*Integrative Biology 200B* *University of California, Berkeley*
*"Ecology and Evolution"* *Spring 2011*
*by Nat Hallinan, updated by Nick Matzke*

# Lab 13: Coalescence

Today we are going to use *Mesquite* to investigate coalescence. We will simulate a coalescent process for several alleles on a population tree. We will compare these simulated allele trees to the population trees and calculate several values that indicate something about the nature of the coalescent. We will also look at the effect of generations and population size on the coalescence process. Finally we will compare an actual gene tree for a group of alleles to several population trees.

## Establishing an Association

The first thing that we are going to do is establish an association between populations and the alleles that they contain. This is a necessary step for any comparison between two groups of taxa. In this case one group of taxa is the populations and the other is the alleles found in those populations. It could just as well be a pair of mutualists, organisms and their parasites, species and duplicated gene families that they contain or even geographic areas and the species that reside in them.

Open the file **Coalesence example 1** in *Mesquite*. Open the tree. As you can see, there is not a lot of information here, just a tree with three taxa. Now we are going to create a set of alleles for the taxa in this tree. Select **Taxa&Trees > New Block of Taxa.** Call the block **Alleles** to separate it from the **Populations** block, and let's start with **7** alleles. Let's assign two alleles each to populations 1 and 2, and three alleles to population 3. Rename the alleles Allele 1a, Allele 1b, Allele 2a, etc.

To establish the association select **Taxa&Trees > New Association**. Select **Populations** as the **first block**. In general it is easier if you select your containing taxa as your first block. Give the association a creative name. The **Populations Taxa** block will appear at the front of your screen, but now it will have a column for associated alleles and an association editor on the right for creating your associations. Select **Population 1**, go to the Association editor and select the ***two alleles from population 1***, then press **the arrow** at the top of the association editor. Those two alleles should appear in the **Alleles column** next to **population 1**. Do the same for populations 2 and 3. You can use the plus and minus buttons to edit the associations, if you make any mistakes.

Save your work and open the nexus file in a text editor. You will see four blocks in addition to the *Mesquite* Block: Two taxa blocks, one for the populations and one for the genes; a tree block for the population tree; and an association block explaining the association between the populations and their alleles. This block is in the standard format for associations in Nexus files. It is not as widely accepted as other aspects of the nexus file format, so be sure to check your documentation about file formats, if you are using another program.

## Simulating Coalescence on the Population Tree

Now we are going to let *Mesquite* simulate a coalescent process for these eight genes on the population tree. This works by considering the chance of any two alleles coming together in

a single haploid population as 1/Ne per generation. It goes back generation by generation calculating the probability of each allele lineage coalescing until they have all come together, although by default it actually uses an exponential approximation of that process.

Select **Taxa&Trees > Multi Tree Window**. Select **Alleles**, because we will be making trees of Alleles on the Taxon tree. Then choose **Simulated Trees : Coalescence Contained within Current Tree**. Let's start with 10000 as our Ne. Just make 20 trees to keep the amount that you have to look at for this one lab down. Look at the trees. Use **Multi-tree > Number of columns** and **Multi-tree > Number of rows** to adjust the number of rows and columns and change the window size so that you can easily view many trees at once. How often are the alleles of each species monophyletic? How often does the gene topology reflect the true topology of the populations?

## Comparing the Depth of the Coalescence

In the simulated trees window select **Drawing> Branches Proportional to Length** and **Drawing > Tree Form > Curvogram**. You can now clearly visualize when the different simulations completely coalesce. As you can see, some come together shortly before the initial split in lineages 200 generations ago (in Mesquite for the coalescence, branch length equals generations), but some diverged much longer ago.

Let's look at how the coalescence depth varies among our different simulations. In the **Population tree window** select **Analysis > New Bar & Line Chart for > Trees, alleles, Simulated trees: Coalescence contained within current tree, 10000** for population size,**Tree depth** and let's do **1000** trees for the simulation. There you go. There are no simulations that coalesce until after 200 generations ago. Where is the peak? What is the oldest coalescence that your simulation found?

