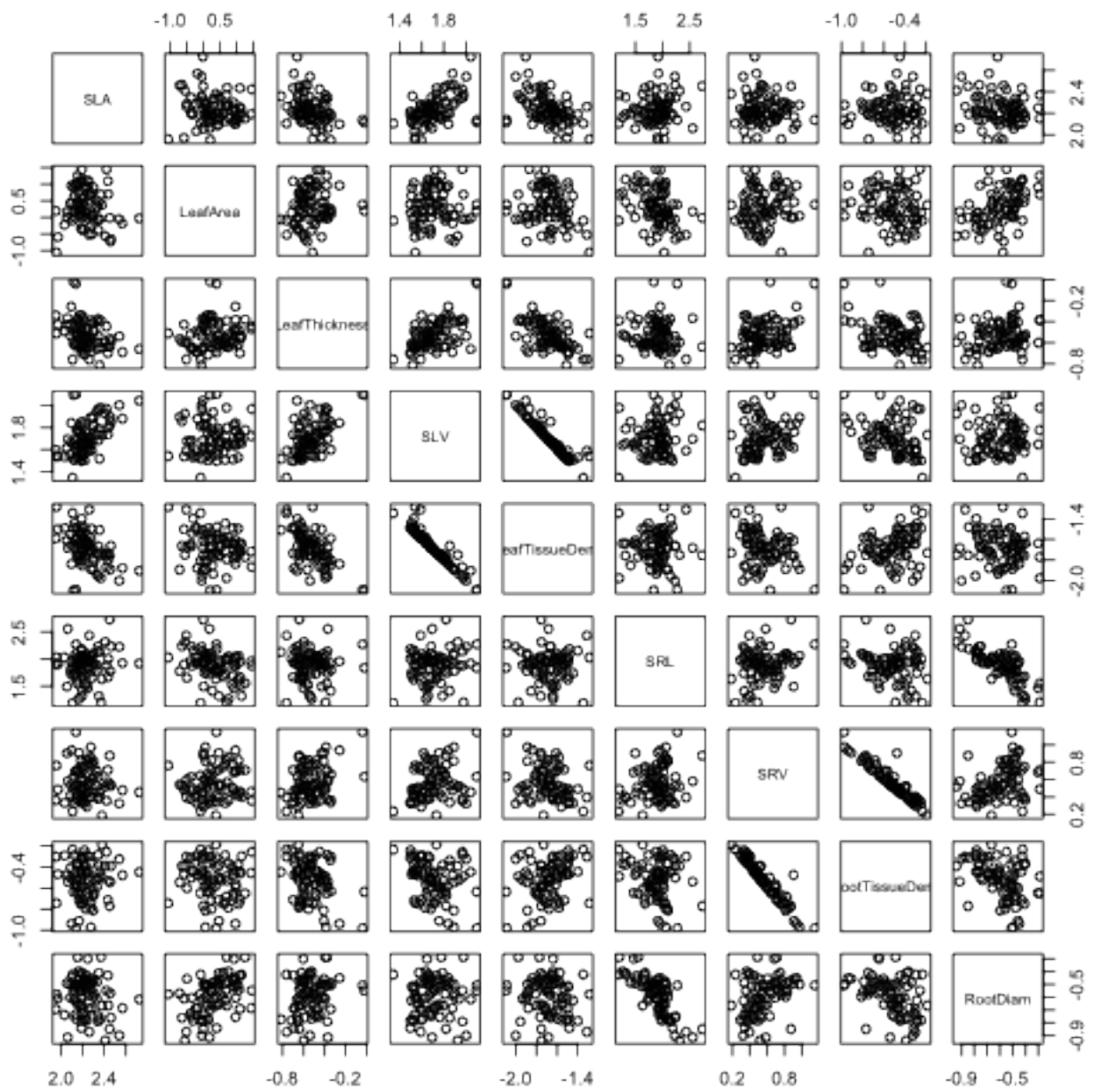
```
# replace filename with file.choose() to open interactive window
traits <- read.csv("grassland_data/species.traits.csv", header =
TRUE, row.names = 1)
# take a peek at the data
head(traits)
```

```
##                            SLA LeafArea LeafThickness    SLV
LeafTissueDens
## Achillea_millefolium   140.3   9.2754        0.4163  59.57
0.018085
## Allium_textile         137.7   2.4454        0.9147 125.69
0.008137
## Amelanchier_alnifolia  156.1  14.0649        0.2900  45.45
0.022841
## Androsace_occidentalis 257.2   0.2747        0.2535  84.22
0.017706
## Antennaria_neglecta    171.0   1.7320        0.2810  48.14
0.020920
## Antennaria_parvifolia  193.9   0.3172        0.2467  47.64
0.021048
##                              SRL    SRV RootTissueDens RootDiam
## Achillea_millefolium    74.15  5.039          0.2554   0.3124
## Allium_textile         187.85 14.014          0.1050   0.3108
## Amelanchier_alnifolia   20.88  2.519          0.5040   0.3761
## Androsace_occidentalis 207.46  3.292          0.4071   0.1149
## Antennaria_neglecta    124.73  6.711          0.1594   0.2750
## Antennaria_parvifolia   44.94  4.004          0.2504   0.3496
```

```
# plot the data
pairs(traits)
```




```
# some variables look skewed - log transform all variables
traits <- log10(traits)
# plot the transformed data
pairs(traits)
```



Metadata

We have some information about the samples, including the habitat and site they were collected from, and a few basic environmental variables such as slope and moisture regime.

```
# replace filename with file.choose() to open interactive window
metadata <- read.csv("grassland_data/plot.metadata.csv", header =
TRUE, row.names = 1)
# take a peek at the data
head(metadata)
```

```
##            habitat    site slope aspect slope.position
rel.moisture
## mix-O-1 Mixedgrass Onefour     0    270              3.0
1
## mix-O-2 Mixedgrass Onefour    20    130              1.5
2
## mix-O-3 Mixedgrass Onefour     5     90              1.0
2
## mix-O-4 Mixedgrass Onefour     5     40              2.0
1
## mix-O-5 Mixedgrass Onefour     5    130              2.0
1
## mix-O-6 Mixedgrass Onefour     1     90              3.0
1
```

Phylogeny

If you have a phylogeny in the commonly used Newick or Nexus format it can be imported into R with the read.tree or read.nexus functions.

```
# replace filename with file.choose() to open interactive window
phy <- read.tree("grassland_data/grassland.phylogeny.newick")
class(phy)
```

```
## [1] "phylo"
```

```
phy
```

```
##
## Phylogenetic tree with 76 tips and 68 internal nodes.
##
## Tip labels:
##  Antennaria_neglecta, Antennaria_parvifolia, Erigeron_glabellus,
Erigeron_pumilus, Heterotheca_villosa,
Symphyotrichum_falcatum_var._falcatum, ...
## Node labels:
##  , , , , , , ...
##
## Rooted; includes branch lengths.
```

Our phylogeny is a special object of type `phylo`. The `phylo` format itself is documented at the **ape** homepage (http://ape.mpl.ird.fr/). A `phylo` object is a special type of `list` object - it has different elements such as tip labels and edge lengths, and R knows how to summarize and plot a `phylo` object due to the way it is defined by the **ape** package.

```
# list the elements of our phylogeny
names(phy)
```

```
## [1] "edge"        "Nnode"       "tip.label"   "edge.length"
"node.label"
## [6] "root.edge"
```

```
# what are the first few tip labels?
phy$tip.label[1:5]
```

```
## [1] "Antennaria_neglecta"   "Antennaria_parvifolia"
"Erigeron_glabellus"
## [4] "Erigeron_pumilus"      "Heterotheca_villosa"
```

```
# how many tips does our phylogeny have?
Ntip(phy)
```

```
## [1] 76
```

```
# plot our phylogeny (the cex argument makes the labels small
enough to
# read)
plot(phy, cex = 0.5)
```

Cleaning and matching data sets

Our workspace contains the community, trait, phylogeny, and metadata that we will need for our analyses.

```
ls()
```

```
## [1] "comm"     "metadata" "phy"      "traits"
```

The data sets we are using today have already been cleaned up so that they contain the same species and the same samples, but often when we are working with our own data, there will be mismatches among different types of data. For example, our community data might only contain a subset of the species in our phylogeny, or there might be some species for which we have trait information but no phylogenetic information. For some analyses, R will assume that species are in the same order in both the community data set and the phylogeny. Sometimes there might be a typo in the labels for a dataset, and we will want to catch those before proceeding.

There are several functions in **picante** that are designed to make sure different data sets match with one another. We should check that our phylogeny and community contain the same species, and that they are in the same order. The `match.phylo.comm` takes a community object and a phylo object, reports any species that are not present in both data sets, and outputs a version of each object in the same order and containing the same species.

```
# check for mismatches/missing species
combined <- match.phylo.comm(phy, comm)
# the resulting object is a list with $phy and $comm elements.
replace our
# original data with the sorted/matched data
phy <- combined$phy
comm <- combined$comm
```

We should do the same matching for our trait data.

```
combined <- match.phylo.data(phy, traits)
# the resulting object is a list with $phy and $data elements.
replace our
# original data with the sorted/matched data
phy <- combined$phy
traits <- combined$data
```

We should also check whether our community data and metadata are in the same order.

```
all.equal(rownames(comm), rownames(metadata))
```

```
## [1] TRUE
```

```
# they all match - if they didn't we could sort them to the same
order sort
# metadata rows to be in the same order as community rows
metadata <- metadata[rownames(comm), ]
```

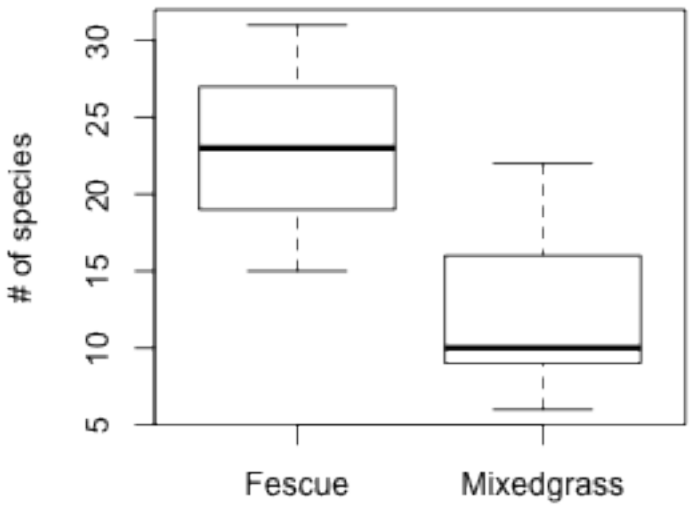
We're done! All of our data are now ready for analysis. In each of the sections below we will explore different ways of analyzing the biodiversity of plants in these grasslands.

