Lab5: Diversification analysis with BAMM

Comparative Biology and Macroevolution

May 17, 2019

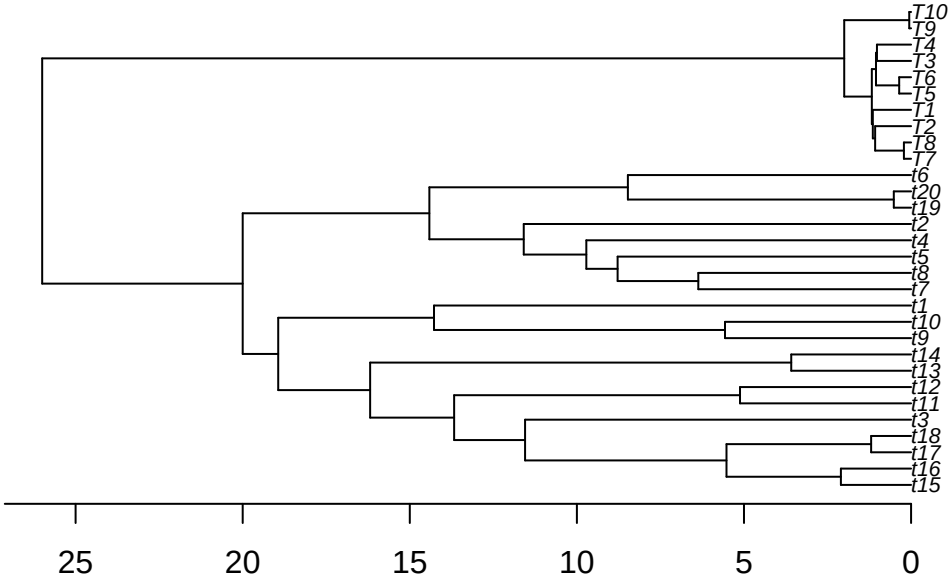
```
library(BAMMtools)
```

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

```
library(coda)
```

Classwork

```
tree <- read.tree("tree.tre")
tree <- ladderize(tree)
plot(tree, cex = 0.7)
axisPhylo()
```



Introduction

Today we will be analyzing speciation rates of this simple tree with BAMM and *BAMMtools*. All of the analyses we do today could be repeated for extinction or net diversification rates.

BAMM can be used to model speciation-extinction rates and phenotypic evolutionary rates across phylogenetic trees. To run a diversification analysis on your dataset, you need the following (easiest if all in the same directory):

1. A time-calibrated phylogenetic tree
2. A control file
3. The BAMM program

The steps for BAMM / *BAMMtools* analysis are:

1. Prepare BAMM control file
2. Run BAMM (a command-line program)
3. Import BAMM results into R
4. Check quality of BAMM results
5. Analyze BAMM results with *BAMMtools*

Prepare BAMM control file

We will use *BAMMtools* to calculate the priors from our data. Then we will use these priors to prepare the control file.

BAMM can run parallel threads- use the following commands in Terminal to get CPU count:

```
nproc # Linux
sysctl -n hw.ncpu # Mac
msinfo32 # Windows
```

Verify BAMM assumptions. Your phylogenetic tree must be ultrametric, it must be fully bifurcating (no polytomies), and all branch lengths must be greater than 0. BAMM checks for these things, but it's also good to do a quick check in R using the ape package. You can do this as follows:

```
is.ultrametric(tree)
```

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

```
is.binary.tree(tree)
```

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

```
min(tree$edge.length)
```

```
## [1] 0.02910036
```

We can now create the control file in R.

```
# estimate priors and create control files
priors <- setBAMMpriors(tree, outfile = NULL)
generateControlFile(file = "controlfile.txt", params = list(
    treefile = "tree.tre",
    globalSamplingFraction = "1", # This is 1 for complete sampling
    seed = sample(1:1000000, 1),
    overwrite = "0",
    expectedNumberOfShifts = "1",
    lambdaInitPrior = as.numeric(priors["lambdaInitPrior"]),
    lambdaShiftPrior = as.numeric(priors["lambdaShiftPrior"]),
    muInitPrior = as.numeric(priors["muInitPrior"]),
    numberOfGenerations = "5000000",
    mcmcWriteFreq = "1000",
    eventDataWriteFreq = "1000",
    printFreq = "1000",
    acceptanceResetFreq = "1000",
    outName = "classwork",
    numberOfChains = "2",# set to number of CPUs
    deltaT = "0.01"))
```

Run BAMM

1. Place these files in a folder:
    - BAMM program (not necessary if installed with Homebrew)
    - BAMM control file
    - tree (time-calibrated, binary, and ultrametric)
2. Use command line program (*Terminal* for Mac / *Command Prompt* (?) for PC)
    - set working directory using `cd ~/directory/`
    - run BAMM using `bamm -c controlfile.txt` or `./bamm -c controlfile.txt`

