Empirical analyses

Daniel S. Caetano and Luke J. Harmon

April 2, 2018

This script will reproduce the analyses of the empirical data for Centrarchidae fishes and for the anole lizards as described in the manuscript. The analyses depend on our 'ratematrix' package that can be easily installed from github using a function from the package 'devtools'. Please install 'devtools' in order to install 'ratematrix'.

IMPORTANT NOTE: This script will not be updated after the publication of the article. It is possible that some of the usage have changed. Please check the github repository (https://github.com/Caetanods/ratematrix) for an updated version of the package and for working tutorials.

Package installation

```
library( devtools )
install_github("Caetanods/ratematrix")
```

```
## Skipping install of 'ratematrix' from a github remote, the SHA1 (53bd2465) has not changed since las
##   Use `force = TRUE` to force installation
```

Load dependencies

```
library( ratematrix )
library( phytools )
```

```
## Loading required package: ape
```

```
## Loading required package: maps
```

Make analyses for the Centrarchidae

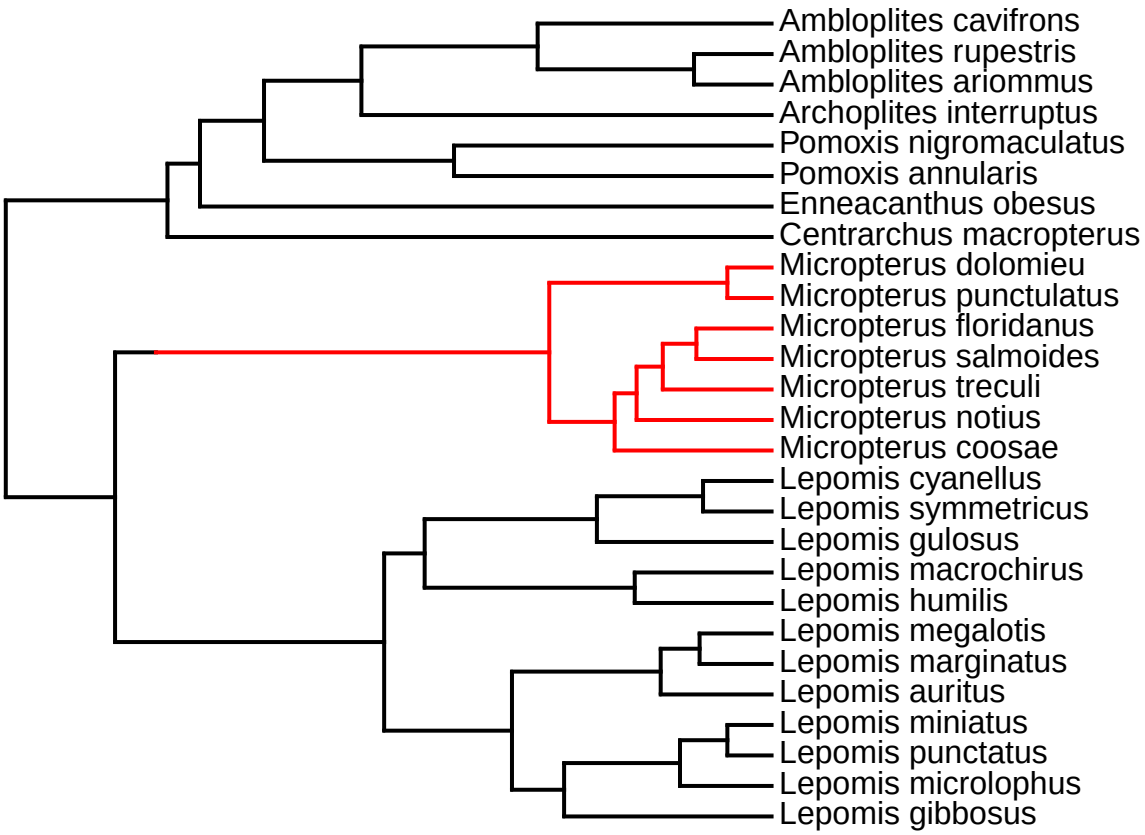
The data for the Centrarchidae

```
data("centrarchidae")
```

Here using a fixed regime mapped to the phylogenetic tree:

```
plotSimmap(centrarchidae$phy.map)
```

```
## no colors provided. using the following legend:
##       0       1
## "black"   "red"
```