Visualizing and summarizing biodiversity data

Community richness and diversity

At a most basic level, we can ask about the overall taxonomic diversity of these grasslands. How many plant species are there? Do habitats differ in species richness?

```
# compare species richness between fescue and mixedgrass habitats
boxplot(specnumber(comm) ~ metadata$habitat, ylab = "# of species")
```

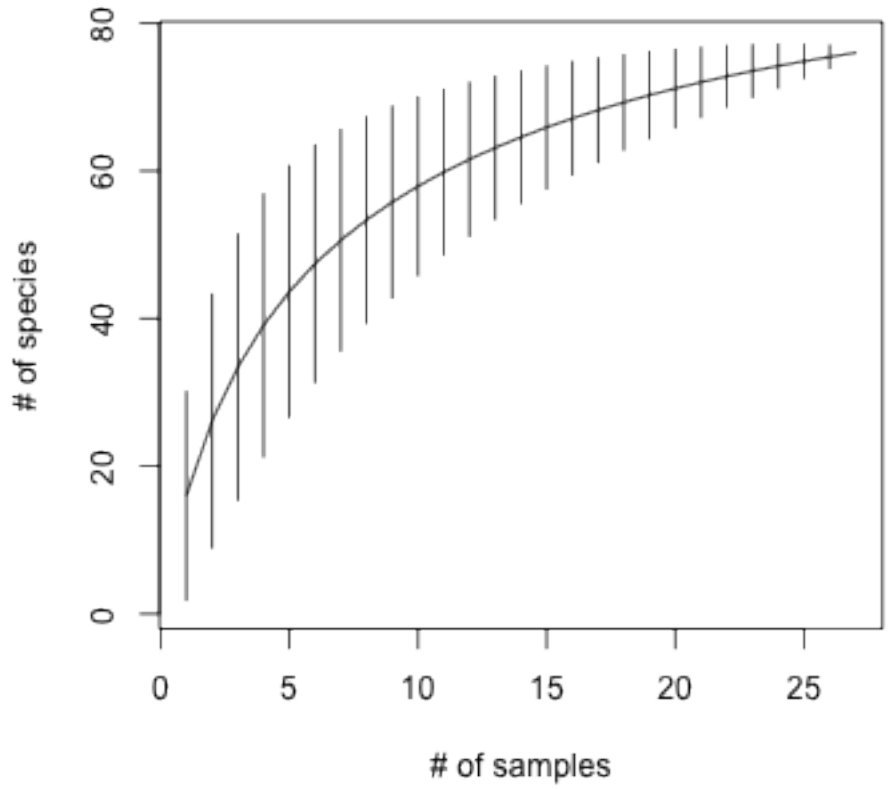



```
# statistical test of difference
t.test(specnumber(comm) ~ metadata$habitat)
```

```
##
##  Welch Two Sample t-test
##
## data:  specnumber(comm) by metadata$habitat
## t = 5.137, df = 17.18, p-value = 7.972e-05
## alternative hypothesis: true difference in means is not equal to
0
## 95 percent confidence interval:
##   6.274 15.008
## sample estimates:
##     mean in group Fescue mean in group Mixedgrass
##                    22.70                    12.06
```

Did we do a good job of sampling the diversity that is out there? We can look at a collector's curve to assess this.

```
# plot species accumulion curve across samples
plot(specaccum(comm), xlab = "# of samples", ylab = "# of species")
```



Multivariate community analysis

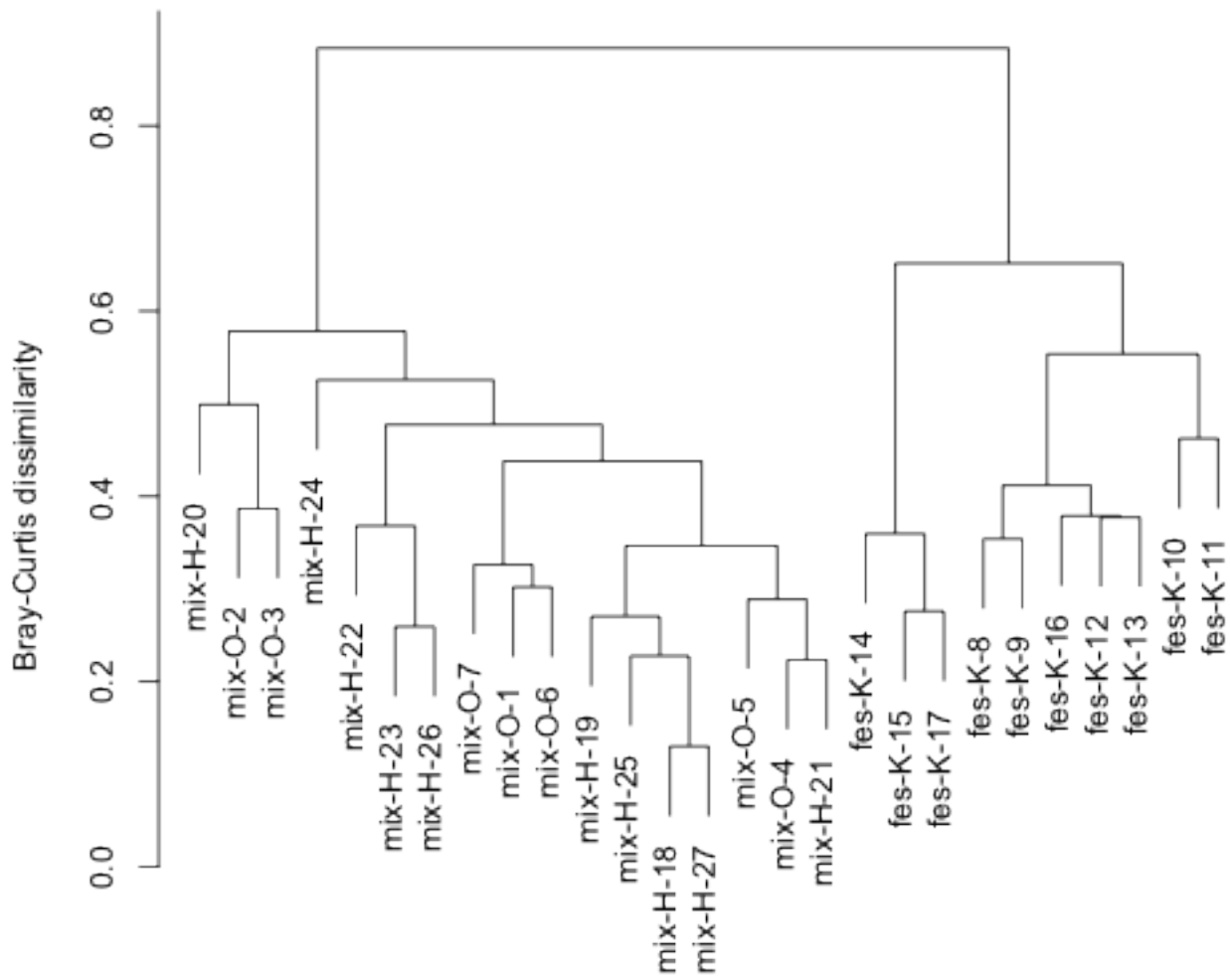
How does the composition of plant communities vary across different samples? How are habitat type and environmental variables related to plant community composition?

We can use multivariate ordination methods to explore community structure in more detail. These methods are available in the **vegan** package, which also includes excellent documentation and tutorials for these methods. The book “Numerical Ecology in R” by Borcard et al. gives a great overview of multivariate analysis methods.

Hierarchical clustering

We can cluster together plots based on their overall community composition. We will calculate Bray-Curtis dissimilarity among all the samples, an abundance-weighted measure of how similar two communities are in terms of their species composition. We will then cluster together communities that are similar using an agglomerative hierarchical clustering algorithm.

```
# calculate Bray-Curtis distance among samples
comm.bc.dist <- vegdist(comm, method = "bray")
# cluster communities using average-linkage algorithm
comm.bc.clust <- hclust(comm.bc.dist, method = "average")
# plot cluster diagram
plot(comm.bc.clust, ylab = "Bray-Curtis dissimilarity")
```



It looks like mixedgrass and fescue habitats contain different plant community types - the two main clusters separate fescue samples from all other samples.

