RERConverge: walk-through of binary trait analysis

Wynn Meyer

10/17/2017

This walk-through provides instructions for implementing the RERConverge package to identify genes whose evolutionary rates shift in association with change in a binary trait. For information on how to download and install RERConverge, see the README on github.

As an example, we will run the code in the `runMarineSub` script in the vignettes folder.

Reading in gene trees

To run RERconverge, you will first need to supply a file containing **gene trees** for all genes to be included in your analysis. This is a tab delimited file with the following information on each line:

Gene_name Newick_tree

An example file is provided in *data/mammal62aa_meredplus_wCM.trees*, which you can view in any text editor.

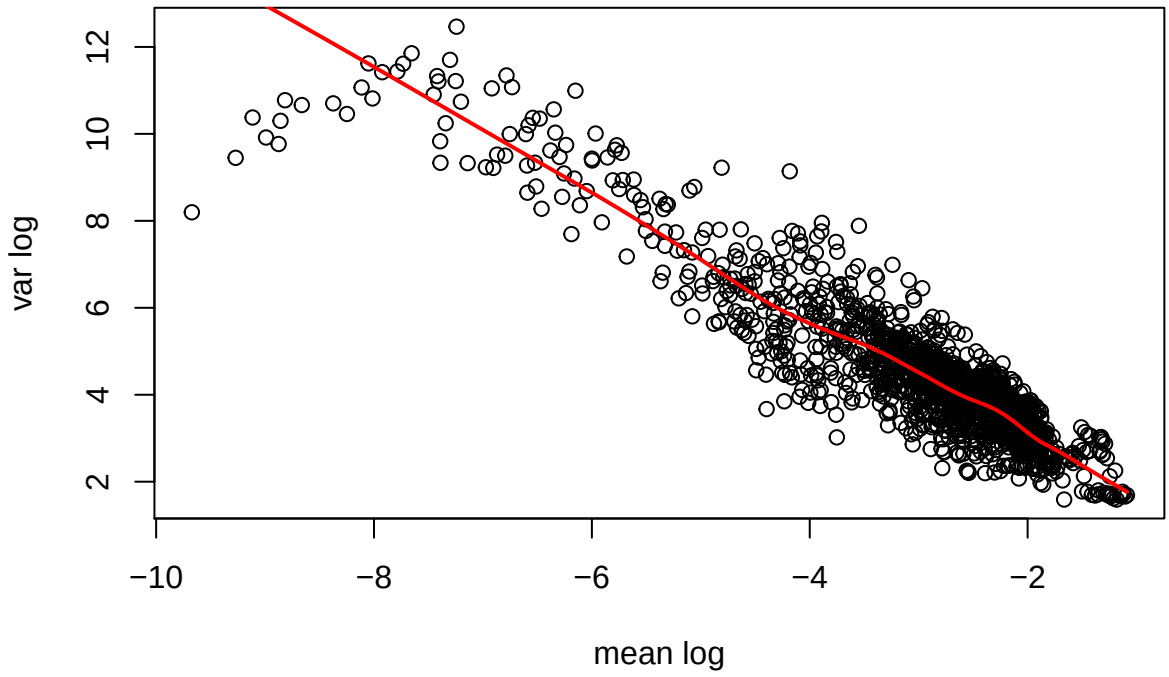
In R, load the RERConverge library and read in the gene trees. The `readTrees` function takes quite a while to read in trees for all genes, so we will limit ourselves to the first 200 using `max.read` (this will still take a minute or so, so be patient):

```
#library(devtools)
#install_github("nclark-lab/RERconverge")
#library("RERconverge")
#This is a temporary alternative to using library
#(these are the latest versions of the scripts).
source('../R/enrichmentFuncs.R')
source('../R/plottingFuncs.R')
source('../R/projection_coevo.R')
source('../R/RERfuncs.R')
source('../R/coordsfn.R')
```

```
mamTrees=readTrees("../data/mammal62aa_meredplus_wCM.trees", max.read = 200)
```

```
## max is  62
```

```
## estimating master tree branch lengths from 32 genes
```



First, the code tells us that the maximum number of tips in the gene trees is 62. It then reports that it will use the 32 genes in this set that have data for all 62 species to estimate a **master tree**. The tree topology for the master tree (but not its branch lengths) will be used for subsequent analyses.

