Computational Phylogenetics, Spring 2017

Final Assignment
Anna E. Hiller
May 5th 2017

Divergence Time Estimation using *Diglossa* (Aves: Thraupidae) Informative Prior on the Root Age and Universal Molecular Clock

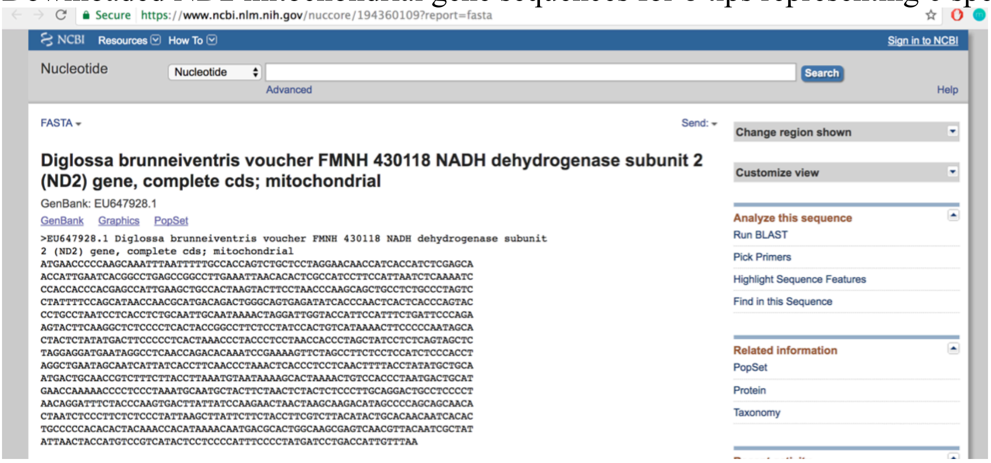
I chose to run a divergence time analysis to date a tree because I want to apply phylogenetic methods to phylogeographic analyses. One of the key components of phylogeography is understanding *when* speciation events occur (e.g., splits between taxa). Because there are often no fossils of recent taxa in birds, especially Passerines, molecular clock estimates are used instead. I choose a group I am interested in working on (the *Diglossa* tanagers). In the future I want to use this same approach to estimate phylogeographic divergence times and then compare to estimates of divergence times obtained via estimation of population genetic parameters.

Overview of Dating

Given a phylogenetic tree with branch lengths (rate*time) and a calibration to constrain the node (either a fossil calibration, or a node from a previous phylogeny, can also use tips), you can estimate divergence times for other nodes (e.g., you can date speciation events).

Obtained Data

Downloaded ND2 mitochondrial gene sequences for 8 tips representing 6 species.



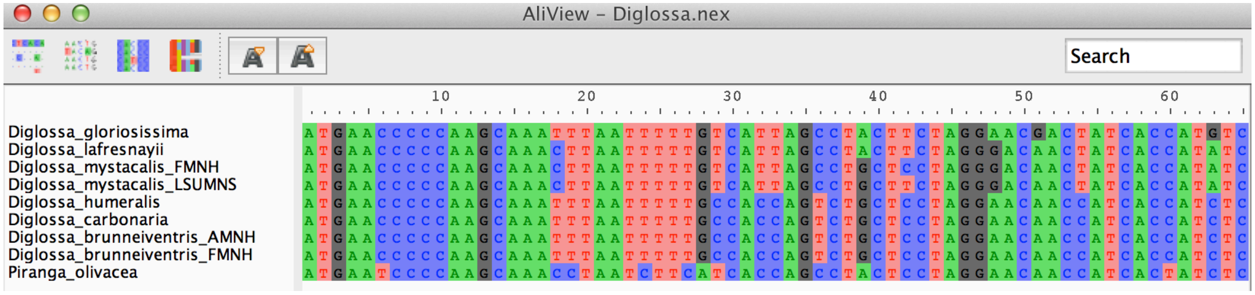
Diglossa brunneiventris voucher FMNH 430118 NADH dehydrogenase subunit 2 (ND2) gene, complete cds; mitochondrial

GenBank: EU647928.1

GenBank Graphics PopSet

```
>EU647928.1 Diglossa brunneiventris voucher FMNH 430118 NADH dehydrogenase subunit
2 (ND2) gene, complete cds; mitochondrial
ATGAACCCCCAAGCAAATTTAATTTTTGCCACCAGTCTGCTCCTAGGAACAACCATCACCATCTCGAGCA
ACCATTGAATCACGGCCTGAGCCGGCCTTGAAATTAACACACTCGCCATCCTTCCATTAATCTCAAAATC
CCACCACCCACGAGCCATTGAAGCTGCCACTAAGTACTTCCTAACCCAAGCAGCTGCCTCTGCCCTAGTC
CTATTTTCCAGCATAACCAACGCATGACAGACTGGGCAGTGAGATATCACCCAACTCACTCACCCAGTAC
CCTGCCTAATCCTCACCTCTGCAATTGCAATAAAACTAGGATTGGTACCATTCCATTTCTGATTCCCAGA
AGTACTTCAAGGCTCTCCCCTCACTACCGGCCTTCTCCTATCCACTGTCATAAAACTTCCCCCAATAGCA
CTACTCTATATGACTTCCCCCTCACTAAACCCTACCCTCCTAACCACCCTAGCTATCCTCTCAGTAGCTC
TAGGAGGATGAATAGGCCTCAACCAGACACAAATCCGAAAAGTTCTAGCCTTCTCCTCCATCTCCCACCT
AGGCTGAATAGCAATCATTATCACCTTCAACCCTAAACTCACCCTCCTCAACTTTTACCTATATGCTGCA
ATGACTGCAACCGTCTTTCTTACCTTAAATGTAATAAAAGCACTAAAACTGTCCACCCTAATGACTGCAT
GAACCAAAAACCCCTCCCTAAATGCAATGCTACTTCTAACTCTACTCTCCCTTGCAGGACTGCCTCCCCT
AACAGGATTTCTACCCAAGTGACTTATTATCCAAGAACTAACTAAGCAAGACATAGCCCCAGCAGCAACA
CTAATCTCCCTTCTCTCCCTATTAAGCTTATTCTTCTACCTTCGTCTTACATACTGCACAACAATCACAC
TGCCCCCACACACTACAAACCACATAAAACAATGACGCACTGGCAAGCGAGTCAACGTTACAATCGCTAT
ATTAACTACCATGTCCGTCATACTCCTCCCCATTTCCCCTATGATCCTGACCATTGTTTAA
```