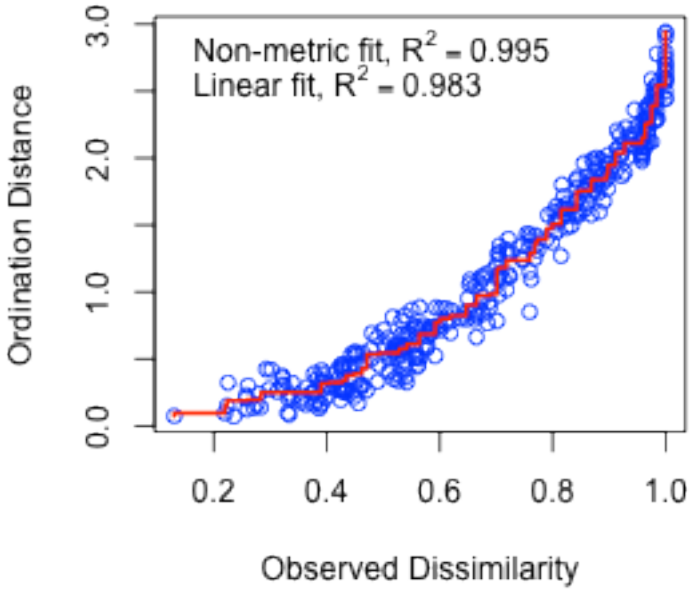
Ordination

There are numerous ordination methods available in R. For now, let's use non-metric multidimensional scaling to visualize the multivariate structure of these communities.

```
# The metaMDS function automatically transforms data and checks
solution
# robustness
comm.bc.mds <- metaMDS(comm, dist = "bray")
```

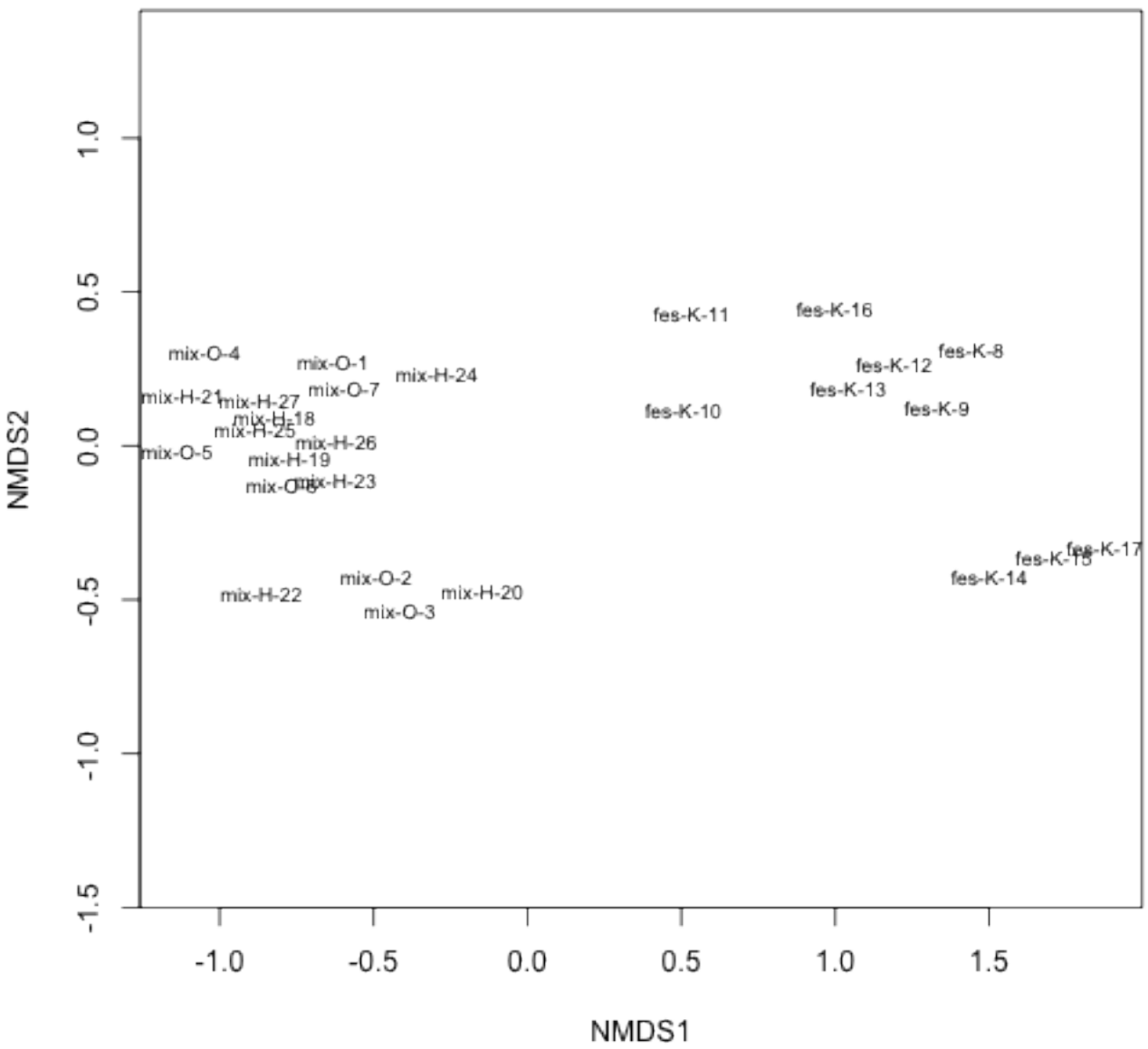
```
## Run 0 stress 0.07174
## Run 1 stress 0.08563
## Run 2 stress 0.08269
## Run 3 stress 0.0791
## Run 4 stress 0.07174
## ... New best solution
## ... procrustes: rmse 0.0009911  max resid 0.004479
## *** Solution reached
```

```
# Assess goodness of ordination fit (stress plot)
stressplot(comm.bc.mds)
```

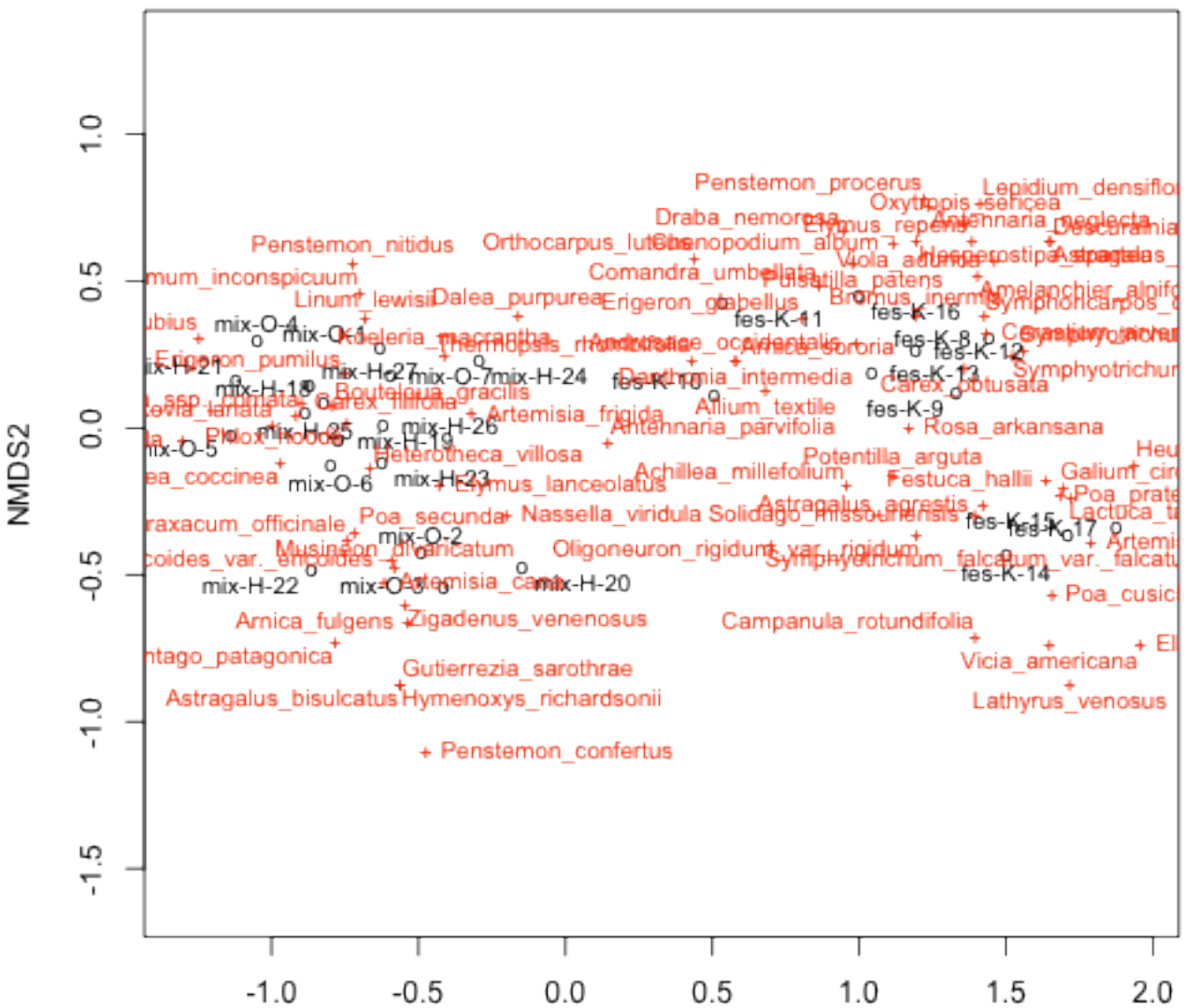


We can plot the ordination results in a variety of different ways.

