

# Phylogenetic Diversity - Traits

*Z620: Quantitative Biodiversity, Indiana University*

*February 17, 2017*

## OVERVIEW

Up to this point, Quantitative Biodiversity has focused on aspects of taxonomic diversity. However, measures of taxonomic diversity overlook the evolutionary relatedness (i.e., phylogenetics) of individuals in a community. In this Handout, we introduce concepts and tools that will allow you to integrate phylogenetic information to test and better understand how evolutionary history influences patterns of biodiversity.

After completing this exercise you will be able to:

1. create phylogenetic trees to view evolutionary relationships from sequence data
2. map the traits of taxa onto phylogenetic trees
3. test for phylogenetic signal within trait distributions and trait-based patterns of biodiversity
4. control for the phylogenetic relatedness of taxa to identify correlations among

## BACKGROUND

Functional traits are morphological, behavioral, or physiological characteristics of an individual that affect its performance or fitness under a set of environmental conditions. We can think of functional traits as falling into two general categories. First, “response traits” influence how an individual contends with features of its environment (Lavorel and Garnier 2002). For example, waxy leaves are generally thought to aid in plant tolerance to desiccation. Second, “effect traits” are characteristics of an individual that can alter the environment (Lavorel and Garnier 2002). For example, the ability for a plant and its associated microorganisms to fix atmospheric nitrogen can influence soil fertility. It is often assumed that more closely related (i.e., more recently diverged) taxa have a greater degree of trait similarity than more distantly related taxa that diverged further back in time. Therefore, it is possible to detect trait-environment and trait-function relationships simply because an assemblage is comprised of closely related taxa. We can not only test for this so-called **phylogenetic signal**, but also correct for it. In order to explore the relationship between the evolutionary relatedness of taxa and the distribution of functional traits among those taxa, we must introduce another fundamental data structure: the **phylogeny**. A phylogeny is an evolutionary tree constructed by comparing the divergence in gene or protein sequences among a group of organisms. Constructing a phylogeny relies on the assumption that more distantly related organisms have accumulated more differences in their sequences since the time they last shared a common ancestor, while the sequences of more closely related taxa would be more similar. In the following sections, we will provide an introduction to constructing phylogenetic trees. From this, you will gain evolutionary insight into patterns and processes of biodiversity (e.g., diversification, adaptation, etc.).

## 1) SETUP

### A. Retrieve and Set Your Working Directory

```
rm(list = ls())  
getwd()  
setwd("~/GitHub/QB-2017/Week6-PhyloTraits")
```

## B. Load Packages

Several R packages have been developed for conducting phylogenetic analyses (<http://goo.gl/DtU16j>). We will describe the packages in greater details in the sections below. The `require()` function in R returns `TRUE` if the package was successfully loaded or `FALSE` if the package failed to load. This `for` loop loads each package and installs the package when `require()` returns `FALSE`.

```
package.list <- c("ape", "seqinr", "phylobase", "adephylo", "geiger",
  "picante", "stats", "RColorBrewer", "caper", "phylolm", "pmc",
  "ggplot2", "tidyr", "dplyr", "phangorn", "pander")
for (package in package.list) {
  if (!require(package, character.only = TRUE, quietly = TRUE)) {
    install.packages(package)
    library(package, character.only = TRUE)
  }
}
```

## 2) DESCRIPTION OF DATA

We will address question related to functional traits and phylogenetics using a data set that contains information on 39 strains of freshwater bacteria. The bacteria were isolated from two lakes that greatly differ in their biogeochemical properties, especially nutrient content. One lake has extremely low concentrations of phosphorus (i.e., oligotrophic) while the other lake is adjacent to a bird sanctuary and receives high inputs of phosphorus (i.e., eutrophic). We hypothesize that these two environments contain taxa with different traits that influence their ability to use different types of phosphorus. To test this hypothesis, we extracted DNA from all of bacterial strains and amplified the 16S rRNA gene, which is a commonly used “marker” for delineating microbial taxa and examining phylogenies of bacteria. In addition, we measured the maximum growth rate  $\mu_{max}$  of each strain on 18 different phosphorus resources. With this information, we can measure the **niche breadth** of each strain and consider this a functional trait. In the following sections, we take a phylogenetic approach to mapping phosphorus resource use onto a phylogenetic tree while testing for ecological trade-offs related to resource use.

## 3) SEQUENCE ALIGNMENT

One of the first steps in performing a phylogenetic analysis is to align the sequences that have been obtained from your samples. The sequences can represent nucleotides (DNA or RNA) or proteins (amino acids). Computer algorithms are used to arrange sequences relative to one another based on conserved (i.e., non-variable) regions. The non-conserved or “variable” regions are then used for making phylogenetic inference, including tree construction. There are two different ways to align sequences. Global alignment relies on algorithms that attempt to align the entire span of sequences. Local alignment is typically a preferable approach and uses algorithms that initiate alignment by identifying regions in a sequence that have a high degree of similarity. In the example below, we will be using a local alignment approach.

### A. Examining a FASTA-File

We will read in a FASTA-formatted file into the `muscle` alignment program. FASTA is a very common text-based format that is used across platforms and disciplines. The files contain sequences (e.g., CTAAG) and names that are associated with samples. Let’s take a second to look at the `p.isolates.fasta` file. Open the terminal and change directory to the data folder. You can then type `less p.isolates.fasta` at the command line. You can use [Ctrl+V] and [Ctrl+Y] to scroll towards the bottom and top of the file, respectively. When you are done, you can close the file by hitting [Ctrl+X]. Do not save any inadvertent changes to `p.isolates.fasta`.

## B. Performing an Alignment

R is a great data analysis language, but it is not an all-inclusive environment for conducting phylogenetic analyses. Fortunately, the R language and R Studio environment can be made to interact with other software and computing languages. Here we will use the `muscle` program in the command language interpreter `Bash`, the language that you are using on the lab Macs. Using `Bash`, we will call the program `muscle`, read in our FASTA file, and create a FASTA-aligned output file (with an `.afa` extension).

```
muscle -in ./data/p.isolates.fasta -out ./data/p.isolates.afa
```

## C. Visualizing the Alignment

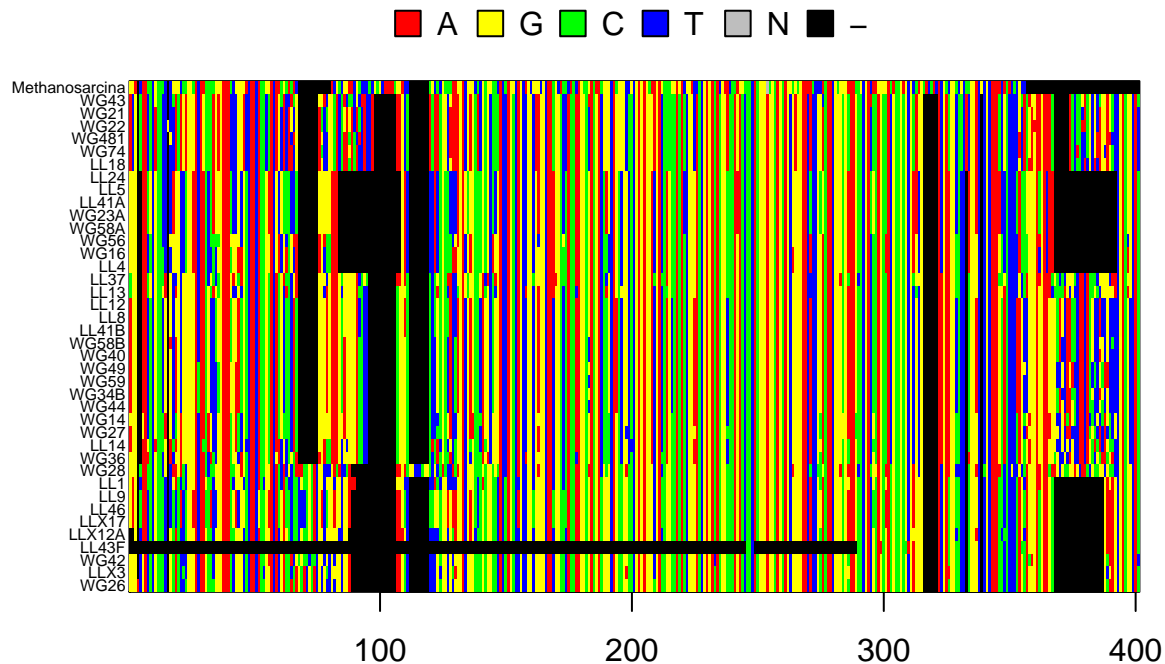
After performing the alignment, it is good practice to view your aligned sequence file. From this, you should be able to visualize conserved and variable regions among your sequences. This may not be practical for databases with a large number of long sequences. In our dataset, however, the alignment only contains sequences for 39 bacterial taxa plus a distantly related reference sequence that we will use as an **outgroup**. To visualize our alignment, first, we will read in the FASTA-aligned file using the `read.alignment` function in the `seqinr` package. Then, using the `ape` package, we will convert the alignment file into a DNABin object, which is a bit-level coding scheme that allows R to store and manipulate sequence data. Last, we will visualize the alignment by color-coding nucleotides different colors, where red = adenine, yellow = guanine, green = cytosine, blue = thymidine, grey = ambiguous call, and black = alignment gaps. (Alignment gaps are sites that are introduced during an alignment to increase matching and can reflect insertion and deletions.)

```
# Read Alignment File {seqinr}
read.aln <- read.alignment(file = "./data/p.isolates.afa", format = "fasta")

# Convert Alignment File to DNABin Object {ape}
p.DNABin <- as.DNABin(read.aln)

# Identify Base Pair Region of 16S rRNA Gene to Visualize
window <- p.DNABin[, 100:500]

# Command to Visualize Sequence Alignment {ape}
image.DNABin(window, cex.lab = 0.50)
```



```
# Optional Code Adds Grid to Help Visualize Rows of Sequences
#grid(ncol(window), nrow(window), col = "lightgrey")
```

An inspection of the figure suggests that our alignment is good because many vertical bars are aligned with the same color, indicating they share the same nucleotide at that site. If this visualization showed a jumbled mess, we may want to realign our sequences using different parameters/programs, etc.

