Introduction to Phylogenetic Diveristy

Presented by Maxwell J. Farrell

May 19 2018

Contents

Presented as part of the *IDEAS Computational Modeling Workshop 2018* http://ideas.ecology.uga.edu/computational-summer-workshop.html

The material presented here is largely comprised of workshop materials graciously published by leading researchers on phylogenetic community ecology. The original materials are:

- Simon Joly's course on Comparative Methods
- Steven Kembel's workshop on Biodiversity in R
- Unpublished code by William D. Pearse

Libraries

```
require(ape)
require(picante)
require(phytools)
require(geiger)
```

Phylogenetic Diversity

Phylogenetic diversity is a measure of biodiversity that incorporates phylogenetic difference between species. There are several different definitions of phylogenetic diversity and several ways to estimate it. Here we will only see a few method and instead focus on the concept.

The concept of phylogenetic diversity is to provide more information than a simple species diversity index. It is often applied to species within geographic areas, but can also be used to quantify host specificity. For instance, two parasites might be able to infect the same number of species, but one might infect species that are more evolutionary distant from one another than the other parasite. For example, a given virus might infect six species: three bats and three rodents, whereas the other also infects six species, but infects two bats, two rodent, a primate, and an ungulate. Clearly, the latter can infect species that represent a more evolutionary diverse set of hosts. PD was proposed to incorporate these notions in conservation biology. Since then, these concepts have also been used in community ecology, and for quantifying host specificity.

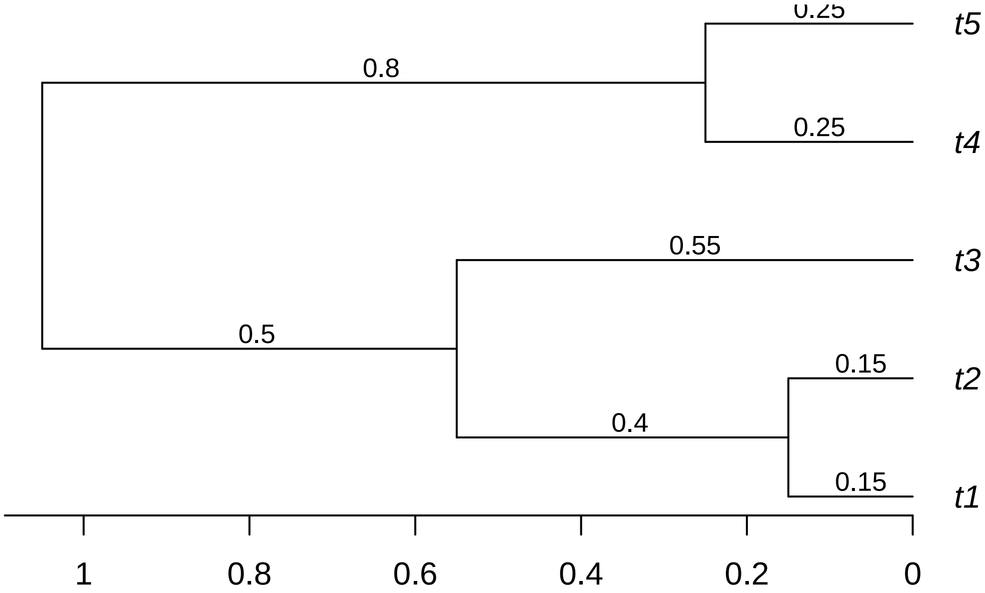
Here is a conceptual diagram of some ways to think about host specificity from Poulin et al. (2011):

Faith's PD

The original definition of phylogenetic diversity (PD) was proposed by Faith (1992). Formally, Phylogenetic diversity is defined as the *Sum of all branch lengths in the portion of a phylogenetic tree connecting the focal set of species.*

Faith's PD can be calculated using the `picante` package in `R`. `picante` uses community matrices to calculate PD. These standardly consist in a matrix in which species are in columns and communities (or region for biogeography applications) are in rows. To show an example of application, we will use the example provided with the `picante` package.

```
phy <- "(((t1:0.15,t2:0.15):0.4,t3:0.55):0.5,(t4:0.25,t5:0.25):0.8);"
phy <- read.tree(text=phy)
plot(phy, label.offset=0.05)
edgelabels(c(0.5,0.4,0.15,0.15,0.55,0.8,0.25,0.25),adj=c(0.3,-0.3),frame="none",bg="",cex=0.8)
axisPhylo() # put up a scale bar
```



We can calculate PD by summing the branch lengths in a tree:

```
data(phylocom)

phy$edge.length
```

```
## [1] 0.50 0.40 0.15 0.15 0.55 0.80 0.25 0.25
```

```
pd_tree <- sum(phy$edge.length)
pd_tree
```

```
## [1] 3.05
```

But often we are interested in PD of communities across sites, or in the case of phylogenetic diversity, information about species across communities. For this workshop we will look at phylogenetic diversity as a measure of host specificity using the Global Mammal Parasite Database 2.0, and the Fritz mammal supertree.

```
# GMPD 2.0
gmpd <- read.csv("data/GMPD_datafiles/GMPD_main.csv", as.is=T)

# Removing parasites not reported to species
Sys.setlocale('LC_ALL','C')
```

