Tree Simulation assignment

Gaurav Kandlikar

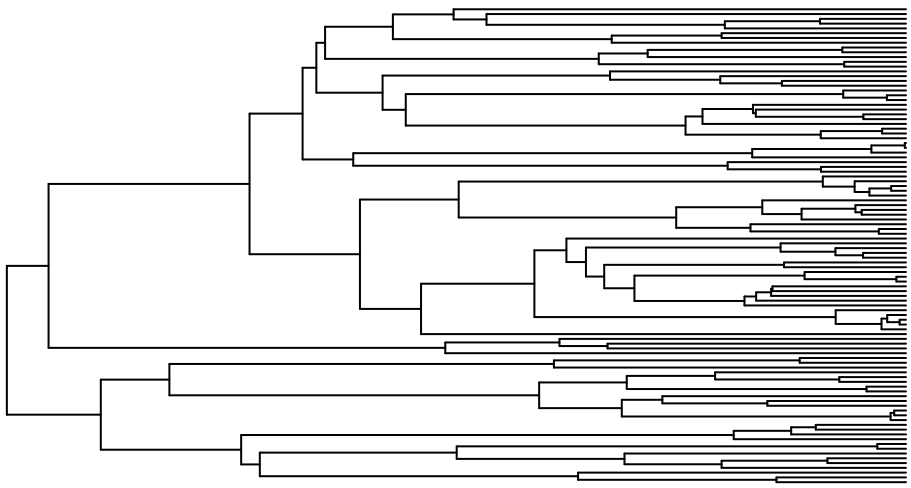
November 5, 2015

getting set up

```
library(geiger); library(phytools); library(diversitree); library(DDD); library(magrittr)

source("~/grad/courses/UCLA/eeb_200a_evolution/alfaro_docs/assignments/tree_sim/alfaro_docs/rabosky_fun
```

Using `simulateTree` in `phytools` to simulate tree under birth-death model

```
pars <- c(10, 0) # First lambda, then mu- these are the true parameters underlying the tree

tt <- simulateTree(pars = pars, max.taxa = 100)

plot(tt, show.tip.label = F, main = bquote(paste("True parameters: ", lambda == .(pars[1]), ", ", mu ==
```



We know the paramters underlying the birth-death process here, so we can check how well we can estimate them froom the tree. For this we use the `made.bd()` function of `diversitree`. This is the description:

```
Prepare to run a constant rate birth-death model on a phylogenetic tree. This fits the Nee et al. 1994 
```

We then use `fitDiversitree()` from `rabosky_functions.r`

```
fit <- make.bd(tt) %>% fitDiversitree()

# Extract parameter estimates
fit$pars
```

```
##  lambda      mu
## 10.7751  0.0010
```

We can do a null model to get a 95% confidence interval on these estimates:

```
reps <- 1000
pars <- c(10, 0) # First lambda, then mu- these are the true parameters underlying the tree

lambdas <- numeric(reps)
mus <- numeric(reps)

for (i in 1:reps) {
  fit <- fitDiversitree(make.bd(simulateTree(pars = pars, max.taxa = 100)))
  estimates <- fit$pars
  lambdas[i] <- estimates["lambda"]
  mus[i] <- estimates["mu"]
}

mean(lambdas); mean(mus)
```

```
## [1] 10.87514
```

```
## [1] 1.641939
```

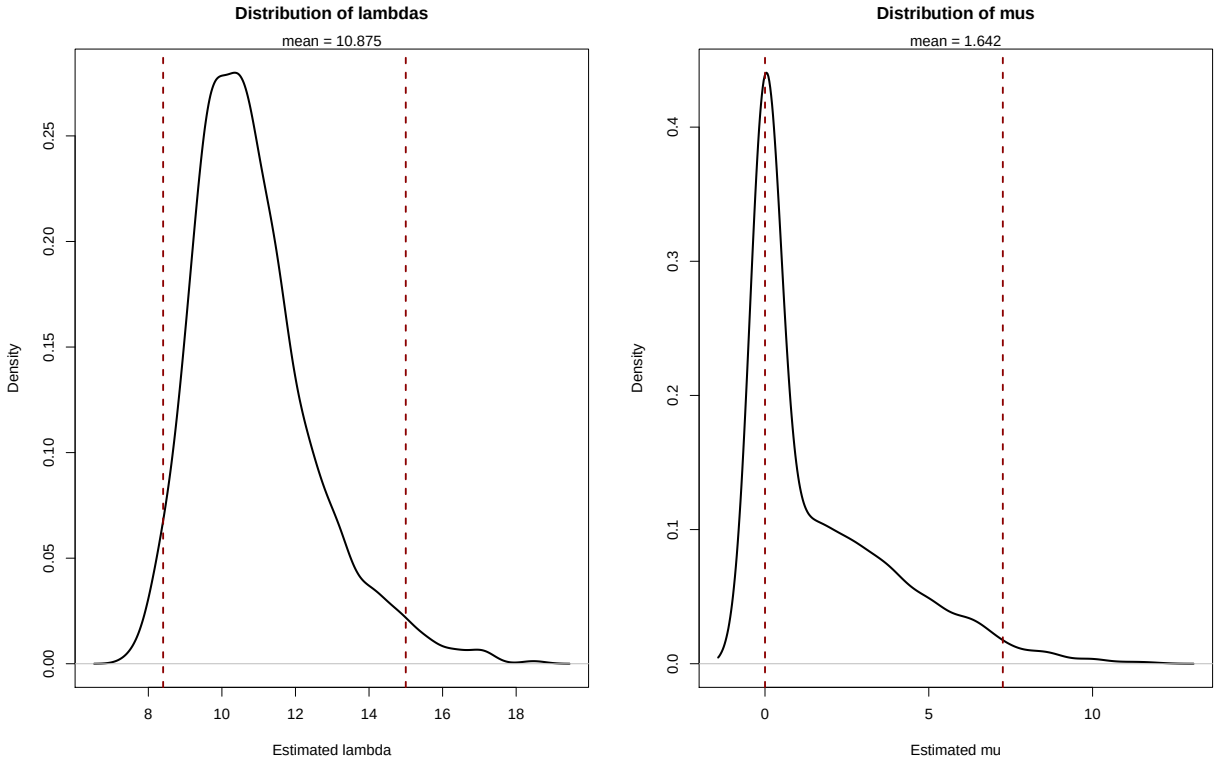
```
# Quantiles
mu_lines <- quantile(mus, probs = c(0.025, 0.975))
lambda_lines <- quantile(lambdas, probs = c(0.025, 0.975))

lambda_lines; mu_lines
```

```
##      2.5%     97.5%
##  8.402472 14.999207
```

```
##     2.5%    97.5%
## 0.001000 7.260158
```

```
# Plot
par(mfrow = c(1,2))
# Lambda plot
plot(density(lambdas), main = "Distribution of lambdas", xlab = "Estimated lambda", lwd = 2.2)
abline(v = lambda_lines, lwd = 2, col = "darkred", lty = 2)
# text(x = c(lambda_lines[1]-1, lambda_lines[2]+1), y = 0.285, labels = round(lambda_lines, 3))
# text(x = 16, y = 0.285, labels = round(mean(lambdas), 3))
mtext(side = 3, text = paste("mean =", round(mean(lambdas), 3)))
```

```
# Mu plot
plot(density(mus), main = "Distribution of mus", xlab = "Estimated mu", lwd = 2.2)
abline(v = mu_lines, lwd = 2, col = "darkred", lty = 2)
# text(x = c(mu_lines[1]-1, mu_lines[2]+1), y = 0.51, labels = round(mu_lines, 3))
# text(x = 10, y = 0.51, labels = round(mean(mus), 3))
mtext(side = 3, text = paste("mean =", round(mean(mus), 3)))
```




```
dev.off()
```

```
## null device
##           1
```

```
# Coefficients of variation- to see how tight the spread is
lambda_cv <- sd(lambdas)/mean(lambdas)
mus_cv <- sd(mus)/mean(mus)

lambda_cv; mus_cv
```

```
## [1] 0.1507231
```

```
## [1] 1.355052
```

What does this imply about what we can learn from empirical studies of molecular phylogenies about diversification?