The figure here is a log-scale plot of the variance versus the mean for all possible branch lengths in the tree, summarized across all genes in the dataset. Where the relationship between mean and variance starts to break down for very small values (in this case very large negative values) can be an indicator of the point at which values are too close to zero to provide accurate information for analyses. In our example plot, this happens around e^{-7}. We use approximately this value as our cutoff to exclude short branches in further analyses.

Estimating relative evolutionary rates (RER)

The next step is to estimate **relative evolutionary rates**, or RER, for all branches in the tree for each gene. Intuitively, a gene's RER for a given branch represents how quickly or slowly the gene is evolving on that branch, relative to its overall rate of evolution throughout the tree. For more details about how RER are computed, see (Chikina, Robinson, and Clark 2016) and (Partha et al. 2017).

We will use the `getAllResiduals` function to calculate RER. The input variable `useSpecies` is a vector that can be used to specify a subset of species to use in the analysis; here we will use the full set of tip labels in the master tree. We will also filter any branches shorter than 0.001 using `cutoff` (see above for estimating a reasonable cutoff from the mean-variance plot). Here is the basic method, with the recommended settings:

```
mamRERlogW=getAllResiduals(mamTrees,useSpecies=mamTrees$masterTree$tip.label,
                           transform = "log", weighted = T, cutoff=0.001)
```

While this is running, it will print out i=N for a series of numbers. This shows how many genes have had RER estimated.

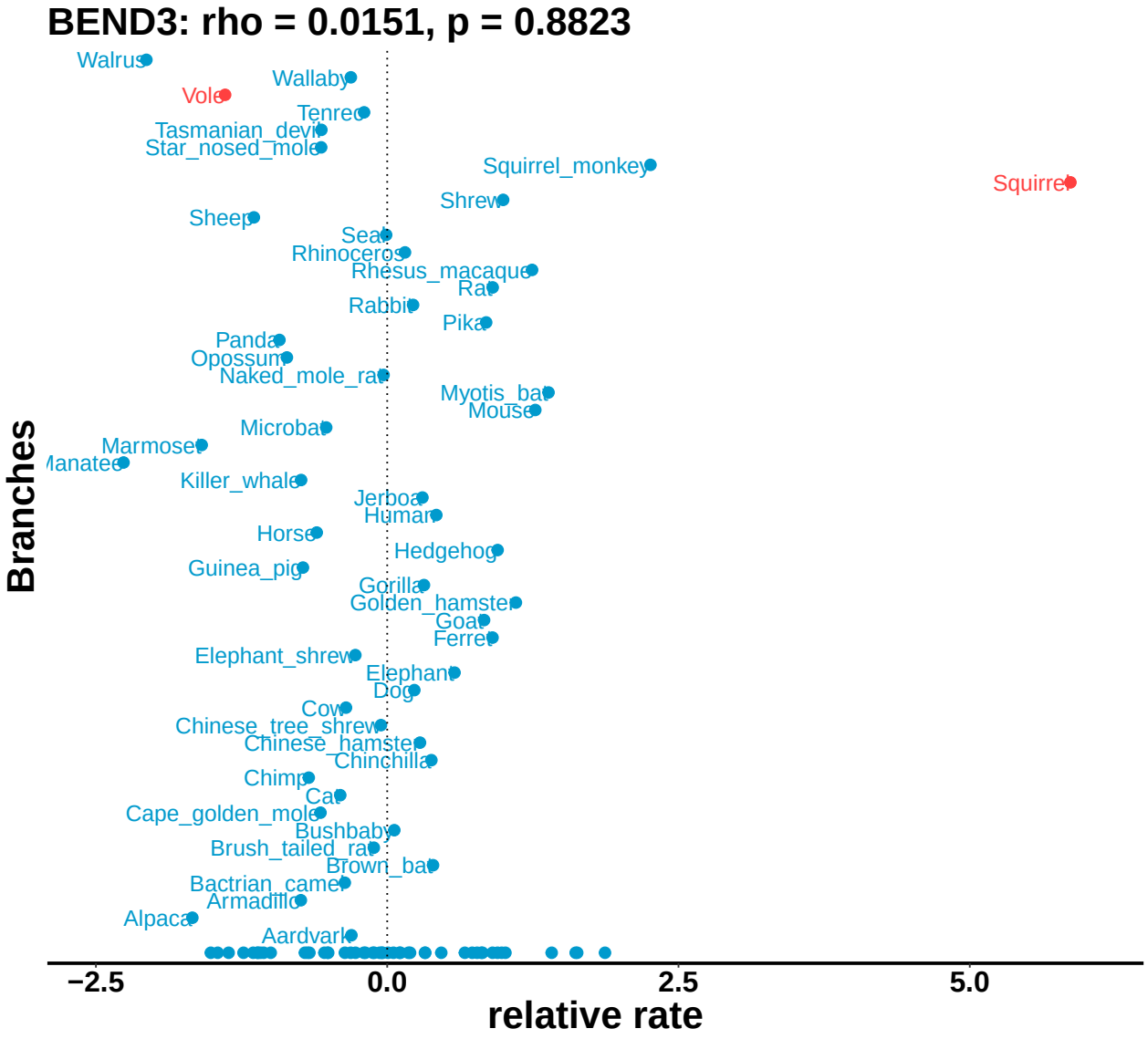
In many cases, it is helpful to scale the raw RER to account for the variance across genes within the genome. This is helpful in estimating the empirical significance of particular genes in downstream analyses. Here is the function to scale RER:

```
mamRERlogWs=scale(mamRERlogW)
```