Import BAMM results into R

Import BAMM results with `getEventData()`. This R object is a complex data structure that will be used for most of the analyses in *BAMMtools*

```
edata <- getEventData(tree, eventdata = "classwork_event_data.txt", burnin = 0.1)
```

```
## Reading event datafile:  classwork_event_data.txt
##      ..........
## Read a total of 5000 samples from posterior
##
## Discarded as burnin: GENERATIONS <  499000
## Analyzing  4501  samples from posterior
##
## Setting recursive sequence on tree...
##
## Done with recursive sequence
```

```
summary(edata)
```

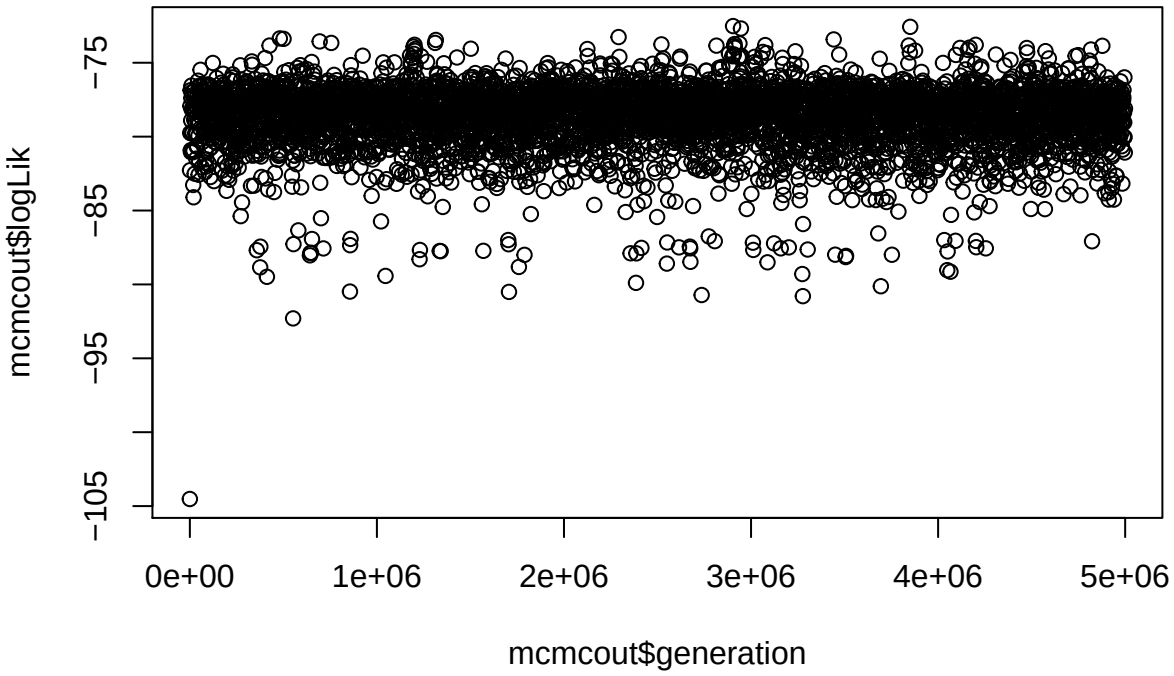
```
##
## Analyzed 4501 posterior samples
## Shift posterior distribution:
##
##          0    0.01100
##          1    0.72000
##          2    0.20000
##          3    0.05000
##          4    0.01100
##          5    0.00360
##          6    0.00089
##
## Compute credible set of shift configurations for more information:
##  See ?credibleShiftSet and ?getBestShiftConfiguration
```

Check quality of BAMM results

Assess MCMC convergence

Plot the log-likelihood trace of your MCMC output file. This should look fuzzy. If it looks jagged you should increase `numberOfGenerations = %%%%` in your control file and rerun BAMM

```
mcmcout <- read.csv("classwork_mcmc_out.txt")
plot(mcmcout$logLik ~ mcmcout$generation)
```



Test for convergence of the MCMC chains with the *coda* package for R

```
burnstart <- floor(0.1 * nrow(mcmcout)) # Discard the first 10% of samples as burnin
postburn <- mcmcout[burnstart:nrow(mcmcout), ]
effectiveSize(postburn$N_shifts)
```

```
##     var1
## 3011.603
```

```
effectiveSize(postburn$logLik)
```

```
##     var1
## 2954.491
```

We want the above numbers to be >200... if not, increase `numberOfGenerations = %%%%` in your control file and rerun BAMM.