As we see in `lambda_cv` and `mus_cv` which are the coefficients of variation (`sd/mean`), the estimate of speciation rate is much tighter than that of extinction rate. This suggests that emperical studies based on

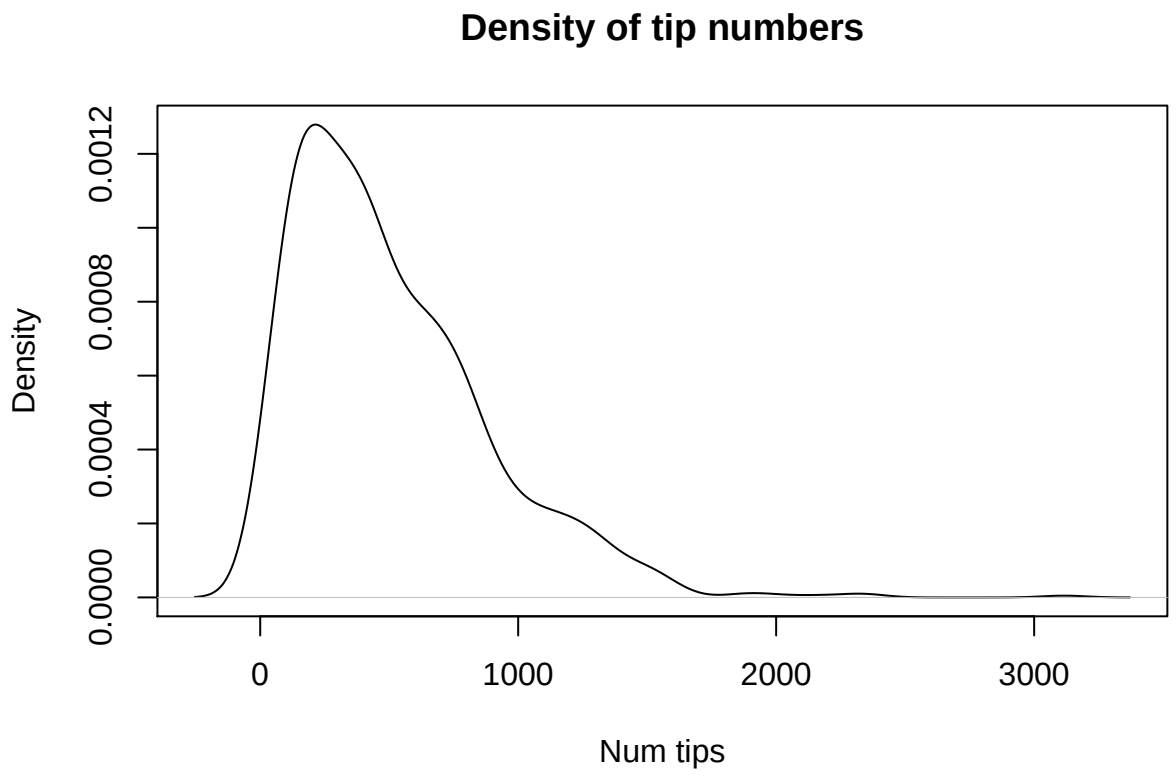
molecular phylogenies *can* calculate some sort of extinction rate, but this rate is not necessarily to be believed-having fossil data may be invaluable in these cases.

Exercise 2

Simulate 100 trees under a constant rate of birth and death. Extract the number of species from each tree. Create a histogram of the resulting distribution. How much stochasticity is associated with the outcome of a birth-death process? What does this suggest about our ability to intuitively identify clades that have undergone exceptional speciation?

I will use `pbtree()` in `phytools` to make trees with constant birth death rates:

```
reps <- 1000
num_tips <- numeric(reps)
for (i in 1:reps) {
  tt <- pbtree(b = .5, d = .05, t = 12)
  num_tips[i] <- length(tt$tip.label)
}
par(mfrow = c(1,1))
plot(density(num_tips), main = "Density of tip numbers", xlab = "Num tips")
```



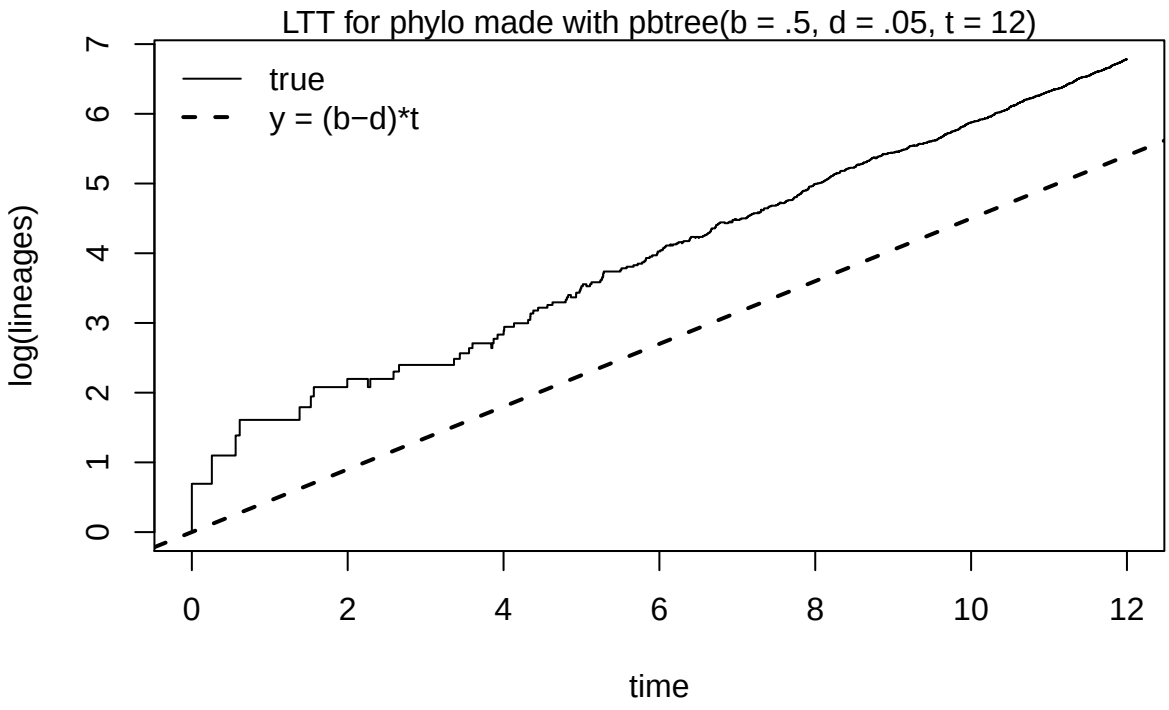

```
mean(num_tips); quantile(num_tips, c(0.025, 0.975))
```

```
## [1] 508.935
```

```
##     2.5%    97.5%
##   41.975 1401.100
```

There is clearly a wide range of final number of species derived from the same underlying birth/death rates- here the range is from 0(!) to 1500. This means that the estimates of lambda and mu that we can calculate from a molecular phylogeny can in turn lead to a wide range of topologies. We can use `ltt()` to make a lineage-through-time plot

```
ltt(tt)
mtext(side = 3, text = "LTT for phylo made with pbtree(b = .5, d = .05, t = 12)")
# This will plot the last tree made in the loop above
# Should the slope of this line be equal to (b-d)?
# N(t) = N(0)*e^((b-d)*t)
# Yes, it should be- since if we take a log then the (b-d) becomes multiplicative

abline(a = c(0, 0.45), lwd = 2, lty = 2)
legend("topleft", lwd = c(1,2), lty = c(1,2), legend = c("true","y = (b-d)*t"), bty = "n")
```