Prepare the prior distributions. Here using marginal uniform prior for the evolutionary rate matrix and a uniform prior for the vector of phylogenetic means (root value).

```
par.mu <- rbind( c(-10,10), c(-10,10) )
par.sd <- rbind( c(0,10), c(0,10) )
```

The 'makePrior' function will produce a prior object given the parameters.

```
prior <- makePrior(r=2, p=2, par.mu=par.mu, par.sd=par.sd)
```

Now we will run four independent MCMC chains, both starting from the prior distribution. Then we check for convergence using Gelman R (potential scale reduction factor). This test checks if the variance between the chains is smaller than the variances within chains. Larger variances within each chain compared with among chains is an indicative that the multiple chains are sampling from the same posterior distribution.

IMPORTANT NOTE: This part of the analysis will take a some time to run.

```
## ## Uncomment to run.
## outname <- paste("centrarchidae_rep_", 1:4, sep="")
## sample.prior <- lapply(1:4, function(x) samplePrior(n=1, prior=prior, sample.sd=TRUE
##                                          , rebuild.R=FALSE) )
## for( i in 1:4 ){
##     ratematrixMCMC(data=centrarchidae$data, phy=centrarchidae$phy.map, gen=1000000,
##                    chunk=1000, v=50, w_sd=0.5, w_mu=0.2, prior=prior,
##                    start=sample.prior[[x]], outname=outname[x])
## }
## ## The 'ratematrixMCMC' function writes the output of the MCMC to the working directory.
##        Now we need to read the results back to file.
## handle <- lapply(list.files(pattern='^centrarchidae_rep.*.rds'), readRDS )
```

```
## mcmc <- lapply(handle, readMCMC)
```

If you decided to run the MCMC again, then remember to work with the object `mcmc` rather than the other objects we refer to below. The lines below assume that you will load the results we provide as a online supplement.

As an alternative to run the MCMC analysis again you can use the file made available from the same Dryad repository. Here we will assume that the file `centrarchida_MCMC_results.RData` is in the current working directory. Otherwise, please change the pat to the file below:

```
load("centrarchidae_MCMC_results.RData")
```

Check for convergence.

```
checkConvergence(centrarchidaeMCMC$post[[1]], centrarchidaeMCMC$post[[2]]
                 , centrarchidaeMCMC$post[[3]], centrarchidaeMCMC$post[[4]])
```

```
## $gelman
## $gelman$diag_root
## Potential scale reduction factors:
##
##               Point est. Upper C.I.
## Gape_width             1       1.01
## Buccal_length          1       1.00
##
##
## $gelman$narrow_diet
## Potential scale reduction factors:
##
##     Point est. Upper C.I.
## 1,1       1.01       1.02
## 1,2       1.00       1.00
## 2,1       1.00       1.00
## 2,2       1.02       1.02
##
##
## $gelman$wide_diet
## Potential scale reduction factors:
##
##     Point est. Upper C.I.
## 1,1          1       1.00
## 1,2          1       1.00
## 2,1          1       1.00
## 2,2          1       1.01
##
##
##
## $ess
##          root narrow_diet wide_diet
## var1 1539.055    4443.481  5694.913
## var2 3501.621    5040.962  6143.263
## var3       NA    5040.962  6143.263
## var4       NA    3789.770  5923.386
```