Edited the nexus file and visualized in AliView.

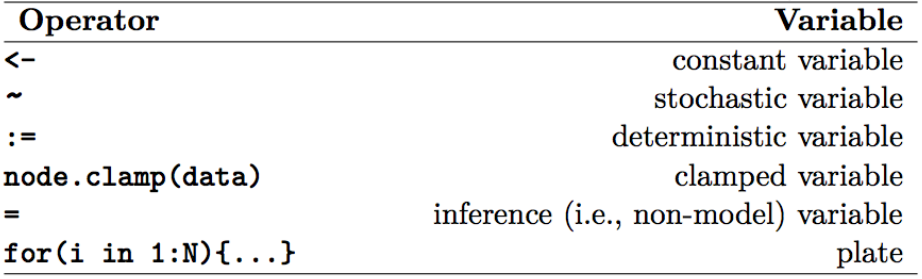


```
                              10        20        30        40        50        60
Diglossa_gloriosissima        ATGAACCCCCAAGCAAATTTAATTTTTGTCATTAGCCTACTTCTAGGAACGACTATCACCATGTC
Diglossa_lafresnayii          ATGAACCCCCAAGCAAACTTAATTTTTGTCATTAGCCTACTTCTAGGGACAACTATCACCATATC
Diglossa_mystacalis_FMNH      ATGAACCCCCAAGCAAACTTAATTTTTGTCATTAGCCTGCTCCTAGGGACAACTATCACCATATC
Diglossa_mystacalis_LSUMNS    ATGAACCCCCAAGCAAACTTAATTTTTGTCATTAGCCTGCTTCTAGGGACAACTATCACCATATC
Diglossa_humeralis            ATGAACCCCCAAGCAAATTTAATTTTTGCCACCAGTCTGCTCCTAGGAACAACCATCACCATCTC
Diglossa_carbonaria           ATGAACCCCCAAGCAAATTTAATTTTTGCCACCAGTCTGCTCCTAGGAACAACCATCACCATCTC
Diglossa_brunneiventris_AMNH  ATGAACCCCCAAGCAAATTTAATTTTTGCCACCAGTCTGCTCCTAGGAACAACCATCACCATCTC
Diglossa_brunneiventris_FMNH  ATGAACCCCCAAGCAAATTTAATTTTTGCCACCAGTCTGCTCCTAGGAACAACCATCACCATCTC
Piranga_olivacea              ATGAATCCCCAAGCAAACCTAATCTTCATCACCAGCCTACTCCTAGGAACAACCATCACTATCTC
```

Overview of Key Code

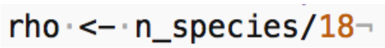
Types of functions in RevBayes

Table 1: Rev assignment operators, clamp function, and plate/loop syntax.



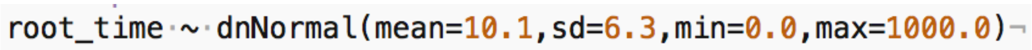
Operator	Variable
`<-`	constant variable
`~`	stochastic variable
`:=`	deterministic variable
`node.clamp(data)`	clamped variable
`=`	inference (i.e., non-model) variable
`for(i in 1:N){...}`	plate

I used a GTP+G model of sequence evolution and the default settings for the model of constant-rate birth-death processes.



```
rho <- n_species/18
```

Rho represents nTaxa / totalTaxa, the probability of sampling species at the present. For this dataset I changed the value to be 6 / 18, since there are 18 described *Diglossa* species. I am curious how altering this value effects the results though, because Diglossa is a highly polymorphic group with many subspecies that could potentially be elevated to species rank.

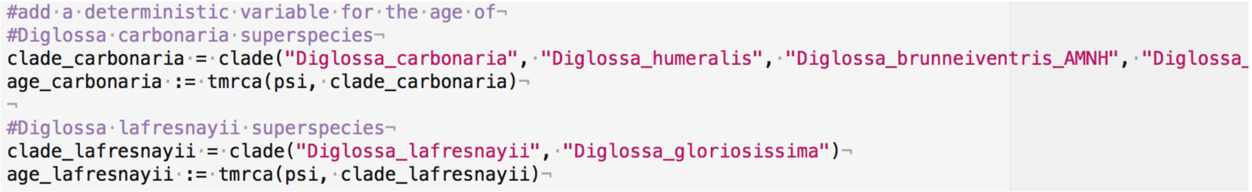


```
root_time ~ dnNormal(mean=10.1,sd=6.3,min=0.0,max=1000.0)
```

Root time is the informative prior used to condition the root age and inform our dating. I used the estimated date for the split between the *Diglossa lafresnayii* and *Diglossa carbonaria* species complexes as the prior, taken from:

> Mauck III, W. M., & Burns, K. J. 2009. Phylogeny, biogeography, and recurrent evolution of divergent bill types in the nectar-stealing flowerpiercers (Thraupini: Diglossa and Diglossopis). Biological Journal of the Linnean Society. 98:14-28.