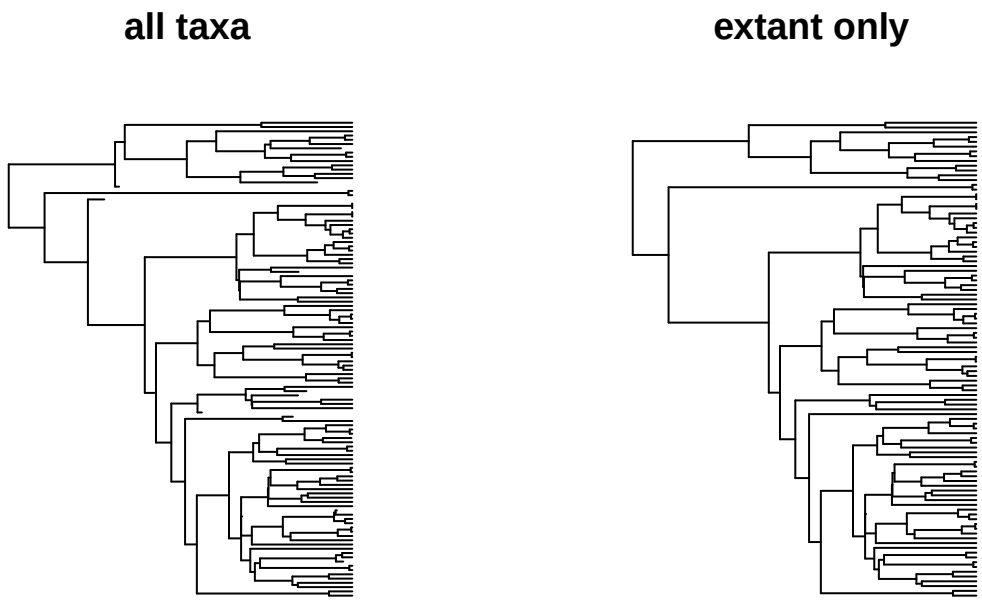


Let us now consider some trees where extinct taxa are analyzed

Simulation where extinct taxa are analysed. We use the function `sim.bdtree` from `geiger`.

```
# Part of Alfaro's chunk- not sure if needed here
# pars <- c(10, 5)
# tt <- simulateTree(pars, max.taxa=100)
```

```
# sim.bdtree:
ttEx <- sim.bdtree(b = 10, d = 1, stop = "taxa", t = 100, n = 100 )

livingOnly <- drop.fossil(ttEx) # Subset to extant taxa only

par(mfrow = c(1,2))
plot(ttEx, show.tip.label = F, main = "all taxa")
plot(livingOnly, show.tip.label = F, main = "extant only")
```

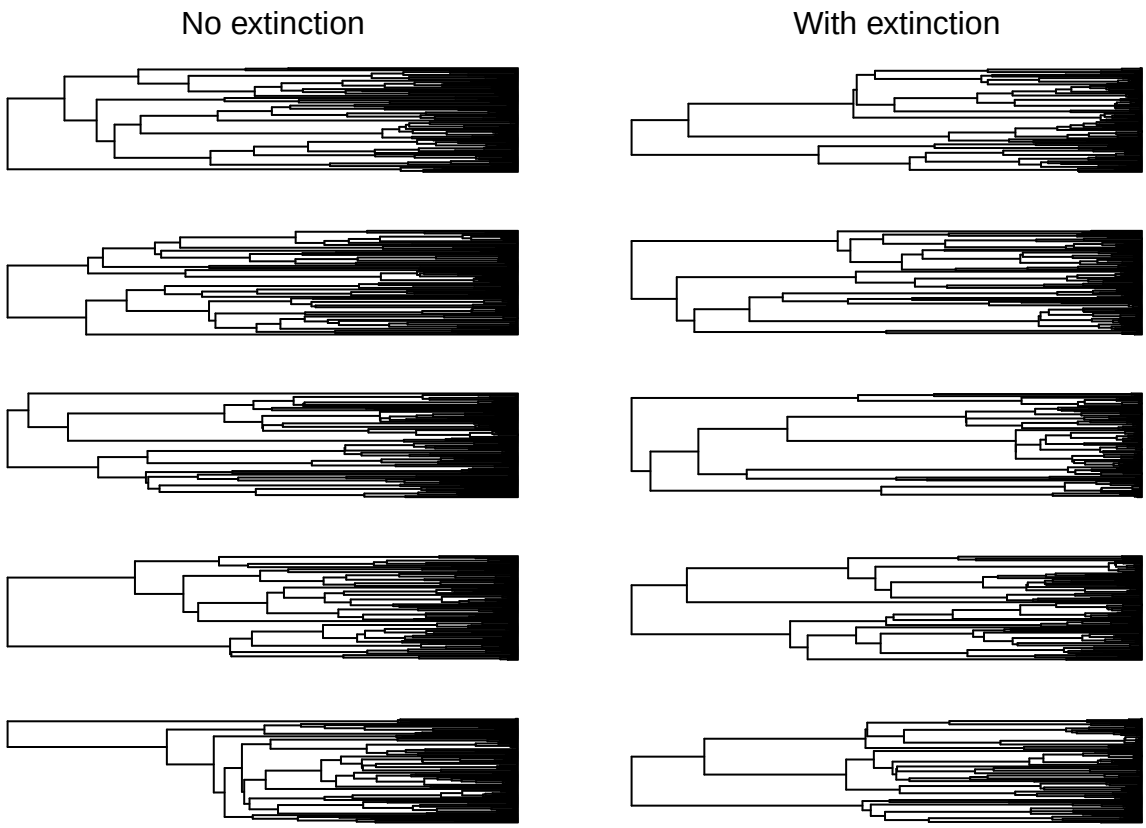


Exercise 3. What does extinction do to the shape of the tree?

Simulate 5 trees with and without extinction that have similar net diversification rates. Can you say anything about the general shape of the trees that have been simulated with extinction?

```
reps <- 5
gammas <- matrix(nrow = reps, ncol = 2)
colnames(gammas) <- c("no_Ex", "Ex")
par(mfrow = c(reps ,2), mar = c(1,1,1,1), oma = c(0,0,2,0))

for(i in 1:reps) {

  # Make a tree with no extinction- b = 0
  tt_noEx <- sim.bdtree(b = 4, d = 0, stop = "taxa", t = 100, n = 100)
```

```
  # Make a tree with extinctions- b != 0. Note that net div rate = 4
  ttEx <- sim.bdtree(b = 10, d = 6, stop = "taxa", t = 100, n = 100, extinct = F)
  ttEx <- drop.fossil(ttEx)

  # Save gammas in matrix
  gammas[i, "no_Ex"] <- gammaStat(tt_noEx)
  gammas[i, "Ex"] <- gammaStat(ttEx)

  # Plot
  plot(tt_noEx, show.tip.label = FALSE)
  if(i == 1) (mtext(side = 3, line = 1, text = "No extinction"))
  plot(ttEx, show.tip.label = FALSE)
  if(i == 1) (mtext(side = 3, line = 1, text = "With extinction"))
}
```

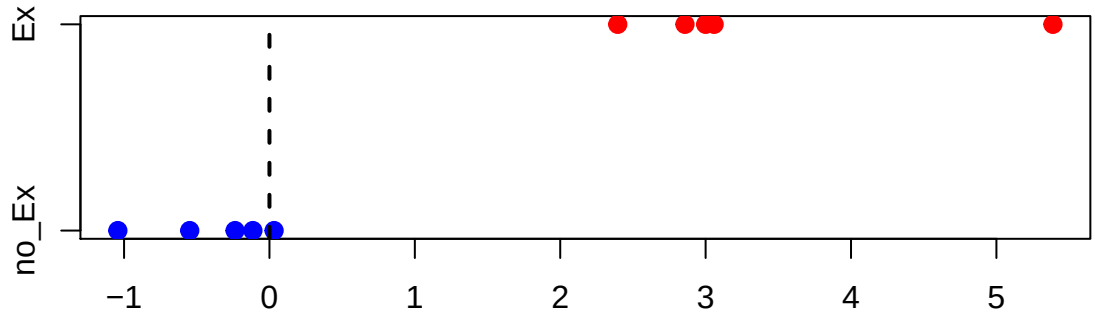



```
dev.off()
```

```
## null device
##           1
```