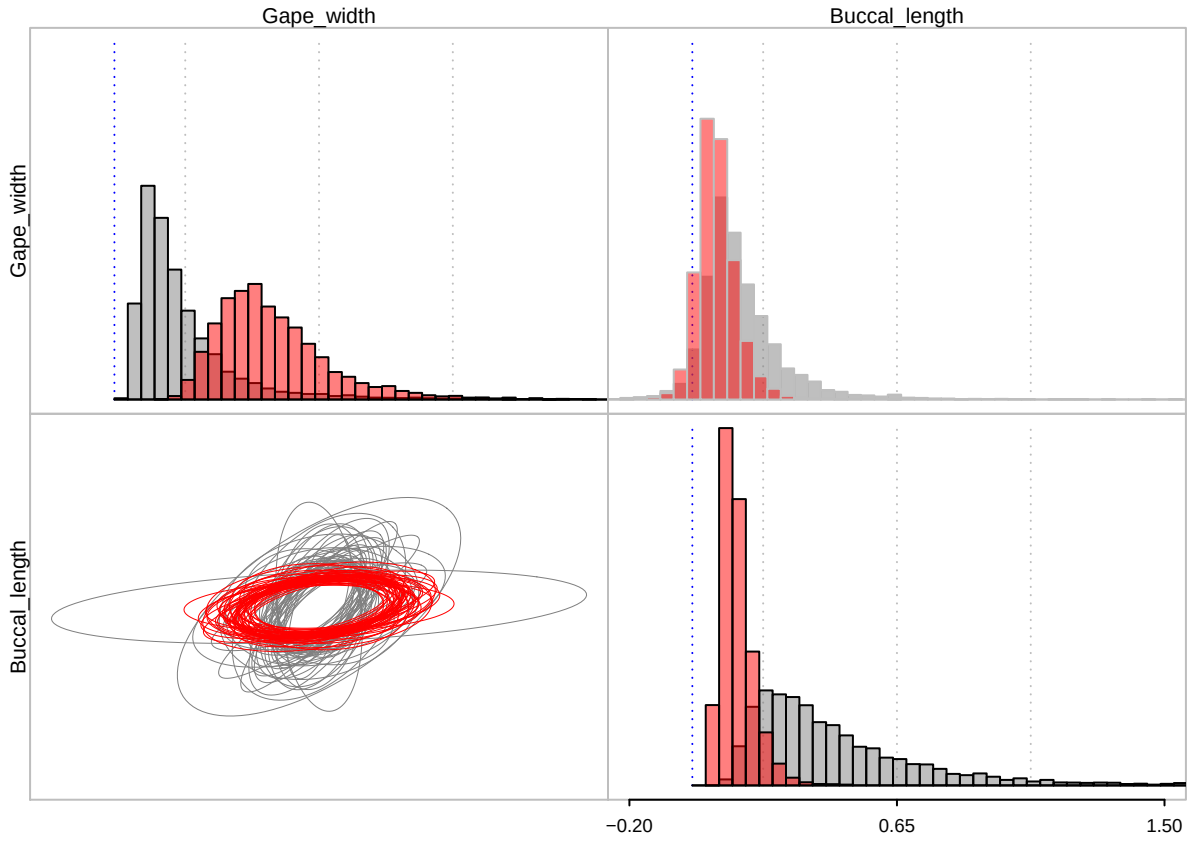
We can merge the posteriors to form a larger one.

```
post.mcmc <- mergePosterior(centrarchidaeMCMC$post[[1]], centrarchidaeMCMC$post[[2]]
                , centrarchidaeMCMC$post[[3]], centrarchidaeMCMC$post[[4]])
```

Plot the combined posterior distribution for the evolutionary rate matrix for each regime:

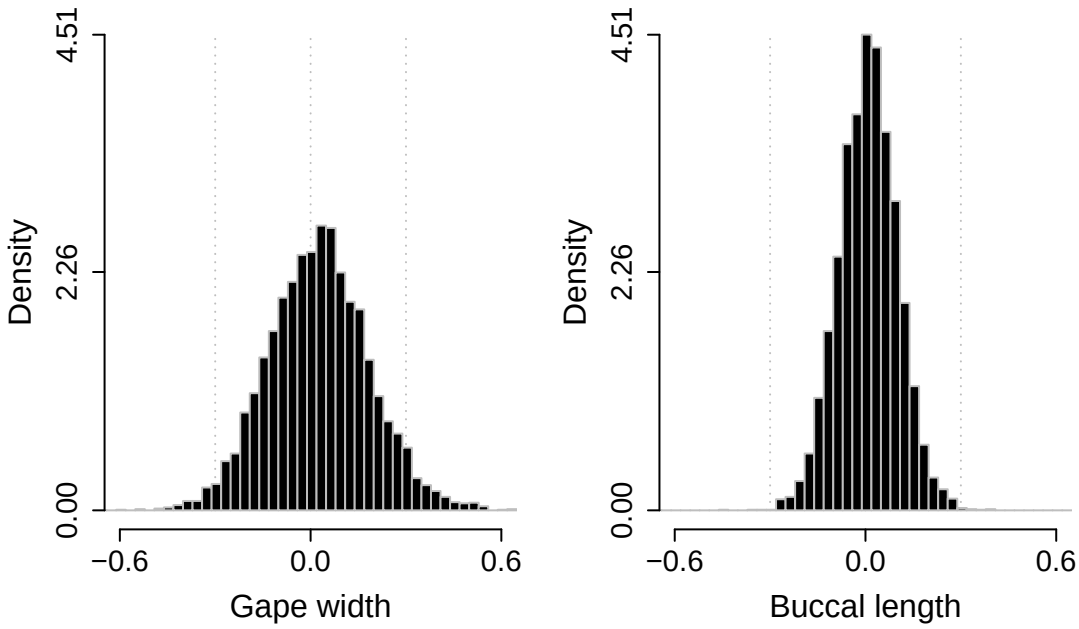
```
plotRatematrix(post.mcmc, set.xlim = c(-0.2,1.5), show.zero = TRUE,
               colors=c("#808080", "#ff0000"), alphaOff = 0.5, alphaDiag = 0.5)
```

```
## Plotting multiple regimes.
## Table with regimes and colors (names or HEX):
## narrow_diet   wide_diet
##     #808080     #ff0000
```