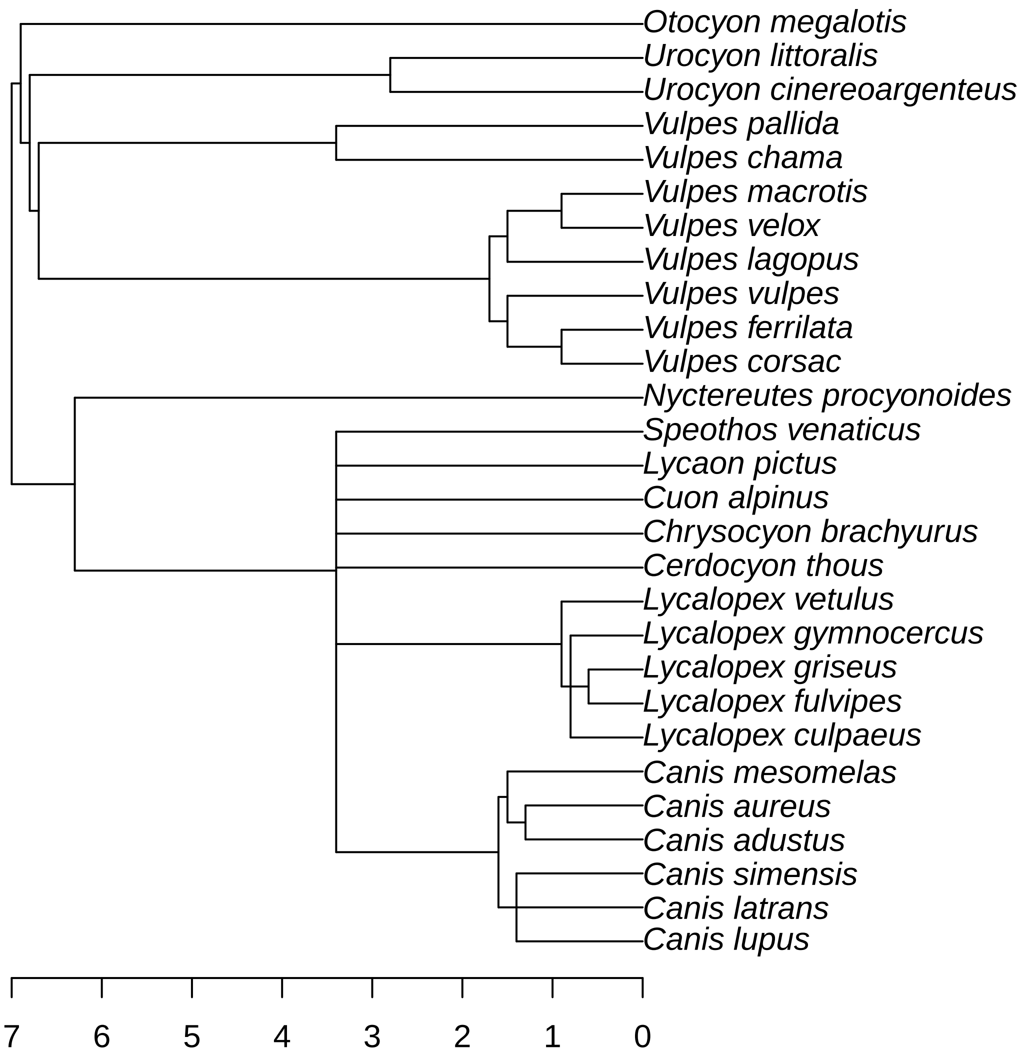
```r
gmpd <- gmpd[grep("sp[.]",gmpd$ParasiteCorrectedName, invert=TRUE),]
gmpd <- gmpd[grep("ABOLISHED",gmpd$ParasiteCorrectedName, invert=TRUE),]
gmpd <- gmpd[grep("no binomial name",gmpd$ParasiteCorrectedName, invert=TRUE),]
gmpd <- gmpd[grep("not identified to genus",gmpd$ParasiteCorrectedName, invert=TRUE),]
gmpd <- gmpd[grep("SPLITTED in ICTV",gmpd$ParasiteCorrectedName, invert=TRUE),]
gmpd <- gmpd[grep("Diphyllobothrium sp",gmpd$ParasiteCorrectedName, invert=TRUE),]

# Removing hosts with no binomial name reported
gmpd <- gmpd[grep("no binomial name",gmpd$HostCorrectedName, invert=TRUE),]

# Tree
fritz_tree <- read.nexus("data/Fritz_2009.tre")[[1]]

# Binomial cleaning & matching
gmpd$HostCorrectedName <- gsub(" ", "_", gmpd$HostCorrectedName)
species.to.exclude <- fritz_tree$tip.label[!(fritz_tree$tip.label %in%
                                                   gmpd$HostCorrectedName)]
# Subsetting tree
gmpd_tree <- drop.tip(fritz_tree,species.to.exclude)

# Community matrix
com <- table(gmpd$HostCorrectedName, gmpd$ParasiteCorrectedName)
```

Have a look at the information in the GMPD, as well as com.

CHALLENGE

Create a subset of the GMPD that includes only species in the Canidae, prune the tree, and re-create the host-parasite association matrix (community matrix).

Make a host-parasite association matrix presence-absence (1/0).