I then back-calculated (see code) the standard deviation based on the number of tips and 95% confidence interval given in the paper.



```
#add a deterministic variable for the age of
#Diglossa carbonaria superspecies
clade_carbonaria = clade("Diglossa_carbonaria", "Diglossa_humeralis", "Diglossa_brunneiventris_AMNH", "Diglossa_
age_carbonaria := tmrca(psi, clade_carbonaria)

#Diglossa lafresnayii superspecies
clade_lafresnayii = clade("Diglossa_lafresnayii", "Diglossa_gloriosissima")
age_lafresnayii := tmrca(psi, clade_lafresnayii)
```

Also, I added in deterministic variables for certain ages I was interested in.

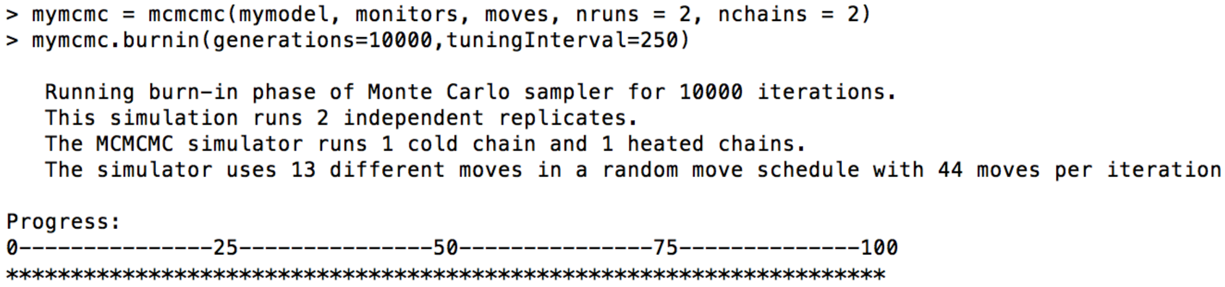
This analysis used a global molecular clock rate, which assumes a constant rate of substitution across the tree (vs. a relaxed molecular clock which allows for rate variation across lineages). Because the analysis was based on an informative prior, the clock rate is estimated from the data.

I then created a model of sequence evolution, and attach the sequence data to the tip nodes using the .clamp function (used for observed data).

Executed the Analysis

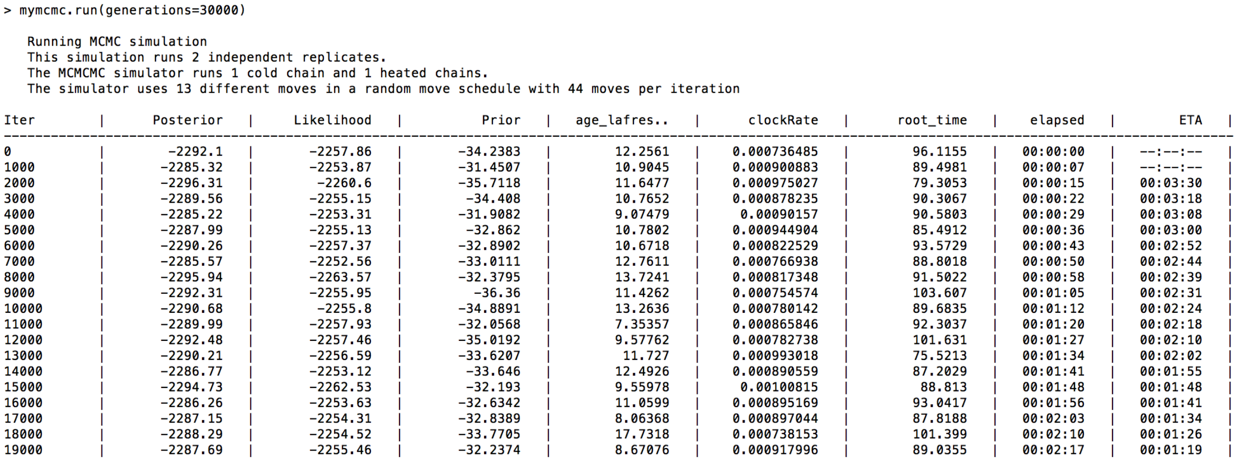
I set up monitors (mni), which record the states of the Markov Chain, created the MCMC object, and ran the analysis. Note that I used 2 replicate runs and two chains (1 cold, 1 heated) to make sure I got good estimates.

Burnin



```
> mymcmc = mcmcmc(mymodel, monitors, moves, nruns = 2, nchains = 2)
> mymcmc.burnin(generations=10000,tuningInterval=250)

   Running burn-in phase of Monte Carlo sampler for 10000 iterations.
   This simulation runs 2 independent replicates.
   The MCMCMC simulator runs 1 cold chain and 1 heated chains.
   The simulator uses 13 different moves in a random move schedule with 44 moves per iteration

Progress:
0---------------25---------------50---------------75--------------100
*******************************************************************
```