Plot the root value:

```
plotRootValue(post.mcmc, set.xlab = c("Gape width", "Buccal length"), set.cex.lab = 1,
              set.cex.axis = 1.5, set.xlim = c(-0.6,0.6))
```



Now we can compute the summary statistics based on the posterior distribution. First statistics compute the overlap in density between regimes for each of the elements of the rate matrix. This is an 'overall' test and captures the average difference between regimes without distinction between evolutionary correlations or rates of evolution.

```
testRatematrix(chain=post.mcmc, par="all")
```

```
##            mat #1 x #2
## test value   0.5806667
```

Now check for differences in the structure of correlation between rate matrix regimes.

```
testRatematrix(chain=post.mcmc, par="correlation")
```

```
##            mat #1 x #2
## test value       0.464
```

Finally, check for differences in the rates of evolution of each trait between rate matrix regimes.

```
testRatematrix(chain=post.mcmc, par="rates")
```

```
##            mat #1 x #2
## test value   0.9156667
```

Compare with MLE results

In this analyses we want to compare the results of the 'ratematrix' approach based on summary statistics with the formal model selection using Maximum likelihood estimates (MLE) and likelihood-ratio tests.

```
library(mvMORPH)
```

```
## Loading required package: corpcor
```

```
## Loading required package: subplex
```

```
## ##
## ## mvMORPH package (1.0.9)
```

```
## ## Multivariate evolutionary models
```

```
## ##
## ## See the tutorials: browseVignettes("mvMORPH")
```

```
## ##
## ## To cite package 'mvMORPH': citation("mvMORPH")
## ##
```

The MLE for the simple model, with a single rate matrix regime.

```
mle.m1 <- mvBM(tree = centrarchidae$phy, data = centrarchidae$data, model="BM1")
```

```
## successful convergence of the optimizer
## a reliable solution has been reached
##
## -- Summary results for multiple rate BM1 model --
## LogLikelihood:    26.71139
## AIC:      -43.42279
## AICc:     -42.17279
## 5 parameters
##
## Estimated rate matrix
## ______________________
##               Gape_width Buccal_length
## Gape_width     0.3393671     0.1003002
## Buccal_length  0.1003002     0.1628531
##
## Estimated root state
## ______________________
##          Gape_width Buccal_length
## theta: 3.521413e-15  4.159433e-15
```

The MLE for the complex model, with two rate matrix regimes.

```
mle.mm <- mvBM(tree = centrarchidae$phy, data = centrarchidae$data, model="BMM")
```

```
## successful convergence of the optimizer
## a reliable solution has been reached
##
## -- Summary results for multiple rate BMM model --
## LogLikelihood:    31.95391
## AIC:      -47.90782
## AICc:     -44.70782
## 8 parameters
##
## Estimated rate matrix
## ______________________
## , , 0
##
```

```
##               Gape_width Buccal_length
## Gape_width    0.41369670    0.07862978
## Buccal_length 0.07862978    0.11215498
##
## , , 1
##
##               Gape_width Buccal_length
## Gape_width     0.1298543     0.1668617
## Buccal_length  0.1668617     0.3108325
##
##
## Estimated root state
## ----------------------
##        Gape_width Buccal_length
## theta: 0.07188519    0.01002995
```

Calculate the likelihood ratio test:

```
pchisq(-2 * ( mle.m1$LogLik - mle.mm$LogLik ), df=3, lower.tail=FALSE)
```

```
## [1] 0.01486277
```