## Visualizing the Comparison between Population Trees and Allele Trees

We can make a pretty tree that shows the coalescence process within the population tree. In the **Population Tree window** select **Drawing > Tree Form > Contained Gene (or Other) Tree. Use the branch lengths**, and **autoresolve Polytomies** (although you won't have any Polytomies from the simulation). Choose **Simulated trees: Coalescence contained within current tree, 10000**. Now doesn't that look cool? You can see how the simulated gene tree fits together with the population tree. You can scan through the different simulations using the arrows in the box on the bottom right.

Many of the controls normally found in the Drawing and Tree menus will now be found in the Contained menu. Select **Contained > Display Contained Tree**. A new window will appear showing the simulated tree that you are viewing. In this window use the ***second*** **drawing menu** to set **branches proportional to lengths** and **tree form > curvogram**. Scan through the different examples. Can you tell which simulations would require allele losses by looking at the reconstruction within the population tree?

## Quantifying the Comparison between Population Trees and Allele Trees

*Mesquite* can also calculate two values that compare allele trees to populations. The first is **s**. The population tree is not used to calculate s, so it does not technically compare the two trees. First you consider each population as a character state for the alleles. Then trace that character on the allele tree using parsimony. You can calculate s as the minimum number of steps in that character. Thus the smaller s is, the more monophyletic the alleles for each

population are. If the populations have been separated for a long time it can be used to calculate migration rates.

To see a histogram of s for simulated trees select **analysis > new bar & line chart for > trees, alleles, simulated trees: coalescence contained within current tree, 10000, s of Slatkin&Maddison, 1000 trees**. What values does this graph have? What value predominates?

**Deep coalescence** directly compares the allele tree and the population tree. First it makes a best fit of the population tree to the allele tree. It then counts the number of gene lineages present on each branch – 1 and adds them all together. Thus it is a measure of incomplete lineage sorting. Select **analysis > new bar & line chart for > trees, alleles, simulated trees: coalescence contained within current tree, 10000, Deep coalescence (gene tree)**. Deep coalescence (gene tree) compares values for different allele trees on a single population tree, while Deep coalescence (species tree) does the opposite. Hit OK a bunch of times and simulate 1000 trees. All these numbers look pretty high. Let's see how changing the parameter values effects the calculations.

## Effect of Branchlength on Coalescence

Remember that coalescence simulations calculate the probability of two genes coming together generation by generation. Therefore a decrease in the number of generations along a branch will decrease the chance that two alleles coalesced on that branch and instead will tend to move the coalescent event to earlier branches. In *Mesquite* branch lengths are equal to generations, when doing a coalescence simulation.

Use the little curved arrow at the top of the tabs for the population tree and the three charts to **pop out as separate window**, and arrange them so that you can see everything at once. First let's try altering branch lengths throughout the tree and see what that does. In the **Population Tree** window select **Tree > Alter Transform Branchlengths > Scale All Branchlengths**. Let's multiply all these branchlengths by 10. How does that affect tree depth? s? Deep coalescence? If you're not sure why try comparing the different simulations under the short and the long branch trees. What happens if you multiply these branch lengths by another ten? Is the effect clearer now?

What if you just change one branch? Choose the **Adjust branch length tool**. Click on the internal branch connecting population 2 and 3 to the root. Change the value from 10,000 to 50 and hit the arrow. Does this effect s or deep coalescence more? Can you tell why? What effect does it have on the allele tree topologies? Does it effect conclusions about the actual population tree?

## Effect of Population Size on Coalescence

The other major parameter in the coalescence model is population size. The chance of two alleles coalescing is inversely proportional to the effective population size. So the larger the population size the further back in time the coalescent event will occur. First reset the length of the interior branch to 10,000.

There are two ways to adjust the population size. Every simulation uses a default value of Ne for its calculations. In our case we went for 10,000 in all of out simulations. You can set this value to whatever you want before running a simulation, or you can change it after the fact by going to **Contained > Coalescence Simulation > Set Ne**. The problem with this method for our current analysis is that we ran every simulation separately, so we would have to go to each

window one at a time to change the value of Ne. I think that there is a way to have each window showing the same set of simulations, but I could not figure out how.

*Mesquite* also provides a way of manipulating population size by altering the tree. Ne for each branch is actually the default population size for the simulation times the width of that branch. In this way different branches can have different population sizes. To change the population size for the entire tree, go to **Tree > Set all lineage widths**. Let's start with **0.1** to cut the population size down to 1000. What effect did that have on your calculated values and tree topologies? Do you know why?