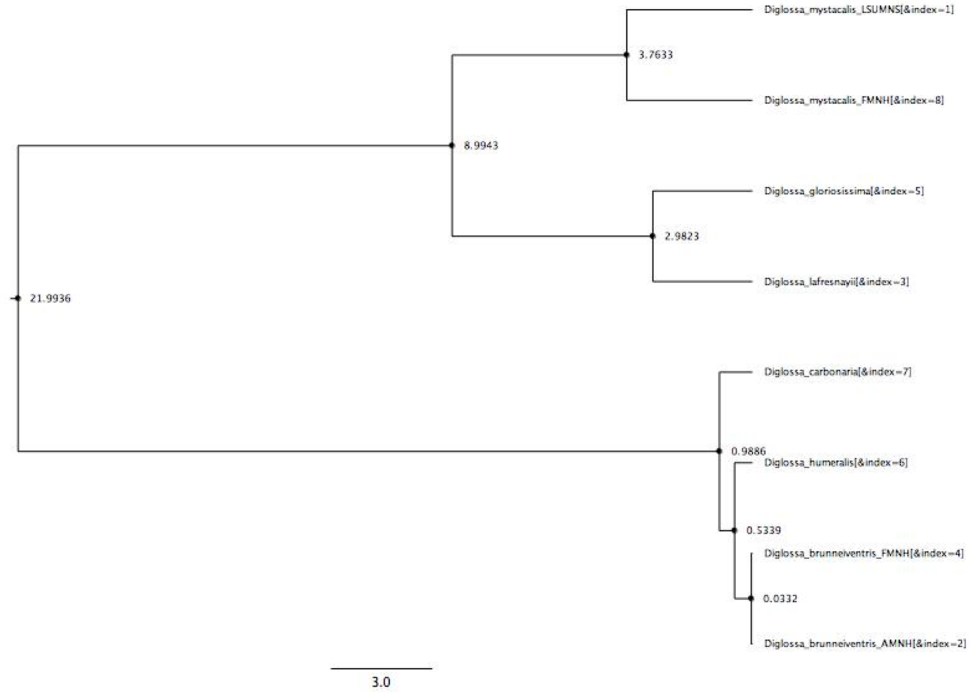
MCMC



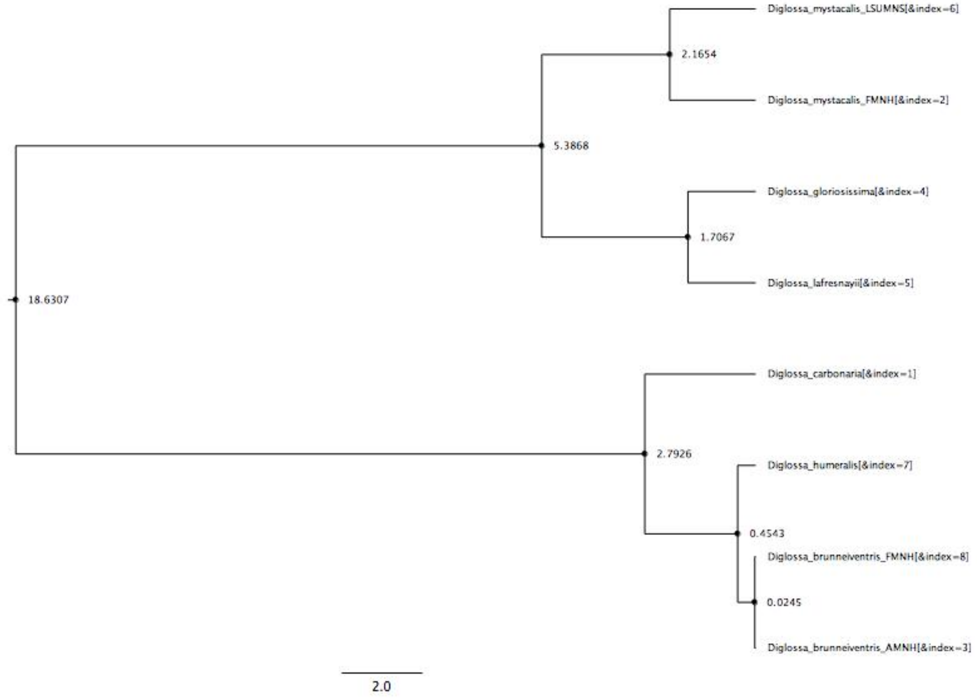
```
> mymcmc.run(generations=30000)

   Running MCMC simulation
   This simulation runs 2 independent replicates.
   The MCMCMC simulator runs 1 cold chain and 1 heated chains.
   The simulator uses 13 different moves in a random move schedule with 44 moves per iteration

Iter        |      Posterior   |     Likelihood   |          Prior   |   age_lafres..   |      clockRate   |      root_time   |    elapsed   |        ETA   |
-------------------------------------------------------------------------------------------------------------------------------------------------------------
0           |        -2292.1   |       -2257.86   |       -34.2383   |        12.2561   |    0.000736485   |        96.1155   |   00:00:00   |   --:--:--   |
1000        |       -2285.32   |       -2253.87   |       -31.4507   |        10.9045   |    0.000900883   |        89.4981   |   00:00:07   |   --:--:--   |
2000        |       -2296.31   |        -2260.6   |       -35.7118   |        11.6477   |    0.000975027   |        79.3053   |   00:00:15   |   00:03:30   |
3000        |       -2289.56   |       -2255.15   |        -34.408   |        10.7652   |    0.000878235   |        90.3067   |   00:00:22   |   00:03:18   |
4000        |       -2285.22   |       -2253.31   |       -31.9082   |        9.07479   |     0.00090157   |        90.5803   |   00:00:29   |   00:03:08   |
5000        |       -2287.99   |       -2255.13   |        -32.862   |        10.7802   |    0.000944904   |        85.4912   |   00:00:36   |   00:03:00   |
6000        |       -2290.26   |       -2257.37   |       -32.8902   |        10.6718   |    0.000822529   |        93.5729   |   00:00:43   |   00:02:52   |
7000        |       -2285.57   |       -2252.56   |       -33.0111   |        12.7611   |    0.000766938   |        88.8018   |   00:00:50   |   00:02:44   |
8000        |       -2295.94   |       -2263.57   |       -32.3795   |        13.7241   |    0.000817348   |        91.5022   |   00:00:58   |   00:02:39   |
9000        |       -2292.31   |       -2255.95   |         -36.36   |        11.4262   |    0.000754574   |        103.607   |   00:01:05   |   00:02:31   |
10000       |       -2290.68   |        -2255.8   |       -34.8891   |        13.2636   |    0.000780142   |        89.6835   |   00:01:12   |   00:02:24   |
11000       |       -2289.99   |       -2257.93   |       -32.0568   |        7.35357   |    0.000865846   |        92.3037   |   00:01:20   |   00:02:18   |
12000       |       -2292.48   |       -2257.46   |       -35.0192   |        9.57762   |    0.000782738   |        101.631   |   00:01:27   |   00:02:10   |
13000       |       -2290.21   |       -2256.59   |       -33.6207   |         11.727   |    0.000993018   |        75.5213   |   00:01:34   |   00:02:02   |
14000       |       -2286.77   |       -2253.12   |        -33.646   |        12.4926   |    0.000890559   |        87.2029   |   00:01:41   |   00:01:55   |
15000       |       -2294.73   |       -2262.53   |        -32.193   |        9.55978   |     0.00100815   |         88.813   |   00:01:48   |   00:01:48   |
16000       |       -2286.26   |       -2253.63   |       -32.6342   |        11.0599   |    0.000895169   |        93.0417   |   00:01:56   |   00:01:41   |
17000       |       -2287.15   |       -2254.31   |       -32.8389   |        8.06368   |    0.000897044   |        87.8188   |   00:02:03   |   00:01:34   |
18000       |       -2288.29   |       -2254.52   |       -33.7705   |        17.7318   |    0.000738153   |        101.399   |   00:02:10   |   00:01:26   |
19000       |       -2287.69   |       -2255.46   |       -32.2374   |        8.67076   |    0.000917996   |        89.0355   |   00:02:17   |   00:01:19   |
```