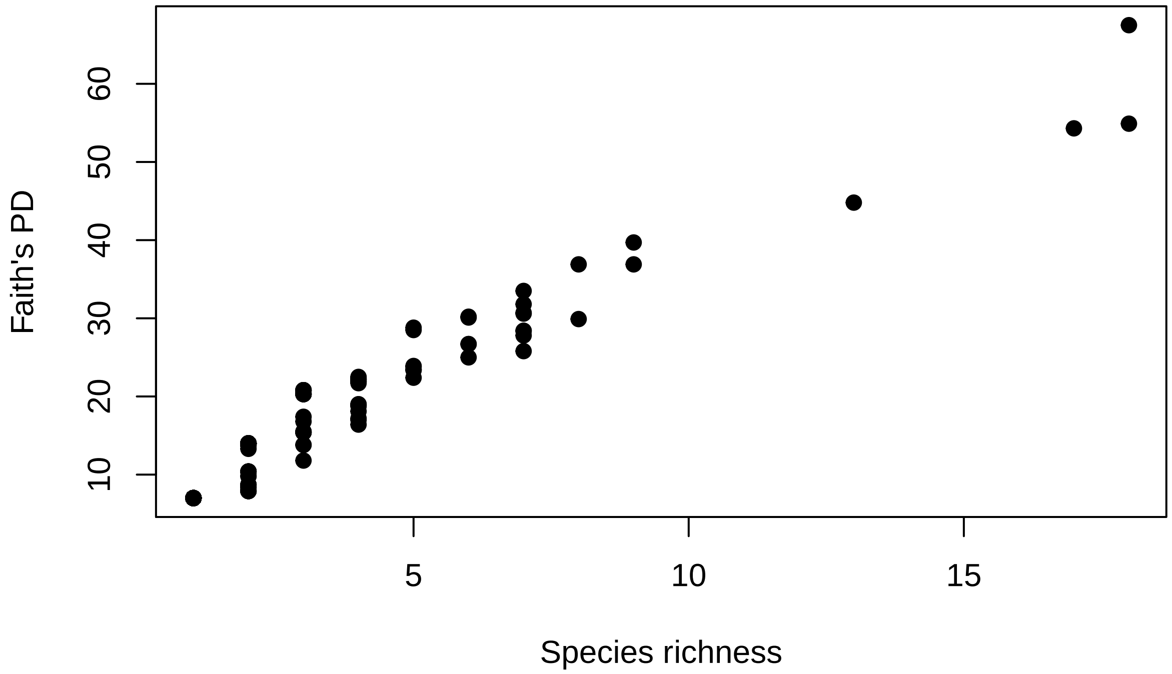


We can calculate for each parasite using the pd function in picante.

```
# pd expects rows to be "sites", and columns to be "species"
# in our case, since we are interested in host PD, our parasites are "sites"
# and we need to transpose our community matrix.
com <- t(com)

# Now calculate PD
com.pd <- pd(com, canid_tree, include.root=TRUE)
head(com.pd)
```

```
##                                    PD SR
## Acanthocheilonema dracunculoides  7.0  1
## Acanthocheilonema reconditum      7.0  1
## Alaria alata                     21.9  4
## Alaria americana                  8.4  2
## Alaria arisaemoides               7.0  1
## Alaria canis                     13.8  2
```

```
# Compare PD and species richness
plot(com.pd$PD ~ com.pd$SR, xlab = "Species richness", ylab = "Faith's PD", pch=19)
```



The function `pd` results in a two column matrix. The first column is Faith's PD and the second is the species richness, that is the number of species in the sample.

By default, PD includes the root in the calculations. If you want to exclude the root, you can indicate `include.root=FALSE` in the function. Also, note that the PD values are function of the tree length. If you change the units, for instance from nucleotide substitions per site per year to years, the results will be different.