#### 4) MAKING A PHYLOGENETIC TREE

Once you have aligned your sequences, the next step is to construct a phylogenetic tree. Not only is a phylogenetic tree effective for visualizing the evolutionary relationship among taxa, but as you will see later, the information that goes into a phylogenetic tree is needed for downstream analysis. There are many algorithms, theories, and tools that are used for making phylogenetic trees. It is not possible to cover all of these topics in Quantitative Biodiversity, but we try to highlight some of the major concepts in the following sections, including the table below:

Tree Method	Properties
UPGMA	Hierarchical clustering technique. This technique progressively groups the most similar pairs of sequences. Sometimes used as a guide tree for more sophisticated phylogenetic analyses, but prone to distorting tree topology and assumes constant rates of evolution. UPGMA can be implemented in R with the <i>upgma()</i> function in the <i>panghorn</i> package.
Neighbor Joining	Starts with an unresolved tree in a "star network". Algorithm then identifies the two most similar taxa and creates a node. This process is iteratively applied to the remaining sequences. Can serve as a starting point for assessing evolutionary models with more sophisticated methods like maximum likelihood. Neighbor joining can be implemented in the R package <i>ape</i> using the <i>bionj()</i> function.
Maximum Parsimony	A character-based, non-parametric approach that involves identifying one of many possible trees based on the minimum number of evolutionary "steps" (i.e. the most parsimonious tree). No underlying evolutionary model. Maximum parsimony can be executed using the <i>parsimony()</i> function in the package <i>panghorn</i> . The function <i>optim.parsimony</i> performs tree rearrangements using nearest-neighbor interchanges (NNI) to optimize parsimony scores.
Maximum Likelihood	Parametric approach to identifying the relative probability associated with data fitting a given model of character evolution. One of the most commonly used approaches for building trees from sequence data. Statistically sound and flexible for accommodating realistic models of sequence evolution. Maximum likelihood can be implemented in R using the <i>pml()</i> function in the <i>panghorn</i> package. Parameters can be optimized using the <i>optim.pml()</i> function and evolutionary models can be compared with the <i>model.tests()</i> function
Bayesian	Uses likelihood function and prior probabilities to create a posterior probability of a tree fitting a given model of evolution. Computationally intensive; often uses Markov chains Monte Carlo (MCMC). Unlike other methods, parameters are not constant; rather, they come from probabilistic distributions. Along with maximum likelihood, considered one of the more preferable ways of generating trees. R has some packages (e.g., <i>ips</i> ) the wrap popular Bayesian software such as MrBayes

## A. Neighbor Joining Trees

As a method, nearest neighbor joining takes a distance matrix that specifies the distance between each pair of taxa as input. The algorithm starts with a tree where all taxa are equally related, finds and joins the taxa with the lowest distance into a new node, re-calculates the distance between all the remaining taxa and the new node, and repeats this process until all the taxa are accounted for.

Neighbor joining is commonly used for making a "guide tree" that can incorporate more sophisticated models of evolution. The first step in making a neighbor joining tree is to create a distance matrix. We will do this with the `dist.dna` function in the `ape` package. In the R chunk below, the "model = raw" argument means that `dist.dna` will estimate distances based on the proportion of sites that differ between pairs of sequences. The "pairwise.deletion = FALSE" argument means that an entire site (i.e., column) is deleted if there is a

missing observation (i.e., a gap) for one or more of the sequences (i.e., rows).

```
# Create Distance Matrix with "raw" Model {ape}  
seq.dist.raw <- dist.dna(p.DNABin, model = "raw", pairwise.deletion = FALSE)
```

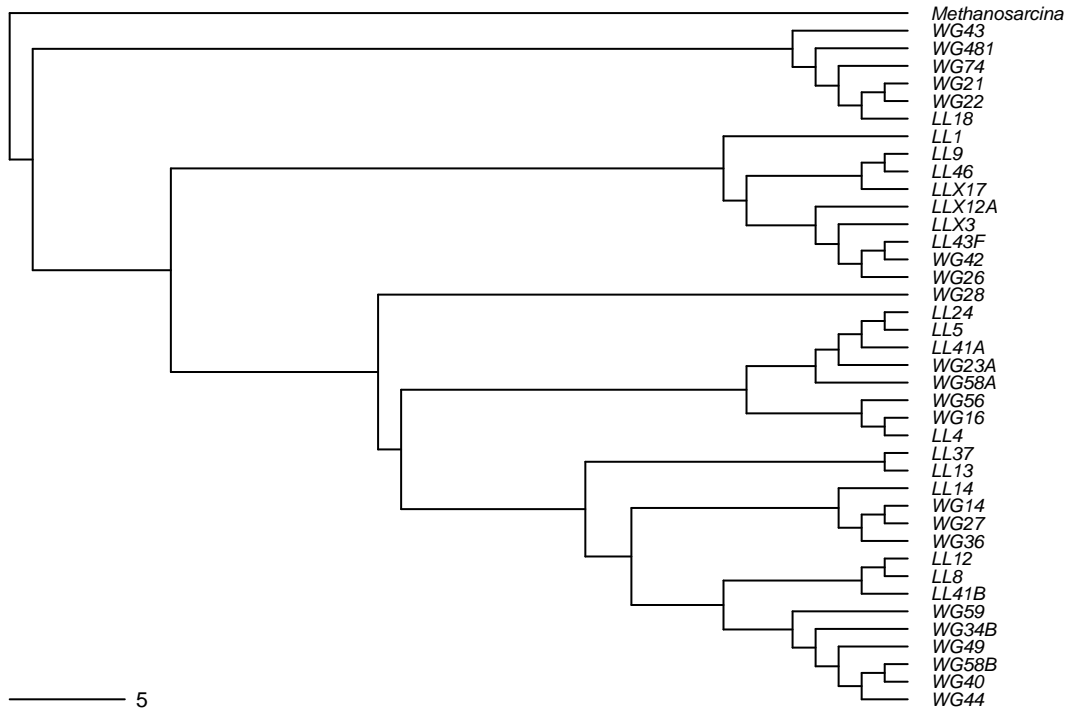
After calculating the distances between our sequences, we are now ready to make a neighbor joining tree using the `bionj` function in `ape`. In the process, we will use *Methanosarcina* as a reference to determine the evolutionary relationship among the members of our tree (i.e. an **outgroup**). By using an outgroup we have a point of comparison for the rest of our taxa (i.e. the **ingroup**), also known as **rooting** the tree. Ideally, the outgroup should be a taxon that you know all of your taxa diverged from farther back in the past than when they diverged from each other.

To store phylogenies we'll be using the `phylo` class, the basic phylogenetic data structure used in R. A `phylo` object stores four objects: 1) the tip labels (i.e. taxa) of the tree, 2) the number of nodes (i.e. the inferred most recent common ancestor) on the tree, 3), the names of the edges (i.e. the evolutionary distance between nodes) on the tree, and 4) and the length of the edges

At this point, we can plot our tree.

```
# Neighbor Joining Algorithm to Construct Tree, a 'phylo'  
# Object {ape}  
nj.tree <- bionj(seq.dist.raw)  
  
# Identify Outgroup Sequence  
outgroup <- match("Methanosarcina", nj.tree$tip.label)  
  
# Root the Tree {ape}  
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)  
  
# Plot the Rooted Tree {ape}  
par(mar = c(1, 1, 2, 1) + 0.1)  
plot.phylo(nj.rooted, main = "Neighbor Joining Tree", "phylogram",  
  use.edge.length = FALSE, direction = "right", cex = 0.6,  
  label.offset = 1)  
add.scale.bar(cex = 0.7)
```

## Neighbor Joining Tree



## B) SUBSTITUTION MODELS OF DNA EVOLUTION

You now have a distance matrix that represents the shared evolutionary history of your taxa. However, these are raw estimates of phylogenetic distance and they do not account for the fact that multiple substitutions may have occurred at the same site over time (e.g., A -> T -> G -> C would be counted as one change of base), something that is likely common if you are dealing with taxa that diverged hundreds of millions of years ago. In addition, raw distance measures do not take into account substitution biases for one particular nucleotide over the other (e.g., A-T biased mutation). Various models have been developed to account for multiple substitutions and estimate the rates that nucleotides change over time. In the following table, we describe a few of the more commonly used models that correct for multiple substitutions as well as the probability of one nucleotide transitioning to a different nucleotide. We will then correct the raw distances of your matrix.