Calculate the difference in the correlation between the regimes:

```
cov2cor( mle.mm$sigma[,,1] ) ## Micropterus, narrow diet regime.
```

```
##               Gape_width Buccal_length
## Gape_width     1.0000000     0.3650369
## Buccal_length  0.3650369     1.0000000
```

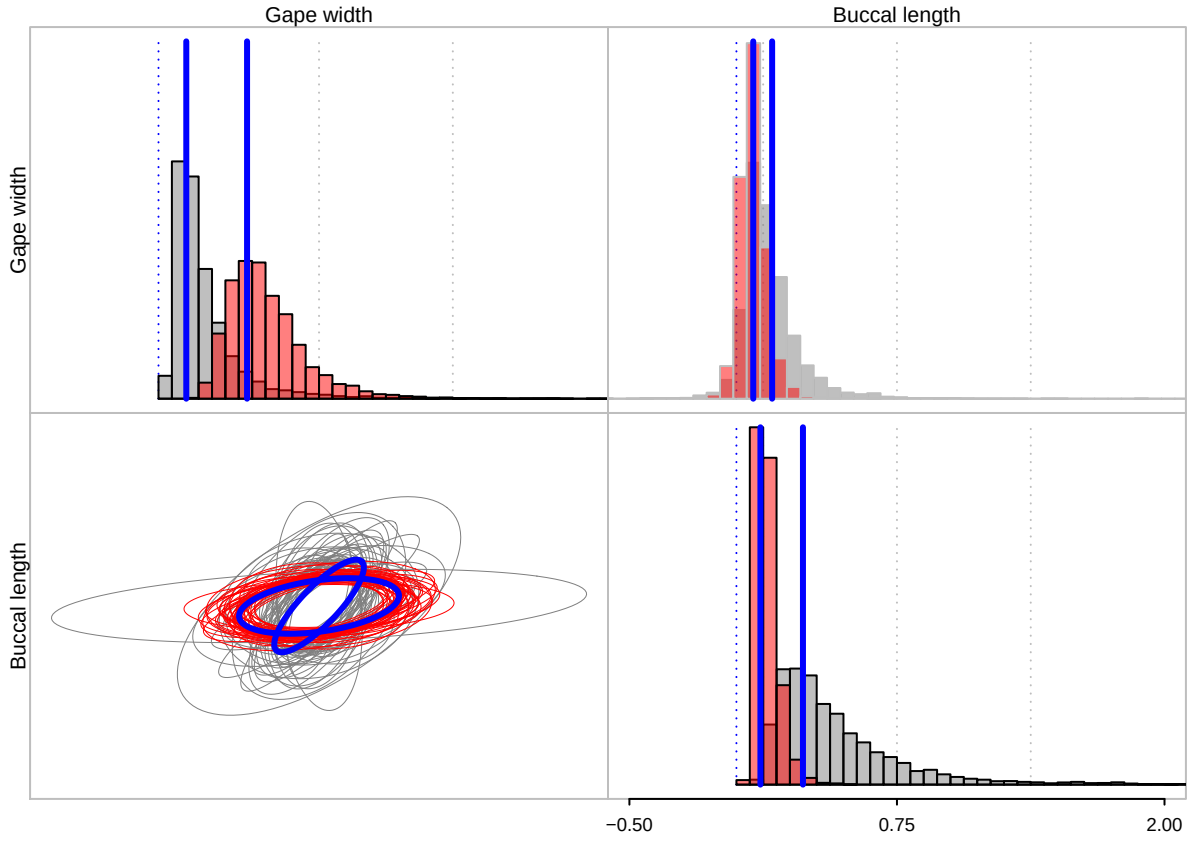
```
cov2cor( mle.mm$sigma[,,2] ) ## Rest of the clade, wide diet regime.
```

```
##               Gape_width Buccal_length
## Gape_width     1.0000000     0.8305491
## Buccal_length  0.8305491     1.0000000
```

Plot the MLE estimate for the single rate and two rates model on top of the posterior distribution that we estimated using the `ratematrix` package.

```
line.mat <- list( as.matrix(mle.mm$sigma[,,1]), as.matrix(mle.mm$sigma[,,2]) )
plotRatematrix(post.mcmc, set.xlim = c(-0.5,2), show.zero = TRUE,
               colors=c("#808080", "#ff0000"), alphaOff = 0.5, alphaDiag = 0.5,
               point.matrix = line.mat, point.color = c("blue","blue"), point.wd = 3
               , set.leg=c("Gape width", "Buccal length"))
```

```
## Plotting multiple regimes.
## Table with regimes and colors (names or HEX):
## narrow_diet   wide_diet
##     #808080     #ff0000
```