Outputs

Tree from Run 1

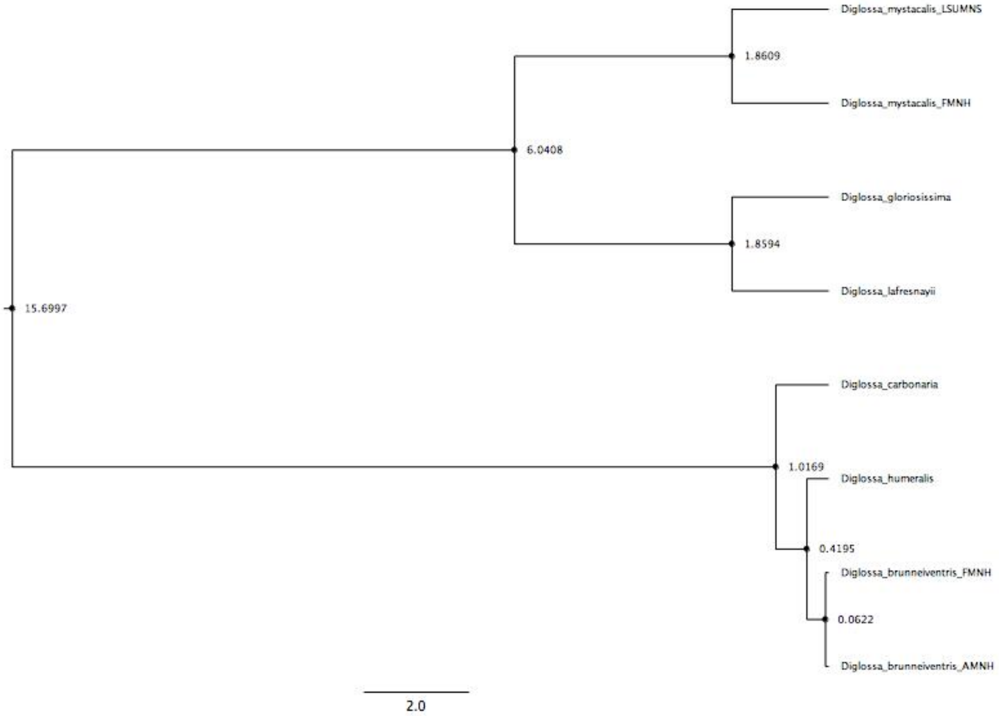

Diglossa_mystacalis_LSUMNS[&index=1]
3.7633
Diglossa_mystacalis_FMNH[&index=8]
8.9943
Diglossa_gloriosissima[&index=5]
2.9823
Diglossa_lafresnayii[&index=3]
21.9936
Diglossa_carbonaria[&index=7]
0.9886
Diglossa_humeralis[&index=6]
0.5339
Diglossa_brunneiventris_FMNH[&index=4]
0.0332
Diglossa_brunneiventris_AMNH[&index=2]
3.0

Tree from Run 2


Diglossa_mystacalis_LSUMNS[&index=6]
2.1654
Diglossa_mystacalis_FMNH[&index=2]
5.3868
Diglossa_gloriosissima[&index=4]
1.7067
Diglossa_lafresnayii[&index=5]
18.6307
Diglossa_carbonaria[&index=1]
2.7926
Diglossa_humeralis[&index=7]
0.4543
Diglossa_brunneiventris_FMNH[&index=8]
0.0245
Diglossa_brunneiventris_AMNH[&index=3]
2.0