Analyze BAMM results with *BAMMtools*

Analysis of rate shifts

Bayes factors can be used to select the best model out of a set of candidate models used in MCMC simulations. Similar to AIC, Bayes factors penalize a model for complexity. In this case we will use Bayes factors to choose the best model for the number of speciation rate shifts in our clade. The model with the highest Bayes factor is the best model for our data.

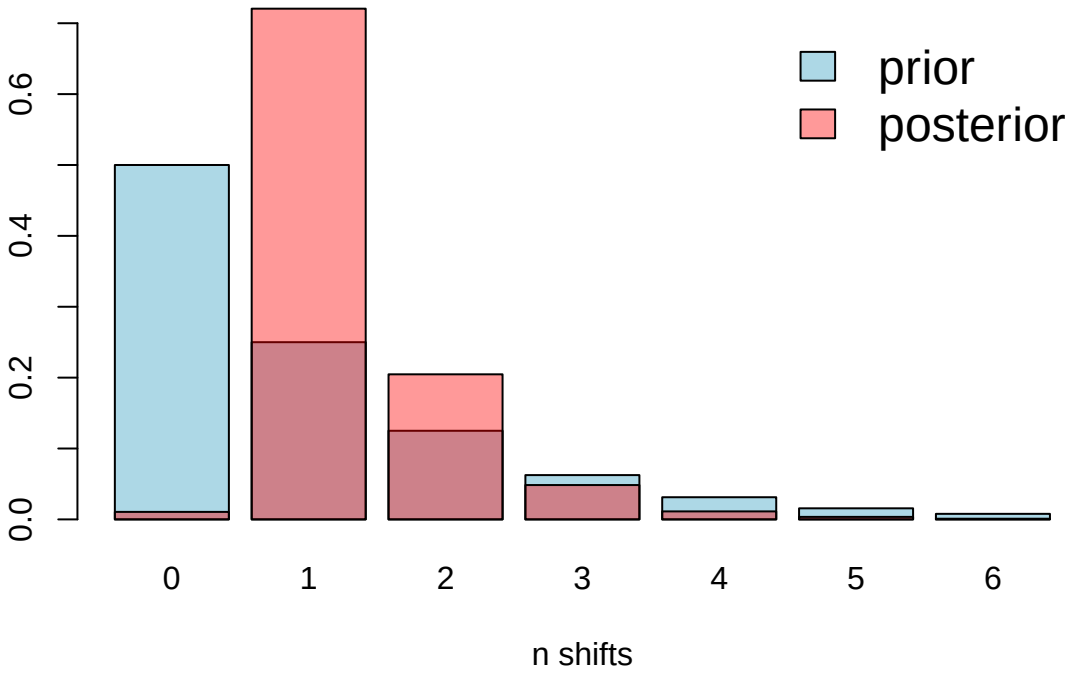
```
computeBayesFactors("classwork_mcmc_out.txt", expectedNumberOfShifts = 1, burnin = 0.1)
```

```
##        0          1          2          3          4           5
## 0   1.00 0.00772320 0.01358696 0.02802691 0.06127451  0.09765625
## 1 129.48 1.00000000 1.75923913 3.62892377 7.93382353 12.64453125
## 2  73.60 0.56842756 1.00000000 2.06278027 4.50980392  7.18750000
```

```
## 3  35.68 0.27556379 0.48478261 1.00000000 2.18627451  3.48437500
## 4  16.32 0.12604263 0.22173913 0.45739910 1.00000000  1.59375000
## 5  10.24 0.07908557 0.13913043 0.28699552 0.62745098  1.00000000
## 6   5.12 0.03954279 0.06956522 0.14349776 0.31372549  0.50000000
##            6
## 0  0.1953125
## 1 25.2890625
## 2 14.3750000
## 3  6.9687500
## 4  3.1875000
## 5  2.0000000
## 6  1.0000000
```

Next we will plot histograms of the prior and posterior probability distributions

```
plotPrior("classwork_mcmc_out.txt", expectedNumberOfShifts = 1)
```