What if we have one lineage with a much bigger population size than its sister lineage? Choose the **Adjust lineage widths** tool (horizontal ruler with an arrow). Click on the branch leading to taxon 3. Set it back to one and click the arrow. How does this affect your calculated values? What about your topologies (you may have to remove the **branches proportional to branchlengths** in the **multi-tree window** in order to see this)? In particular is the tree correct?

**Question 1:** Which species' alleles usually come out as paraphyletic? Why?

## Bottlenecks

One really cool trick that you can do in *Mesquite* is set up bottlenecks for only a portion of a branch. First return all the branch widths to0.1. Now select the **insert node tool** (its picture has three arrows: to top right, to bottom left, from bottom right). Click on the root branch to insert a new node. Iif you do not see a root branch add one by clicking the adjust branch length tool on the root.. Choose the **Adjust lineage widths** tool again and click on the upper of the two root branches. Set it to 0.01. Set the branch length for this bottle neck to 80. Now you have set it up so that there was a contraction in the population to 100 individuals for 80 generations before the division at the base of your tree. What effect did this have on your calculated values, especially tree depth? Do you know why it had that affect on tree depth?

Try setting up some other scenarios. What if the population fluctuated right before the divergence? Try setting up two more periods before the first bottleneck: one with normal population sizes for 200 years, preceded by another 80 year bottleneck. Do you see two peaks in your tree depth histogram?

**Question 2:** What if there was no bottleneck at the root, but instead there was one at the base of the population 2 lineage after splitting from population 3? What effect would that have on the topology of the simulated allele trees?

## Comparing Population Trees Using a Gene Tree

All these simulated trees are very good for getting a handle on what the effects of population parameters are on allele distributions among those populations. However, when looking at real data with real alleles you can usually derive a reasonable estimation of what the true gene tree is. There are two ways to take advantage of this that I am aware of. One method is to take a known gene tree and a known species tree. Coalescence can then be simulated on the gene tree under different combinations of branch lengths and population sizes. Statistics, such as s and deep coalescence, can then be calculated for these simulations and for your actual gene tree in order to estimate what set of branch lengths and population sizes were likely to produce the gene tree in question. This can also be done using likelihood models.

Another option is to use a tree of alleles to compare different species trees. It might be reasonable to suspect that the true species tree is the one that has the smallest deep coalescence count. Open the file **Coalescence Example 3**. (I stole this example from the *Mesquite* examples.) As you will see, there are three species trees and one gene tree in this example. Page through the different species trees. First trace the gene tree in the species trees. Select **Drawing > Tree Form > Contained Gene (or Other) Tree**. This time you want to select **Treat Contained as Unrooted**. For simulated gene trees you know where the actual root of the gene tree is, so it is appropriate to use that root; however, in practice it is very difficult to find the true root of a gene tree, so we will pick the rooting that minimizes the incompleteness of lineage sorting. Use **stored trees**, because this time we are using a tree that we already have.

You will see the gene tree traced in the taxon tree. In the bottom right the contained tree window gives two measures of fit for these trees, deep coalescence and birth death. Deep coalescence is an appropriate measure for allele trees, while birth death is an appropriate measure for trees of duplicated genes. In any case they should both be minimized and that minimum will correspond under almost all circumstances. Page through the taxon trees.

**Question 3.** Which tree is the best fit for the gene tree?

*Mesquite* is also capable of searching through all possible trees to find the "best" tree under any parameter value you choose. *Mesquite* is designed for flexibility, not efficiency. Thus, it can use a variety of different criteria to evaluate trees, but its tree searching algorithms are not fast, and it should not be used if another program is available. Select **Taxa&Trees > Make New Trees Block from > Tree Search > Heuristic, Species** (cause we are making a new tree of species). Then select **Deep Coalescence (Species Tree).** (Why did we choose species tree this time?) Click **treat contained as unrooted, OK, stored trees, SPR rearranger.** Leave max trees at 100. This may take a minute.

View the new tree block. How many trees did you get? Why do we get so many trees? Is there one taxon that really jumps around a lot? Why? Trace the gene tree in these trees. How do the deep coalescence values for these trees compare to those we calculated for our original three trees?