We can visualize the RERs for any given gene using the `plotRers` function. Here is an example.

```
masterrooted <- root(mamTrees$masterTree,
                     outgroup=c("Platypus","Wallaby","Tasmanian_devil","Opossum"))
wspmr <- masterrooted$tip.label[masterrooted$edge[,2]]
wmvole <- which(wspmr=="Vole")
wmsquirrel <- which(wspmr=="Squirrel")
colMaster <- c(rep("black",length(masterrooted$edge)))
colMaster[wmvole] <- "blue"
colMaster[wmsquirrel] <- "red"
tb3 <- root(mamTrees$trees$BEND3,outgroup=c("Wallaby","Tasmanian_devil","Opossum"))
wspbr <- tb3$tip.label[tb3$edge[,2]]
wbvole <- which(wspbr=="Vole")
wbsquirrel <- which(wspbr=="Squirrel")
coltb3 <- c(rep("black",length(tb3$edge)))
coltb3[wbvole] <- "blue"
coltb3[wbsquirrel] <- "red"
par(mfrow=c(1,2))
plot(masterrooted, main = "Average tree", edge.col=colMaster, cex=0.8) #plot average tree
plot(tb3, main = "BEND3 tree",edge.col=coltb3, cex=0.8) #plot individual gene tree
```


```
par(mfrow=c(1,1))
phenvExample <- foreground2Paths(c("Vole","Squirrel"),mamTrees)
plotRers(mamRERlogWs,"BEND3",phenv=phenvExample) #plot RERs
```



The upper left plot is a tree with branch lengths representing the average rates across all genes. The upper right plot is the same tree, but with branch lengths representing rates specifically for the BEND3 gene. The plot below these represents the estimated RERs for terminal branches. Notice how the RER for vole is negative; this is because the branch leading to vole in the BEND3 tree is shorter than average. On the other hand, the RER for squirrel is positive because the branch leading to squirrel in the BEND3 tree is longer than average.

Reading in or generating trait trees

Now we will associate variation in these RERs with variation in traits across the tree. To do so, we first need to provide information about which branches of the tree have the trait of interest (**foreground branches**). There are three possible ways to do this:

1) Provide a binary tree file. This should be a file in Newick format with branch lengths zero for background branches and one for foreground branches. An example is provided in *data/MarineTreeBin.txt*.

```
marineb=read.tree("../data/MarineTreeBinCommonNames.txt")
marinebrooted = root(marineb,outgroup=c("Platypus", "Wallaby","Tasmanian_devil","Opossum"))
```

This tree file has all extant marine species, plus the dolphin-killer whale ancestor, as foreground.