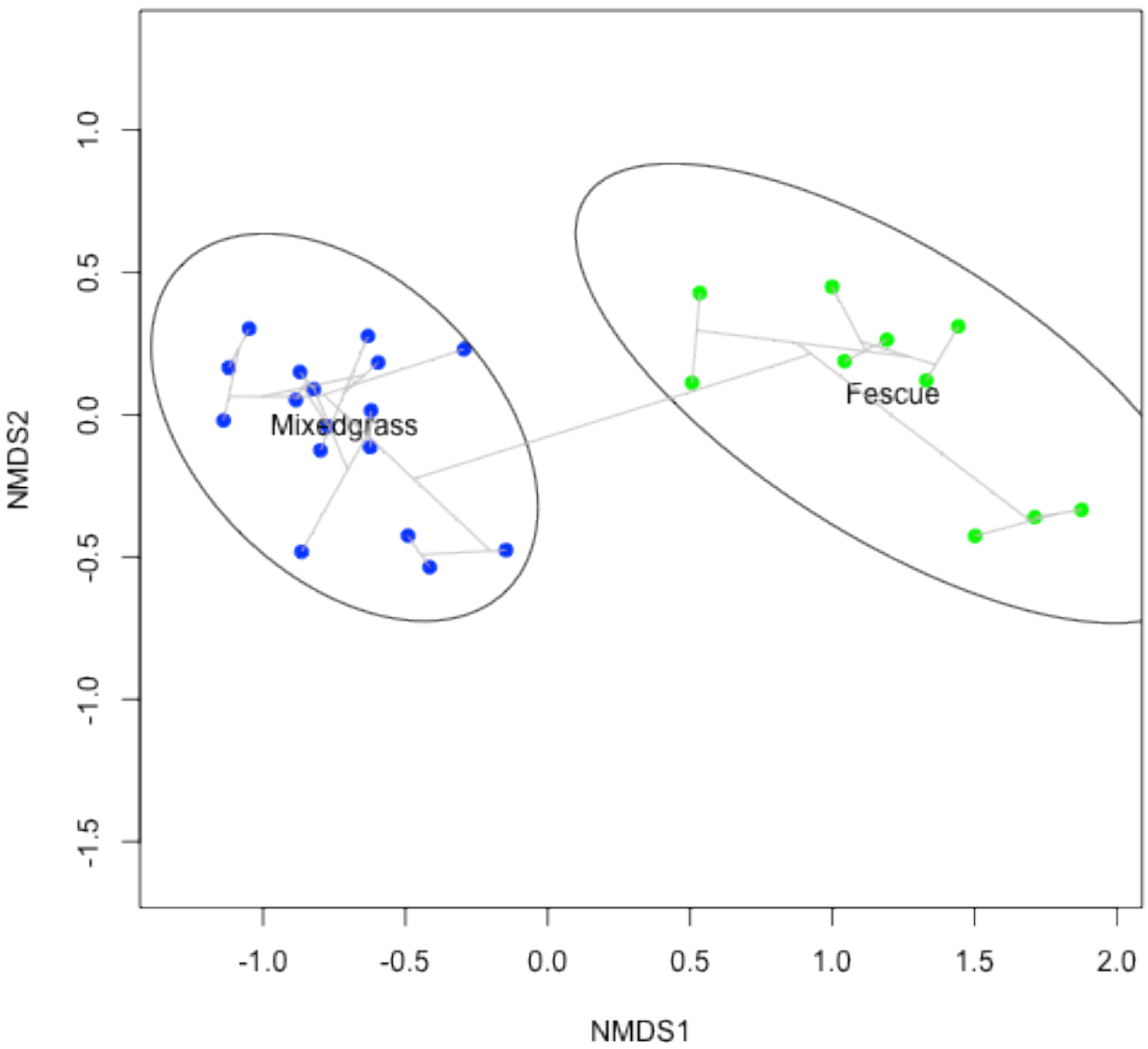
```
# plot site scores as text
ordiplot(comm.bc.mds, display = "sites", type = "text")
```




```
# automated plotting of results - tries to eliminate overlapping
labels
ordipointlabel(comm.bc.mds)
```

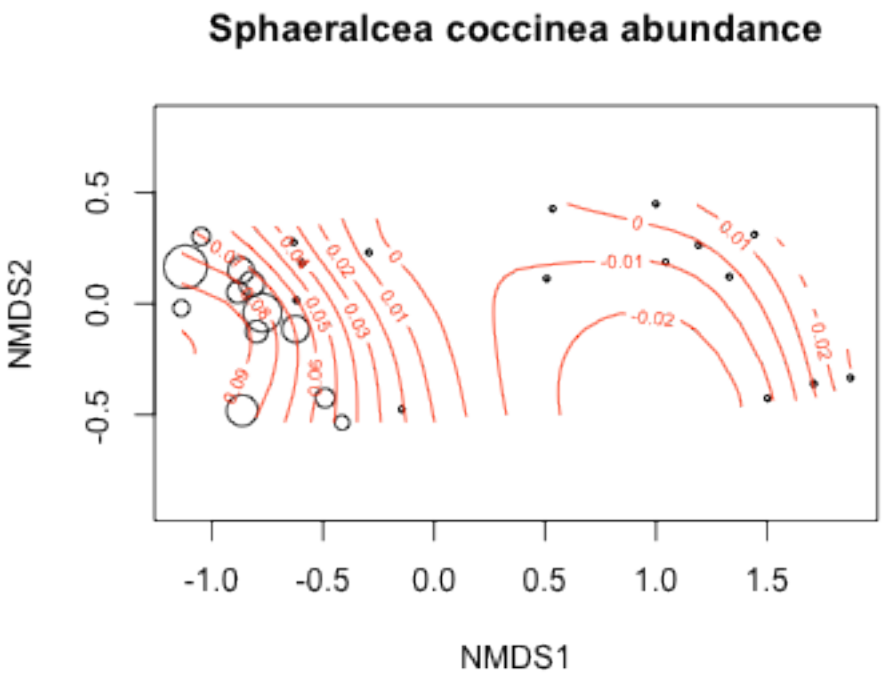



```
# ordination plots are highly customizable set up the plotting area
but
# don't plot anything yet
mds.fig <- ordiplot(comm.bc.mds, type = "none")
# plot just the samples, colour by habitat, pch=19 means plot a
circle
points(mds.fig, "sites", pch = 19, col = "green", select =
metadata$habitat ==
    "Fescue")
points(mds.fig, "sites", pch = 19, col = "blue", select =
metadata$habitat ==
    "Mixedgrass")
# add confidence ellipses around habitat types
ordiellipse(comm.bc.mds, metadata$habitat, conf = 0.95, label =
TRUE)
# overlay the cluster results we calculated earlier
ordicluster(comm.bc.mds, comm.bc.clust, col = "gray")
```



We can also visualize the abundance of species. The `ordisurf` function fits a smooth surface to estimates of species abundance.

```
# plot Sphaeralcea abundance. cex increases the size of bubbles.
ordisurf(comm.bc.mds, comm[, "Sphaeralcea_coccinea"], bubble =
TRUE, main = "Sphaeralcea coccinea abundance",
    cex = 3)
```

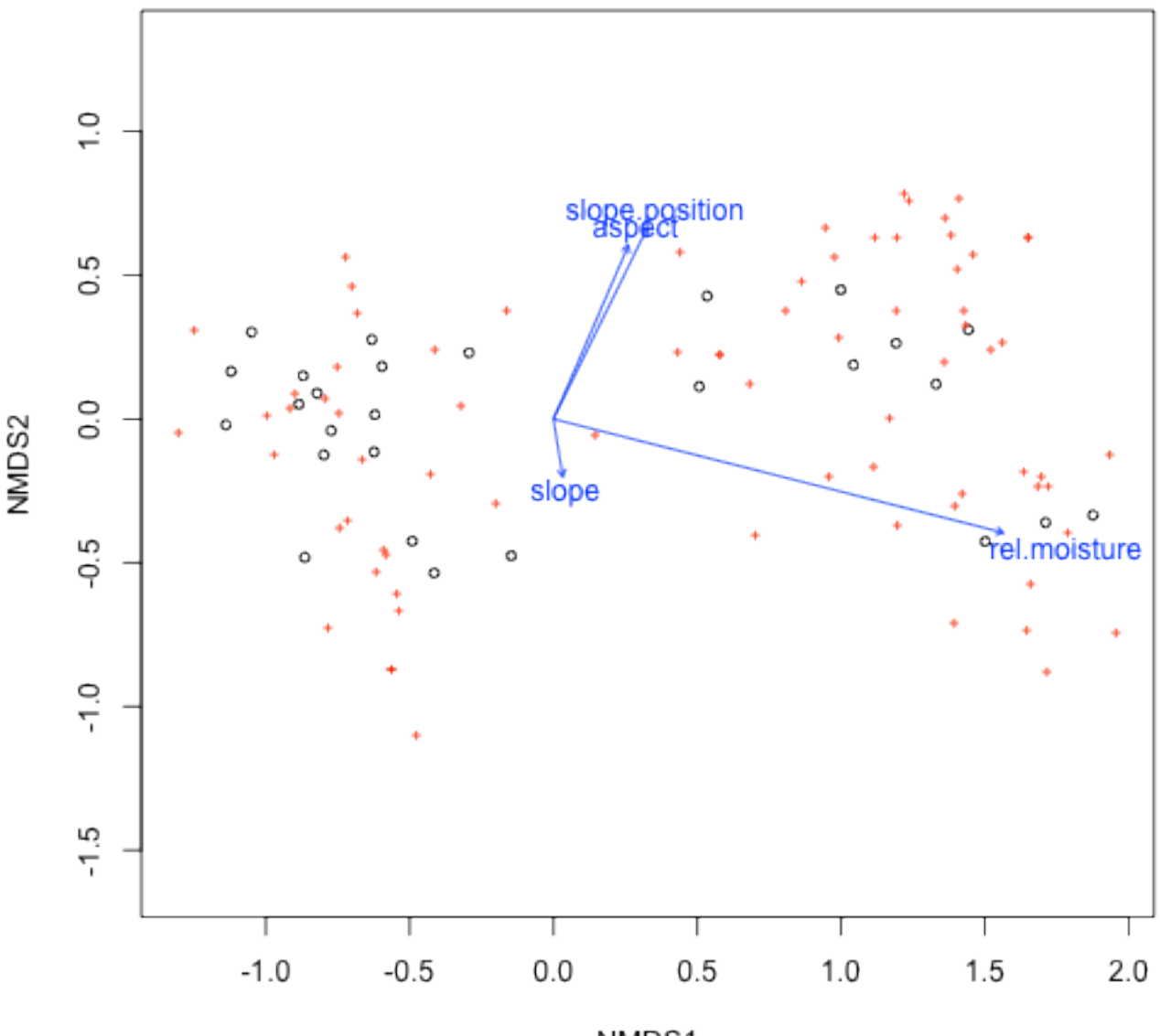



```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(x1, x2, k = 10, bs = "tp", fx = FALSE)
## <environment: 0x106add840>
##
## Estimated degrees of freedom:
## 4.3  total = 5.3
##
## REML score: -43.6
```

Adding environmental and trait data to ordinations

How are environmental variables correlated with the ordination axes?

```
ordiplot(comm.bc.mds)
# calculate and plot environmental variable correlations with the
axes use
# the subset of metadata that are environmental data
plot(envfit(comm.bc.mds, metadata[, 3:6]))
```



It is also possible to do a constrained ordination such as constrained correspondence analysis (CCA) or redundancy analysis (RDA), where trait or environmental data are incorporated directly into the ordination. These methods are implemented in the functions `cca` and `rda` in **vegan**.