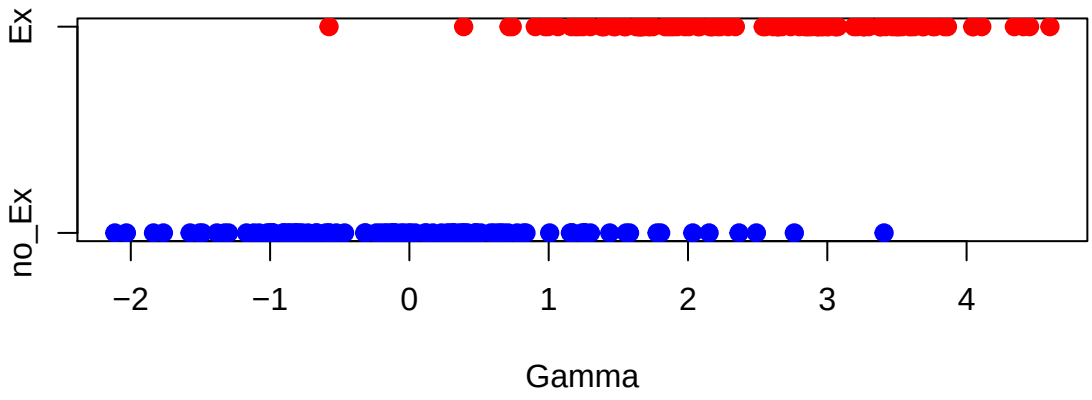
Eyeballing, it seems that the trees simulated with no-extinctions are "pulled left" - that is, have more early nodes than do the extinction trees. We can run numbers on this - I predict `gamma` of the no-extinction trees to be more negative than that of the no-extinction trees.

```
stripchart(as.data.frame(gammas), pch = 19, col = c ("blue", "red"), cex = 1.2)
abline(v = 0, lty = 2, lwd = 2)
```



The sample of 5 above is suggestive, but not conclusive- 5 is a small number of reps! Let's do it with more to confirm the pattern:

```
reps <- 100
gammas <- matrix(nrow = reps, ncol = 2)
colnames(gammas) <- c("no_Ex", "Ex")

for(i in 1:reps) {

  # Make a tree with no extinction
  tt_noEx <- sim.bdtree(b = 4, d = 0, stop = "taxa", t = 100, n = 100)

  # Make a tree with extiontion and drop extinct taxa
  ttEx <- sim.bdtree(b = 10, d = 6, stop = "taxa", t = 100, n = 100, extinct = F)
  ttEx <- drop.fossil(ttEx)

  gammas[i, "no_Ex"] <- gammaStat(tt_noEx)
  gammas[i, "Ex"] <- gammaStat(ttEx)

}

stripchart(as.data.frame(gammas), pch = 19, col = c ("blue", "red"), cex = 1.2, xlab = "Gamma")
```




```
t.test(gammas[,1],gammas[,2])
```

```
##
##  Welch Two Sample t-test
##
## data:  gammas[, 1] and gammas[, 2]
## t = -16.204, df = 197.23, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.748922 -2.152414
## sample estimates:
##  mean of x  mean of y
## 0.05592812 2.50659611
```

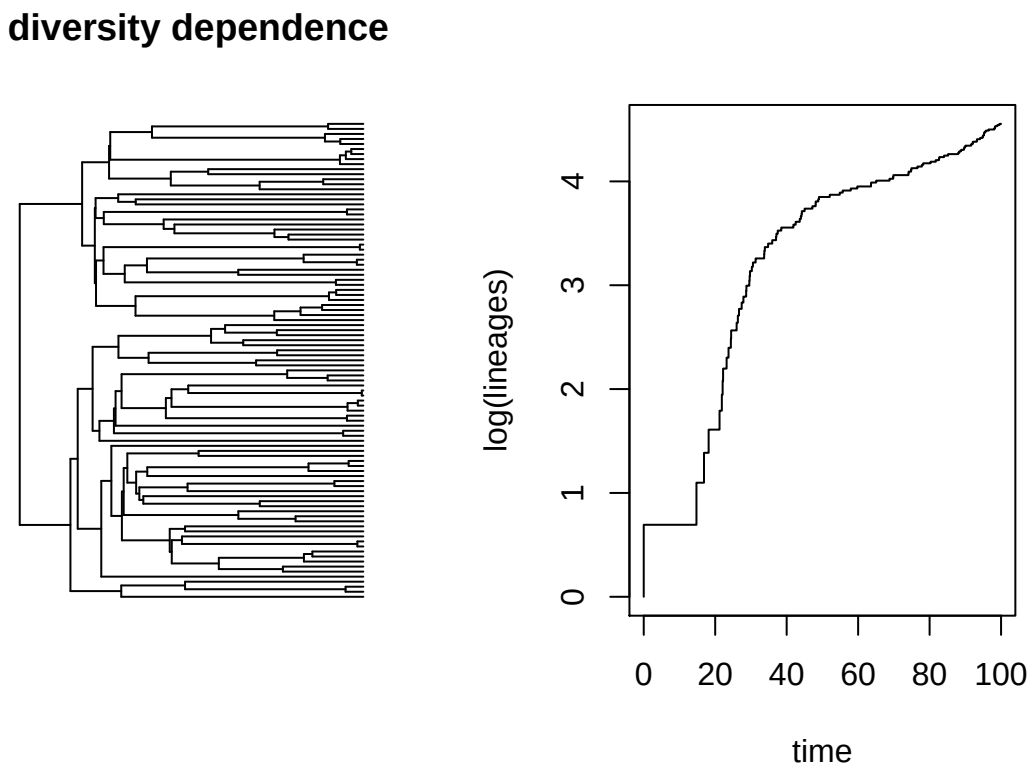
Simulating trees under a density dependent model!

From the help file of `dd_sim`:

```
dd_sim(pars, age, ddmodel = 1)
Vector of parameters:
  pars[1] corresponds to lambda (speciation rate)
  pars[2] corresponds to mu (extinction rate)
  pars[3] corresponds to K (clade-level carrying capacity)
  age Sets the crown age for the simulation
```

```
# Simulate
ddTree <- dd_sim(c(0.12,0.02,100),100)

str(ddTree)
# Subset to tree of extant taxa
ddLivingOnly <- ddTree[[1]]

# Plot phylo and LTT
```

```
par(mfrow = c(1,2))
plot(ddLivingOnly, show.tip.label = F, main = "diversity dependence")
ltt(ddLivingOnly)
```




```
# Check number of taxa in the tree- should be ~100 since that is clade carrying capacity
length(ddLivingOnly$tip.label)
```

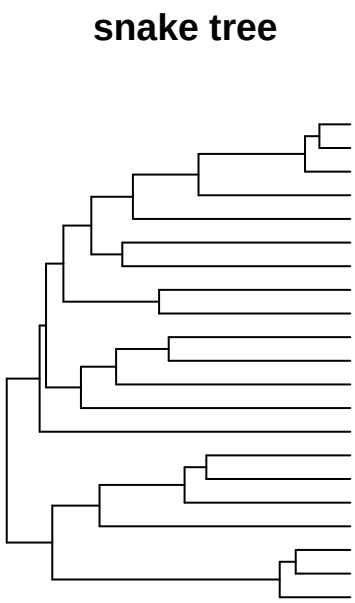
```
## [1] 95
```

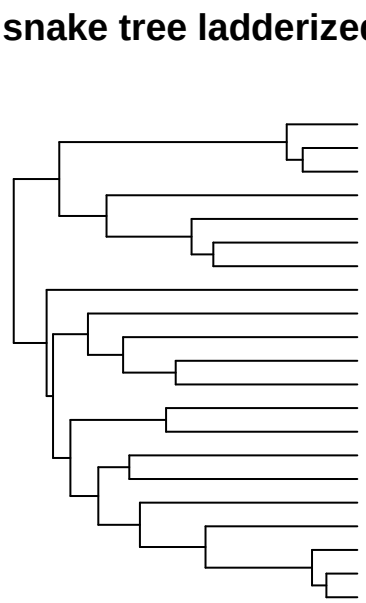
Exercise 4

Does the diversity dependent tree or lineage through time plot look different than the pure birth tree?