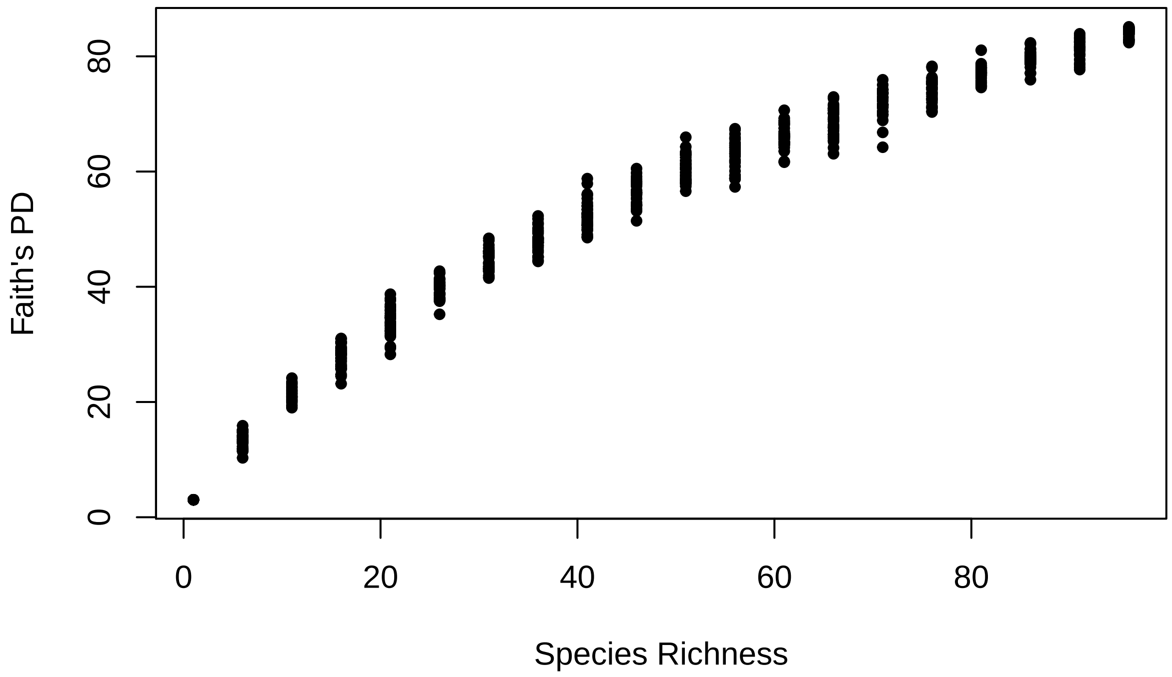
CHALLENGE

Try running pd with `include.root=FALSE` and see what happens. Why does this happen?

Faith's PD is highly correlated with species richness, this isn't just a property of the data at hand. We can conduct a simulation to show this:

```
set.seed(100)
n.spp <- 100
simtree <- sim.bdtree(n=n.spp)
richness <- seq(1,100,5)
reps <- 30
comm <- matrix(0, nrow=length(richness)*reps, n.spp)
x <- 1
for(i in seq_along(richness)){
    for(j in seq_len(reps)){
        spp <- sample(seq_along(simtree$tip.label),richness[i])
        comm[x,spp] <- 1
        x <- x+1
    }
}
colnames(comm) <- simtree$tip.label
```

```
pd.calc <- pd(comm, simtree)

plot(pd.calc$PD ~ (pd.calc$SR), ylab="Faith's PD", xlab="Species Richness", pch=20)
```



So we may need to turn to other measures of phylogenetic diversity if we want to compare phylogenetic diversity while taking this into account into account.

Phylogenies in community ecology

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.

In the last 15 years, phylogenetic information has been increasingly used in the field of community ecology. The whole field really started with the seminal publication of Webb et al. (2002). These types of questions are addressed by the measures of community phylogenetic structure such as MPD and MNTD.

> MPD is the *mean pairwise distance* between all species in each community
>
> MNTD is the *mean nearest taxon distance*, that is the the mean distance separating each species in the community from its closest relative.

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.

```
phydist <- cophenetic(canid_tree)

com.mpd <- mpd(com, phydist, abundance.weighted=FALSE)
head(com.mpd)
```

```
## [1]       NA       NA 11.73333  2.80000       NA 13.60000
```

```r
com.mntd <- mntd(com, phydist, abundance.weighted=FALSE)
head(com.mntd)
```

```
## [1]    NA    NA  8.25  2.80    NA 13.60
```

Note! `mpd` and `mtnd` return `NA` when there is only one species in the community.

CHALLENGE

Using simulation, determine the relationships of both MTND and MNTD with species richness.

MPD is generally thought to be more sensitive to deeper branches in the tree, namely 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.

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.

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.

SES Metrics

Once MPD and MNTD statistics are estimated, we need to compare them with that expected with some null model of phylogeny or community randomization. The resulting *standardized effect size* (SES) metrics describe the difference between phylogenetic distances in the observed communities versus null communities generated with some randomization method, divided by the standard deviation of phylogenetic distances in the null data:

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

Two very similar statistics can be found in the literature, NRI and NTI. They can be obtained from SES_{MPD} and SES_{MNTD} using the following formulas:

$$SES_{MPD} = -1 \times NRI$$

$$SES_{MNTD} = -1 \times NTI$$

Several different null models can be used to generate the null communities that we compare observed patterns to. 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 of `picante`.

Here is an example with the Canid data and using the `phylogeny.pool`, which randomizes community data matrix by drawing species from pool of species occurring in the distance matrix (phylogeny pool) with equal probability.

```r
ses.mpd.result <- ses.mpd(com, phydist, null.model = "phylogeny.pool",
abundance.weighted = FALSE, runs = 999)
head(ses.mpd.result)
```

```
##                                  ntaxa  mpd.obs mpd.rand.mean mpd.rand.sd
## Acanthocheilonema dracunculoides     1       NA           NaN          NA
## Acanthocheilonema reconditum         1       NA           NaN          NA
## Alaria alata                         4 11.73333      10.76974    2.212872
## Alaria americana                     2  2.80000      10.74354    4.131837
## Alaria arisaemoides                  1       NA           NaN          NA
## Alaria canis                         2 13.60000      10.92352    4.166562
##                                  mpd.obs.rank  mpd.obs.z mpd.obs.p runs
## Acanthocheilonema dracunculoides           NA         NA        NA  999
## Acanthocheilonema reconditum               NA         NA        NA  999
## Alaria alata                              540  0.4354507     0.540  999
## Alaria americana                           40 -1.9225209     0.040  999
## Alaria arisaemoides                        NA         NA        NA  999
## Alaria canis                              443  0.6423705     0.443  999
```

