Feb. 28, 2014**. Character evolution – Phylogenetic signal and character correlations using parsimony**

Assigned reading:
Maddison WP, Slatkin M (1991) Null models for the number of evolutionary steps in a character on a
     phylogenetic tree. Evolution 45: 1184-1197.
Donoghue, M. J. 1989. Phylogenies and the analysis of evolutionary sequences, with examples from the
     seed plants. Evolution 43:1137-1156.

Other particularly useful reading:
Maddison WP, Maddison DR (2000) MacClade 4 (pdf manual), Sunderland MA: Sinauer, chapters 3, 4
     and 22 (equivalent to chapters 4, 5 and 17 in published MacClade 3 manual).
Maddison, W.P. A method for testing the correlated evolution of two binary characters: are gains or
     losses concentrated on certain branches of a phylogenetic tree? Evolution 44, 539-557 (1990).

     In this portion of the class, we are interested in examining how discrete-state characters evolve on a tree individually and together. These are characters that meet the 'discrete-state' criteria for taxonomic characters. We will focus on binary traits (those with just two states); multistate characters are often generalizations of the binary case but can get more complicated, especially for the study of trait correlations. Other quantitatively varying characters can also be studied, but they require different methods that we'll cover later.
     When we focus on the question of character evolution, we will assume that we have already obtained a phylogeny (with possible uncertainties). As we map the history of particular characters, we are not reevaluating the underlying phylogenetic hypothesis. Some have argued that you must not include a character for phylogeny reconstruction and then also map its ancestral states to test evolutionary hypotheses. What do you think?
     Mapping of character states on a tree is an essential starting point for comparative methods related to phenotypic traits. The analysis of ancestral/derived states is central to phylogenetic approaches to questions of adaptation, divergent/convergent evolution, and correlated evolution. A typical adaptive hypothesis posits that trait X is an adaptation to selective pressure Y. For example: tough, evergreen leaves, termed sclerophylls, are an adaptation to semi-arid environments. This hypothesis may be tested in various ways, but one of the key predictions is that lineages with sclerophyll leaves *first* encountered semi-arid environments, and *then* evolved sclerophylly. If they evolved sclerophylly *before* encountering the hypothesized selective environment, then this trait must have evolved (and thus be an adaptation) in response to some other factor. This is one type of question that would lead you to conduct a comparative test.

I. Ancestral states
     The simplest approach to mapping ancestral states is the parsimony principle: we seek the reconstruction that requires the fewest evolutionary 'steps' or transitions between character states. We don't know that evolution proceeded in this way, but the basic argument for parsimony is that we should not assume any additional evolutionary events, beyond the minimum number necessary to explain the observed patterns. After calculating the ancestral states for a character,

we can then test a number of hypotheses regarding the number of steps, their distribution on a tree, the polarity of character change, and the sequences or associations of change in two or more characters.

The exact algorithm for calculating ancestral states by parsimony depends on the type of character and the assumed 'cost' of transition among different states:
1) Unordered traits have 2 or more states, and all transitions require only a single step;
2) Ordered traits have 3 or more states on an ordinal scale, and the number of steps is equal to the difference in state values (i.e., from 1 to 3 requires 2 steps);
3) Dollo traits: change is only allowed from ancestral to derived state (reversals require 'infinite' steps);
4) Arbitrary step matrices allow user to assign any desired cost structure to transitions.

In many cases, most parsimonious reconstructions (MPR) can be estimated visually, though it is easy to miss alternative equally parsimonious reconstructions when there are several gains and losses. To illustrate the mechanics of parsimony reconstruction, the formal algorithm for Fitch parsimony for unordered states is provided on the last page. Note: In contrast with the maximum likelihood methods we will study later in the class, branch lengths are not incorporated in parsimony algorithms. One can calculate a branch length after the fact based on the number of characters that change on a branch, but they are not used in the process of determining the MPR.