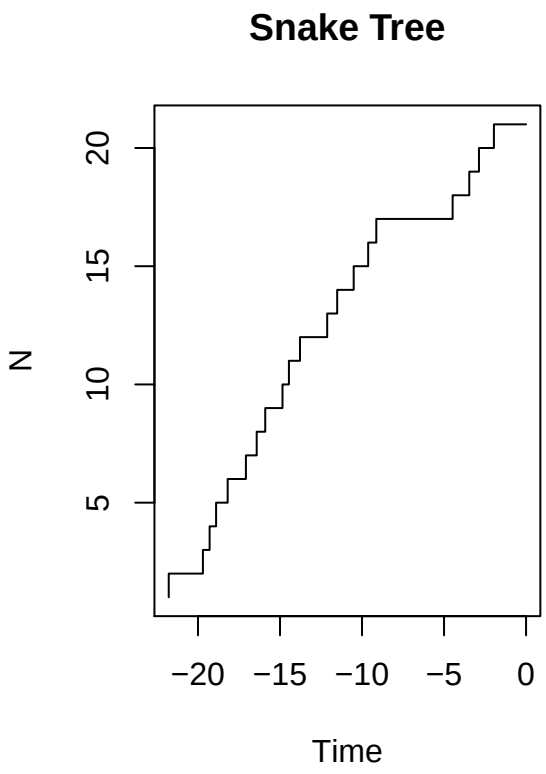
The two LTTs are not identical- and nor should they be. LTT of a tree made under a DD model should show a saturating curve as the clade diversity approaches the user-specified clade carrying capacity, whereas the pure birth / birth-death graphs should show a continued increase increase.

```
snake_tree <- read.tree(file = "~/grad/courses/UCLA/eeb_200a_evolution/alfaro_docs/assignments/tree_sim/
par(mfrow = c(1,2))
plot(snake_tree, show.tip.label = F, main = "snake tree")
plot(ladderize(snake_tree), show.tip.label = F, main = "snake tree ladderized")
```






```
# make an ltt
ltt.plot(snake_tree, main = "Snake Tree")
```



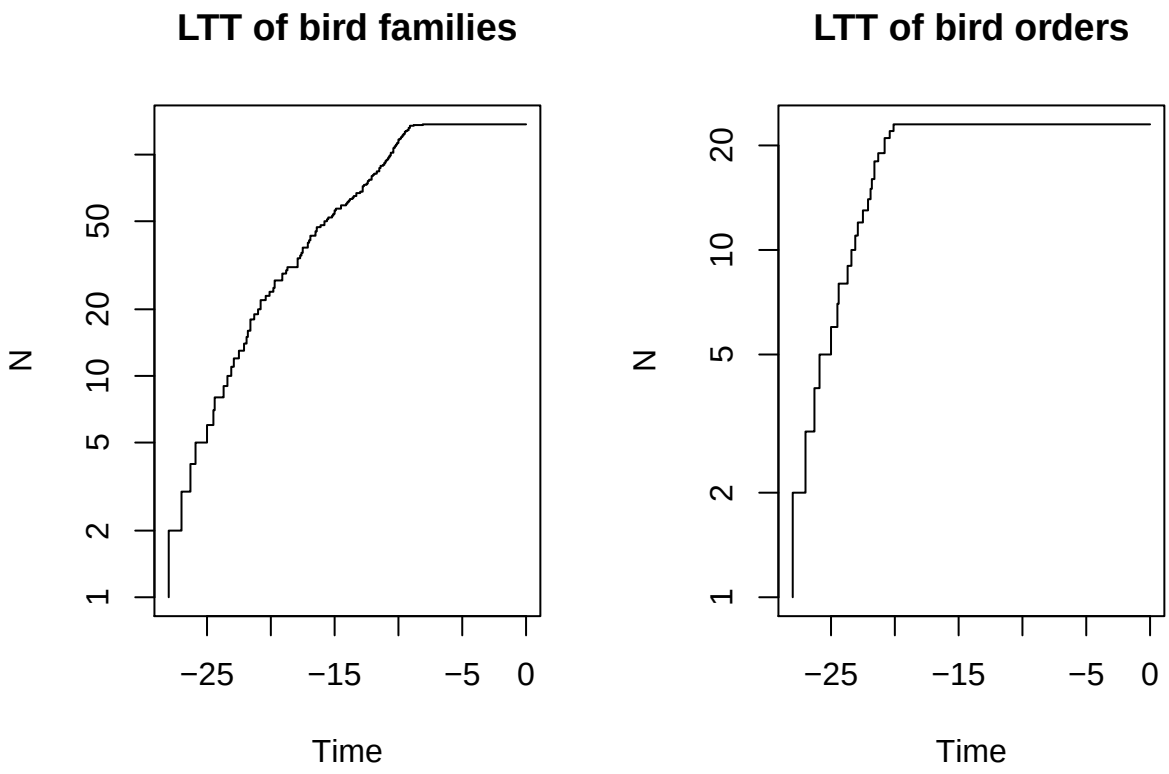
We can import in family and order data and make LTTs too:

```
data("bird.families")
data("bird.orders")
class(bird.families); class(bird.orders)
```

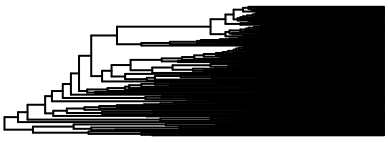
```
## [1] "phylo"
```

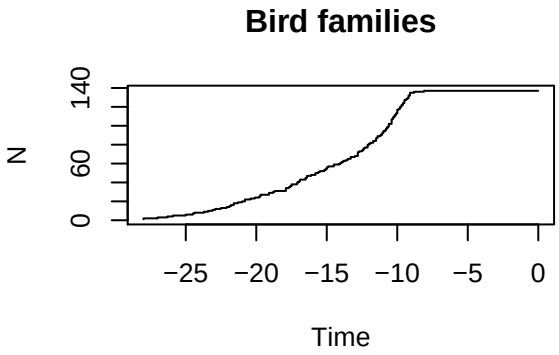
```
## [1] "phylo"
```

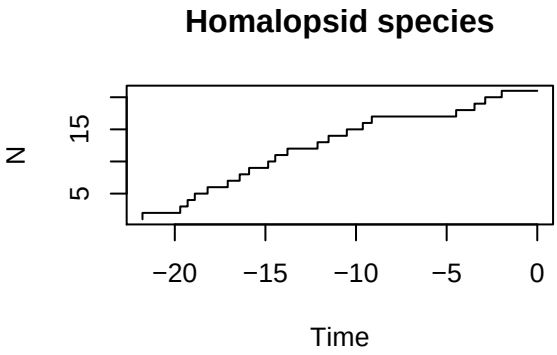
```
par(mfrow = c(1,2))
ltt.plot(bird.families, log = "y", main = "LTT of bird families")
ltt.plot(bird.orders, log = "y", main = "LTT of bird orders")
```