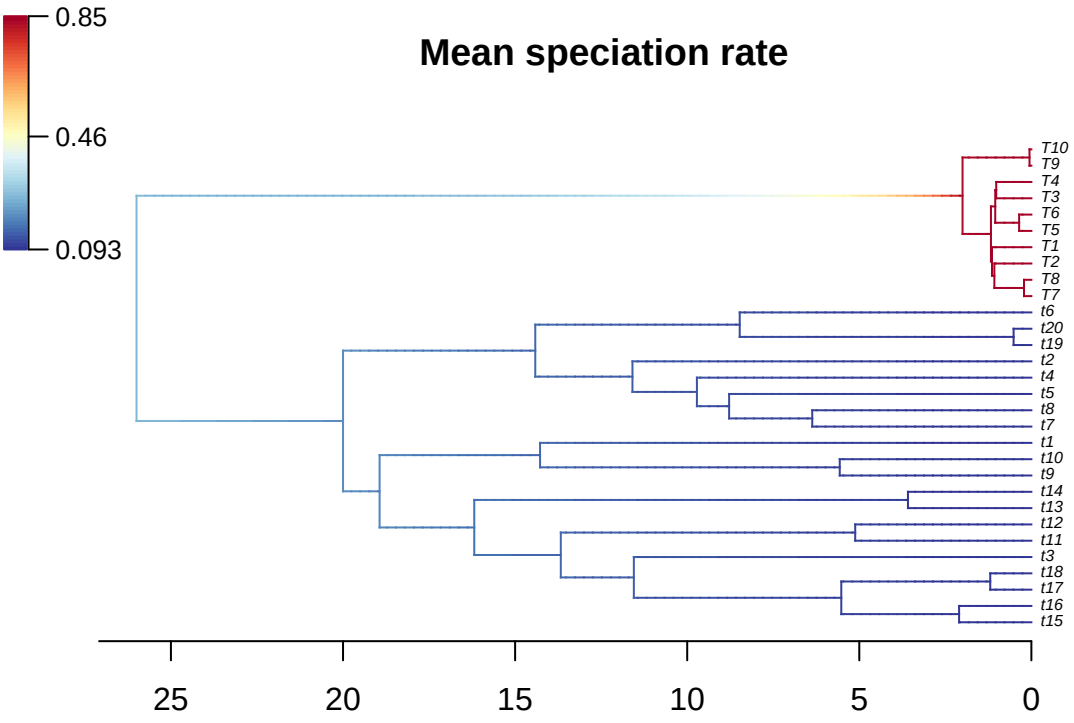


The difference between the prior mode of 0 rate shifts and the posterior mode of 1 rate shift is additional evidence for a model with a single rate shift.

Analysis of rates

One of the most exciting results of BAMM is the phylorate plot. This phylorate plot displays mean, model-averaged diversification rates (*i.e.*, speciation `"s"`, extinction `"e"`, or net diversification `"netdiv"`) on branches of your tree with colors

```
s <- plot.bammdata(edata, spex = "s", labels = T, font = 3, cex = 0.5)
title(main = "Mean speciation rate", sub = "time before present")
addBAMMlegend(s, location = "topleft", nTicks = 1)
axisPhylo()
```



Be careful not to interpret different colors as evidence for the number of rate shifts in your clade.

An important concept to understand here is that BAMM simulates a posterior distribution of distinct ***shift configurations*** on a tree. The 95% credible set is the set of distinct shift configurations that account for 95% of the probability of the data. The `credibleShiftSet()` function searches the posterior distribution of distinct rate shifts and identifies those configurations with the highest probability. We can plot these distinct rate shift configurations and compare the frequency of alternative confgurations in the posterior distribution

```
css <- credibleShiftSet(edata, expectedNumberOfShifts = 1, threshold = 5, set.limit = 0.95)
css$number.distinct # this is number of distinct shift configurations in the data
```

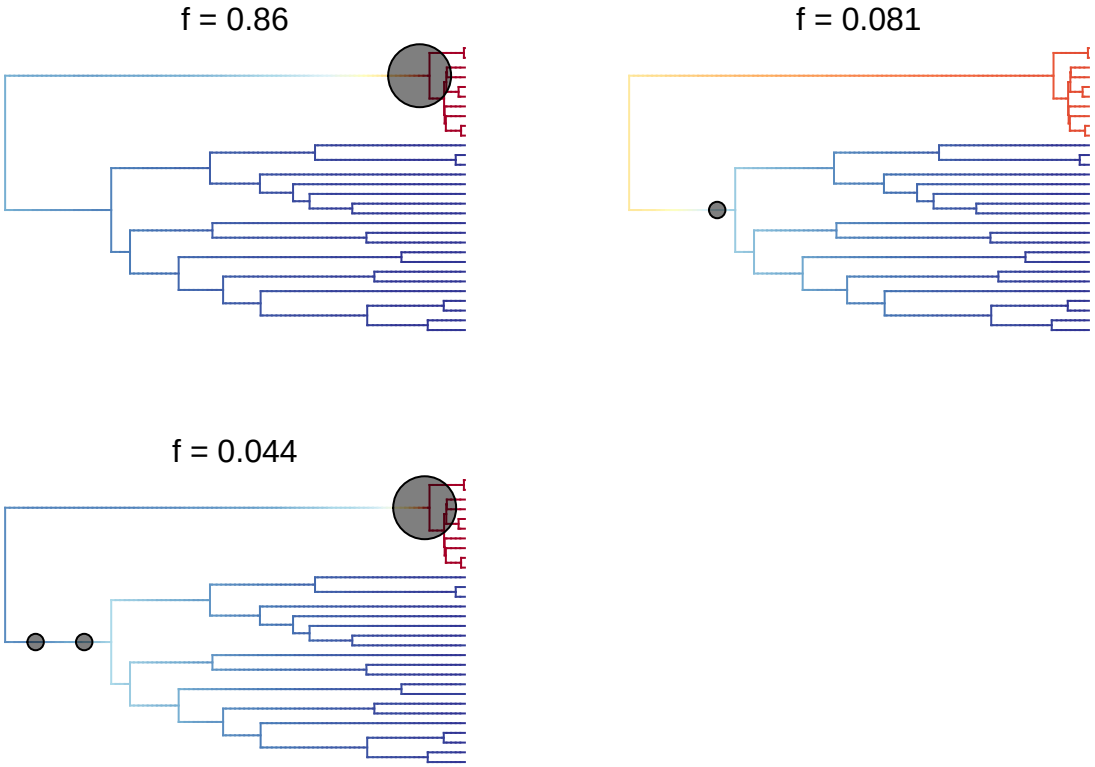
```
## [1] 3
```

```
summary(css)
```

```
##
##  95 % credible set of rate shift configurations sampled with BAMM
##
## Distinct shift configurations in credible set:  3
##
## Frequency of 3 shift configurations with highest posterior probability:
##
##
##    rank     probability cumulative  Core_shifts
##          1 0.85914241  0.8591424          1
##          2 0.08087092  0.9400133          1
##          3 0.04376805  0.9837814          2
```

```
sss <- plot.credibleshiftset(css, border = F)
```

```
## Omitted 0 plots
```