Evolutionary Model	Properties
<b>Jukes-Cantor model (JC69)</b>	Simplest model. Assumes all nucleotides occur at equal frequencies and that these nucleotides can mutate from one to another with equal probability.
<b>Felsenstein model (F81)</b>	Builds from JC69 by allowing nucleotide frequencies to vary.
<b>Kimura model (K80)</b>	Assumes equal frequencies of nucleotides, but recognizes that transition mutations (e.g., purine to a purine [A -> G] or pyrimidine to pyrimidine [C -> T]) occur with higher probability than transversion mutations (e.g., purine to pyrimidine or vice-versa).
<b>Felsenstein model (F84)</b>	Assumes different rates of base transitions and transversions while allowing for differences in base frequencies.
<b>Tamura model (T92)</b>	Similar to K80 but accounts for G + C content.
<b>General Time Reversible model (GTR)</b>	Captures the probabilities associated with nucleotide reversions (e.g., T -> C -> T). All substitution rates are different. Does not assume equal base frequencies.

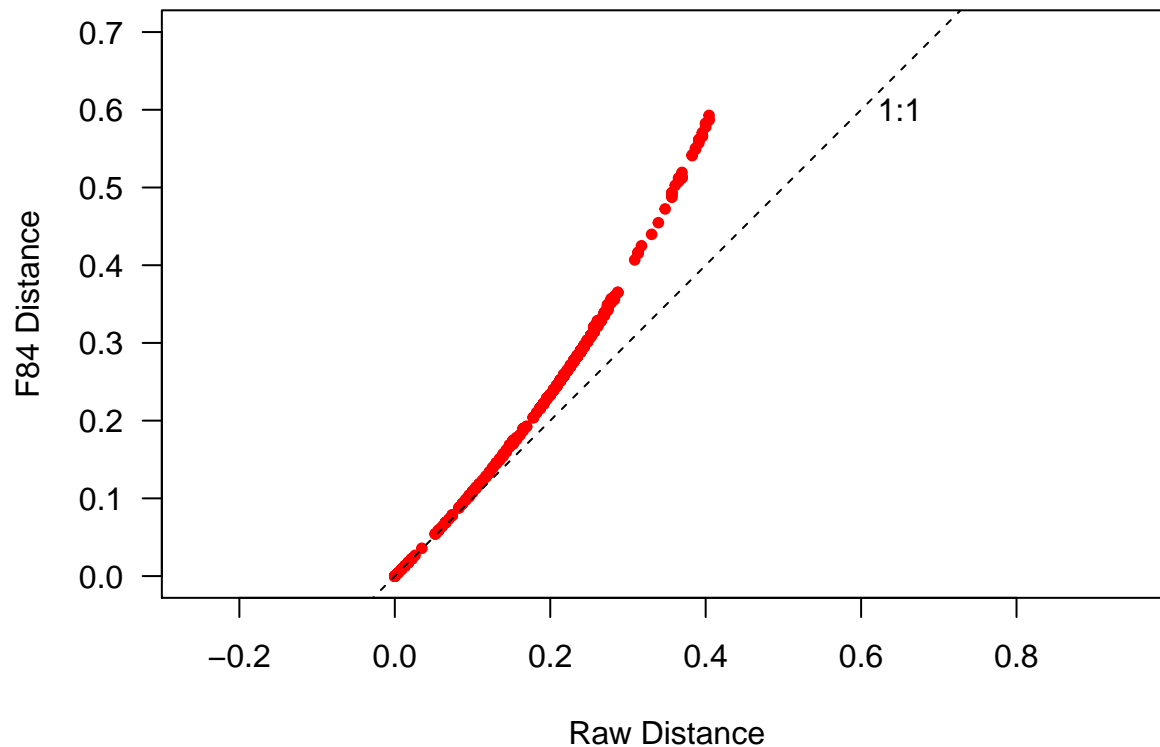
Fortunately, the `dist.dna` function that we used above for generating the neighbor joining tree can easily create distance matrices for a large number of DNA substitution models. Let's create a distance matrix among our bacterial isolates based on the Felsenstein (F84) substitution model using the `dist.dna` function in `ape`.

```
# Create distance matrix with "F84" model {ape}
seq.dist.F84 <- dist.dna(p.DNABin, model = "F84", pairwise.deletion = FALSE)
```

Now, let's compare the "raw" and "F84" distance matrices. First, we will make a **saturation plot**, which allow us to visualize the effect of our DNA substitution model compared to a distance matrix that does not include this feature.

```
# Plot Distances from Different DNA Substitution Models
par(mar = c(5, 5, 2, 1) + 0.1)
plot(seq.dist.raw, seq.dist.F84,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0, 0.7), ylim = c(0, 0.7),
     xlab = "Raw Distance", ylab = "F84 Distance")
abline(b = 1, a = 0, lty = 2)
text(0.65, 0.6, "1:1")
```





We can see from this figure that the F84 distance tends to be greater than the raw distance, so it is **correcting** for multiple substitutions

Second, we will assess how the “raw” and “F84” distance matrices affect tree topology by plotting each tree next to each other (also known as a **cophylogenetic plot**). This will give us a visual sense of whether or not the two trees are in agreement with one another.

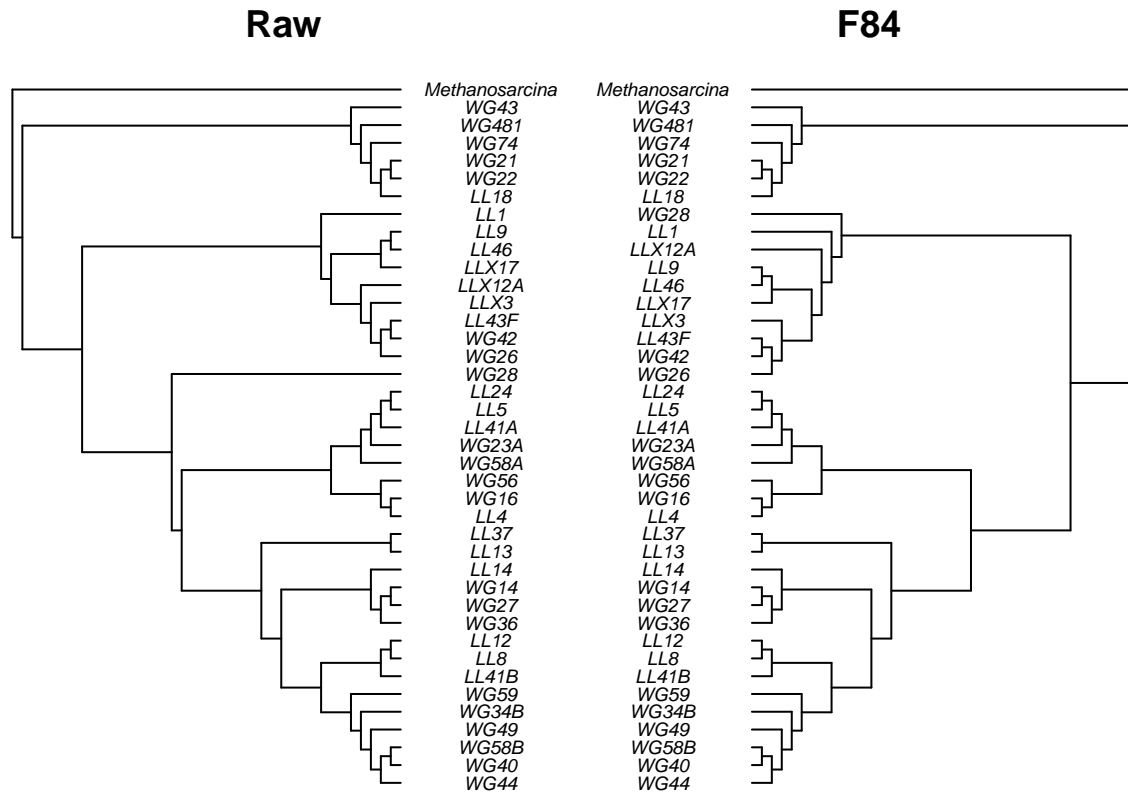
```
# Make Neighbor Joining Trees Using Different DNA Substitution Models {ape}
raw.tree <- bionj(seq.dist.raw)
F84.tree <- bionj(seq.dist.F84)

# Define Outgroups
raw.outgroup <- match("Methanosarcina", raw.tree$tip.label)
F84.outgroup <- match("Methanosarcina", F84.tree$tip.label)

# Root the Trees {ape}
raw.rooted <- root(raw.tree, raw.outgroup, resolve.root=TRUE)
F84.rooted <- root(F84.tree, F84.outgroup, resolve.root=TRUE)

# Make Cophylogenetic Plot {ape}
layout(matrix(c(1,2), 1, 2), width = c(1, 1))
par(mar = c(1, 1, 2, 0))
plot.phylo(raw.rooted, type = "phylogram", direction = "right", show.tip.label=TRUE,
           use.edge.length = FALSE, adj = 0.5, cex = 0.6, label.offset = 2, main = "Raw")

par(mar = c(1, 0, 2, 1))
plot.phylo(F84.rooted, type = "phylogram", direction = "left", show.tip.label=TRUE,
           use.edge.length = FALSE, adj = 0.5, cex = 0.6, label.offset = 2, main = "F84")
```



### C) QUANTIFYING PHYLOGENETIC SIMILARITY

Sometimes we will want to quantify the difference between two phylogenies rather than simply plotting them side-by-side. The package `ape` includes two ways to calculate this difference: 1) the branch length score (BLS) (Kuhner and Felsenstein, 1994) and the symmetric difference (Robinson and Foulds 1981). The symmetric difference is simply the number of partitions (i.e. the number of ways to group the taxa in our tree) that are in one tree and not the other. This is a very simple calculation, but it is extremely sensitive to small differences between the trees. The branch length score between trees  $T$  and  $T'$  can be described as  $BLS(T, T') = \sum_{i=1}^N (t_i - t'_i)^2$ , where  $t_i$  is the branch length if the  $i$ th partition is in both trees and is zero if it isn't. This function should look familiar, since it is just the squared Euclidean distance between two trees and it ranges from values of 0 to 1. The history, appropriateness, and limitations of these methods are nicely discussed in chapter 30 of Felsenstein (2004).