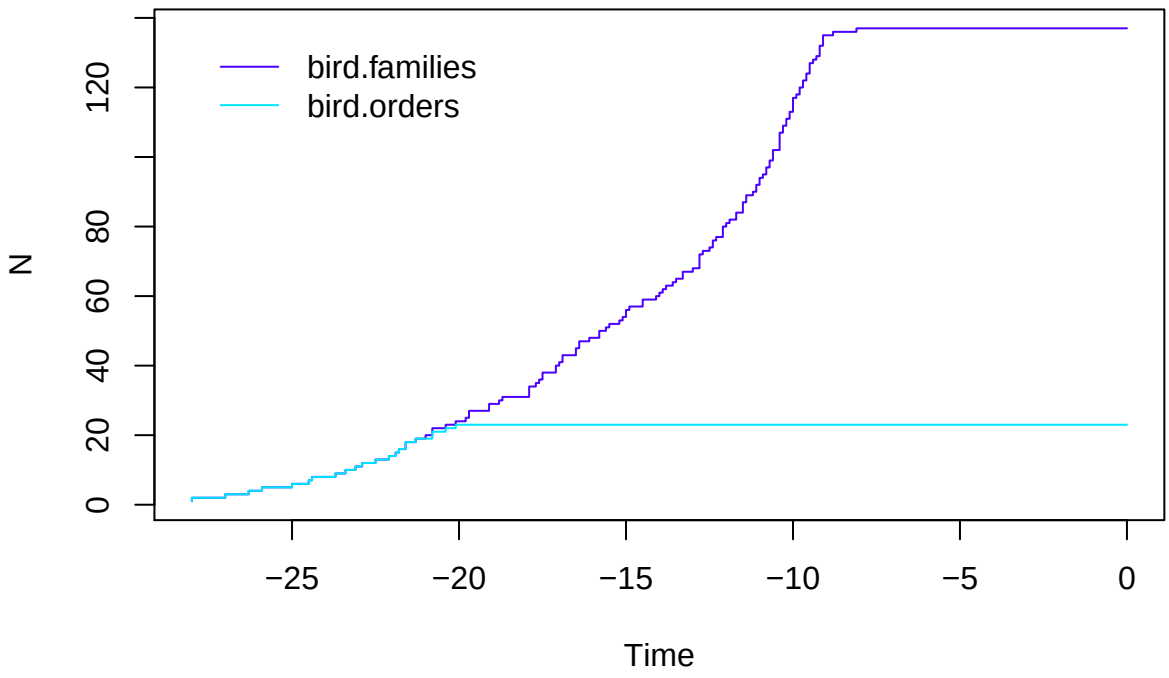

```
layout(matrix(1:4, 2, 2))
plot(bird.families, show.tip.label = FALSE)
ltt.plot(bird.families, main = "Bird families")
plot(snake_tree, show.tip.label = FALSE)
ltt.plot(snake_tree, main = "Homalopsid species")
```








```
# multiple LTTs in one plot
par(mfrow = c(1,1))
mltt.plot(bird.families, bird.orders)
```

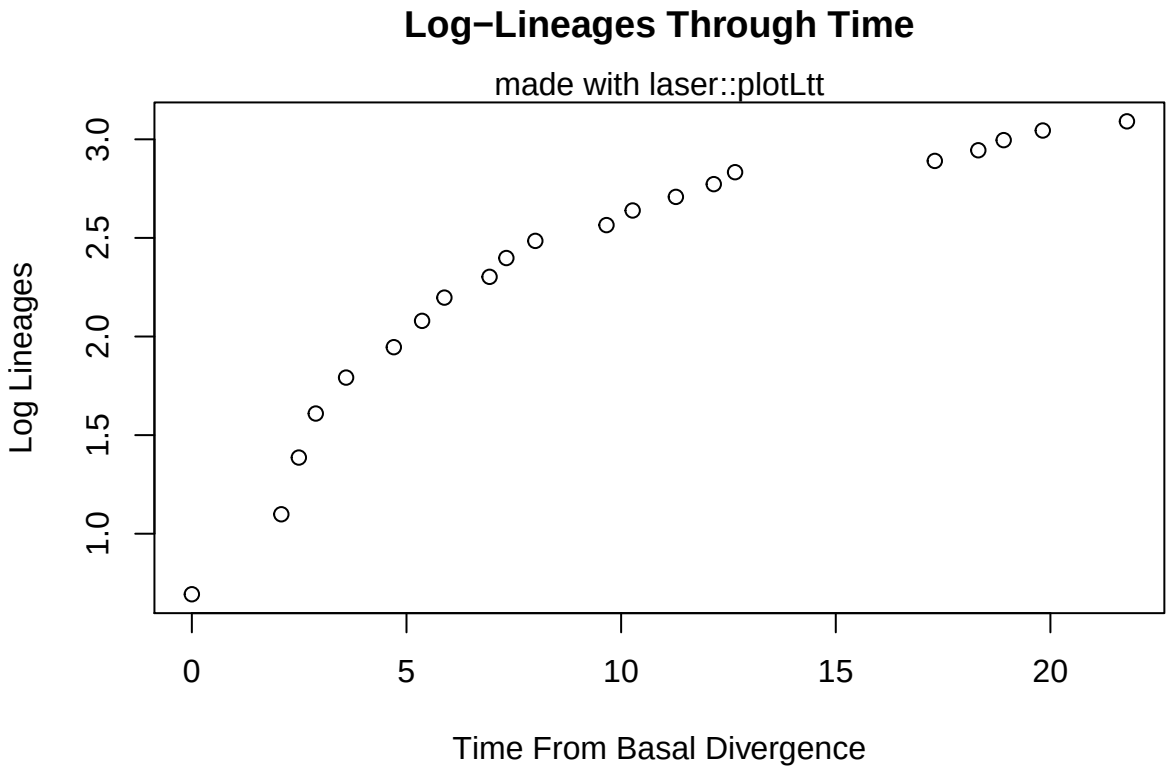


We can also do this in `laser`

```
library(laser)
snake_times <- getBtimes("~/grad/courses/UCLA/eeb_200a_evolution/alfaro_docs/assignments/tree_sim/alfar
laser::plotLtt(snake_times)
```

```
##  [1] 0.6931472 1.0986123 1.3862944 1.6094379 1.7917595 1.9459101 2.0794415
##  [8] 2.1972246 2.3025851 2.3978953 2.4849066 2.5649494 2.6390573 2.7080502
## [15] 2.7725887 2.8332133 2.8903718 2.9444390 2.9957323 3.0445224 3.0910425
```

```
mtext(side = 3, text = "made with laser::plotLtt")
```




```
# Alt we can use the tree we already have in memory
# Commenting this out to keep it clean- but the next two lines are
# equivalent to the two lines above this
# snake_times2<-getBtimes(string = write.tree(snake_tree))
# laser::plotLtt(snake_times2)
```

The Gamma Statistic

Gamma is a statistic calculated from a phylogeny with branch lengths that describes the distribution of waiting times (splitting events) on the tree. **Trees generated under a pure birth model are expected to have a gamma value of 0**. If the gamma value of an empirical phylogeny is very different from 0, this indicates that the distribution of waiting times from the tree is *unlikely to have resulted from a constant-rate pure-birth process*. Negative gamma values indicate that the waiting times on the tree are more concentrated towards the root then expected under a pure birth model while positive gamma values indicate that waiting times are more concentrated towards the tips. Typically, *negative gamma values are interpreted as evidence that the rate of diversification in a tree was fastest early in the history of the clade* and has slowed through time. Since a slowing of diversification rate through time is predicted by several macroevolutionary scenarios of key innovation and adaptive radiation, *negative gamma values are often interpreted as evidence supporting the adaptive radiation of a clade*. Positive gamma values are generally not considered to be strong evidence for increasing diversification towards the present because it is difficult to disentangle the confounding influences of increasing speciation rate, decreasing extinction rate, and the 'pull of the present' (a tendency to underestimate extinction rates in younger taxa because they have not had enough time to go extinct).