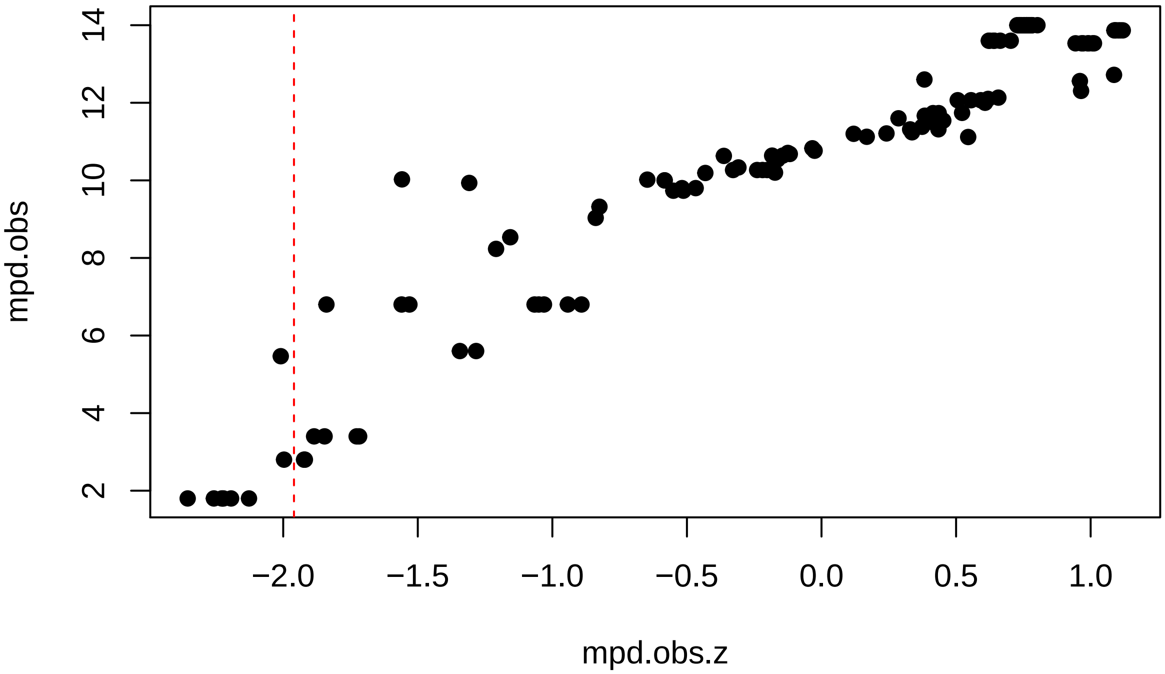
The output includes the following columns:

- 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 11
- 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, or a greater phylogenetic distance among co-occurring species than expected. Negative SES values and low quantiles (mpd.obs.p < 0.05) indicate phylogenetic clustering, or small phylogenetic distances among co-occurring species than expected.

We can separate the significant and non-signifcant mpd values using the critical values of a z-test.

```r
with(ses.mpd.result, plot(mpd.obs~mpd.obs.z, pch=19))
abline(v=-1.96, col="red", lty=2)
abline(v=1.96, col="red", lty=2)
```



Another null model is the `sample.pool` Randomize community data matrix by drawing species from pool of species occurring in at least one community (sample pool) with equal probability. In this case the regional species pool for the randomization test is not the full phylogeny, but instead the subset of species that occur in at least one community.

Depending on the null model and the species pool used, we can change our perceptions of phylogenetic clustering and overdispersion.

CHALLENGE

Re-do the above analysis for parasites documented to infect the Canidae, but expand the species pool for your null model to include all species in Carnivora.

How many parasites now show phylogenetically over or under dispersed host communities?

Since the MPD and MNTD functions can use any distance matrix as input, we could easily calculate trait diversity measures by substituting a trait distance matrix for the phylogenetic distance matrix.

CHALLENGE

Write a function to calculate the maximum observed phylogenetic distance among hosts, and modify one of the ses.X functions to calculare it's standard effect size.

Phylogenetic β diversity

Diveristy can be divided in three components: α diversity, which represents the diversity ofberved at a given site, β diversity, which represents the diversity among sites, and γ diversity, which represents the total diversity. The PD statistic can provide a measure of the α or γ diversity. However, it is sometimes of interest to measure the amount of variation between sites, that is the β diversity.

Several measures of β phylogenetic diversity have been proposed (Swenson 2011), although several of these are very similar (Sewnson 2011). We will see a popular statistic here, that is *Phylosor*, which converges to the Sorensen Index when there is no phylogenetic information. *Phylosor* is defined as

$$Phylosor = \frac{BL_{k_1k_2}}{(BL_{k1} + BL_{k2}) \times \frac{1}{2}}$$

where $BL_{k_1k_2}$ is the total length of the branches shared between community k_1 and k_2, BL_{k_1} and BL_{k_2} are the total branch lengths found in communities k_1 and k_2 respectively.

It can be estimated using the `phylosor` function in the `picante` package.

CHALLENGE

This metric can sometimes be slow to calculate, so calculate it only for *Taenia sp.*.