Trait evolution

Phylogenetic signal

The idea of phylogenetic niche conservatism (the ecological similarity of closely related species) has attracted a lot of attention recently, for example in the widely used framework of inferring community assembly processes based on knowledge of community phylogenetic structure plus the phylogenetic conservatism of traits. (Webb et al. 2002).

Phylogenetic signal is a quantitative measure of the degree to which phylogeny predicts the ecological similarity of species. The K statistic is a measure of phylogenetic signal that compares the observed signal in a trait to the signal under a Brownian motion model of trait evolution on a phylogeny (Blomberg et al. 2003). K values of 1 correspond to a Brownian motion process, which implies some degree of phylogenetic signal or conservatism. K values closer to zero correspond to

a random or convergent pattern of evolution, while K values greater than 1 indicate strong phylogenetic signal and conservatism of traits. The statistical significance of phylogenetic signal can be evaluated by comparing observed patterns of the variance of independent contrasts of the trait to a null model of shuffling taxa labels across the tips of the phylogeny. These tests are implemented in the `Kcalc`, `phylosignal`, and `multiPhylosignal` functions.

Let's measure phylogenetic signal in these data.

```
# one way to do it - apply the Kcalc function to each column of the
# data.frame
apply(traits, 2, Kcalc, phy)
```

```
##            SLA       LeafArea  LeafThickness            SLV
LeafTissueDens
##         0.2563         0.4231         0.2419          0.3311
0.3299
##            SRL            SRV RootTissueDens       RootDiam
##         0.2290         0.2699         0.2545          0.3151
```

```
# another way to do it with significance testing we have to convert
the tree
# to be dichotomous before calculating P-values
multiPhylosignal(traits, multi2di(phy))
```

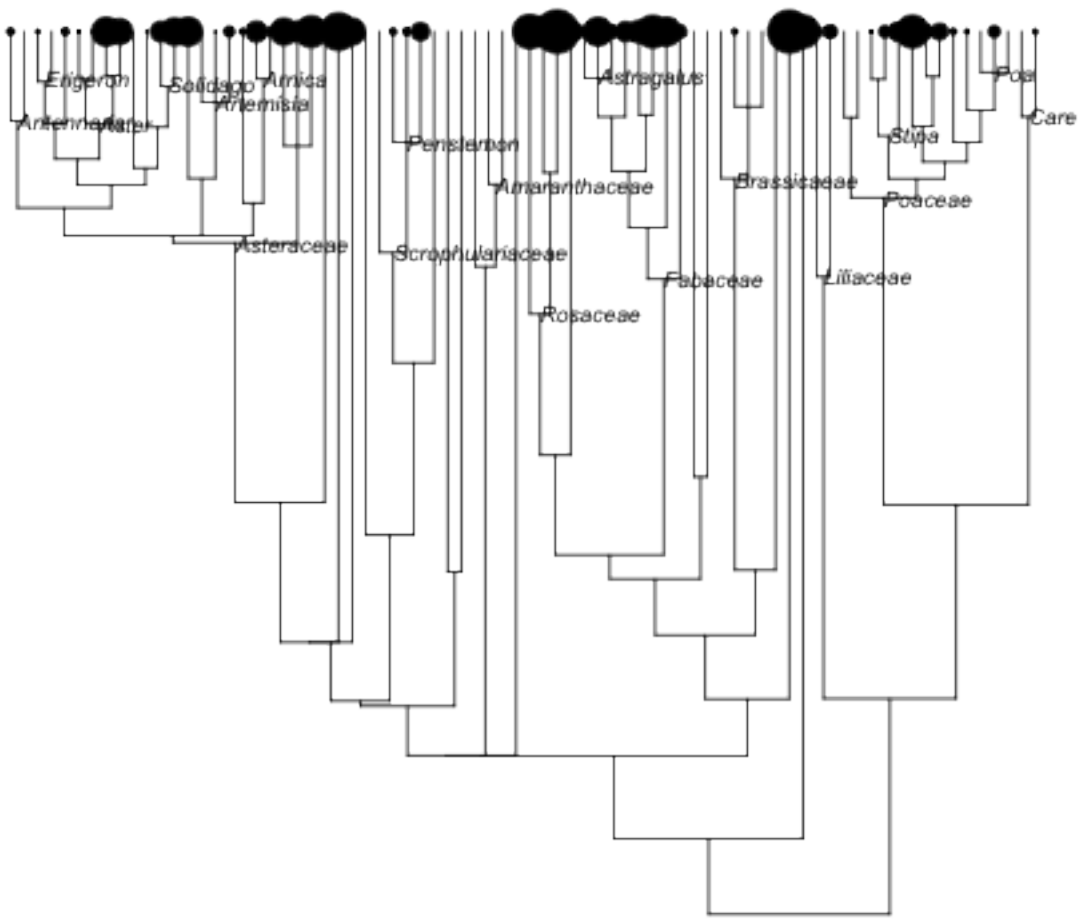
```
##                      K PIC.variance.obs PIC.variance.rnd.mean
## SLA             0.2563        0.0006068             0.0008364
## LeafArea        0.4231        0.0055508             0.0132445
## LeafThickness   0.2419        0.0006669             0.0008837
## SLV             0.3311        0.0005729             0.0010691
## LeafTissueDens  0.3299        0.0006689             0.0012487
## SRL             0.2290        0.0028409             0.0036477
## SRV             0.2699        0.0011065             0.0016683
## RootTissueDens  0.2545        0.0010609             0.0015218
## RootDiam        0.3151        0.0005510             0.0009877
##                 PIC.variance.P PIC.variance.Z
## SLA                      0.037         -1.572
## LeafArea                 0.001         -4.011
## LeafThickness            0.079         -1.330
## SLV                      0.001         -3.234
## LeafTissueDens           0.001         -3.047
## SRL                      0.083         -1.352
## SRV                      0.008         -2.180
## RootTissueDens           0.011         -2.051
## RootDiam                 0.001         -3.195
```

In the output, `K` is the K statistic (magnitude of signal vs. Brownian motion), and `PIC.variance.P` is the P-value of the test for non-random signal. Most variables show more phylogenetic signal than expected by chance.

Visualizing trait evolution

We can visualize trait values on the phylogeny by plotting a different color or size of symbol for each trait value. Let's visualize leaf area, the trait with the strongest phylogenetic signal. The cex argument to the tiplabels function adjusts the size of the trait symbols - some tinkering around with the scaling of the symbol sizes is required depending on the trait.

```
# Plot phylogeny facing upwards. Show node labels but not tip
labels. cex
# shrinks labels.
plot(phy, direction = "up", show.tip.label = FALSE, show.node.label
= TRUE,
    cex = 0.7)
# Plot leaf area on the phylogeny. cex argument scales symbol size
by trait
# value.
tiplabels(pch = 19, col = "black", cex = 3 * (traits[,
"LeafArea"]/max(traits[,
    "LeafArea"])))
```



Phylogenetic analysis of trait relationships

Phylogenetic signal means that closely related species have similar traits. This violates the assumption of independence of data points that is inherent in many methods including correlation and regression (Felsenstein 1985). We can account for non-independence due to phylogenetic signal using methods including phylogenetically independent contrasts and phylogenetic generalised least squares (pGLS).

Generalised least squares methods work just like an ANOVA or linear model - we can test for relationships between categorical or continuous values, optionally taking phylogenetic relatedness into account.

Let's test for a relationship between specific root length (SRL) and root tissue density, taking phylogenetic relationships among species into account.

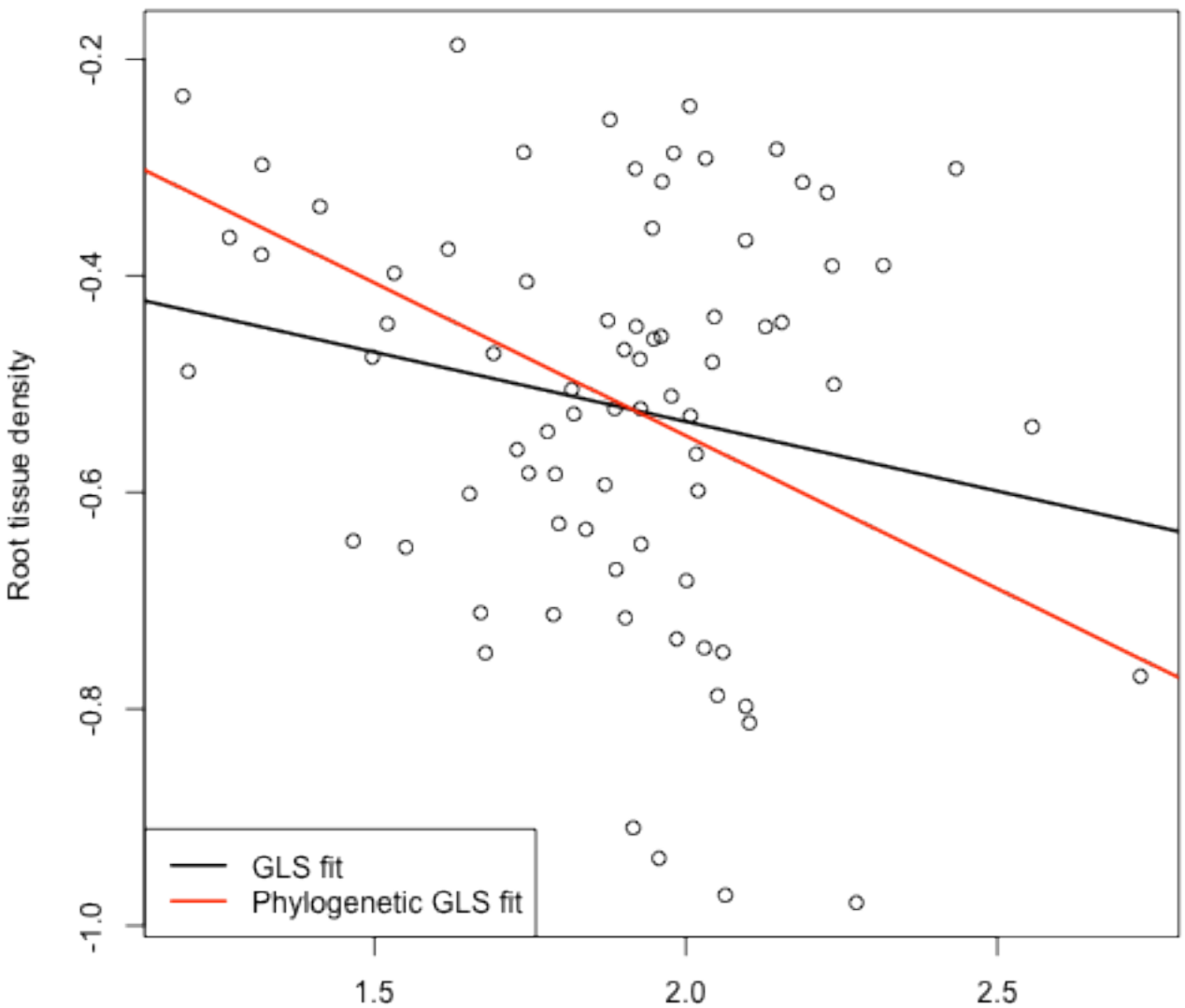
```
# GLS of root tissue density as a function of SRL - non-
phylogenetic model
root.gls <- gls(RootTissueDens ~ SRL, data = traits)
anova(root.gls)
```

```
## Denom. DF: 74
##             numDF F-value p-value
## (Intercept)     1   611.1  <.0001
## SRL             1     3.1  0.0844
```

```
# Phylogenetic GLS - adds effect of phylogeny to the model
root.pgls <- gls(RootTissueDens ~ SRL, correlation =
corBrownian(value = 1,
    phy), data = traits)
anova(root.pgls)
```

```
## Denom. DF: 74
##             numDF F-value p-value
## (Intercept)     1   13.42   5e-04
## SRL             1   20.07  <.0001
```

```
# plot relationship
plot(RootTissueDens ~ SRL, data = traits, xlab = "SRL (specific
root length)",
    ylab = "Root tissue density")
# add model fit lines - coef is the model fit coefficients, lwd
increases
# line width
abline(coef(root.gls), lwd = 2, col = "black")
abline(coef(root.pgls), lwd = 2, col = "red")
legend("bottomleft", legend = c("GLS fit", "Phylogenetic GLS fit"),
lwd = 2,
    col = c("black", "red"))
```