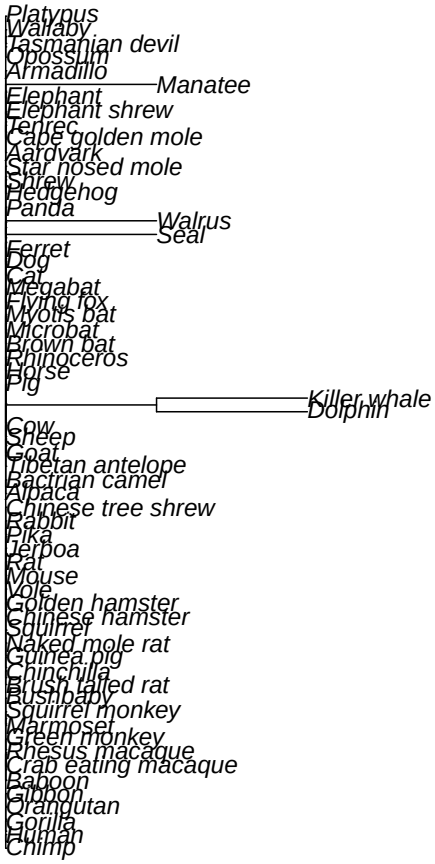
2) Use the interactive branch selection tool. The following should open a plot of the master tree. When the GUI opens, select the marine foreground branches (Walrus, Seal, Killer whale, Dolphin, Killer whale-Dolphin ancestor, and Manatee), and click 'End selection.'

```
marineb2=selectforegroundbranches(mamTrees$masterTree)
```

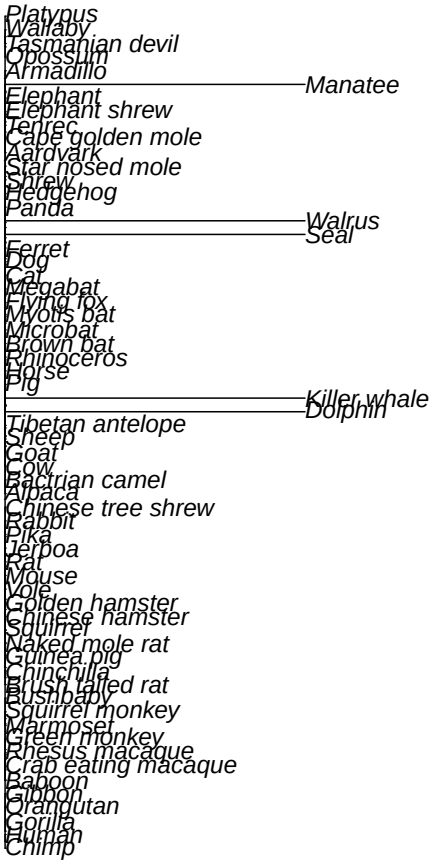
3) Provide a list of species to use as foreground species. This only allows inclusion of extant lineages, so in this case, the Killer whale-Dolphin ancestor will no longer be included.

```
marineextantforeground = c("Walrus","Seal","Killer_whale","Dolphin","Manatee")
wmarine <- which(wspmr %in% marineextantforeground)
marineb3rooted <- masterrooted
marineb3rooted$edge.length <- c(rep(0,length(marineb3rooted$edge.length)))
marineb3rooted$edge.length[wmarine] <- 1
par(mfrow=c(1,2))
plot(marinebrooted, main="Trait tree from file (1) or manual selection (2)")
plot(marineb3rooted, main="Trait tree from foreground list (3)")
```

Trait tree from file (1) or manual selection (2)



Trait tree from foreground list (3)



Some of the genes (like BEND3 above) may not have data for all species, meaning that their phylogeny will be a subset of the full phylogeny. To compare RERs for these genes with trait evolution, we run one of two functions that determine how the trait would evolve along all/many possible subsets of the full phylogeny, generating a set of **paths**. The function `tree2paths` takes a binary tree as input, and the function `foreground2paths` takes a set of foreground species as input.

```
phenvMarine=tree2Paths(marineb, mamTrees)
phenvMarine2=foreground2Paths(marineextantforeground, mamTrees)
```

Correlating gene evolution with trait evolution using RERConverge

Now that we have estimates for the RERs for all genes of interest, as well as a representation of how the trait of interest evolves across the tree, we can use RERConverge to test for an association between relative evolutionary rate and trait across all branches of the tree.

```
corMarineLogWs=getAllCor(mamRERlogWs, phenvMarine)
```

```
## Setting method to Kendall
```

The text displayed shows which correlation method is used to test for association. Here we have used the default, the Kendall rank correlation coefficient, or Tau.