Here we'll calculate branch length score between the raw tree and the tree calculated with F84 distances.

```
# Set method = 'PH85' for the symmetric difference Set method
# = 'score' for the symmetric difference This function
# automatically checks for a root and unroots rooted trees,
# so you can pass it either the rooted or unrooted tree and
# get the same answer.
dist.topo(raw.rooted, F84.rooted, method = "score")

## [1] 0.04387426
```

### D) MAXIMUM LIKELIHOOD TREE

Often a nearest neighbor joined tree does not accurately reconstruct the real phylogeny. While neighbor joining is very fast and has the convenient property of returning the correct tree for a correct distance matrix,

we are rarely in the position where we know the correct distance matrix. In addition, it 1) it only uses a distance matrix, which doesn't take into account specific nucleotide states, 2) it only gives one tree, meaning that it is not a statistical method, and 3) it can be sensitive to the underlying substitution model. Instead, a maximum likelihood (ML) tree is often a superior choice, as it is built on the robust statistical procedure of ML (i.e., the process of finding the parameter value that maximizes the likelihood of the data). In addition, ML has the advantage of being the estimation method that is least affected by sampling error, is fairly robust to the assumptions of a particular model, allows you to compare different trees, and takes into account nucleotide states (as opposed to just distance). However, ML estimation is a computationally intensive process that can require a lot of time.

This is because ML estimation needs to find the tree that gives the data the highest probability, a difficult task that requires the use of optimized algorithms. Because ML estimation can take a long time, we will not ask you to generate a maximum likelihood tree for your assignment. However, below we have supplied code using functions from the package `phangorn` to generate a ML tree.

```
# Requires alignment to be read in with as phyDat object
p.DNABin.phyDat <- read.phyDat("./data/p.isolates.afa", format="fasta", type="DNA")
fit <- pml(nj.rooted, data=p.DNABin.phyDat)
# Fit tree using a JC69 substitution model
fitJC <- optim.pml(fit, TRUE)
# Fit tree using a GTR model with gamma distributed rates.
fitGTR <- optim.pml(fitGTR, model="GTR", optInv=TRUE, optGamma=TRUE,
# rearrangement = "NNI", control = pml.control(trace = 0))
# You can perform model selection with either an ANOVA test or by AIC value
anova(fitJC, fitGTR)
AIC(fitJC)
AIC(fitGTR)

# And you can perform a bootstrap test to see how well-supported the edges are
bs = bootstrap.pml(fitJC, bs=100, optNni=TRUE,
  control = pml.control(trace = 0))
```

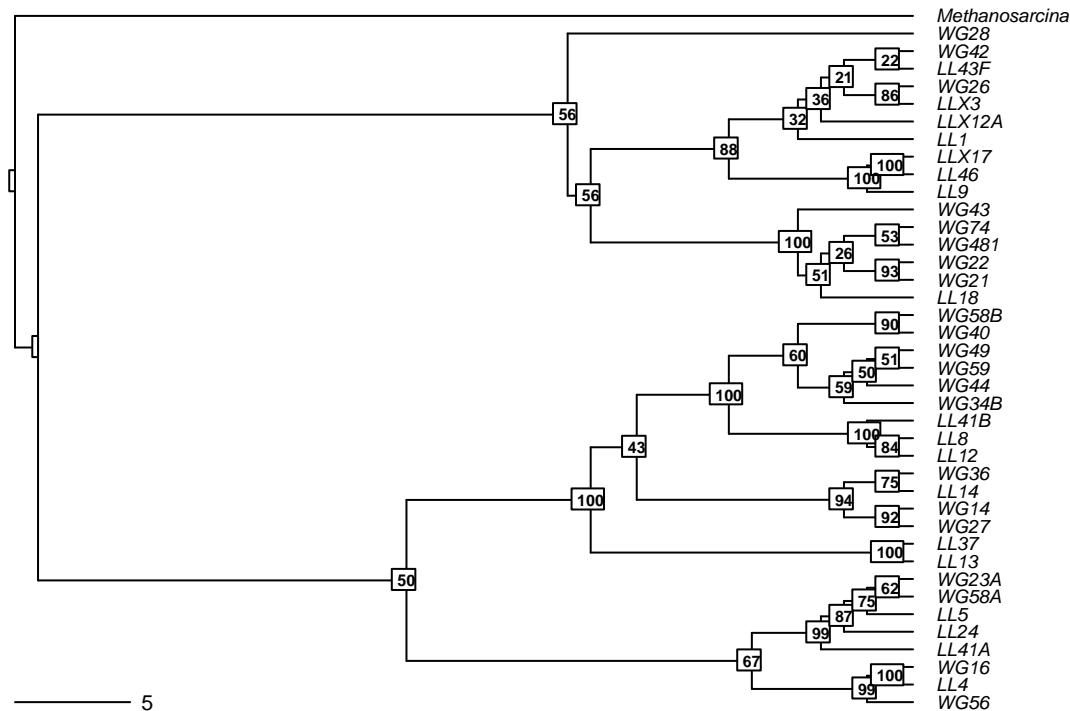
## E) BOOTSTRAP SUPPORT

Because building a phylogenetic tree is very much a statistical exercise we often want to know how confident we are in the placement of each branch in our phylogeny. For example, say that we make a tree today and tomorrow we run the same code, but get a tree with a node in a different place. How do we know which tree is correct? Since we don't know the real tree we'll use the next best thing. We'll re-sample our own data to determine how reliable the tree is. This process is known as **bootstrapping**. To perform bootstrapping, we will select a subset of columns from our alignment randomly with replacement. Next we will use that sample to construct a new tree and compare its topology to the original tree. Each interior branch of the original tree that is different from the bootstrap tree gets a score of 0 while all other interior branches get a score of 1. This procedure is done a large number of times (typically 100 - 10,000, depending on the size of the tree) and it gives us the **bootstrap value** for each interior branch. A general rule of thumb is to treat bootstrap values 95% or higher as operationally correct. Again, because this procedure is computationally intensive you will plot bootstrap support values for a tree that has been previously generated.

```
ml.bootstrap <- read.tree("./data/ml_tree/RAxML_bipartitions.T1")
par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(ml.bootstrap, type = "phylogram", direction = "right",
  show.tip.label = TRUE, use.edge.length = FALSE, cex = 0.6,
  label.offset = 1, main = "Maximum Likelihood with Support Values")
add.scale.bar(cex = 0.7)
nodelabels(ml.bootstrap$node.label, font = 2, bg = "white", frame = "r",
```

cex = 0.5)

## Maximum Likelihood with Support Values



### F) A NOTE ON MAKING YOUR OWN TREES.

R is useful for a lot of things, but it is not optimized for phylogenetic reconstruction and therefore can be extremely slow. If you need to make a phylogeny with more than a dozen or so taxa, then you may need to use packages that don't easily run in an R environment. One of the packages that you will want to check out is the Randomized Axelerated Maximum Likelihood (RAxML) package for generating maximum likelihood trees with bootstrap support (<http://sco.h-its.org/exelixis/web/software/raxml/index.html>). If you're dealing with communities with a large number of taxa (e.g., 16S rRNA amplicon data from microbial communities), programs like RAxML may be too slow. Here you can use the program FastTree to generate approximately-maximum-likelihood phylogenetic trees, sacrificing exactness for speed (<http://www.microbesonline.org/fasttree/>).

To help you with your future tree-building efforts we have included a short script under the folder **Week6-PhyloTraits/bash/** that will allow you to generate a ML tree on the IU computer cluster Mason using the program RAxML. If you are an IU student, postdoc, or faculty member you can register for an account on Mason. See the IU Mason webpage for more information (<https://kb.iu.edu/d/bbhh>).

## 5) INTEGRATING TRAITS AND PHYLOGENY

In the context of biodiversity, many researchers are interested in being able to map traits onto phylogenetic trees. This can be important for addressing questions related to diversification, trait evolution, and ecological interactions. In this section, we will import functional trait data that was collected on the 39 freshwater isolates for which we have 16S rRNA sequences. Specifically, we measured the growth rates of each strain on 18 different forms of phosphorus (e.g., phosphate, DNA, ATP, etc.). With this information we will visualize the functional traits (i.e., growth on different phosphorus) in a phylogenetic context by mapping them onto

the neighbor joining tree constructed under the F84 DNA substitution model (see the table on substitution models) that we created.

## A. Loading Trait Database

```
# Import Growth Rate Data
p.growth <- read.table("./data/p.isolates.raw.growth.txt", sep = "\t", header = TRUE,
                      row.names = 1)

# Standardize Growth Rates Across Strains
p.growth.std <- p.growth / (apply(p.growth, 1, sum))
```

## B. Trait Manipulations

From the section above where we described the data, you will recall that we were interested in answering questions related to trait evolution and ecological interactions. Specifically, we will be testing for a well-known trade-off related to resource use. This trade-off relates to the differing advantages of growing exceptionally well on one or a few resources, versus being able to grow moderately well on many resources, i.e., a **generalist-specialist trade-off** (MacArthur and MacArthur, 1961).