Hint: you need to subset the com to include only Taenia parasites, and prune the tree accordingly

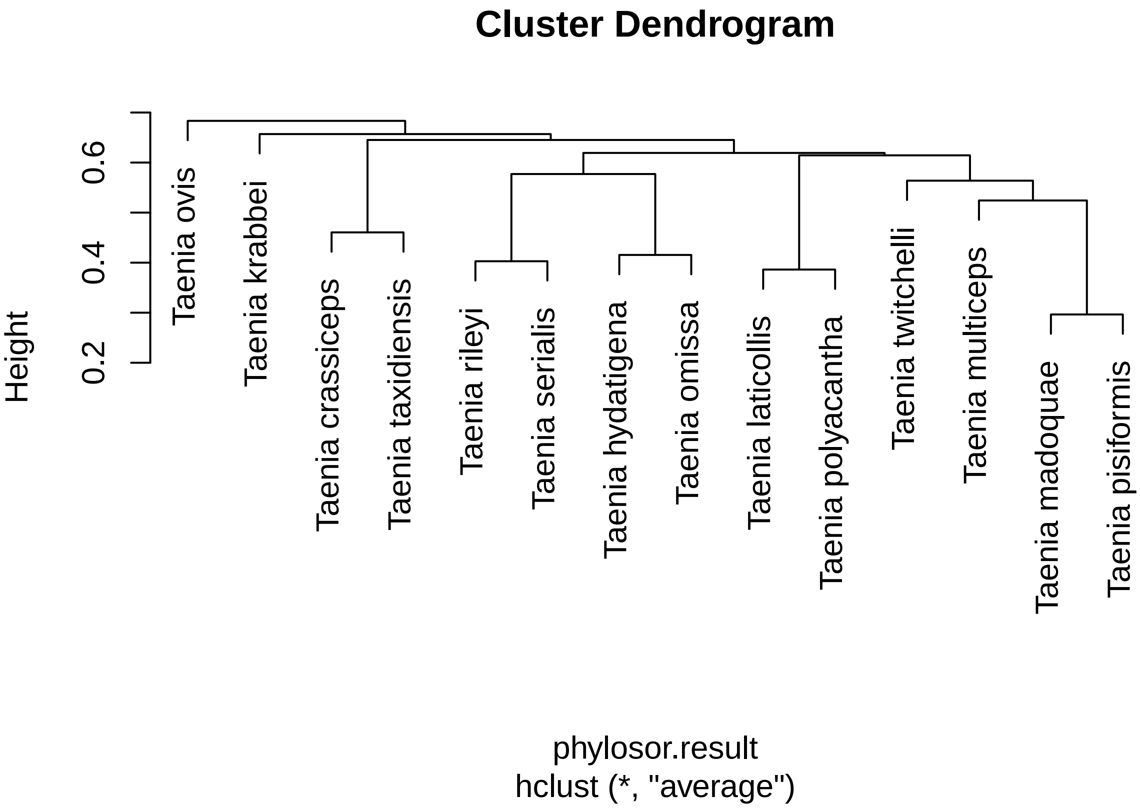
---

This similarity index can be used to represent the similarity between communities using a phenogram.

```r
library(cluster)
```

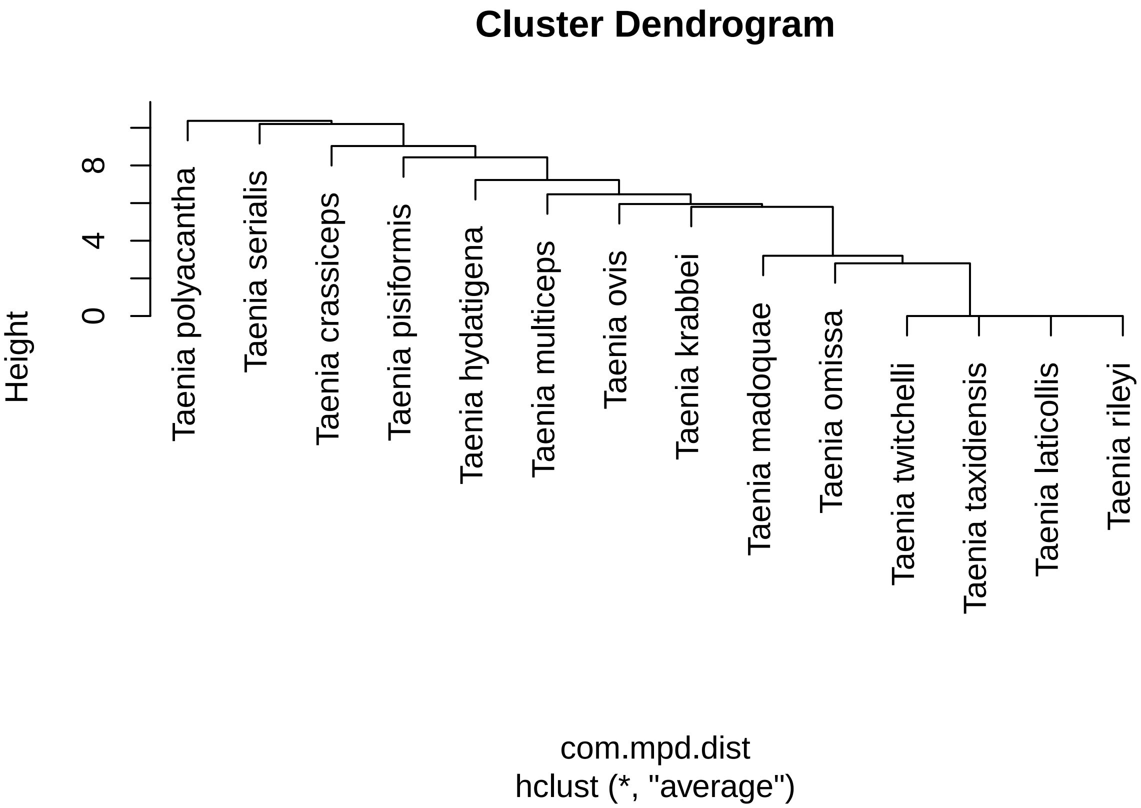
```
##
## Attaching package: 'cluster'

## The following object is masked from 'package:maps':
##
##     votes.repub
```

```r
#UPGMA clustering
phylosor.clusters <- hclust(phylosor.result,method="average")
plot(phylosor.clusters)
```