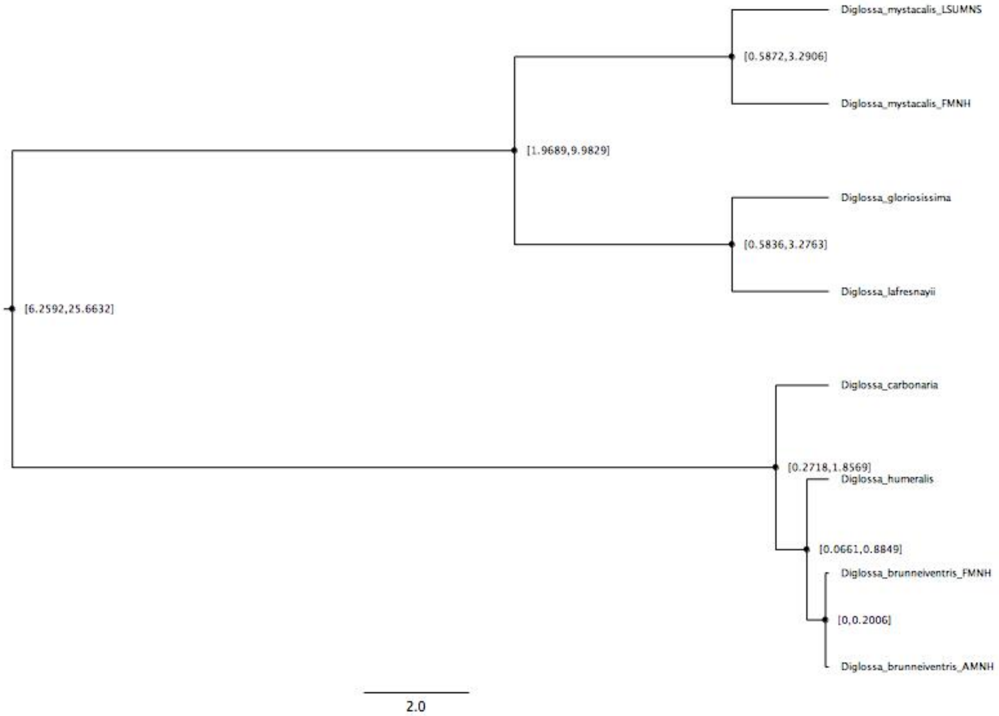
* The topologies and dates look similar across runs. Both recover the short branches in the *D. carbonaria* complex (consistent with published literature).

Consensus Tree



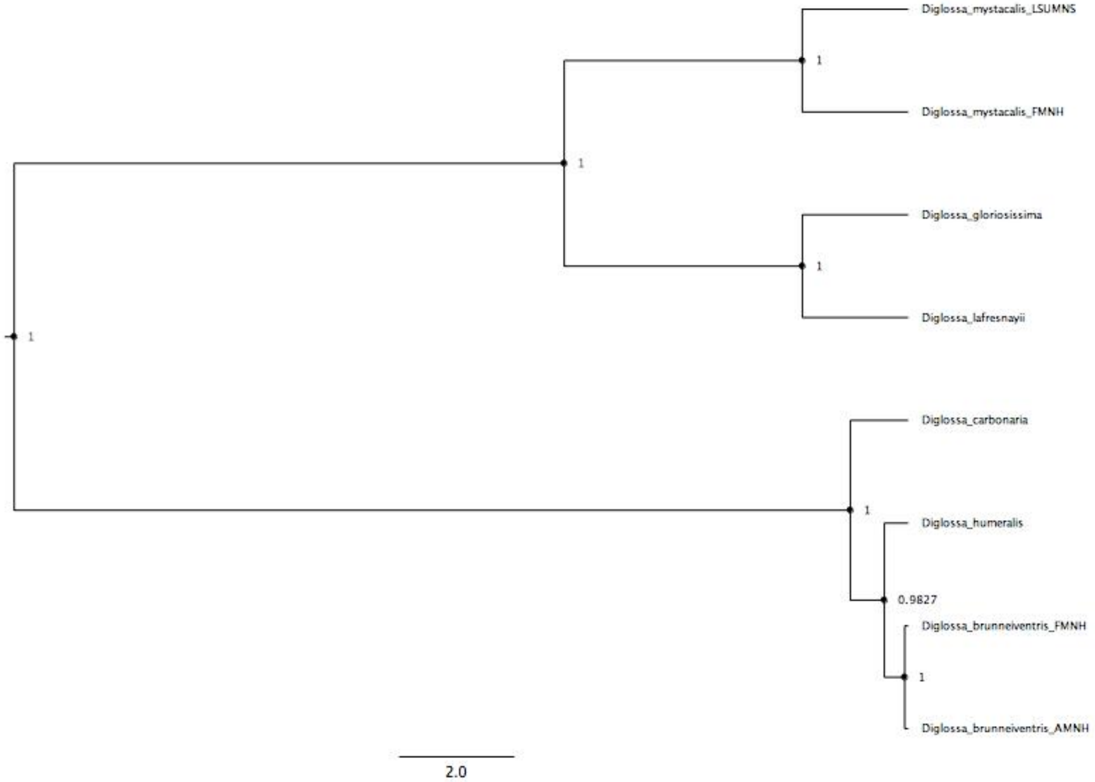
* This tree has a single consensus date on each node.