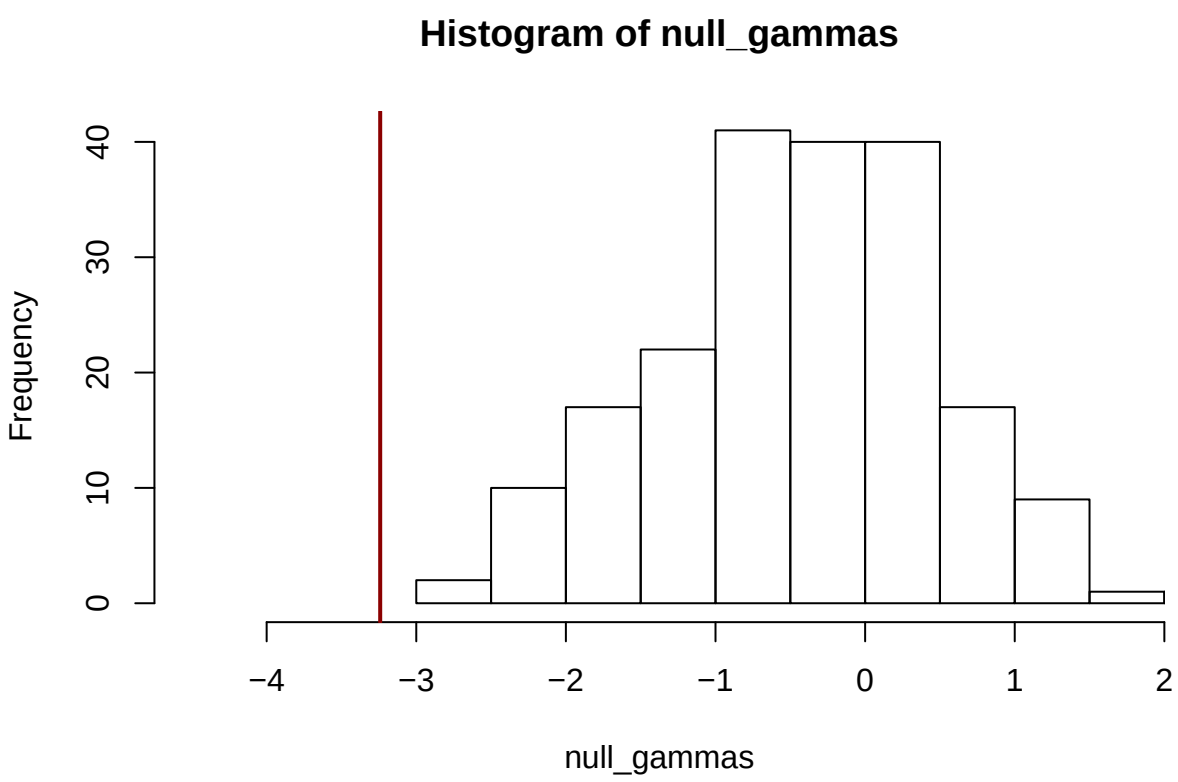
```
snake_gamma <- gammaStat(snake_tree)

# But we need to do MCCR adjustments
# mccrTest(CladeSize, NumberMissing, NumberOfReps, ObservedGamma = NULL, fname=NULL)

snake_mccr <- mccrTest(34, 13, 100, ObservedGamma = snake_gamma)
```

We can do a null model test to check whether the calculated gamma stat is significantly different that what we would expect accounting for the missing taxa.

```
age <- 22
richness <- 34
snakebirth <- log(richness)/age; snakebirth # this is to be lambda
```

```
## [1] 0.1602891
```

```
reps <- 200
null_gammas <- numeric(reps)
for (i in 1:reps){

  sim_tree <- sim.bdtree(snakebirth, d=0, stop = "taxa", n=34)
  pruned <- drop.random(sim_tree, 13) # prune down to the # of taxa in the phylogeny
  null_gammas[i] <- gammaStat(pruned)
}

hist(null_gammas, xlim = c(-4.5,1.75))
abline(v = snake_gamma, lwd = 2, col = "darkred")
```




```
# Calculate a p-value
p_snake_gamma <- (sum(null_gammas <= snake_gamma)+1)/(reps+1)
p_snake_gamma
```

```
## [1] 0.004975124
```

This MCCR test allowed us to reject a model of constant diversification through time in favor of one where early rates where faster than later rates. However this exercise did not attempt to distinguish between different models of diversification.

Fitting diversification models to branching time distributions

Next we will fit a series of models to these branching times to see if any are especially better than the others. The density dependent models are both consistent with the adaptive radiation hypotheses as they model the rate of speciation as a function of the number of species.

```
pb: pure birth
bd: birth-death
DDL: density dependent, logistic
DDX: density dependent, exponential
SPVAR: exponentially declining speciation, constant extinction rate
EXVAR: constant speciation, exponentially increasing extinction rate
BOTHVAR: variable speciation and extinction
```

For each model we will calculate the AIC score and save it to a table. Once all models have been evaluated, we will calculate deltaAIC scores and see if the best model is substantially better than any other models in our pool.

```
# Recall the branching times object
str(snake_times)
```

```
##  num [1:20] 21.8 19.7 19.3 18.9 18.2 ...
```

```
age <- 22
richness <- 34
snakebirth <- log(richness)/age; snakebirth # this is to be lambda
```

```
## [1] 0.1602891
```

```
# We will be the seven models described above
# I will store them in a list:
diversification_models_snakes <- list()

diversification_models_snakes[["pureBirth"]] <- pureBirth(snake_times)
diversification_models_snakes[["birthDeath"]] <- bd(snake_times)
diversification_models_snakes[["ddl"]] <- DDL(snake_times)
diversification_models_snakes[["ddx"]] <- DDX(snake_times)
# the MCCR test above cannot discriminate between early rapid speciation and early slow extinction.
# these models allow those to scenarios to be compared:
diversification_models_snakes[["spvar"]] <- fitSPVAR(snake_times, init = c(2, .2, .1))
diversification_models_snakes[["exvar"]] <- fitEXVAR(snake_times, init = c(.3, .01, 1))
diversification_models_snakes[["bothvar"]] <- fitBOTHVAR(snake_times, init = c(.3, .5, .1, .5))

aics <- lapply(diversification_models_snakes, function(x) as.numeric(x$aic))

sapply(aics, function(x) x-(min(as.numeric(aics))))
```

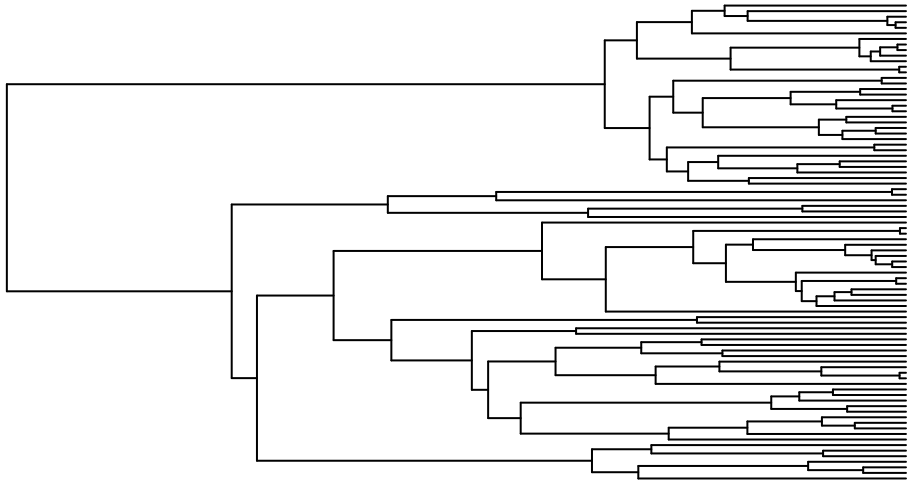
```
##  pureBirth birthDeath        ddl        ddx      spvar      exvar
##  12.446090  14.446090   0.000000   1.759280   4.500428  16.529781
##    bothvar
##   6.570177
```

What would you conclude from this table? How is this different from what you can say after the MCCR test?
The dAIC scores above suggest that the ddl or ddx models (logistic or exponential density dependence) models best explain the data. This suggests that the clade underwent an early adaptive radiation- but we should be wary of drawing this conclusion from molecular data alone and should supplement this dataset with morphological, ecological, and physiological data from the snakes. Adaptive radiation only occurs when species are clearly moving towards unfilled fitness optima, which is not demonstrated by this data set.