To explore the generalist-specialist trade-off using phylogenetic data, we need to create two new variables. First, we need calculate the maximum growth rate ( $\mu_{max}$ ) of each isolate across all phosphorus types. This will help us test the expectation that generalists will have lower maximum growth rates because there is a cost associated with being able to use a lot of different chemical forms of phosphorus. In contrast, we expect that specialists will have a high maximum growth rate on their preferred phosphorus resource.

```
# Calculate Max Growth Rate
umax <- (apply(p.growth, 1, max))
```

Second, we need to come up with a way of quantifying whether a strain is a generalist or a specialist. We will use the niche breadth ( $nb$ ) index from Levins (1968) which is defined as  $\frac{1}{n \cdot \sum p_{xi}^2}$ , where  $n$  is the total number of resources and  $p_{xi}$  is the proportion of observed growth for isolate  $i$  on each resource ( $x$ ).

```
levins <- function(p_xi = ""){
  p = 0
  for (i in p_xi){
    p = p + i^2
  }
  nb = 1 / (length(p_xi) * p)
  return(nb)
}
```

Let's apply our `nb()` function to the isolate growth rate data on the different phosphorus resources.

```
# Calculate Niche Breadth for Each Isolate
nb <- as.matrix(levins(p.growth.std))

# Add Row & Column Names to Niche Breadth Matrix
rownames(nb) <- row.names(p.growth)
colnames(nb) <- c("NB")
```

## C. Visualizing Traits on Trees

Now that we have identified and calculated our traits of interest, we can map those traits onto our phylogenetic tree in order to observe any major patterns. This is similar to the hypothesis-generating visualizations we did in other lessons this semester. Before we start, there are a few things that need to be done. First, we need to be sure that we are dealing with the correct version of the tree, so we will recreate the tree. Second, we need to make sure that this tree is rooted. Last, we need to remove the root, because this isn't part of our traits analysis.

```
# Generate Neighbor Joining Tree Using F84 DNA Substitution Model {ape}
nj.tree <- bionj(seq.dist.F84)

# Define the Outgroup
outgroup <- match("Methanosarcina", nj.tree$tip.label)

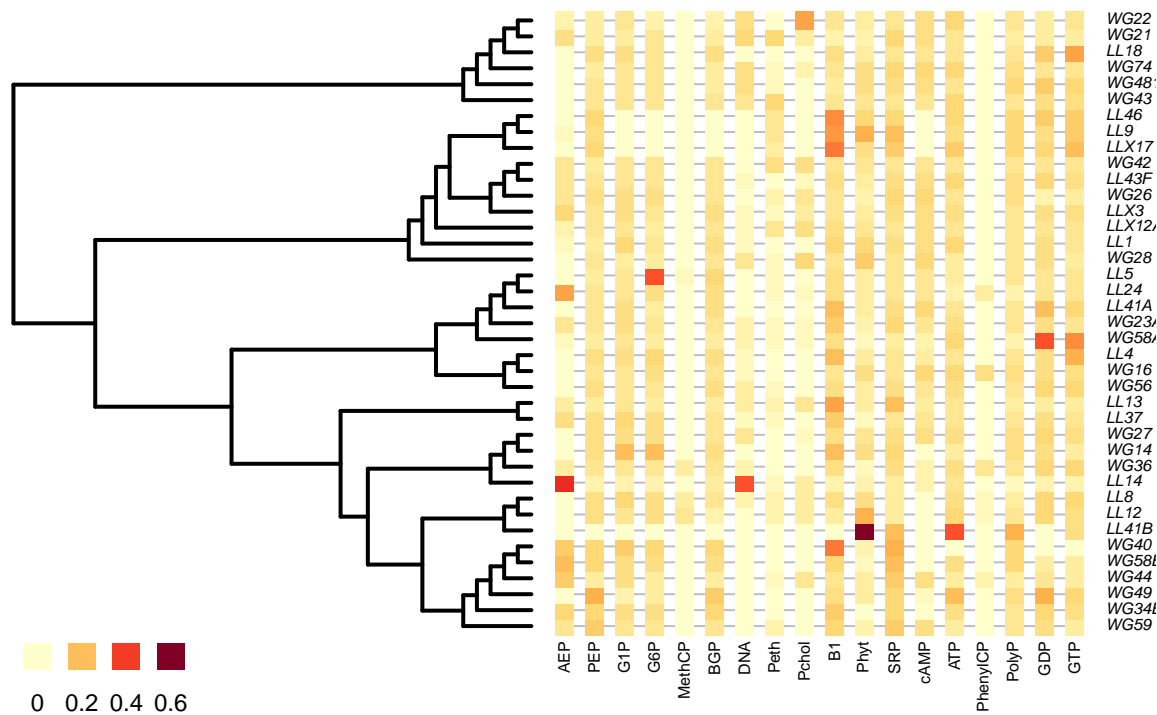
# Create a Rooted Tree {ape}
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)

# Keep Rooted but Drop Outgroup Branch
nj.rooted <- drop.tip(nj.rooted, "Methanosarcina")
```

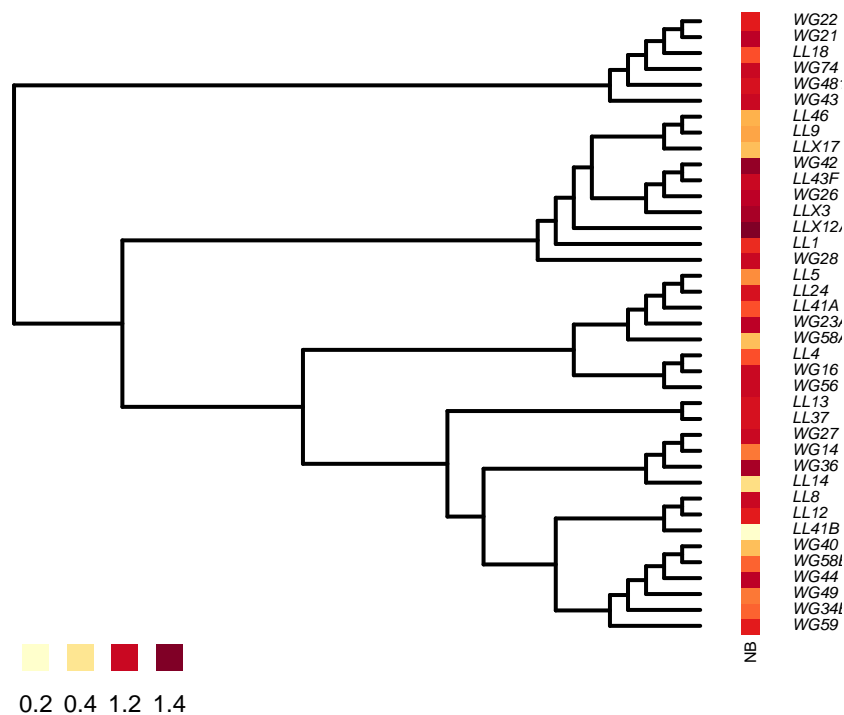
Now we can plot our traits onto the tree. This can be done with either a group of traits, such as the growth rates on different phosphorus sources, or it can be done on a single trait, such as niche breadth. Here we will do both.

```
# Define Color Palette
mypalette <- colorRampPalette(brewer.pal(9, "YlOrRd"))

# Map Phosphorus Traits {adephylo}
par(mar=c(1,1,1,1) + 0.1)
x <- phylo4d(nj.rooted, p.growth.std)
table.phylo4d(x, treetype = "phylo", symbol = "colors", show.node = TRUE,
  cex.label = 0.5, scale = FALSE, use.edge.length = FALSE,
  edge.color = "black", edge.width = 2, box = FALSE,
  col=mypalette(25), pch = 15, cex.symbol = 1.25,
  ratio.tree = 0.5, cex.legend = 1.5, center = FALSE)
```



```
# Niche Breadth
par(mar=c(1,5,1,5) + 0.1)
x.nb <- phylo4d(nj.rooted, nb)
table.phylo4d(x.nb, treetype = "phylo", symbol = "colors", show.node = TRUE,
  cex.label = 0.5, scale = FALSE, use.edge.length = FALSE,
  edge.color = "black", edge.width = 2, box = FALSE,
  col=mypalette(25), pch = 15, cex.symbol = 1.25, var.label=("    NB"),
  ratio.tree = 0.90, cex.legend = 1.5, center = FALSE)
```



## 6) HYPOTHESIS TESTING

An emerging goal of biodiversity research is to understand the effect traits have on the performance of species under different environmental conditions, but also how traits can affect ecosystem functioning. Before attempting to establish relationships among traits, however, it is important to understand the role evolution plays in the distribution of traits. Owing to shared ancestry, we expect species traits to be non-independent with respect to evolutionary history. Therefore, it is important that we test for the presence of this **phylogenetic signal** before drawing conclusions about patterns among traits over a phylogeny. In addition, these patterns of phylogenetic signal provide insight into ecological and evolutionary processes giving rise to the **clustering** or **overdispersion** of traits, which may be associated with environmental filtering and competitive repulsion, respectively. In the following sections, we introduce some methods for testing hypotheses about the distribution of traits with respect to phylogeny.