Similarly, in `picante`, the `comdist` function estimates a β phylodiversity index that is the equivalent to the MPD statistic.

```
com.mpd.dist <- comdist(com_taenia, cophenetic(taenia_hosts), abundance.weighted=TRUE)
plot(hclust(com.mpd.dist,method="average"))
```

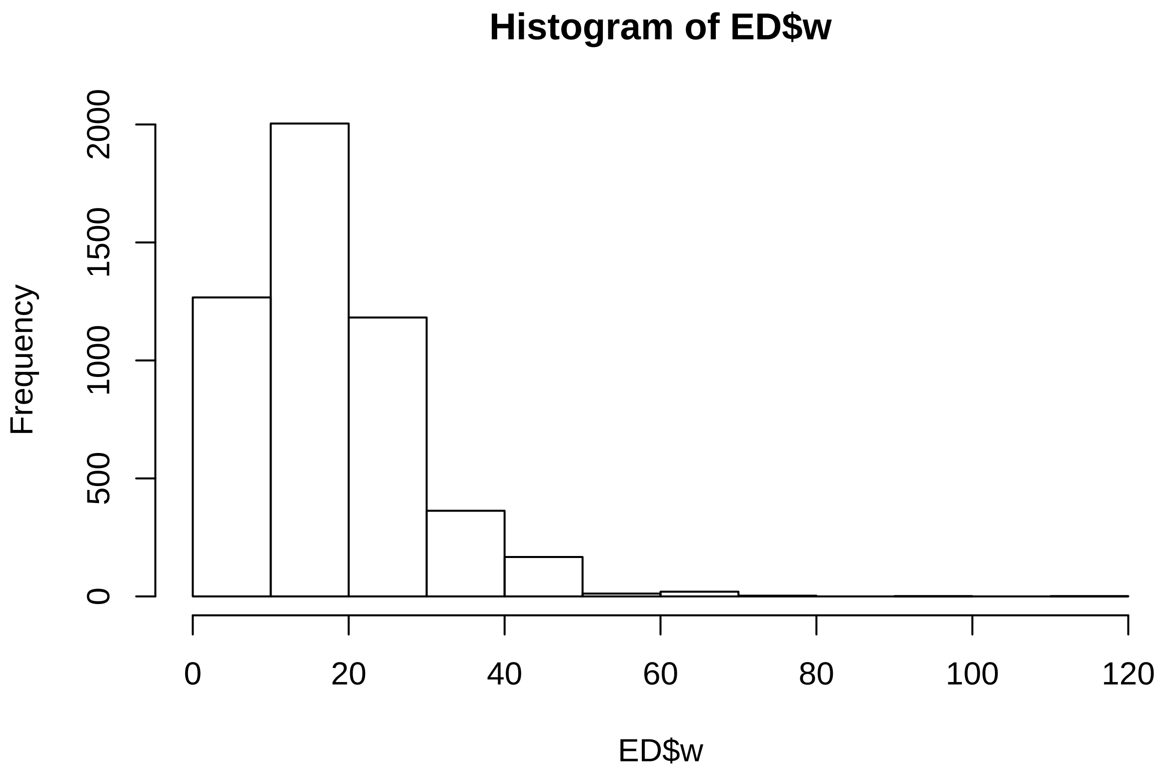


Phylogenetic distinctiveness

Phylogenetic diversity is very present in the conservation literature, but there are other ways phylogenies can be incorporated into statistic that could be useful for conservation. One statistic that is relatively popular is Evolutionary Distinctiveness (ED: Redding and Mooers, 2006). ED represents the evolutionary history that is unique to a species in a given sample. Practically, it is the length of the branch on the tree that links the species to the rest of the tree (i.e., the length of the terminal branch). So species that are on longer branches have higher evolutionary distinctiveness. Another way to look at it is that *living fossils* such as gingkos or coleocanths will have high ED compared to dandelions or finches, respectively.

Lets calculate ED for all Mammals:

```
ED <- evol.distinct(fritz_tree)
hist(ED$w)
```



CHALLENGE

Which species have the highest ED among mammals?

BONUS: Testing host-parasite coevolution

Function parafit tests the hypothesis of coevolution between a clade of hosts and a clade of parasites. Here coevolution is defined as the extent to which the host and parasite phylogenetic trees are congruent, where congruence refers to the degree to which parasites and their hosts occupy corresponding positions in the phylogenetic trees. Perfect congruence is a good indicator of host and parasite cospeciation; a total apsence of conguence indicates random associations in their evolutionary history.

The null hypothesis (H0) of the global test is that the evolution of the two groups, as revealed by the two phylogenetic trees and the set of host-parasite association links, has been independent. Tests of individual host-parasite links are also available as an option.