Exercise 6

Calculate the gamma statistic for the balistoid tree and comment on whether this clade shows evidence for rates that have slowed through time. Fit the laser models to this tree and explain whether balistoid diversification is consistent with density dependent diversification.

```
bali_tree <- read.tree("~/grad/courses/UCLA/eeb_200a_evolution/alfaro_docs/assignments/tree_sim/alfaro_
bali_times <- getBtimes("~/grad/courses/UCLA/eeb_200a_evolution/alfaro_docs/assignments/tree_sim/alfaro_

plot(bali_tree, show.tip.label = F)
```



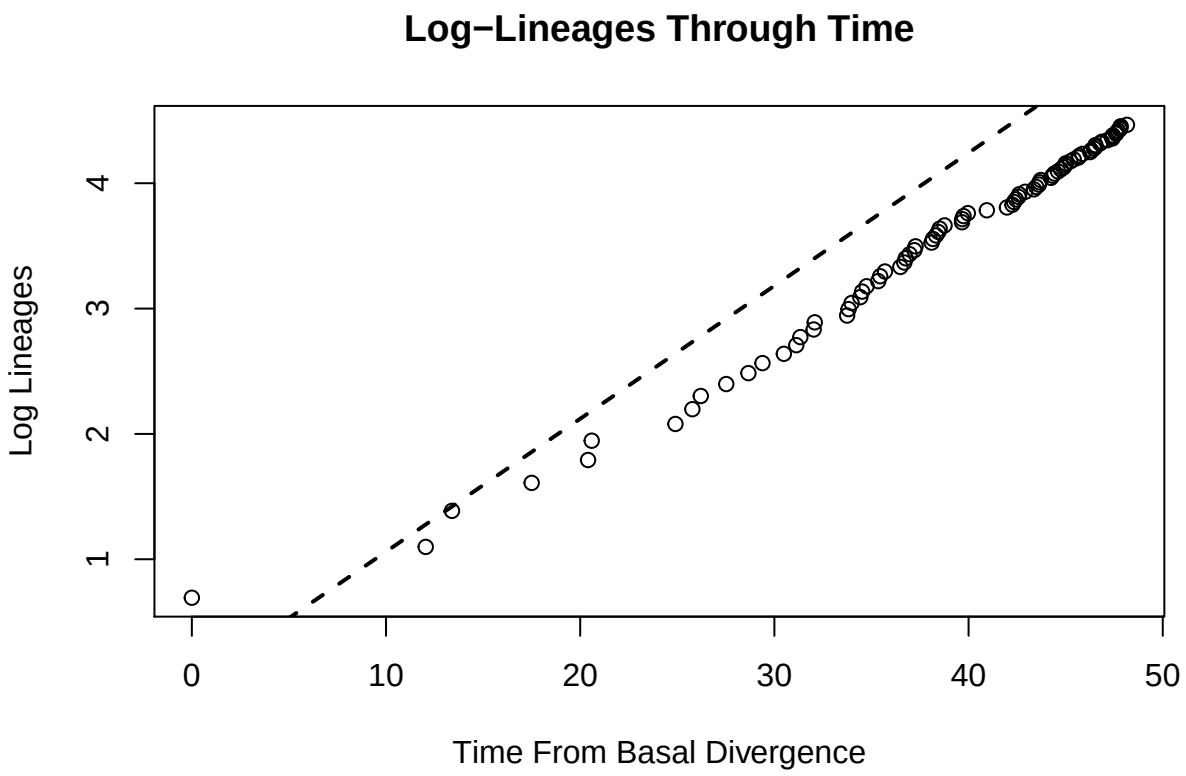
```
plotLtt(bali_times)
```

```
##  [1] 0.6931472 1.0986123 1.3862944 1.6094379 1.7917595 1.9459101 2.0794415
##  [8] 2.1972246 2.3025851 2.3978953 2.4849066 2.5649494 2.6390573 2.7080502
## [15] 2.7725887 2.8332133 2.8903718 2.9444390 2.9957323 3.0445224 3.0910425
## [22] 3.1354942 3.1780538 3.2188758 3.2580965 3.2958369 3.3322045 3.3672958
## [29] 3.4011974 3.4339872 3.4657359 3.4965076 3.5263605 3.5553481 3.5835189
## [36] 3.6109179 3.6375862 3.6635616 3.6888795 3.7135721 3.7376696 3.7612001
## [43] 3.7841896 3.8066625 3.8286414 3.8501476 3.8712010 3.8918203 3.9120230
## [50] 3.9318256 3.9512437 3.9702919 3.9889840 4.0073332 4.0253517 4.0430513
## [57] 4.0604430 4.0775374 4.0943446 4.1108739 4.1271344 4.1431347 4.1588831
## [64] 4.1743873 4.1896547 4.2046926 4.2195077 4.2341065 4.2484952 4.2626799
## [71] 4.2766661 4.2904594 4.3040651 4.3174881 4.3307333 4.3438054 4.3567088
## [78] 4.3694479 4.3820266 4.3944492 4.4067192 4.4188406 4.4308168 4.4426513
## [85] 4.4543473 4.4659081
```

```
age <- 42 # approximate age? from McCord & Westneat 2015
richness <- length(bali_tree$tip.label)
balibirth <- log(richness)/age; balibirth # this is to be lambda
```

```
## [1] 0.1060559
```

```
abline(a = c(0, balibirth), lwd = 2, lty = 2) # check how well pure birth fits this
```




```
diversification_models_bali <- list()

diversification_models_bali[["pureBirth"]] <- pureBirth(bali_times)
diversification_models_bali[["birthDeath"]] <- bd(bali_times)
diversification_models_bali[["ddl"]] <- DDL(bali_times)
diversification_models_bali[["ddx"]] <- DDX(bali_times)
# the MCCR test above cannot discriminate between early rapid speciation and early slow extinction.
# these models allow those to scenarios to be compared:
diversification_models_bali[["spvar"]] <- fitSPVAR(bali_times, init = c(2, .2, .1))
diversification_models_bali[["exvar"]] <- fitEXVAR(bali_times, init = c(.3, .01, 1))
diversification_models_bali[["bothvar"]] <- fitBOTHVAR(bali_times, init = c(.3, .5, .1, .5))

aics_bali <- lapply(diversification_models_bali, function(x) as.numeric(x$aic))

aics_bali
```

```
## $pureBirth
## [1] -28.21652
##
## $birthDeath
## [1] -26.37007
```

```
##
## $ddl
## [1] -26.21649
##
## $ddx
## [1] -26.39073
##
## $spvar
## [1] -24.28635
##
## $exvar
## [1] -24.36991
##
## $bothvar
## [1] -22.28633
```

```
sapply(aics_bali, function(x) x-(min(as.numeric(aics_bali))))
```

```
##  pureBirth birthDeath        ddl        ddx      spvar      exvar
##   0.000000   1.846455   2.000033   1.825791   3.930172   3.846618
##    bothvar
##   5.930197
```

The lowest AIC here is of the `pureBirth` model, but I am hesitant to conclude from this that the lineage has experienced this sort of speciation so far. In the morpho lab, a comment suggests that "# delata AIC or AICC scores >2 are ususally considered to provide positive support"- here, there are many models with dAIC < 2. So, pureBirth, birthDeath, and diversity-dependence all seem to be potentially viable explanations for diversification in this lineage.

```
library(plyr); library(ggplot2); library(apTreeshape)
```

```
##
## Attaching package: 'apTreeshape'
##
## The following object is masked _by_ '.GlobalEnv':
##
##     colless
```

```
numtrees <- 1000
trees <- pbtree(n = 50, nsim = numtrees, ape = F)

foo <- function(tree, metric = "colless") {
  if (metric == "colless") {
    #xx <- as.treeshape(x)  # convert to apTreeshape format
    colless(tree)
    #colless(xx, norm = "yule")  # calculate colless' metric
  } else if (metric == "gamma") {
    gammaStat(tree)
  } else stop("metric should be one of colless or gamma")
}

theme_myblank <- function() {
```

```
  stopifnot(require(ggplot2))
  theme_blank <- ggplot2::theme_blank
  ggplot2::theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
                 panel.background = element_blank(), plot.background = element_blank(),
                 axis.title.x = element_text(colour = NA), axis.title.y = element_blank(),
                 axis.text.x = element_blank(), axis.text.y = element_blank(), axis.line = element_blank
                 axis.ticks = element_blank())
}
```