There is a weak relationship between SRL and root tissue density. The relationship is not significant if we do not take phylogenetic relatedness into account. We see a stronger and significant relationship between SRL and root tissue density after taking phylogenetic relatedness into account.

Phylogenetic and trait diversity

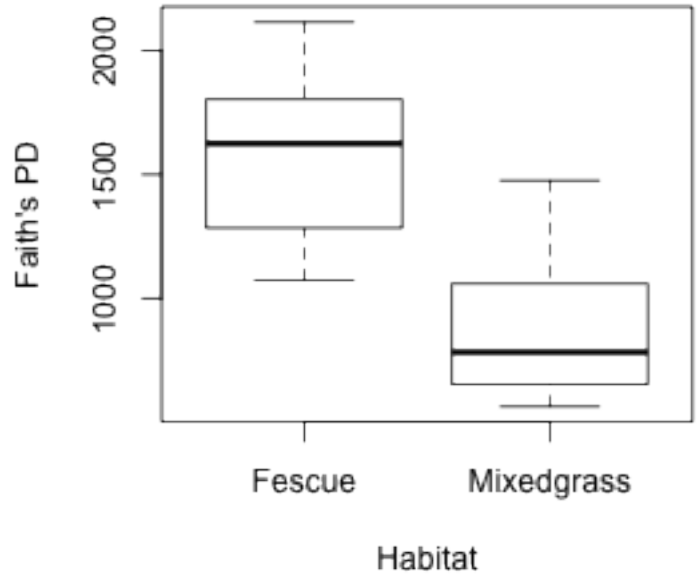
Phylogenetic diversity

One of the earliest measures of phylogenetic relatedness in ecological communities was the phylogenetic diversity (PD) index proposed by Faith. Faith's PD is defined as the total branch length spanned by the tree including all species in a local community, optionally including the root node of the phylogeny. The pd function returns two values for each community, Faith's PD and species richness (SR).

```
# Calculate Faith's PD
comm.pd <- pd(comm, phy)
head(comm.pd)
```

```
##              PD SR
## mix-o-1 1072.4 16
## mix-o-2 1475.5 22
## mix-o-3 1406.2 21
## mix-o-4  564.6  6
## mix-o-5  783.4 10
## mix-o-6 1028.6 13
```

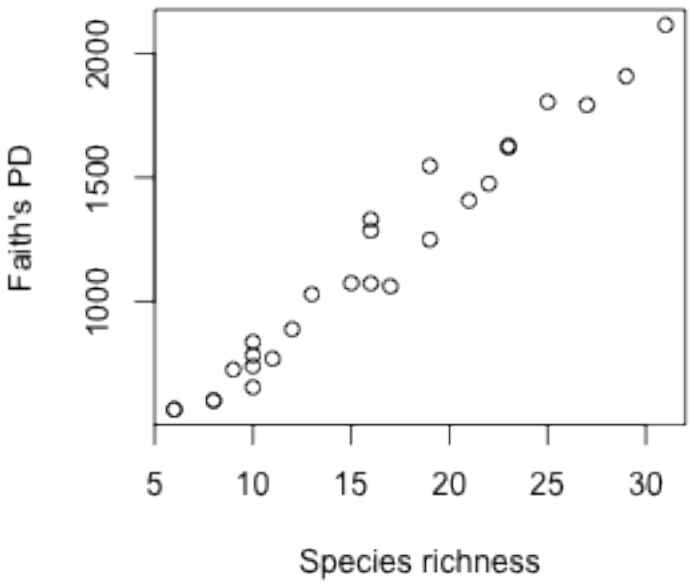
```
# Plot Faith's PD by habitat
boxplot(comm.pd$PD ~ metadata$habitat, xlab = "Habitat", ylab =
"Faith's PD")
```




```
# Test for PD differences among habitats
t.test(comm.pd$PD ~ metadata$habitat)
```

```
##
##  Welch Two Sample t-test
##
## data:  comm.pd$PD by metadata$habitat
## t = 5.716, df = 17.63, p-value = 2.195e-05
## alternative hypothesis: true difference in means is not equal to
0
## 95 percent confidence interval:
##  451.2 976.9
## sample estimates:
##     mean in group Fescue mean in group Mixedgrass
##                   1602.2                   888.2
```

```
# Compare PD and species richness
plot(comm.pd$PD ~ comm.pd$SR, xlab = "Species richness", ylab =
"Faith's PD")
```



Faith's PD is lower in mixedgrass habitats than in fescue habitats. But Faith's PD is highly correlated with species richness, and we already know that there are fewer species in mixedgrass habitats, so we need some way to compare phylogenetic diversity that takes this fact into account.

\(MPD\), \(MNTD\), \(SES_{MPD}\) and \(SES_{MNTD}\)

Another way of thinking about the phylogenetic relatedness of species in a community is to ask 'how closely related are the average pair of species or individuals in a community', and relate the patterns we observe to what we'd expect under various null models of evolution and community assembly. These types of questions are addressed by the measures of community phylogenetic structure such as MPD, MNTD, NRI and NTI described by Webb et al. and implemented in Phylocom.

The function `mpd` will calculate the mean pairwise distance between all species or individuals in each community. Similarly, the `mntd` function calculates the mean nearest taxon distance, the average distance separating each species or individual in the community from its closest heterospecific relative. The `mpd` and `mntd` functions differs slightly from the `pd` function in that they take a distance matrix as input rather than a phylogeny object. A `phylo` object can be converted to a interspecific phylogenetic distance matrix using the `cophenetic` function. Since the mpd and mntd functions can use any distance matrix as input, we can easily calculate trait diversity measures by substituting a trait distance matrix for the phylogenetic distance matrix. We'll return to this idea shortly.

If the community data represent abundance measures, the abundance data can be taken into account. Doing so changes the interpretation of these metrics from the average distance among two randomly chosen species from a community, to the average distance among two randomly chosen individuals in a community.

Measures of 'standardized effect size' of phylogenetic community structure can be calculated for MPD and MNTD by compared observed phylogenetic relatedness to the pattern expected under some null model of phylogeny or community randomization. Standardized effect sizes describe the

difference between average phylogenetic distances in the observed communities versus null communities generated with some randomization method, standardized by the standard deviation of phylogenetic distances in the null data:

$SES_{metric} = \frac{ Metric_{observed} - mean(Metric_{null}) }{sd(Metric_{null})}$

Phylocom users will be familiar with the measures NRI and NTI; SES_{MPD} and SES_{MNTD} are equivalent to -1 times NRI and NTI, respectively. Several different null models can be used to generate the null communities. These include randomizations of the tip labels of the phylogeny, and various community randomizations that can hold community species richness and/or species occurrence frequency constant. These are described in more detail in the help files, as well as in the Phylocom manual. Let's calculate some of these measures of community phylogenetic structure for our example data set. We will ignore abundance information, and use a simple null model of randomly drawing species while keeping sample species richness constant.

```
# convert phylogenety to a distance matrix
phy.dist <- cophenetic(phy)
# calculate ses.mpd
comm.sesmpd <- ses.mpd(comm, phy.dist, null.model = "richness",
abundance.weighted = FALSE,
    runs = 999)
head(comm.sesmpd)
```

```
##         ntaxa mpd.obs mpd.rand.mean mpd.rand.sd mpd.obs.rank
mpd.obs.z
## mix-o-1    16   231.3         238.3      11.564           235
-0.60885
## mix-o-2    22   239.5         237.4       9.533           539
0.22635
## mix-o-3    21   236.5         237.5       9.883           387
-0.09495
## mix-o-4     6   222.5         239.6      24.021           193
-0.71130
## mix-o-5    10   234.2         238.8      17.617           307
-0.26249
## mix-o-6    13   239.4         239.1      12.710           428
0.02488
##         mpd.obs.p runs
## mix-o-1     0.235  999
## mix-o-2     0.539  999
## mix-o-3     0.387  999
## mix-o-4     0.193  999
## mix-o-5     0.307  999
## mix-o-6     0.428  999
```

```
# calculate ses.mntd
comm.sesmntd <- ses.mntd(comm, phy.dist, null.model = "richness",
abundance.weighted = FALSE,
    runs = 999)
head(comm.sesmntd)
```