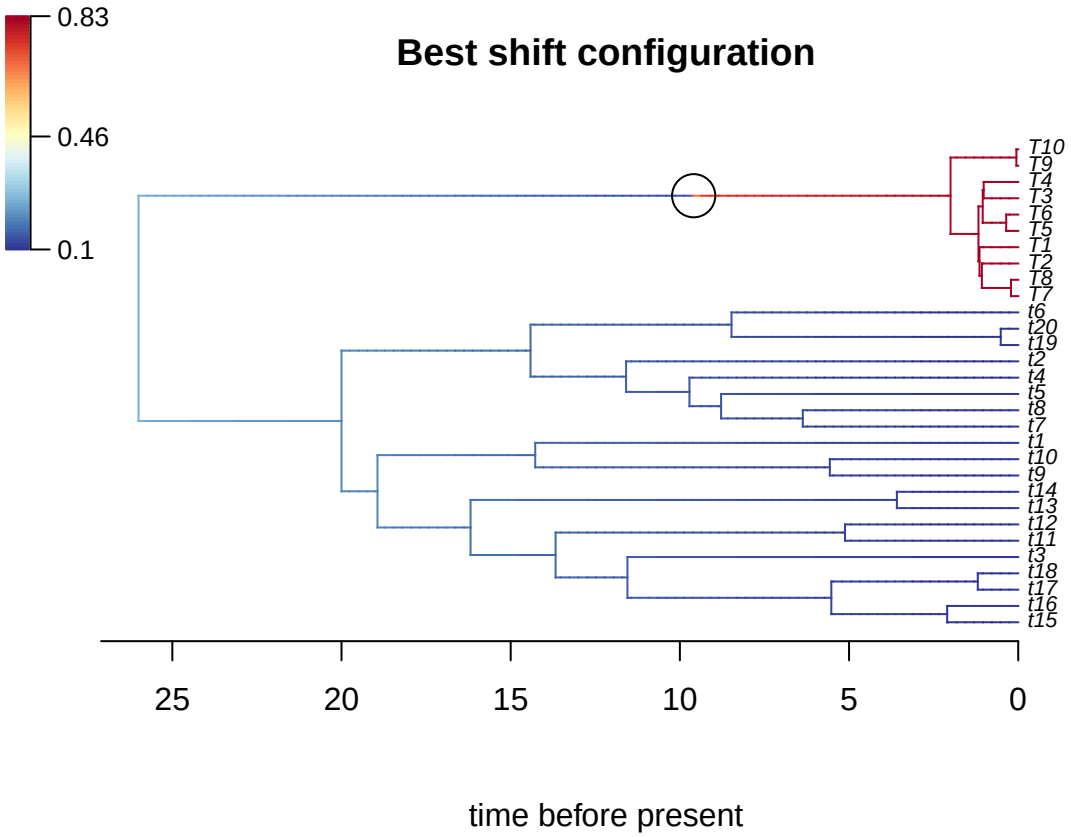
We can see that ~80% of results in the posterior disribution were for a shift configuration where speciation increases leading to clade T, ~11% of the posterior results were for a different shift configuration where speciation rates decrease leading to clade t, and ~6% were for a shift configuration with an increase on the branches leading to both clade T and clade t.