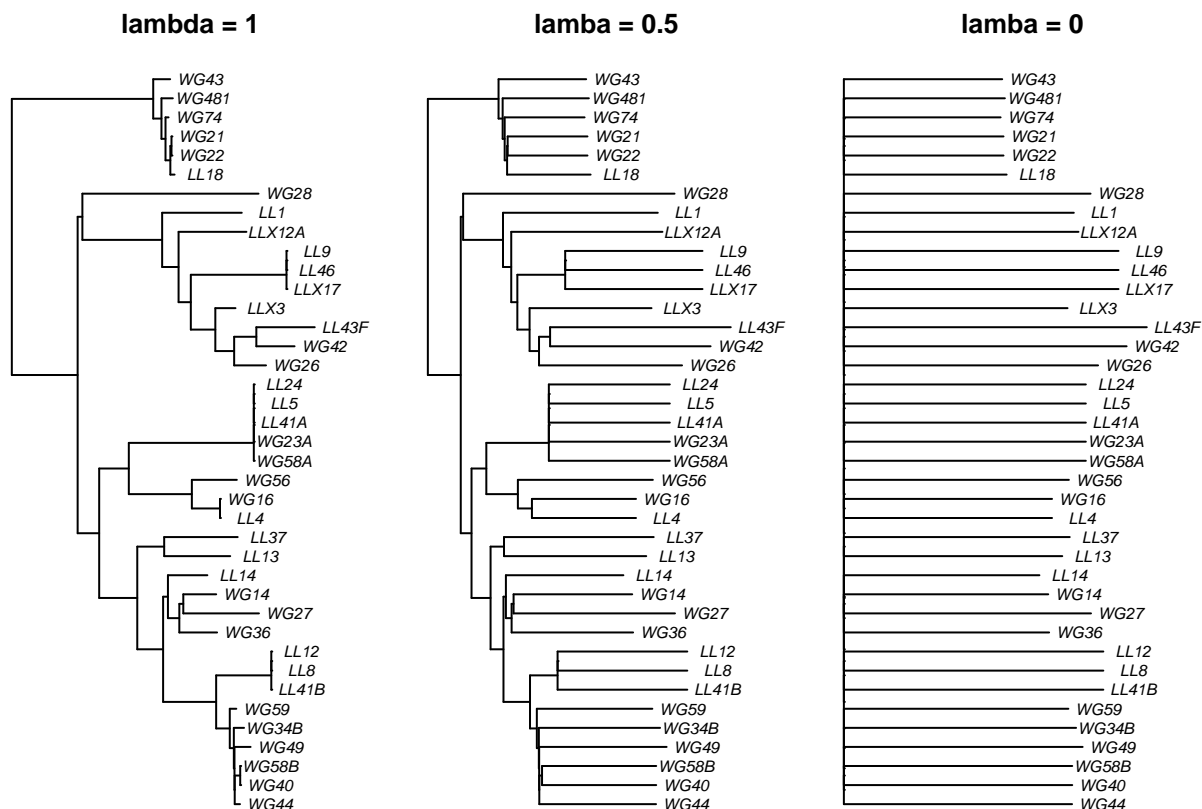
### A) Phylogenetic Signal: Pagel's Lambda

Mark Pagel developed a fairly simple technique to quantify phylogenetic signal, which involves a tree transformation (<http://goo.gl/zSSxoB>). The transformation is accomplished by scaling the off-diagonal elements of the variance-covariance matrix, which describe the tree topology and branch lengths, by the value “lambda”. When lambda equals 1, you have your original, non-transformed branch lengths. When lambda equals 0, you have removed all phylogenetic signal from your tree. Let's look at what happens when we transform our tree with lambda values of 1, 0.5, and 0.

```
# Visualize Trees With Different Levels of Phylogenetic Signal {geiger}
nj.lambda.5 <- rescale(nj.rooted, "lambda", 0.5)
nj.lambda.0 <- rescale(nj.rooted, "lambda", 0)

layout(matrix(c(1,2,3), 1, 3), width = c(1, 1, 1))
par(mar=c(1,0.5,2,0.5)+0.1)
plot(nj.rooted, main = "lambda = 1", cex = 0.7, adj = 0.5)
plot(nj.lambda.5, main = "lambda = 0.5", cex = 0.7, adj = 0.5)
plot(nj.lambda.0, main = "lambda = 0", cex = 0.7, adj = 0.5)
```





Now, let's generate quantitative output that will allow us to assess how much phylogenetic signal is in our data set. For this test, we will compare the phylogenetic signal of niche breadth using trees with lambda values of 1 (untransformed) and 0 (transformed).

```
# Generate Test Statistics for Comparing Phylogenetic Signal {geiger}
fitContinuous(nj.rooted, nb, model = "lambda")
fitContinuous(nj.lambda.0, nb, model = "lambda")
```

From here we're able to get AIC scores. Differences in AIC scores provide maximum likelihood inference regarding the quality of model fit to a given data set. Models with lower AIC values are better. However, if the difference in AIC values between two models ( $\Delta$  AIC) is not greater than at least 2, then the models are considered equivalent.

## B) Phylogenetic Signal: Blomberg's K

Blomberg et al. (2003) derived a statistic that quantifies phylogenetic signal by comparing observed trait distributions on a tree to what would evolve under Brownian (i.e., random) motion (<http://goo.gl/0xU9Et>). Blomberg's K is calculated as the mean squared error of the tip data (i.e., trait) measured from the phylogenetic-corrected mean and the mean squared error based on the variance-covariance matrix derived from the given phylogeny under the assumption of Brownian motion (Münkemüller et al 2012, <http://goo.gl/c08XrA>). A K-value of 1 means that a trait is distributed on the tree according to null expectation of Brownian motion. A K-value > 1 means that traits are **clustered** with closely related species being *more similar* than expected by chance. A K-value of < 1 means that traits are **overdispersed** with closely related species *less similar* than expected by chance.

Let's test for phylogenetic signal using Blomberg's K for the standardized growth of our bacterial isolates on *each phosphorus source* based on our neighbor joining tree.

*Note:* In this particular analysis, we are generating 18 test statistics with associated p-values (one for each

phosphorus resource). This means that we are increasing our probability of rejecting the null hypothesis (ie. “no phylogenetic signal”) when it is in fact true. This is called **false discovery rate** (FDR). We are going to correct for FDR using the Benjamini-Hochberg method (“PIC.P.BH”).

```
# First, Correct for Zero Branch-Lengths on Our Tree
nj.rooted$edge.length <- nj.rooted$edge.length + 10^-7

# Calculate Phylogenetic Signal for Growth on All Phosphorus Resources
# First, Create a Blank Output Matrix
p.phylosignal <- matrix(NA, 6, 18)
colnames(p.phylosignal) <- colnames(p.growth.std)
rownames(p.phylosignal) <- c("K", "PIC.var.obs", "PIC.var.mean",
                             "PIC.var.P", "PIC.var.z", "PIC.P.BH")

# Use a For Loop to Calculate Blomberg's K for Each Resource
for (i in 1:18){
  x <- as.matrix(p.growth.std[,i, drop = FALSE])
  out <- phylosignal(x, nj.rooted)
  p.phylosignal[1:5, i] <- round(t(out), 3)
}

# Use the BH Correction on P-values:
p.phylosignal[6, ] <- round(p.adjust(p.phylosignal[4, ], method = "BH"), 3)
```

And from the output of `p.phylosignal` we can see what traits have statistically significant phylogenetic signal.

In addition, let’s test for phylogenetic signal using Blomberg’s K on *niche breadth*:

```
# Calculate Phylogenetic Signal for Niche Breadth
signal.nb <- phylosignal(nb, nj.rooted)
signal.nb

##           K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## 1 3.427719e-06      49966.78      49873.6      0.517
## PIC.variance.Z
## 1      0.004658949
```

And we can see here that phylogenetic signal is particularly low ( $\sim 10^{-6}$ ) and that niche breadth does not have a statistically significant phylogenetic signal under Blomberg’s K ( $P \approx 0.5$ ).

### C. Calculate Dispersion of a Trait

Another way to calculate the phylogenetic signal of a trait is to use a formal test of dispersion. In ecology, dispersion is commonly used to measure how things (e.g., traits or species) are distributed in space. However, we can use similar approaches to determine how well traits are distributed across a phylogenetic tree. Fritz and Purvis (2010) derived a measure of dispersion ( $D$ ) for categorical traits (<http://goo.gl/SUrm5j>). In our case study, we can use  $D$  to determine if the ability of a bacterial isolate to grow on a specific phosphorus resource is overdispersed or clustered. Here, we will calculate  $D$  using the `phylo.d()` function in the `caper` package.

First, there are a few things we need to do: 1) turn our continuous growth data into categorical data, 2) add a column to our data set for our isolate names, and 3) combine our tree and trait data using the `comparative.data()` command in `caper`.

The output of `phylo.d()` will return the estimated value of  $D$ . Estimates of  $D$  can be negative (clustered) or positive (overdispersed). When  $D$  is close 0, traits are randomly clumped. When  $D$  is close to 1, traits are

dispersed in way that is consistent with Brownian motion.