This generates a table with the following output for each gene:

1) Rho: the correlation between relative evolutionary rate and trait across all branches
2) N: the number of branches in the gene tree
3) P: an estimate of the P-value for association between relative evolutionary rate and trait.

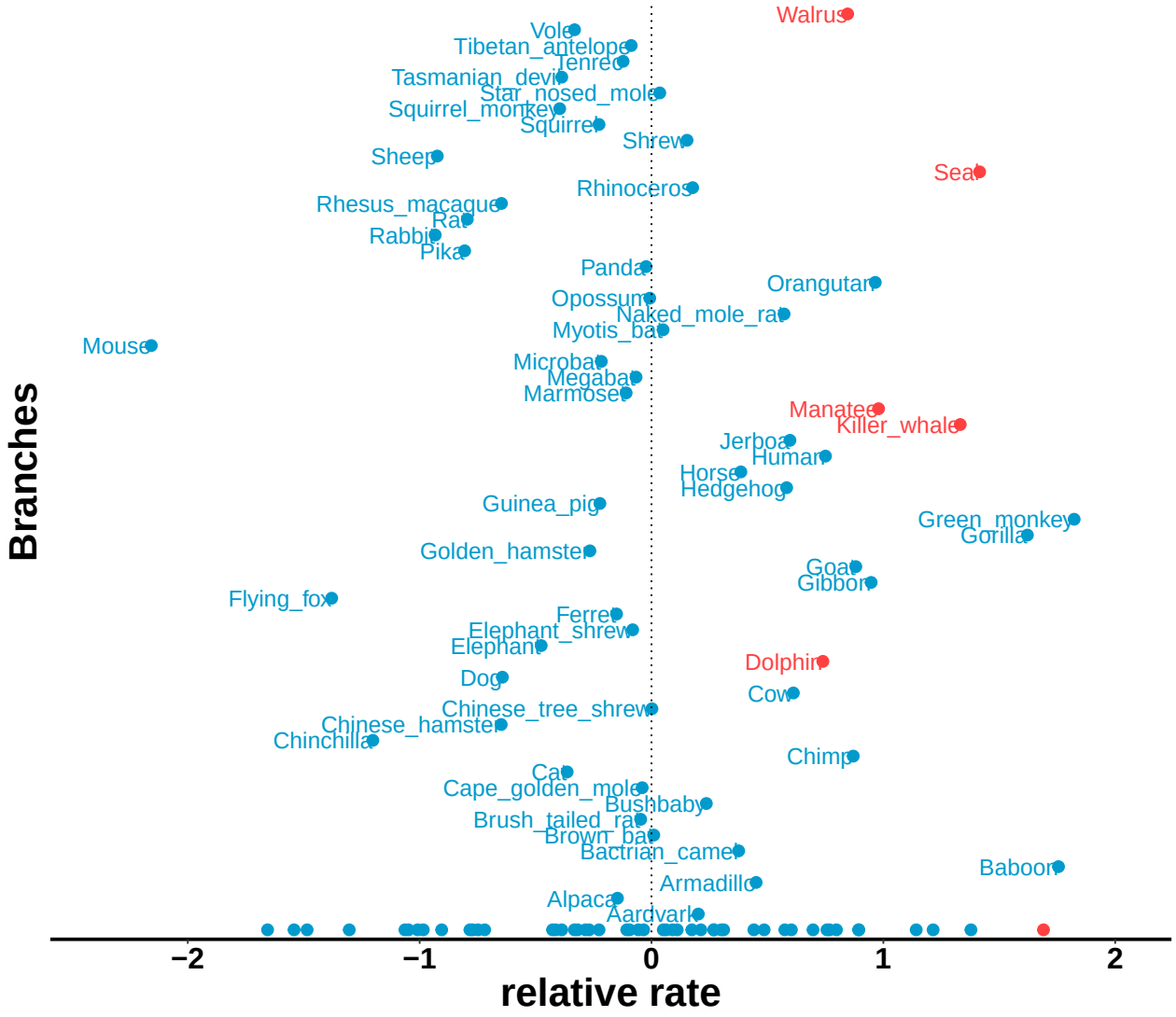
Let's take a look at some of the top genes within this set.

```
head(corMarineLogWs[order(corMarineLogWs$P),])
```