Consensus Tree with 95% Confidence Intervals



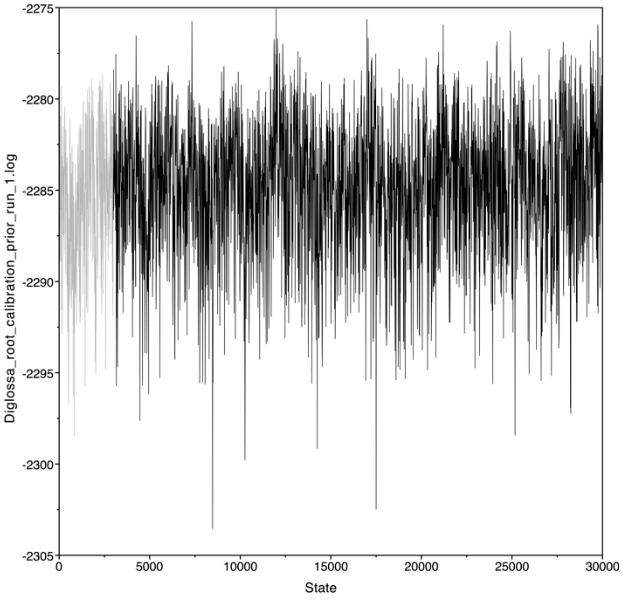
* This tree has the 95% confidence interval for the date on each node, taking uncertainty into account.

Consensus Tree with Posterior Probabilities

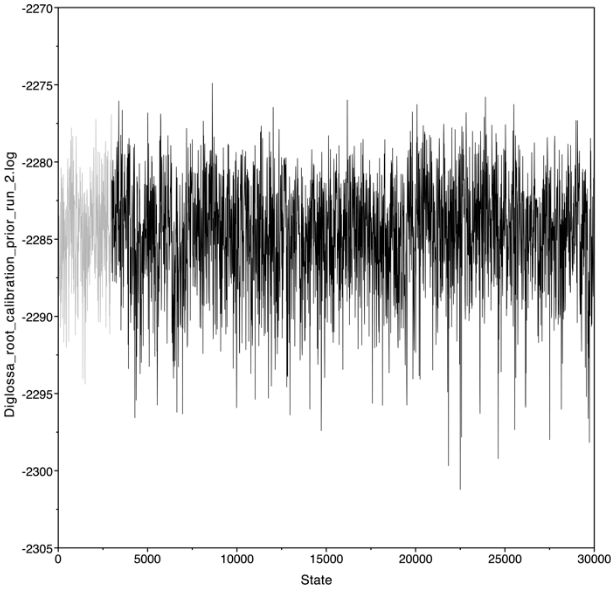


* Posterior Probabilities are close to 1, which corresponds to a high probability that the tree is correct assuming the models are accurate.

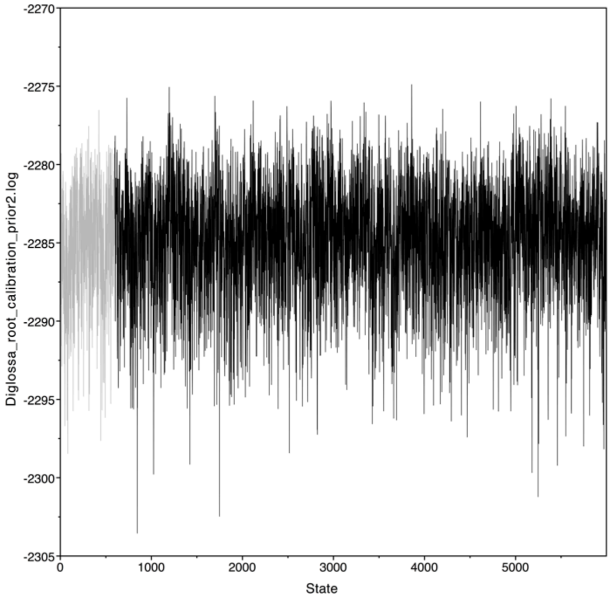
Trace Plot Run 1



Trace Plot Run 2

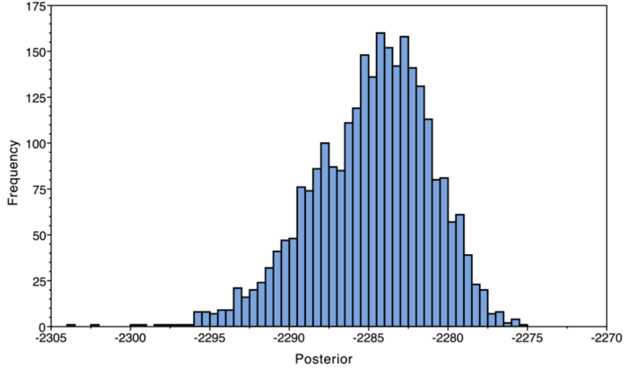


Trace Plot Consensus

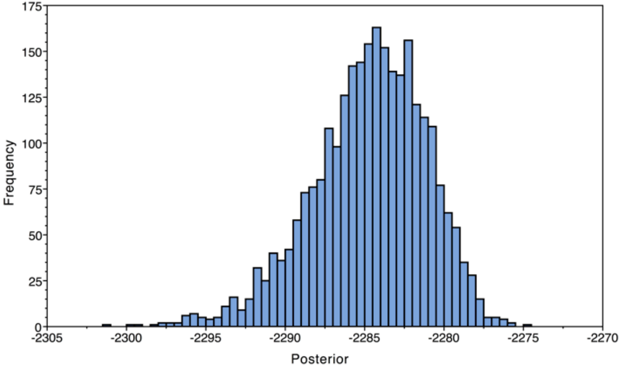


* All three trace plots are very “fuzzy” and linear, so the analysis had good mixing.