Make analyses for the anole lizards

This analyses will use continuous data for 3 traits and we will fit 3 different evolutionary rate matrix regimes to the phylogeny. First load the data from the package:

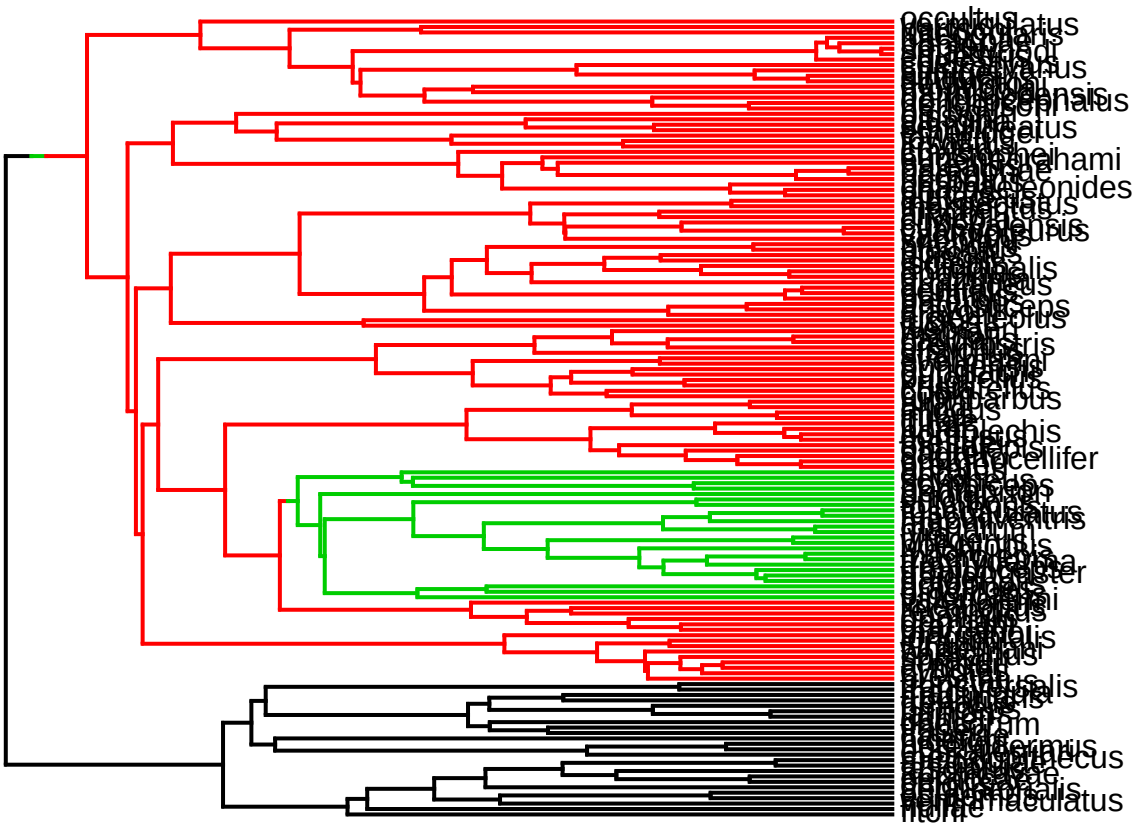
```
data("anoles")
```

Here `anoles` is a list with the data and a list of 100 stochastic maps. We will perform multiple MCMC chains, starting from different points, to estimate the posterior distribution for each of the 3 regimes fitted to the tree.