We just viewed a summary of the distinct rate shift configurations with the highest probabilitiies. Now lets find and plot the single best shift configuration (*i.e.*, the one we see most often in the posterior distribution)

```
plot.new()
best <- getBestShiftConfiguration(edata, expectedNumberOfShifts = 1)
```

```
## Processing event data from data.frame
##
## Discarded as burnin: GENERATIONS <  0
## Analyzing  1  samples from posterior
##
## Setting recursive sequence on tree...
##
## Done with recursive sequence
```

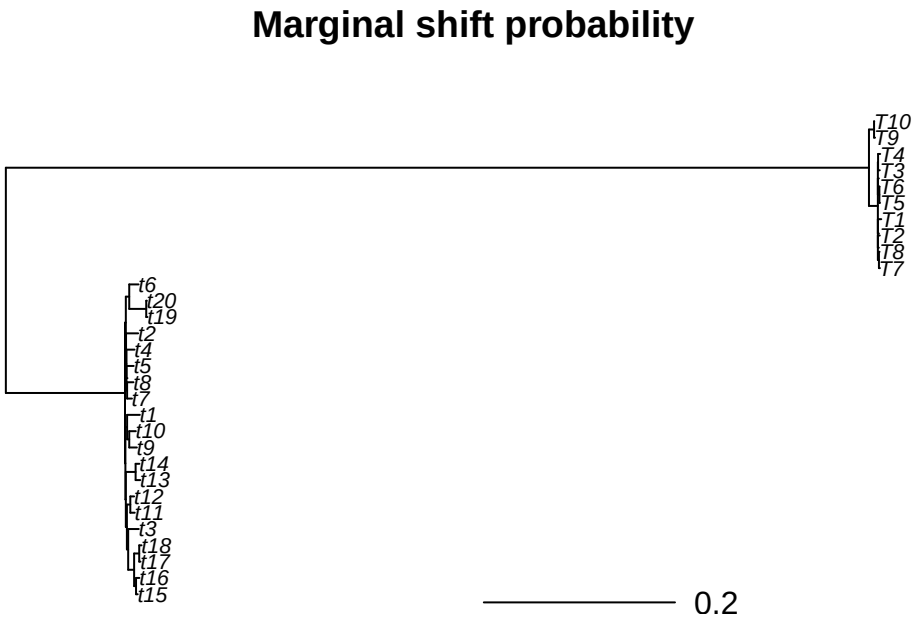
```
ss <- plot.bammdata(best, labels = T, font = 3, cex = 0.7)
title(main = "Best shift configuration", sub = "time before present")
addBAMMlegend(ss, location = "topleft", nTicks = 1)
addBAMMshifts(best, cex = 3, pch = 1)
axisPhylo()
```



This should correspond to the plot in the credible shift set (CSS) with the highest frequency, but will look slightly different because it is just a single sample from the posterior whereas CSS is averaged across samples in the posterior.

Next we will plot a tree with branch lengths scaled by the probability that they contain a rate shift. The longer the branches, the greater probability that there was a rate shift somewhere along that branch

```
par(font = 1)
marg_probs <- marginalShiftProbsTree(edata)
plot.phylo(marg_probs, cex = 0.7)
title(main = "Marginal shift probability")
add.scale.bar(x = 0.5, y = 0.5, font = 1)
```



What branch was the speciation rate more likely to have shifted on?

What was more likely - a rate increase leading to T clade or rate decrease leading to t clade?

IMPORTANT: there is no minimum probability here to be considered 'significant'. For example, a marginal shift probability of 0.5 is very strong evidence that a rate shift occurred somewhere on that branch!

Clade-specific evolutionary rates

We can compute clade-specific marginal distributions of rates with `getCladeRates()`. Below we estimate an overall speciation rate and 90% credible interval, then rates and intervals for the T and t clades separately.

```
global_rates <- getCladeRates(edata)
mean(global_rates$lambda)
```

```
## [1] 0.1790505
```

```
# The speciation rate estimate is 0.19 new species per million years
quantile(global_rates$lambda, c(0.05, 0.95))
```

```
##        5%       95%
## 0.1109176 0.2783375
```

```
# There is 90% probability that the speciation rate of this clade is between 0.11 and 0.29

T_MRCA <- getMRCA(tree, tip = c("T1", "T10"))
T_rates <- getCladeRates(edata, node = T_MRCA)
mean(T_rates$lambda)
```

```
## [1] 0.504537
```

```
quantile(T_rates$lambda, c(0.05, 0.95))
```

```
##        5%       95%
## 0.2468588 0.9153356
```

```
t_MRCA <- getMRCA(tree, tip = c("t1", "t20"))
t_rates <- getCladeRates(edata, node = t_MRCA)
mean(t_rates$lambda)
```

```
## [1] 0.1245441
```