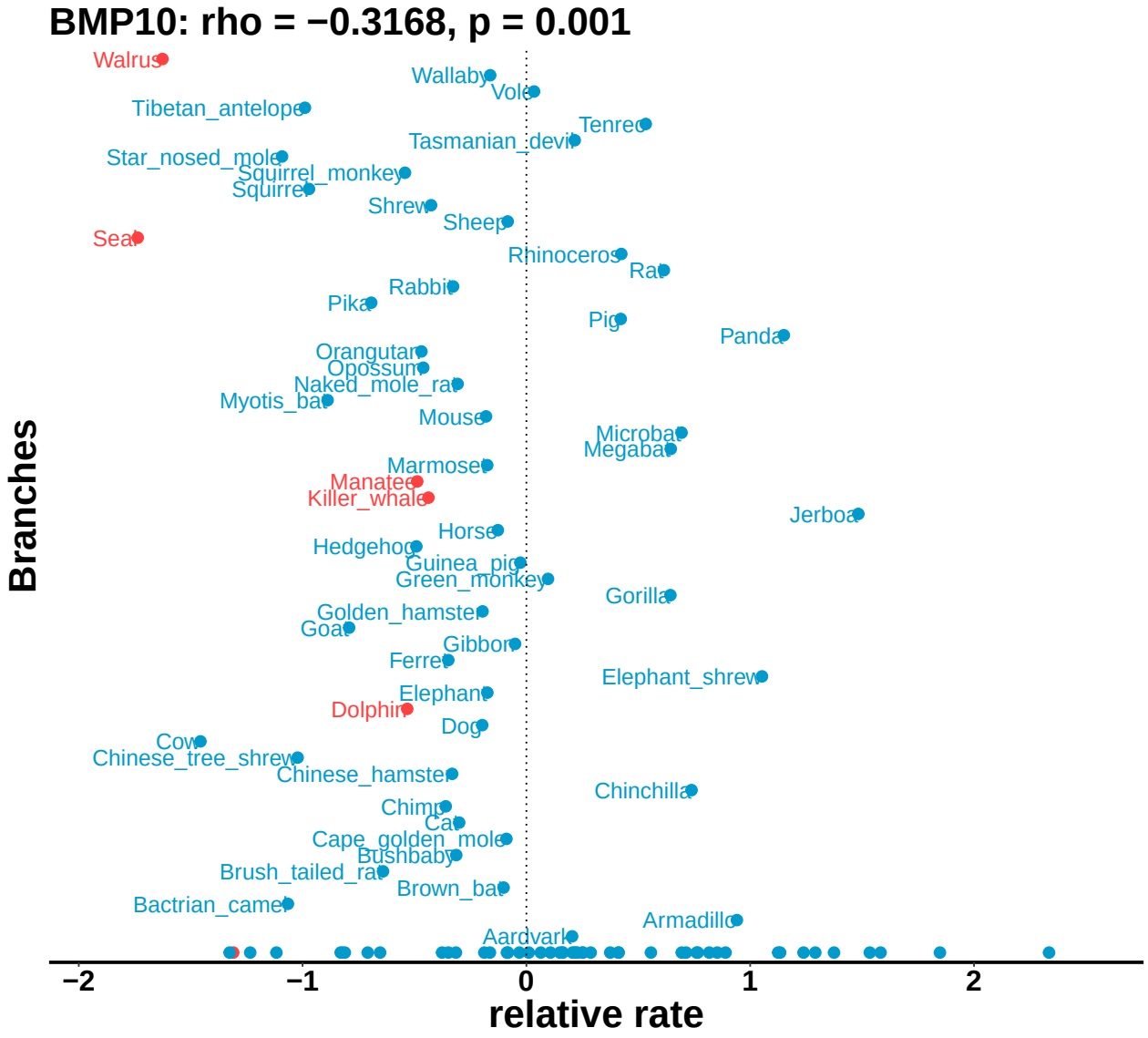
```
##                 Rho   N            P
## ANO2      0.2805028 107 0.0004305835
## BDH1      0.2705129 103 0.0008686652
## BMP10    -0.2598701 105 0.0012360420
## AK124326 -0.3090403  68 0.0021008820
## ATP2A1    0.2343614 103 0.0039151324
## ARHGAP36 -0.2185493  99 0.0083769404
```