First let's plot one of the stochastic maps. Here "island" are the island lineages, "mainland" are the ancestral mainland species whereas "mainland.2" are the mainland species that dispersed back to the mainland.

```
plot( anoles$phy[[1]] )
```

```
## no colors provided. using the following legend:
##     island   mainland mainland.2
##    "black"      "red"   "green3"
```



We can set the prior distribution for this analyses. Note that we will use an informative prior around the mean of the tip data for the root value. We know that this should be good enough and might help convergence. It is easy to change the parameters of the root prior or to set and uniform prior of the root.

```
range.dt <- apply(anoles$data[,1:3], 2, range)
par.mu <- t(range.dt)
par.sd <- cbind(c(0,0,0), sqrt(c(5,5,5)))
prior <- makePrior(r=3, p=3, den.mu="unif", par.mu=par.mu, den.sd="unif", par.sd=par.sd)
```

In this case we will also take samples from the prior distribution in order to use as random starting points for the MCMC:

```
start <- lapply(1:4, function(x) samplePrior(n=1, prior=prior) )
```

Now we will set some parameters for the MCMC chain:

```
out <- paste0("anoles_3_regimes_", 1:4)
## The order of the regimes is: "island", "mainland", "mainland.2"
## island and mainland.2 are the smallest regimes. mainland is the large background regime.
v <- c(15,40,15) ## 3 regimes
w_sd <- matrix(0.02, ncol=3, nrow=3) ## 3 regimes and 3 traits.
w_mu <- (par.mu[,2] - par.mu[,1]) / 2
```

Again, you can uncomment the lines below to run the MCMC analyses. These analyses can take some time to finish.

```
## handle <- list()
## for(i in 1:4){ ## These can be run in parallel to save time.
## handle[[i]] <- ratematrixMCMC(data=anoles$data[,1:3], start=start[[i]], phy=anoles$phy, v=v,
```

```
##                                 w_sd=w_sd, w_mu=w_mu, gen=1000000, prior=prior, outname=out[x])
## }
```

Here we will assume that the file `file` available from the Dryad repository is in the current working directory. Please edit the path if this is not the case. Alternatively, you can use the results of the previous MCMC analyses.

```
load("anoles_MCMC_results.RData")
```

Now we can check the convergence of the posteriors:

```
checkConvergence(anoles.post[[1]], anoles.post[[2]], anoles.post[[3]], anoles.post[[4]])
```

```
## $gelman
## $gelman$diag_root
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## svl        1.01       1.02
## tail       1.01       1.03
## head       1.01       1.02
##
##
## $gelman$island
## Potential scale reduction factors:
##
##     Point est. Upper C.I.
## 1,1          1          1
## 1,2          1          1
## 1,3          1          1
## 2,1          1          1
## 2,2          1          1
## 2,3          1          1
## 3,1          1          1
## 3,2          1          1
## 3,3          1          1
##
##
## $gelman$mainland
## Potential scale reduction factors:
##
##     Point est. Upper C.I.
## 1,1          1          1
## 1,2          1          1
## 1,3          1          1
## 2,1          1          1
## 2,2          1          1
## 2,3          1          1
## 3,1          1          1
## 3,2          1          1
## 3,3          1          1
##
##
## $gelman$mainland.2
## Potential scale reduction factors:
```

```
##
##     Point est. Upper C.I.
## 1,1          1          1
## 1,2          1          1
## 1,3          1          1
## 2,1          1          1
## 2,2          1          1
## 2,3          1          1
## 3,1          1          1
## 3,2          1          1
## 3,3          1          1
##
##
##
## $ess
##          root   island mainland mainland.2
## var1 889.3153 2435.881 2779.104   3681.093
## var2 945.4011 2055.248 2503.431   2899.942
## var3 873.7768 2275.773 2767.240   3254.006
## var4       NA 2055.248 2503.431   2899.942
## var5       NA 2452.135 3035.409   3848.150
## var6       NA 2129.262 2554.172   2887.828
## var7       NA 2275.773 2767.240   3254.006
## var8       NA 2129.262 2554.172   2887.828
## var9       NA 2430.397 2899.497   3550.889
```