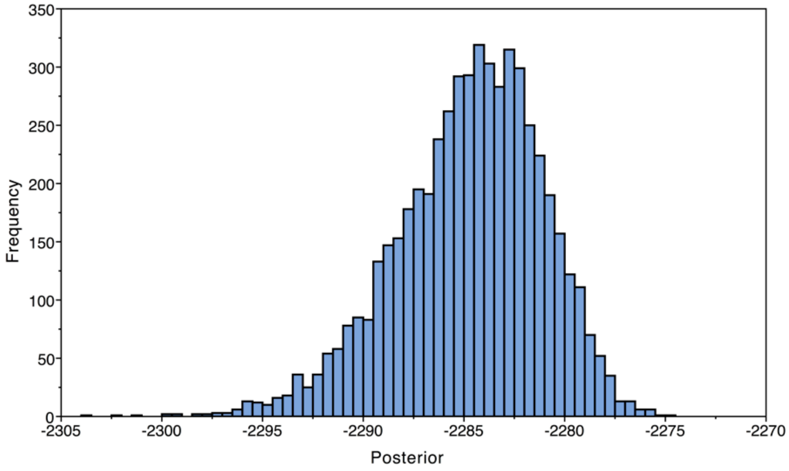
Histogram of Posterior Probabilities Run 1



Histogram of Posterior Probabilities Run 2

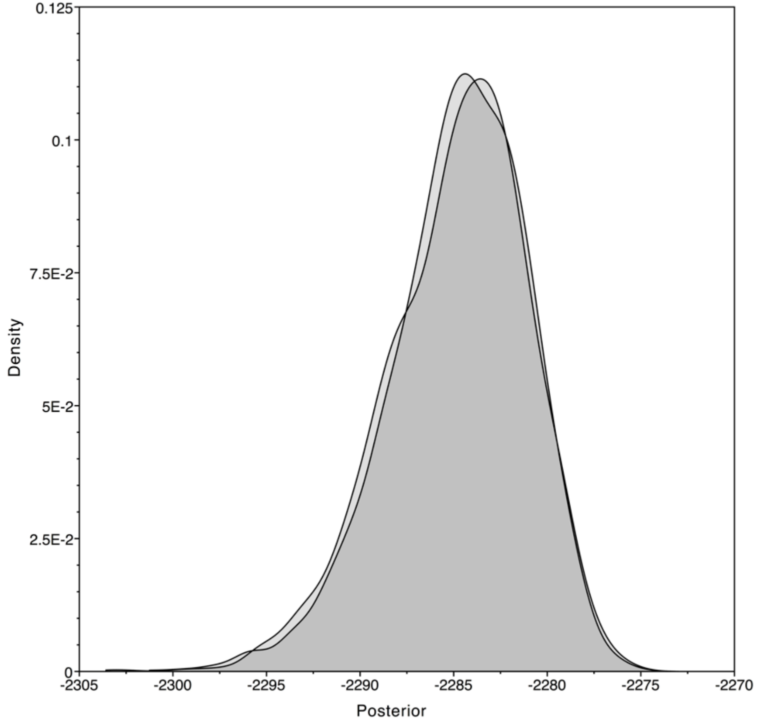


Consensus Run



* Histograms look similar across runs.

Marginal Probability Distribution Runs 1 & 2



* Probability Distributions are similar across runs, curves overlay.

Effective Sample Size (ESS) Values:
Run 1 = 306.6
Run 2 = 452.1
Consensus = 684.8

* All are above 200 so analysis had good mixing.

Glossary

Molecular Clock: a method of using the mutation rate of nucleotide sequences (DNA) or amino acids (proteins) to estimate the time since divergence. Evidence suggests that there is a linear relationship, molecular differences between species pairs are proportional to the time since they diverged. First described by:

> Zuckerkandl, E. and Pauling, L.B. (1962). "Molecular disease, evolution, and genetic heterogeneity". In Kasha, M. and Pullman, B (editors). Horizons in Biochemistry. Academic Press, New York. pp. 189–225.

Model of Sequence Evolution: Markov models of substitution, are matrices of the probability that one amino acid will 'transition' into another (e.g., A -> T). Many are described, Jukes Cantor (1969) model uses fixed values (frequencies and mutation rates) and is the simplest. GTR (Generalized Time-Reversible) uses equilibrium base frequencies (pi) and transition rate parameters (r) and is one of the most complex, but also flexible.

Birth-Death Processes: a continuous-time Markov process where states transitions are birth (increase states by 1) or death (decrease states by 1).

Prior: from "prior beliefs", parameters based on existing knowledge set *before* data are input.

Posterior: from Bayes' theorem, the prior probability of a tree P(A) combined with the likelihood of the data P(B) produce a posterior probability distribution on trees P(A|B). The posterior probability of a tree will indicate the probability of the tree being correct given the data observed and specified models.

Heated Chain: during MCMC sampling, the 'heated' chain traverses a space where the peaks and valleys have been flattened out making them easier to cross. After each iteration the cold chain traverses (unflattened peaks) then accepts or rejects a move based on the space sampled by the heated chain. Genna's Mother and baby robot analogy!

Burn-in: trees generated early in the analysis are discarded. Common method of evaluating nodal support in a Bayesian phylogenetic analysis, by calculating the percentage of trees in the posterior distribution (post-burn-in) that contain the node observed in the 'actual' tree.

Markov Chain Monte Carlo: a method or class of algorithm for sampling from a probability distribution by constructing a Markov chain that has the desired distribution as its equilibrium. The state of the chain after a number of steps is then used as a sample of the desired distribution. It is a way of approximating a distribution.

Effective Sample Size (ESS): the number of effectively independent draws from the posterior distribution that the Markov chain is equivalent to. If ESS is small then the distribution is poor, meaning there is a large standard deviation and bad mixing. ESS<100 is bad, <200 is poor, but >100 is excessive computational time (ref: BEAST documentation)

* In full disclaimer I got many of these definitions from Wikipedia, and then edited to make more phylogenetics specific. This is more for my future use than anything.