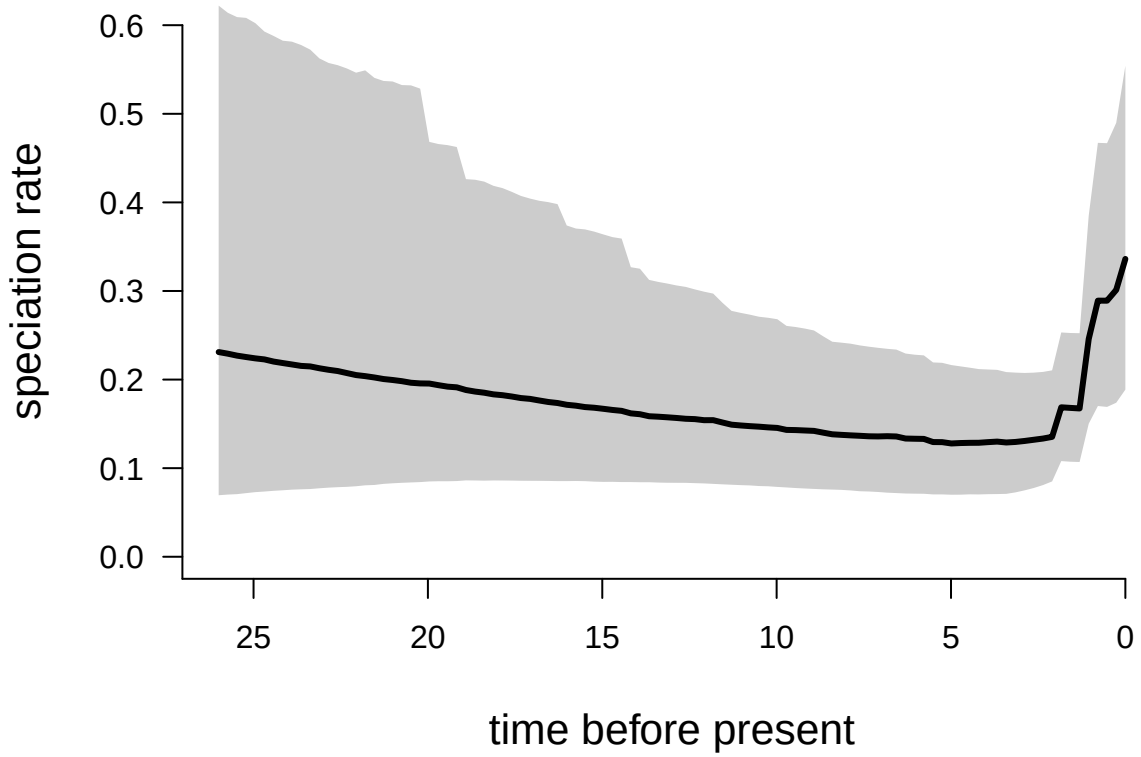
```
quantile(t_rates$lambda, c(0.05, 0.95))
```

```
##         5%        95%
## 0.07260559 0.20485550
```

Rate-through-time analysis

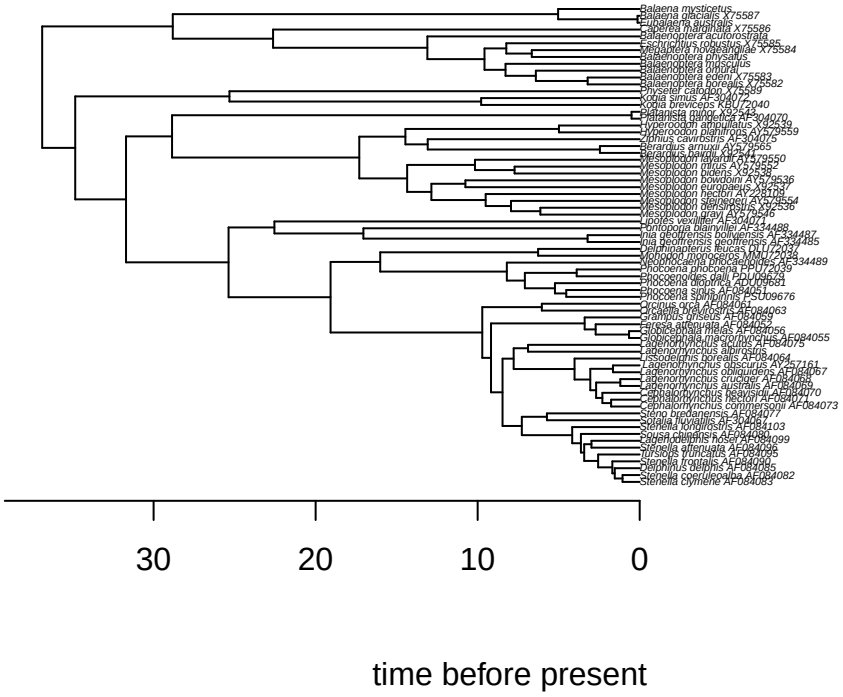
We can plot speciation (default) or extinction rates through time with `plotRateThroughTime()`. This plot can display dynamics in speciation or extinction rates. The black line is the mean speciation rate, and the grey area is the 90% credible interval for the speciation rate

```
par(font = 1)
plotRateThroughTime(edata,
                    ratetype = "speciation",
                    avgCol = "black",
                    intervalCol = "gray80",
                    intervals = c(0.05, 0.95),
                    opacity = 1)
```