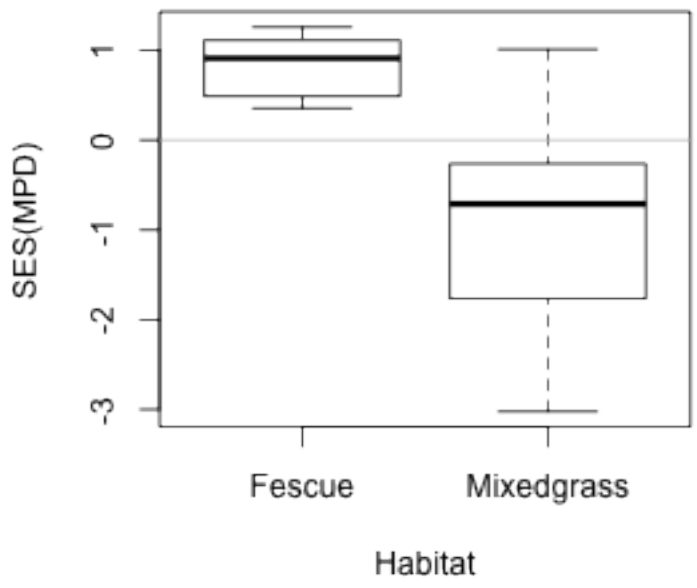
```
##          ntaxa mntd.obs mntd.rand.mean mntd.rand.sd mntd.obs.rank
## mix-o-1     16    94.99          104.67        19.73            328
## mix-o-2     22    97.42           94.64        14.93            584
## mix-o-3     21    98.72           95.77        16.01            559
## mix-o-4      6   136.86          155.24        42.36            333
## mix-o-5     10   107.37          125.04        28.23            262
## mix-o-6     13   118.95          113.24        23.68            607
##          mntd.obs.z mntd.obs.p runs
## mix-o-1     -0.4908      0.328  999
## mix-o-2      0.1863      0.584  999
## mix-o-3      0.1840      0.559  999
## mix-o-4     -0.4339      0.333  999
## mix-o-5     -0.6261      0.262  999
## mix-o-6      0.2413      0.607  999
```

The output includes the following columns:

- `ntaxa` - Number of taxa in community
- `mpd.obs` - Observed mpd in community
- `mpd.rand.mean` - Mean mpd in null communities
- `mpd.rand.sd` - Standard deviation of mpd in null communities
- `mpd.obs.rank` - Rank of observed mpd vs. null communities
- `mpd.obs.z` - Standardized effect size of mpd vs. null communities (equivalent to -NRI)
- `mpd.obs.p` - P-value (quantile) of observed mpd vs. null communities (= mpd.obs.rank / runs + 1)
- `runs` - Number of randomizations

Positive SES values (mpd.obs.z > 0) and high quantiles (mpd.obs.p > 0.95) indicate phylogenetic evenness, while negative SES values and low quantiles (mpd.obs.p < 0.05) indicate phylogenetic clustering, relative to the null model. MPD is generally thought to be more sensitive to tree-wide patterns of phylogenetic clustering and eveness, while MNTD is more sensitive to patterns of evenness and clustering closer to the tips of the phylogeny.

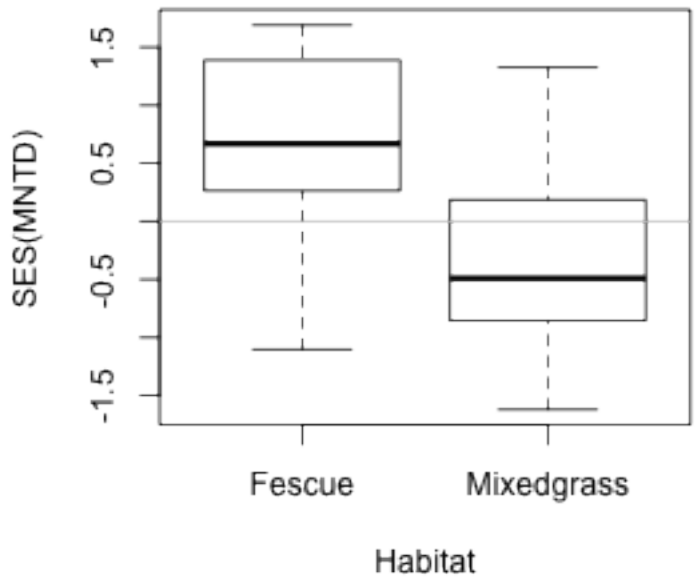
```
# compare ses.mpd between habitats
plot(comm.sesmpd$mpd.obs.z ~ metadata$habitat, xlab = "Habitat",
ylab = "SES(MPD)")
abline(h = 0, col = "gray")
```




```
t.test(comm.sesmpd$mpd.obs.z ~ metadata$habitat)
```

```
##
##  Welch Two Sample t-test
##
## data:  comm.sesmpd$mpd.obs.z by metadata$habitat
## t = 6.058, df = 20.38, p-value = 5.899e-06
## alternative hypothesis: true difference in means is not equal to
0
## 95 percent confidence interval:
##  1.189 2.435
## sample estimates:
##     mean in group Fescue mean in group Mixedgrass
##                   0.8312                  -0.9805
```

```
# compare ses.mntd between habitats
plot(comm.sesmntd$mntd.obs.z ~ metadata$habitat, xlab = "Habitat",
ylab = "SES(MNTD)")
abline(h = 0, col = "gray")
```



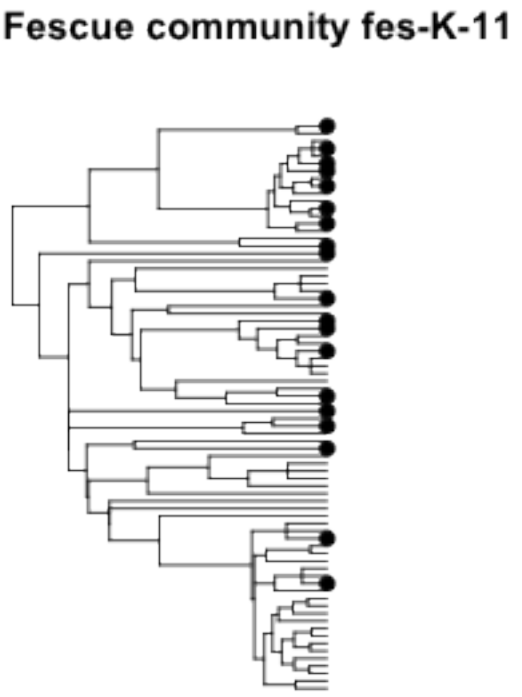

```
t.test(comm.sesmntd$mntd.obs.z ~ metadata$habitat)
```

```
##
##  Welch Two Sample t-test
##
## data:  comm.sesmntd$mntd.obs.z by metadata$habitat
## t = 2.765, df = 14.57, p-value = 0.01474
## alternative hypothesis: true difference in means is not equal to
0
## 95 percent confidence interval:
##  0.2154 1.6798
## sample estimates:
##     mean in group Fescue mean in group Mixedgrass
##                   0.5530                  -0.3946
```

It looks like plant communities from fescue habitats are phylogenetically even (more distantly related than expected by chance, $SES > 0$), and communities from mixedgrass habitats are phylogenetically clustered (more closely related than expected by chance, $SES < 0$).

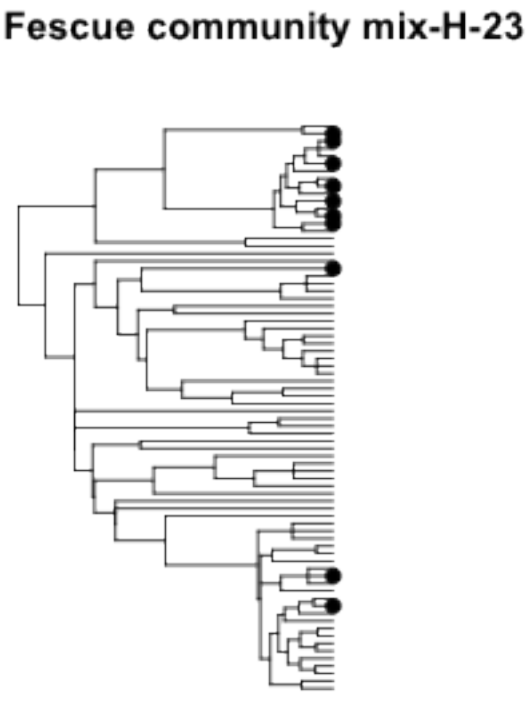
Let's look at the distribution of species from samples in these different habitats on the phylogeny. Fescue community 'fes-K-11' contains species that are phylogenetically even.

```
# plot species present in a fescue community
plot(phy, show.tip.label = FALSE, main = "Fescue community fes-K-
11")
tiplabels(tip = which(phy$tip.label %in% colnames(comm)[comm["fes-
K-11", ] >
    0]), pch = 19)
```



Mixedgrass community 'mix-H-23' contains species that are phylogenetically clumped.