ANO2 and BMP10 are two of the top genes.

```
plotRers(mamRERlogWs,"ANO2",phenv=phenvMarine)
```



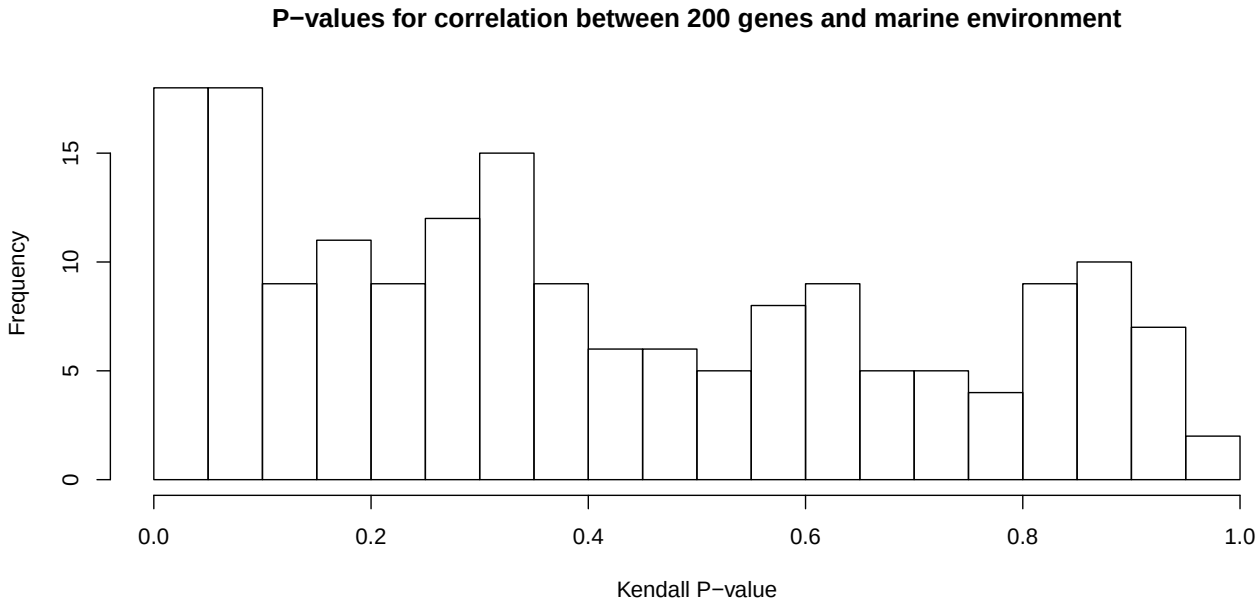

```
plotRers(mamRERlogWs,"BMP10",phenv=phenvMarine)
```



In the ANO2 tree, the marine branches are long, leading to a positive Rho and a low p-value. In contrast, in the BMP10 tree, the marine branches are short. This also yields a low p-value, but with a negative Rho.

To see what the overall pattern of association is across all genes in the set, we can plot a p-value histogram.

```
hist(corMarineLogWs$P, breaks=20, xlab="Kendall P-value",
     main="P-values for correlation between 200 genes and marine environment")
```



There appears to be a slight enrichment of low p-values, but since we have only evaluated the first 200 genes from our ~19,000 gene genome-wide set, we should hold off on drawing conclusions from this.

Conclusion

We've now walked through the basic workflow for RERConverge. For more information about these methods and some results relevant to marine and subterranean adaptation, see (Chikina, Robinson, and Clark 2016) and (Partha et al. 2017).

References

Chikina, Maria, Joseph D Robinson, and Nathan L Clark. 2016. "Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals." *Molecular Biology and Evolution* 33 (9): 2182–92. doi:10.1093/molbev/msw112.

Partha, Raghavendran, Bharesh K Chauhan, Zelia Ferreira, Joseph D Robinson, Kira Lathrop, Ken K Nischal, Maria Chikina, and Nathan L Clark. 2017. "Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling." *eLife* 6 (October). eLife Sciences Publications Limited: e25884. doi:10.7554/eLife.25884.