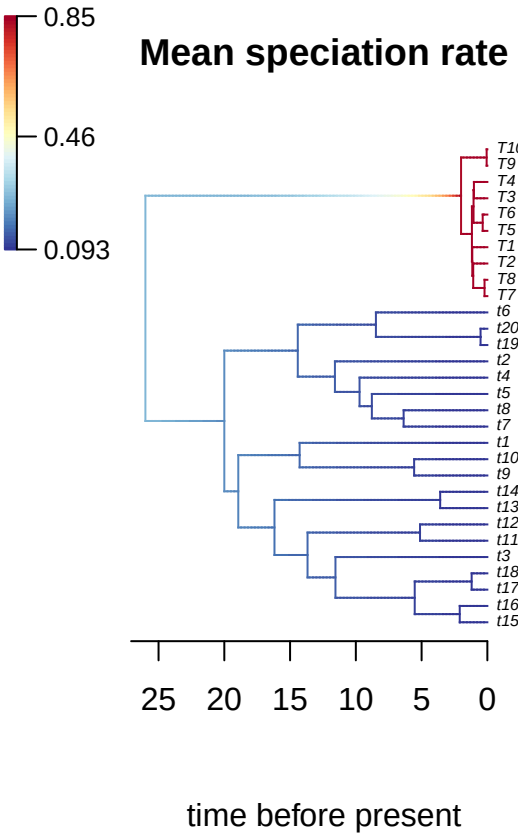
What does the increase in speciation rates at ~2 mya correspond to?

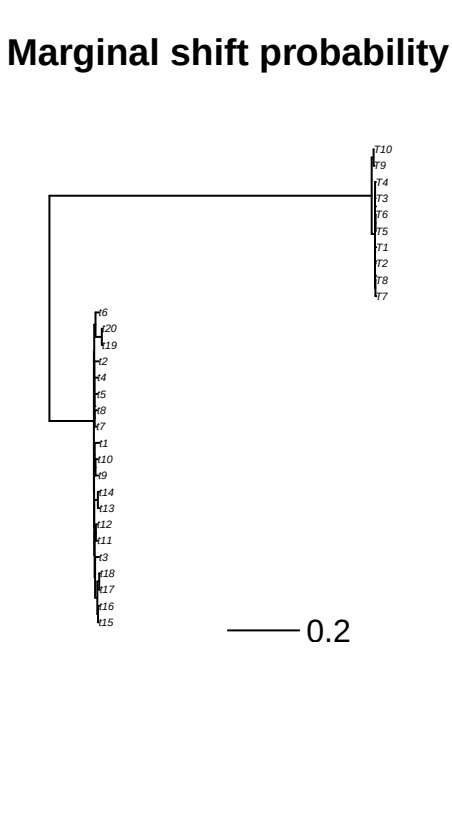
Homework



You will analyze Cetecean speciation rates with BAMM. Standard lab writeup format applies.

1. Use *whaleTree.tre* (Slater *et al.* 2010)
2. Run BAMM
   - Specify the sampling probability of Ceteceans using `globalSamplingFraction = ???`.
   - Run at least 5 million generations. This should take less than an hour
3. Analyze results with *BAMMtools*. Be sure to research the following topics:
   - Number and location(s) of speciation rate shifts in Ceteceans
   - Variation in speciation rates among Cetecean species
   - Dynamics of Cetecean speciation rates
   - Include the following 2 plots side-by-side in same figure
     - Plot mean speciation rates on the tree (*i.e.*, phylorate plot)
     - Plot the tree with branch lengths scaled by marginal shift probability





4. Provide a biological interpretation of your results. Be sure to address the following items in your lab report:
    - How many times has the speciation rate shifted in the Ceteceans?
    - What group(s) of Ceteceans have the highest speciation rates? The lowest rates?
    - Which is the more likely scenario: acceleration of speciation rates in toothed whales or deceleration of speciation rates in baleen whales?
    - What are the dynamics of speciation in Ceteceans?
5. Run BAMM on your project tree. No lab writeup for this part of homework.