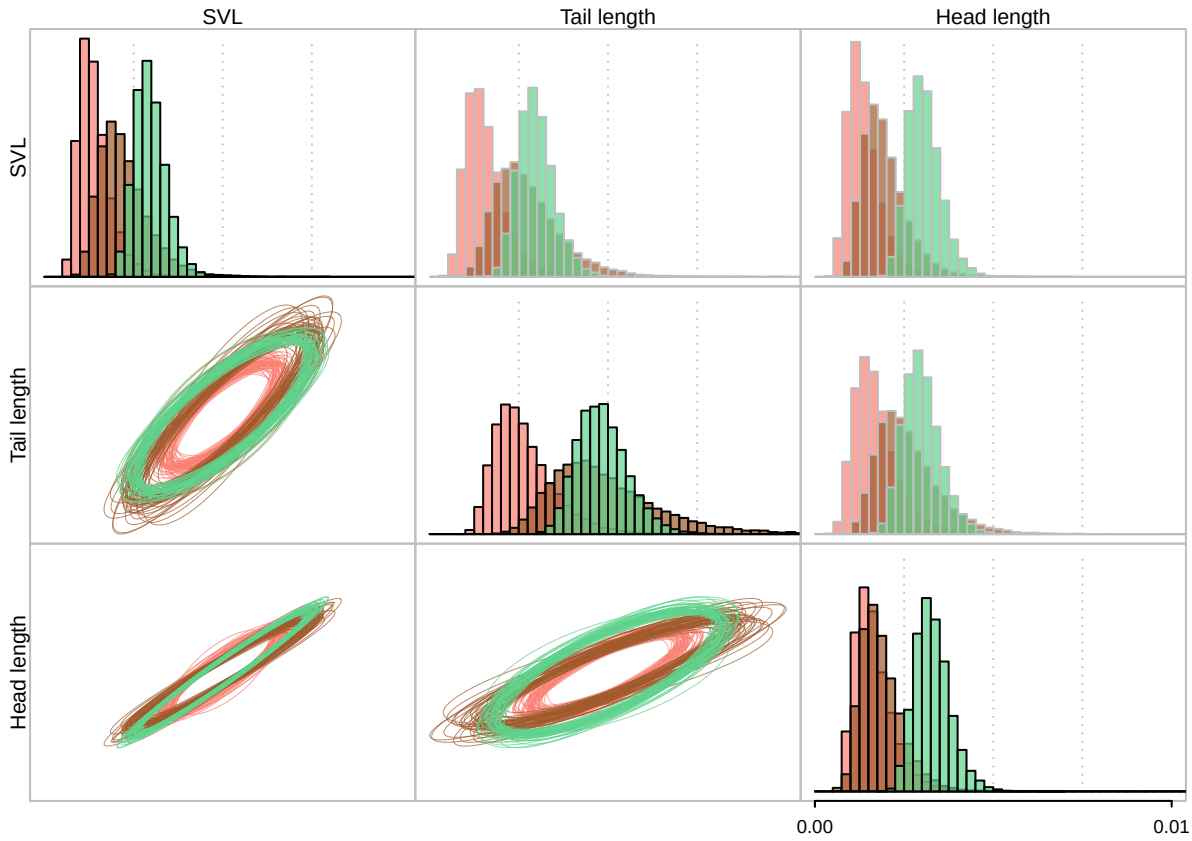
Since the chains converged we can merge them (and give proper names for the regimes):

```
anoles.merged <- mergePosterior(anoles.post[[1]], anoles.post[[2]], anoles.post[[3]],
                                anoles.post[[4]])
names( anoles.merged$matrix ) <- c("mainland.root", "island", "mainland.back")
```

We can plot the posterior distribution:

```
plotRatematrix(anoles.merged, p=c(3,1,2), set.leg=c("SVL","Tail length","Head length"),
               colors=c("#a05a2c","#5fd38d","salmon")[c(3,1,2)],
               alphaOff = 0.7, alphaDiag = 0.7, alphaEll = 0.7, set.xlim = c(0, 0.01))
```

```
## Plotting multiple regimes.
## Table with regimes and colors (names or HEX):
## mainland.root        island mainland.back
##        salmon       #a05a2c       #5fd38d
```



Now we can compute the summary statistics based on the posterior distribution. First statistics compute the overlap in density between regimes for each of the elements of the rate matrix.

```
testRatematrix(chain=anoles.merged, par="all")
```

```
##            mat #1 x #2 mat #1 x #3 mat #2 x #3
## test value      0.5967      0.3126      0.0563
```

Now check for differences in the structure of correlation between rate matrix regimes.

```
testRatematrix(chain=anoles.merged, par="correlation")
```

```
##            mat #1 x #2 mat #1 x #3 mat #2 x #3
## test value      0.5475      0.5671      0.8481
```

Finally, check for differences in the rates of evolution of each trait between rate matrix regimes.

```
testRatematrix(chain=anoles.merged, par="rates")
```

```
##            mat #1 x #2 mat #1 x #3 mat #2 x #3
## test value      0.3037      0.3124      0.0307
```