The `phylo.d()` function uses a permutation test to calculate the probability of  $D$  being different from 1 (no phylogenetic structure; random) or 0 (Brownian phylogenetic structure). Let's use this function to calculate dispersion ( $D$ ) on a few of our phosphorus traits.

```
# Turn Continuous Data into Categorical Data
p.growth.pa <- as.data.frame((p.growth > 0.01) * 1)

# Look at Phosphorus Use for Each Resource
apply(p.growth.pa, 2, sum)

# Add Names Column to Data
p.growth.pa$name <- rownames(p.growth.pa)

# Merge Trait and Phylogenetic Data; Run `phylo.d`
p.traits <- comparative.data(nj.rooted, p.growth.pa, "name")
phylo.d(p.traits, binvar = AEP)
phylo.d(p.traits, binvar = PhenylCP)
phylo.d(p.traits, binvar = DNA)
phylo.d(p.traits, binvar = cAMP)
```

## D. Compare evolutionary models

Often you will want to determine what stochastic model of evolution best describes your data. To do this you will need to compare the fit of each model to your data and determine which provides the best fit. An appropriate statistic to use is the difference in log-likelihood between our two models:  $\delta = -2(\log L_0 - \log L_1)$ . This is a common statistic and, whether or not you knew it, you used  $\delta$  in Week2 when you calculated the fit of different species abundance distribution models using the Akaike information criterion (AIC). However, determining which model describes the evolution of trait data along a phylogeny isn't as simple as performing an AIC comparison, as likelihood-based methods such as these have high error rates that can give us the wrong answer if we don't know the uncertainty of our estimates (Boettiger et al., 2012). To get around this we can use Monte Carlo simulations to estimate the distribution of  $\delta$  and determine which model is appropriate. To do this we'll use the package Phylogenetic Monte Carlo (`pmc`) to determine whether Brownian motion or the Ornstein-Uhlenbeck model describes the evolution of niche breadth. From there we will plot our results using `ggplot2`.

Evolutionary Model	Equation	Properties
<b>Brownian motion</b>	$dY_t = \sigma Y_t dW_t$	Trait evolution proceeds as a random walk through trait space. The rate of trait evolution ( $dY_t$ ) is set by a standard Brownian motion function ( $W_t$ ) and a set constant ( $\sigma$ ). Analogous to random genetic (or ecological) drift.
<b>Brownian motion with a trend</b>	$dY_t = \mu Y_t dt + \sigma Y_t dW_t$	Trait evolution proceeds as a random walk through trait space biased in a particular direction. Analogous to directional evolution at a constant rate ( $\mu$ ) with random genetic drift.
<b>Ornstein-Uhlenbeck</b>	$dY_t = -\alpha(Y_t - \theta)dt + \sigma Y_t dW_t$	Trait evolution proceeds as a random walk through trait toward an optimal trait value ( $\theta$ ). Analogous to stabilizing selection with random genetic drift, where $\alpha$ describes how the strength of stabilizing selection increases with increasing distance from the optimal trait.

As a side note, often measures of phylogenetic signal build off of the simple stochastic models we just examined. For example, Pagel's  $\lambda$  is just the simple Brownian motion model with the internal edges of the phylogeny shortened by  $\lambda$ , reducing the correlations between taxa.

```
# Brownian motion = BM
# Brownian motion + directional selection = trend
# Ornstein-Uhlenbeck = OU
# Pagel's lambda = lambda
is.ultrametric(nj.rooted)
nj.rooted.um <- chronos(nj.rooted)

out <- pmc(nj.rooted.um, nb, "BM", "OU", nboot = 100)

# After running the above (took a long time) I got the following error:
# Warning message:
# In fitContinuous(phy = tree, dat = data, model = model, ..., ncores = 1) :
#   Parameter estimates appear at bounds:
#   alpha

dists <- data.frame(null = out$null, test = out$test)
colnames(dists) <- c("BM", "OU")

#dev.off()

png("./figs/H_test.png", width=480, height=240, res=120)

dists %>%
```

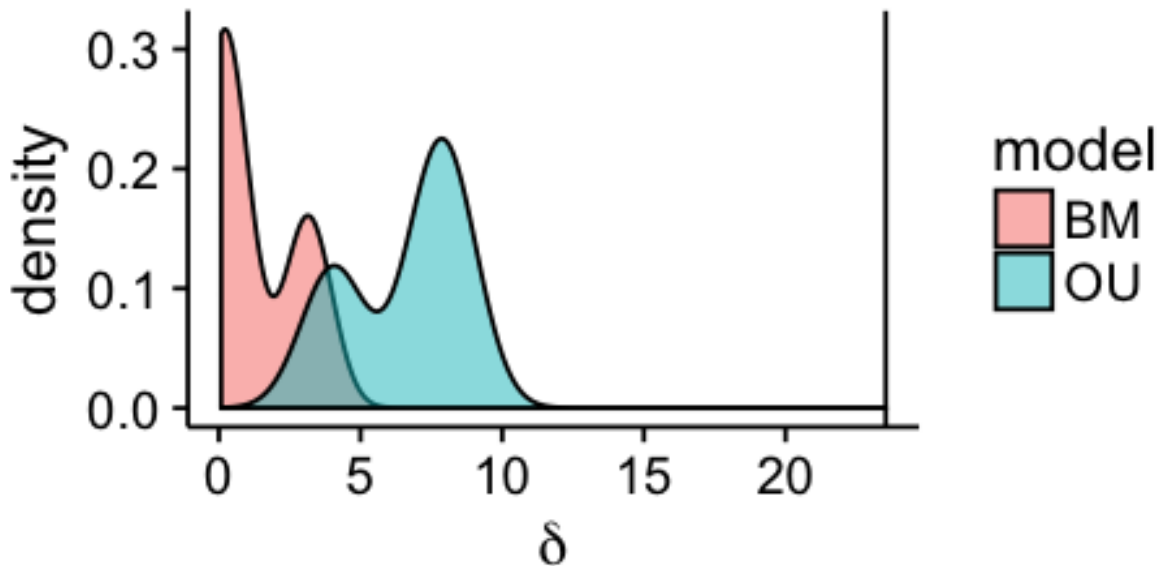


Figure 1: Density plot of the difference in log-likelihood values ( $\delta$ )

```
gather(model, value) %>%
  ggplot(aes(value, fill = model)) +
  geom_density(alpha = 0.5) +
  geom_vline(xintercept = out$lr) +
  xlab(expression(delta)) +
  theme(panel.background = element_blank())
dev.off()
```

Because the Brownian motion and Ornstein-Uhlenbeck models are nested, the proportion of simulated values larger than our observed  $\delta$  for the simpler model provides an approximation for the P-value (i.e. the probability that a difference at least as large would be seen under the null model). Using a traditional alpha of 0.05, we can determine whether we should reject the null hypothesis that Brownian motion (i.e. our simpler model) adequately describes the evolution of niche breadth.

```
length(dists$BM[dists$BM > out$lr]) / length(dists$BM)
```

Where we get a value of zero, suggesting that our P-value is smaller than 0.01.

## 7) PHYLOGENETIC REGRESSION ANALYSIS

You will often want to examine the relationship between two variables. The simplest way to do this is by performing a simple linear regression, a method that you learned in Week 1. However, as outlined above, your samples violate the assumption of independence for regression analysis due to their shared evolutionary history. The way around this issue is to perform a phylogenetic regression. The main difference between the two is that in a simple linear regression the residual errors  $\varepsilon$ , given an independent variable ( $X$ ), are assumed to be independent and identically distributed random variables that follow a normal distribution with a mean of 0 and a variance of  $\sigma^2$ :  $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$ . However, in a phylogenetic regression the variance of the residual errors are described by a covariance matrix ( $V$ ) that takes into account the branch lengths of the underlying

phylogeny:  $\varepsilon | X \sim N(0, V)$ . It is also important to take into account the phylogenetic signal in the residual error when  $V$  is estimated, not just the evolutionary history described by the phylogeny. If phylogenetic signal is not taken into account your phylogenetic regression can have poor statistical performance (Revell, 2010).

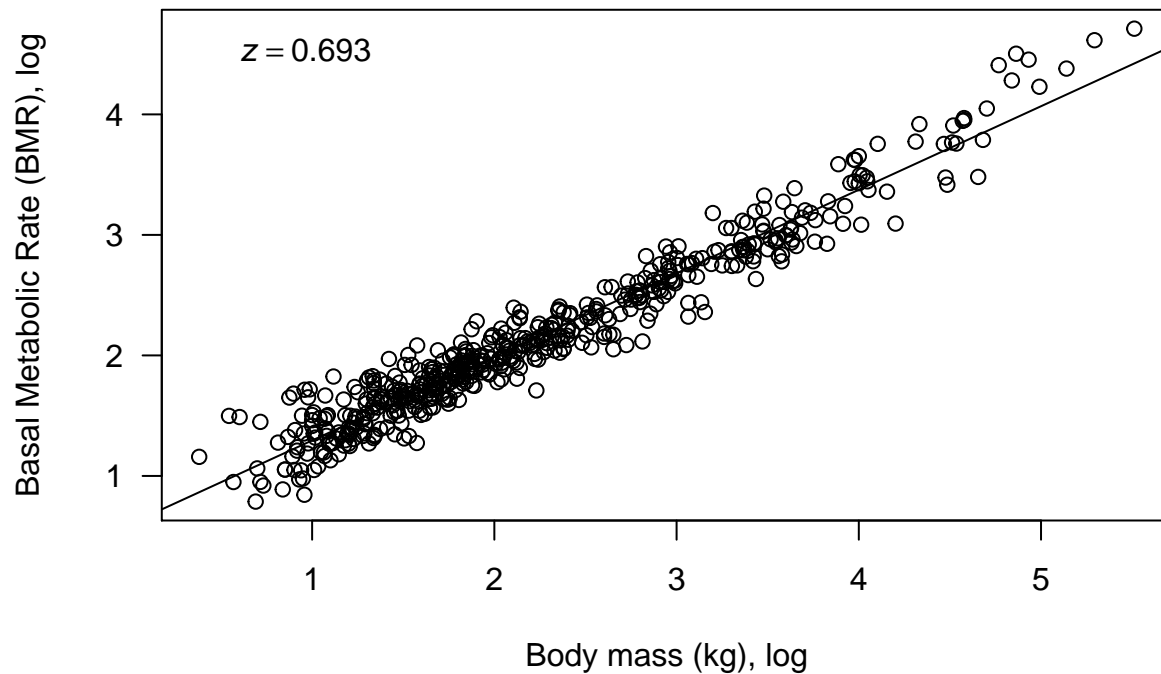
To show how to perform a phylogenetic regression we will need to get a new dataset, since we basically only have one measure from our bacterial dataset (i.e. niche breadth). In the `data` folder we've included a dataset on mammals, which we will use to examine the scaling relationship (i.e. power law) between body mass ( $M$ ) and Basal Metabolic Rate (BMR), the minimal rate of energy expenditure per unit time by endothermic animals at rest. This relationship has significant phylogenetic signal and can be described by the power law  $BMR = cM^z$ , where  $z$  is the scaling exponent and  $c$  is the scaling factor (Capellini et al., 2010). In regression terms this can be described  $\log(BMR) = \log(c) + z\log(M)$ , an equation that you should recognize from Week 4 when we used it to describe the species area relationship (SAR). However, phylogeny is rarely accounted for in foundational power laws such as these. Here we will account for phylogeny in the BMR-mass power law using the super-tree of mammals, a single phylogenetic tree assembled from a combination of smaller phylogenetic trees.