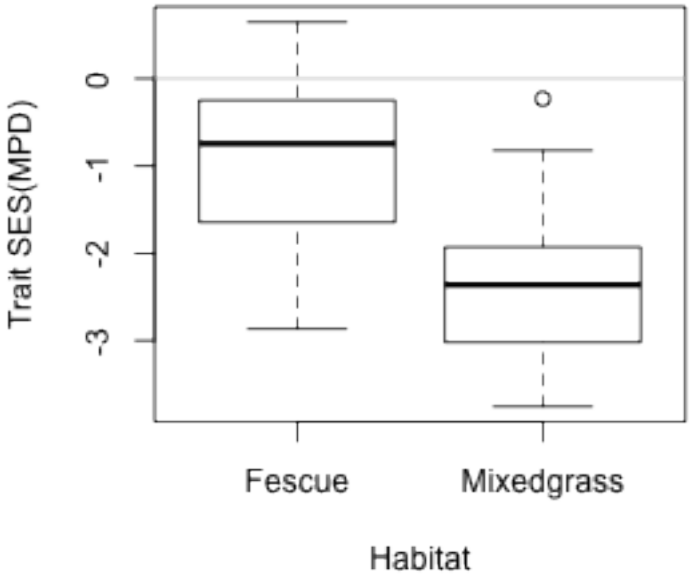
```
# plot species present in a mixedgrass community
plot(phy, show.tip.label = FALSE, main = "Fescue community mix-H-
23")
tiplabels(tip = which(phy$tip.label %in% colnames(comm)[comm["mix-
H-23", ] >
    0]), pch = 19)
```



Trait diversity

We can calculate measures of trait diversity within communities in a manner analogous to the methods we used to calculate phylogenetic diversity. Let's calculate the standardized effect size of functional trait diversity by measuring trait dissimilarity among co-occurring species, and comparing observed trait diversity to a null model.

```
# calculate trait distance - Euclidean distance among scaled trait
values -
# we want the full distance matrix
trait.dist <- as.matrix(dist(scale(traits), method = "euclidean"))
# calculate trait ses.mpd
comm.sesmpd.traits <- ses.mpd(comm, trait.dist, null.model =
"richness", abundance.weighted = FALSE,
    runs = 999)
# compare trait ses.mpd between habitats
plot(comm.sesmpd.traits$mpd.obs.z ~ metadata$habitat, xlab =
"Habitat", ylab = "Trait SES(MPD)")
abline(h = 0, col = "gray")
```



In contrast to the pattern we saw for phylogenetic diversity, trait diversity is lower than expected in both habitats ($SES_{MPD} < 0$), indicating that co-occurring plants have similar leaf and root traits, and this pattern of trait clustering is stronger in mixedgrass habitats.

The `treedive` function in **vegan** calculates a measure of functional trait diversity that is similar to Faith's PD.

Phylogenetic beta-diversity

We can measure patterns of phylogenetic relatedness among communities in a manner similar to the within-community phylogenetic diversity measures described above. The `unifrac` and `phylosor` functions measure the among-community equivalent of Faith's PD, the total unique/shared branch length between communities. The `comdist` and `comdistnt` functions measure the among-community equivalent of MPD and MNTD, the mean pairwise distance or mean nearest taxon distance between pairs of species drawn from two distinct communities.

Let's compare a few different ways of measuring dissimilarity among communities. We've already calculated the Bray-Curtis distance among communities based on shared species (`comm.bc.dist`). Since the Bray-Curtis distance incorporates species abundances, we should use abundance information when calculating phylogenetic and trait diversity as well.

```
# calculate phylogenetic MNTD beta diversity
comm.mntd.dist <- comdistnt(comm, phy.dist, abundance.weighted =
TRUE)
# calculate functional trait MNTD beta diversity
comm.mntd.traits.dist <- comdistnt(comm, trait.dist,
abundance.weighted = TRUE)
# calculate Mantel correlation for taxonomic Bray-Curtis vs.
phylogenetic
# MNTD diversity
mantel(comm.bc.dist, comm.mntd.dist)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = comm.bc.dist, ydis = comm.mntd.dist)
##
## Mantel statistic r: 0.86
##       Significance: 0.001
##
## Upper quantiles of permutations (null model):
##    90%    95%  97.5%    99%
## 0.0769 0.1032 0.1249 0.1594
##
## Based on 999 permutations
```

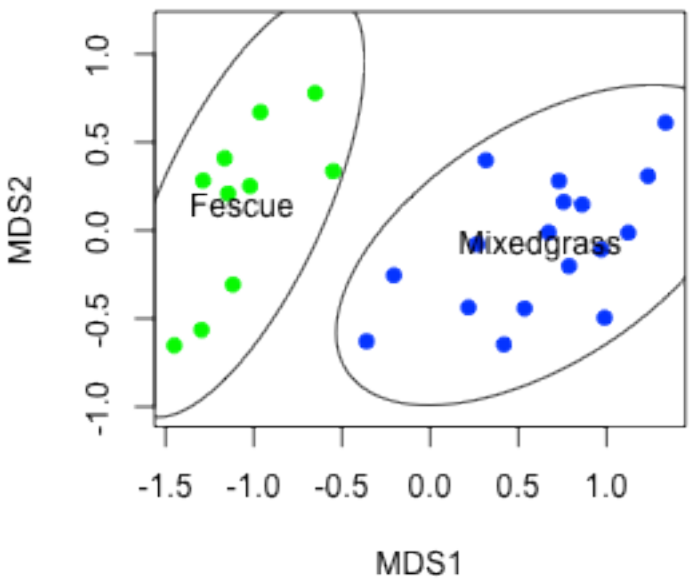
```
# calculate Mantel correlation for taxonomic Bray-Curtis vs. trait
MNTD
# diversity
mantel(comm.bc.dist, comm.mntd.traits.dist)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = comm.bc.dist, ydis = comm.mntd.traits.dist)
##
## Mantel statistic r: 0.952
##       Significance: 0.001
##
## Upper quantiles of permutations (null model):
##    90%    95%  97.5%    99%
## 0.0818 0.1301 0.1652 0.1984
##
## Based on 999 permutations
```

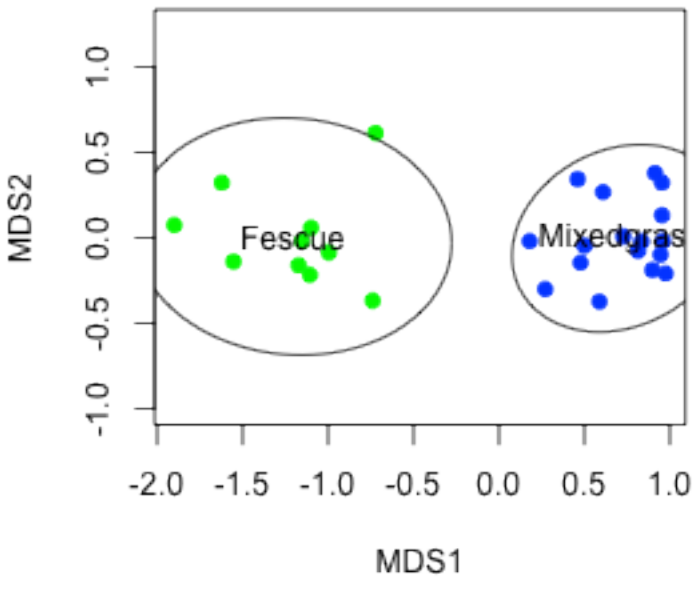
Phylogeny/trait-based ordinations

Since we can calculate phylogeny- and trait-based measures of dissimilarity among samples, we can also perform an ordination of samples based on these metrics. Let's compare phylogeny- and trait-based ordinations with the species-based ordination we performed earlier.

```
# NMDS ordination of phylogenetic distances - use monoMDS since we
only have
# among-sample distances
comm.mntd.mds <- monoMDS(comm.mntd.dist)
# set up the plotting area but don't plot anything yet
mds.fig <- ordiplot(comm.mntd.mds, type = "none")
# plot just the samples, colour by habitat, pch=19 means plot a
circle
points(mds.fig, "sites", pch = 19, col = "green", select =
metadata$habitat ==
    "Fescue")
points(mds.fig, "sites", pch = 19, col = "blue", select =
metadata$habitat ==
    "Mixedgrass")
# add confidence ellipses around habitat types
ordiellipse(comm.mntd.mds, metadata$habitat, conf = 0.95, label =
TRUE)
```




```
# NMDS ordination of trait distances - use monoMDS since we only
have
# among-sample distances
comm.mntd.traits.mds <- monoMDS(comm.mntd.traits.dist)
# set up the plotting area but don't plot anything yet
mds.fig <- ordiplot(comm.mntd.traits.mds, type = "none")
# plot just the samples, colour by habitat, pch=19 means plot a
circle
points(mds.fig, "sites", pch = 19, col = "green", select =
metadata$habitat ==
    "Fescue")
points(mds.fig, "sites", pch = 19, col = "blue", select =
metadata$habitat ==
    "Mixedgrass")
# add confidence ellipses around habitat types
ordiellipse(comm.mntd.traits.mds, metadata$habitat, conf = 0.95,
label = TRUE)
```



It looks like fescue and mixedgrass habitats are quite distinct regardless of how we quantify their biodiversity - they contain different species, phylogenetically distinct taxa, and the traits of species in the two habitats are distinct.

Testing for multivariate differences among groups

We can quantify the relationship between dissimilarity measures and different explanatory variables using the permutational MANOVA (a.k.a. AMOVA) framework in the `adonis` function in **vegan**. This method allows ANOVA-like tests of the variance in beta diversity explained by categorical or continuous variables.

Let's quantify the degree to which habitat can explain taxonomic, phylogenetic, and trait dissimilarity among grasslands.

```
# Taxonomic (Bray-Curtis) dissimilarity explained
adonis(comm.bc.dist ~ habitat, data = metadata)
```

```
##
## Call:
## adonis(formula = comm.bc.dist ~ habitat, data = metadata)
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model    R2 Pr(>F)
## habitat    1      3.37    3.37    25.1 0.501  0.001 ***
## Residuals 25      3.35    0.13          0.499
## Total     26      6.71                  1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Phylogenetic dissimilarity explained
adonis(comm.mntd.dist ~ habitat, data = metadata)
```

```
##
## Call:
## adonis(formula = comm.mntd.dist ~ habitat, data = metadata)
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model    R2 Pr(>F)
## habitat    1     22814   22814    34.9 0.583  0.001 ***
## Residuals 25     16340     654          0.417
## Total     26     39154                  1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Trait dissimilarity explained
adonis(comm.mntd.traits.dist ~ habitat, data = metadata)
```

```
##
## Call:
## adonis(formula = comm.mntd.traits.dist ~ habitat, data =
metadata)
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model    R2 Pr(>F)
## habitat    1     10.92   10.92      64 0.719  0.001 ***
## Residuals 25      4.27    0.17         0.281
## Total     26     15.18                  1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results support the pattern we can see visually in the ordination diagrams. These habitats are distinct in terms of their taxonomic, phylogenetic, and functional trait diversity.