The method, which is described in detail in Legendre et al. (2002), requires some estimates of the phylogenetic trees or phylogenetic distances, and also a description of the host-parasite associations (H-P links) observed in nature.

```
data(gopher.D)
data(lice.D)
data(HP.links)
```

```
rownames(gopher.D) <- rownames(HP.links)
rownames(lice.D) <- colnames(HP.links)

gopher.tree <- nj(as.dist(gopher.D))
lice.tree <- nj(as.dist(lice.D))

res <- parafit(gopher.D, lice.D, HP.links, nperm=99, test.links=TRUE)
```

```
## n.hosts = 15 , n.parasites = 17
## Computation time = 0.340000  sec
```

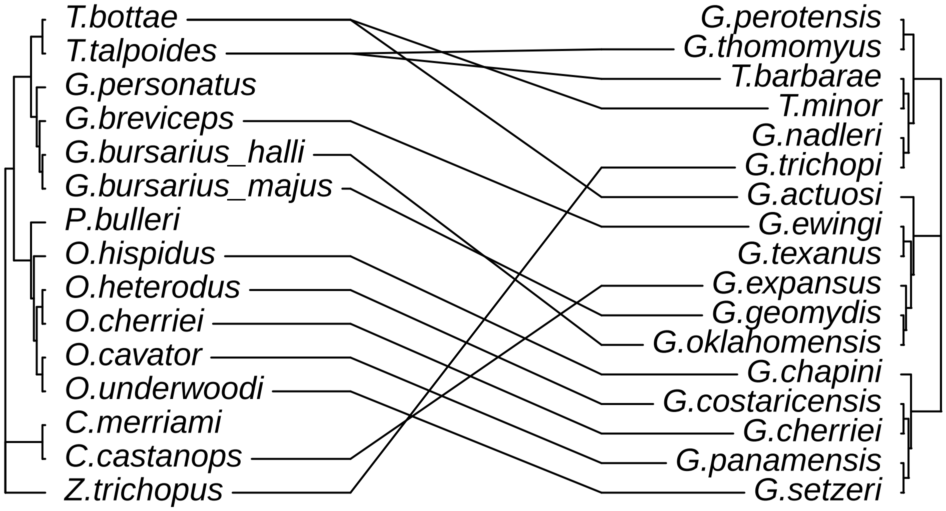
```
res
```

```
##
## Test of host-parasite coevolution
##
## Global test:  ParaFitGlobal = 0.01389872 , p-value = 0.01 ( 99 permutations)
##
## There are 17 host-parasite links in matrix HP
##
## Test of individual host-parasite links ( 99 permutations)
##
##       Host Parasite      F1.stat p.F1    F2.stat p.F2
##  [1,]    1        2 0.0009312199 0.02 0.08323295 0.01
##  [2,]    1        8 0.0011611181 0.08 0.10378138 0.01
##  [3,]    2        1 0.0010501927 0.03 0.09386681 0.01
##  [4,]    2        9 0.0006711532 0.20 0.05998804 0.02
##  [5,]    3        3 0.0017178115 0.01 0.15353895 0.01
##  [6,]    4        7 0.0010412160 0.01 0.09306447 0.01
##  [7,]    5        5 0.0016337474 0.01 0.14602526 0.01
##  [8,]    6        4 0.0019256509 0.01 0.17211575 0.01
##  [9,]    7        6 0.0015769140 0.01 0.14094546 0.01
## [10,]    8       16 0.0012866890 0.02 0.11500498 0.01
## [11,]    9       14 0.0007419528 0.14 0.06631616 0.05
## [12,]   10       17 0.0013933007 0.03 0.12453400 0.02
## [13,]   11       15 0.0015419676 0.02 0.13782193 0.02
## [14,]   12       10 0.0007215688 0.23 0.06449422 0.03
## [15,]   13       11 0.0004458883 0.53 0.03985375 0.15
## [16,]   14       13 0.0006382121 0.38 0.05704375 0.05
## [17,]   15       12 0.0008493063 0.07 0.07591147 0.02
##
##  Number of parasites per host
##       T.talpoides          T.bottae      O.underwoodi        O.hispidus
##                 2                 2                 1                 1
##         O.cavator        O.cherriei       O.heterodus       G.breviceps
##                 1                 1                 1                 1
##      G.personatus G.bursarius_majus G.bursarius_halli       C.castanops
##                 1                 1                 1                 1
##        C.merriami         P.bulleri       Z.trichopus
##                 1                 1                 1
##
##  Number of hosts per parasite
##         T.minor      T.barbarae       G.setzeri      G.cherriei
##               1               1               1               1
##    G.panamensis G.costaricensis       G.chapini     G.thomomyus
##              1               1               1               1
##       G.actuosi      G.expansus    G.perotensis      G.trichopi
##              1               1               1               1
##       G.nadleri       G.texanus G.oklahomensis        G.ewingi
##              1               1               1               1
##      G.geomydis
##              1
```

```
links <- res$link.table[,1:2]
links[,1] <- rownames(gopher.D)[as.numeric(links[,1])]
links[,2] <- rownames(lice.D)[as.numeric(links[,2])]
```

We can then plot the significant cophylogenetic associations:

```
links.sig <- links[res$link.table[,"p.F2"]<0.05,]
links.nonsig <- links[res$link.table[,"p.F2"]>0.05,]

cophyloplot(gopher.tree, lice.tree, assoc=links.sig, gap=10, space=300, length.line=80, rotate=FALSE)
```



Try plotting with `rotate=TRUE`!