First, we will clean the mammal dataset and super-tree.

```
# Input the tree and dataset
mammal.Tree <- read.tree("./data/mammal_best_super_tree_fritz2009.tre")
mammal.data <- read.table("./data/mammal_BMR.txt", sep = "\t",
  header = TRUE)
# Select the variables we want to analyze
mammal.data <- mammal.data[, c("Species", "BMR_.ml02.hour.",
  "Body_mass_for_BMR_.gr.")]
mammal.species <- array(mammal.data$Species)
# Select the tips in the mammal tree that are also in the
# dataset
pruned.mammal.tree <- drop.tip(mammal.Tree, mammal.Tree$tip.label[-na.omit(match(mammal.species,
  mammal.Tree$tip.label))])
# Select the species from the dataset that are in our pruned
# tree
pruned.mammal.data <- mammal.data[mammal.data$Species %in% pruned.mammal.tree$tip.label,
  ]
# Turn column of Species names into rownames
rownames(pruned.mammal.data) <- pruned.mammal.data$Species
```

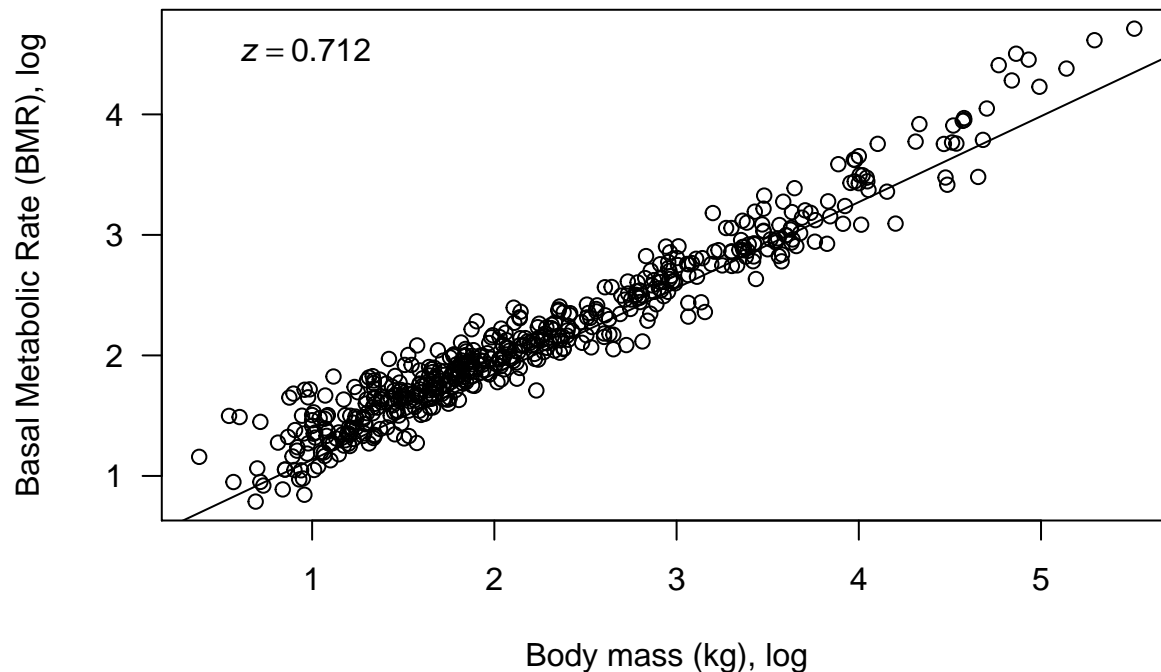
Now let's look at the relationship between mass and BMR:

```
# Run a simple linear regression
fit <- lm(log10(BMR_.ml02.hour.) ~ log10(Body_mass_for_BMR_.gr.),
  data = pruned.mammal.data)
plot(log10(pruned.mammal.data$Body_mass_for_BMR_.gr.), log10(pruned.mammal.data$BMR_.ml02.hour.),
  las = 1, xlab = "Body mass (kg), log", ylab = "Basal Metabolic Rate (BMR), log")
abline(a = fit$coefficients[1], b = fit$coefficients[2])
b1 <- round(fit$coefficients[2], 3)
eqn <- bquote(italic(z) == .(b1))
# plot the slope
text(0.5, 4.5, eqn, pos = 4)
```



Now let's correct for phylogeny using the `phylolm` function from the package `phylolm`.

```
# Run a phylogeny-corrected regression with no bootstrap
# replicates
fit.phy <- phylolm(log10(BMR_.ml02.hour.) ~ log10(Body_mass_for_BMR_.gr.),
  data = pruned.mammal.data, pruned.mammal.tree, model = "lambda",
  boot = 0)
plot(log10(pruned.mammal.data$Body_mass_for_BMR_.gr.), log10(pruned.mammal.data$BMR_.ml02.hour.),
  las = 1, xlab = "Body mass (kg), log", ylab = "Basal Metabolic Rate (BMR), log")
abline(a = fit.phy$coefficients[1], b = fit.phy$coefficients[2])
b1.phy <- round(fit.phy$coefficients[2], 3)
eqn <- bquote(italic(z) == .(b1.phy))
text(0.5, 4.5, eqn, pos = 4)
```



and we see that we get a power law of approximately  $BMR = 0.42M^{0.71}$ . So by accounting for how shared evolutionary history impacts the relationship between mass and BMR (i.e. the non-independence of our data) we actually get an higher slope, suggesting a stronger relationship between mass and BMR than expected based off of a simple linear regression.

## REFERENCES

- Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? Measuring the power of comparative methods. *Evolution* 66:2240–2251.
- Bartoszek, K., J. Pienaar, P. Mostad, S. Andersson, and T.F. Hansen. 2012. A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology* 314:204–215.
- Butler, M., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist* 164:683–695.
- Capellini, I., C. Venditti, and R. Barton. 2010. Phylogeny and metabolic scaling in mammals. *Ecology* 91:2783–2793.
- Cadotte, M. W., and T.J. Davies. 2016. *Phylogenies in Ecology: A Guide to Concepts and Methods*. Princeton.
- Felsenstein, J. 1985. Phylogenies and the comparative method. *The American Naturalist* 125:1–15.
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer.
- Fritz, S. A. and A. Purvis. 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* 24:1042–1051.
- Goldman, N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36:182–198.
- Grafen, A. 1989. The Phylogenetic Regression. *Phil. Trans. R. Soc. Lond. B.* 326:119–157.
- Hall, P. and S. Wilson. 1991. Two guidelines for bootstrap hypothesis testing. *Biometrics*. 47:757–762.
- Hansen, T. F., and K. Bartoszek. 2012. Interpreting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology* 61:413–425.



- Hansen T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* 50:1404–1417.
- Kuhner, M. K. and J. Felsenstein. 1994. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11:459–468.
- Ho, L. T. and C. Ané. 2014. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Syst Biol* 63:397–408.
- Lavorel, S., and Garnier, E. (2002) Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional Ecology* 16:545–556
- Losos, J. B. 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters* 11:995–1003.
- Lynch, M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45:1065–1080.
- MacArthur, R. H., and J. W. MacArthur. 1961. On Bird Species Diversity. *Ecology* 42:594–598
- Münkemüller, T., S. Lavergne, B. Bzeznik, S. Dray, T. Jombart, K. Schiffrers, and W. Thuiller. 2012. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution* 3:743–756.
- Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.
- Paradis, E. 2012. Analysis of Phylogenetics and Evolution with R. Springer.
- Revell, L., L. Harmon, and D. Collar. 2008. Phylogenetic Signal, Evolutionary Process, and Rate. *Systematic Biology* 57:591–601.
- Revell, L.J. and D. C. Collar. 2009. Phylogenetic analysis of the evolutionary correlation using likelihood. *Evolution* 63:1090–1100.
- Revell, L. J. 2010. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* 1:319–329.
- Robinson, D. R. and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131–147.