II. Is a trait 'conserved'? Null models for the number of changes on a tree

After reconstructing the evolutionary history of a trait, you might wish to test a hypothesis about whether the trait has undergone an 'unusually' small or large number of transitions between alternative states. This is the question of phylogenetic signal: if the trait exhibits a small number of transitions, then it will be conserved over large portions of the tree and close relatives will tend to exhibit the same state (high phylogenetic signal = conserved trait). If the trait exhibits a large number of changes, then close relatives will frequently have different states (low phylogenetic signal = divergent/convergent trait).

Testing the hypothesis of phylogenetic signal requires that we compare the observed number of transitions to some null hypothesis, and ask whether the data is improbable, and thus significantly different from, the null. In this case, an appropriate null model could be constructed by asking how many changes would be expected in the trait if there was no phylogenetic signal, i.e. states of the trait were randomly arranged among the taxa so close relatives share the same state no more often than expected by chance (Maddison and Slatkin 1991). This is an example of a non-parametric test (i.e., there is no elegant solution that can be derived analytically from the underlying statistical parameters). The significance of non-parametric tests can be solved using Monte Carlo methods, which simply means brute force computer simulation under the null model.

The basic logic of a Monte Carlo test is as follows:
I.   Choose a test statistic, T, in this case the number of evolutionary steps on a tree.
II.   Calculate T for the observed data; we'll call this T_0.
III.   Choose a null model to obtain random permutations of the data in which the pattern of interest is removed, but all other relevant aspects of the data are maintained. In this case, we can permute (= sample without replacement) the states of the character across the tips of the tree. What aspect of the data is randomized using this null? What aspects of the

data are kept constant?

IV. Conduct 99, 199, 999, 1999, 9999, 19999 randomizations using the null model in III, calculating the test statistic for each of the null data sets, T_{1-9999}. The number of randomizations depends on the complexity of the null model, the speed of your computer, the precision you desire for the significance value, and your patience. The total number of observations, **N**, will be the number of randomizations plus 1 for the observed data. In some cases, it will be possible to enumerate all possible permutations of a data set under the null model. In this case the total number of possible permutations is N, one of which is the observed value, and significance can be calculated as an *exact test*.

V. Add the observed value to the set of null values, and then sort the entire distribution to determine the rank of T_0, which we'll call **R**. In the case of ties, T_0 should be placed closest to the middle of the rankings, i.e. at the highest rank if it is less than the median and at the lowest rank if it is greater than the median. So if $T_0 = 5$ (bold case below) in a null distribution based on 9 randomized data sets:

4, 5, 5, **5**, 6, 6, 7, 7, 8, 9: R = 4 out of 10

or 1, 2, 3, 3, 4, **5**, 5, 5, 6, 7: R = 6 out of 10

VI. Now, to calculate a significance value you have to decide between two alternative tests. If you are only interested in testing the hypothesis that T_0 is lower than the null expectation *OR* that it is greater than the null expectation, you are conducting a one-tailed test. For example, you may want to know if a trait exhibits high phylogenetic signal, but you don't really care if it exhibits lower signal than random. This would correspond to a test that the rank of T_0 is significantly lower than the median. On the other hand, if you want to test for T_0 deviating from the null hypothesis in either direction, it is a two-tailed test. The p-value for significance under these alternatives is:

One-tailed low: $P \leq R/N$

One-tailed high: $P \leq (N+1-R)/N$

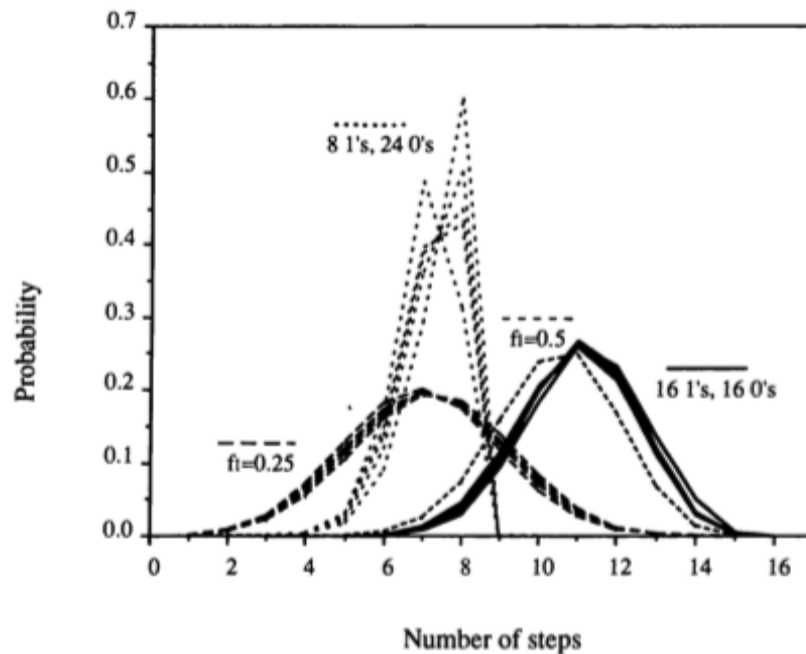Two-tailed: $\min(2R/N, 2*(N+1-R)/N)$

FIG. 2. Probability distribution of the number of steps on a tree of 32 taxa, for various null models of random trees and characters. There are four groups of curves, two with randomly generated characters (probability of taxon having state 1 f_1 = 0.25 or 0.5) and two with characters with fixed numbers of 0s and 1s (with 8 1s and 24 0s, or with 16 1s and 16 0s). Within each group are five separate curves that are not very different from each other, for (1) equiprobable random trees, (2) random joining trees, (3), random partition trees, (4) a fixed asymmetrical tree, and (5) a fixed symmetrical tree. There is only a single curve for one group, f_1 = 0.5, because all trees give the exact same curve.
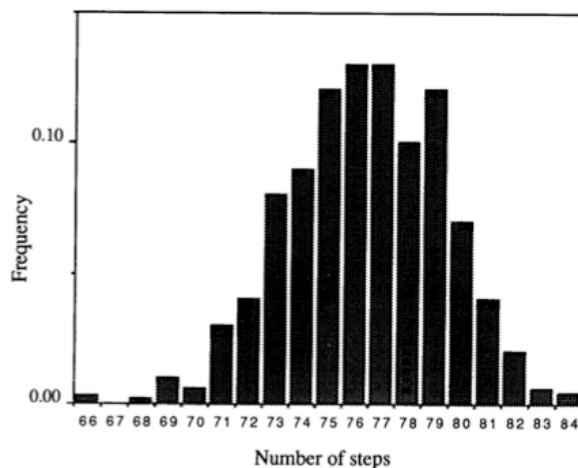

FIG. 4. Probability distribution for the number of intercontinental migration events required by trees relating the 134 human mitochondrial samples of Cann et al. (1987), under the null hypothesis that humans are panmictic. Based on a sample of 500 random-joining trees generated by MacClade. The number of migration events required by their published tree, 37–47, is much less than that expected under panmixis.

III. Correlated evolution of two binary traits

Maddison's (1990) concentrated changes test was one of the first tests introduced to test for correlations in the evolution of discrete traits, in a phylogenetic context. The test examines the question: are changes in character B concentrated on portions of the phylogeny where character A has a particular state, more than expected by chance? For example: does dioecy evolve from hermaphroditism more often than expected in lineages that have fleshy (vs. dry) fruits? Does larval gregariousness in butterflies evolve more often than expected in lineages with warning coloration?

The test proceeds in several steps:

1) first reconstruct the evolution of character A on the tree. This is considered fixed – an independent factor that may influence evolution of character B, but is not dependent upon it.

2) character B is mapped on the tree, and the total number of evolutionary transitions (gains and losses), and the number of transitions that occur against each background in trait A, are tabulated;

3) by exact calculation, or simulation, the number of possible ways to arrange the same number of changes on the given phylogeny are calculated, along with the number of arrangements which involve as many or more changes against the relevant background of character A.

4) The significance of the observed pattern is calculated as the proportion of possible arrangements with *as many or more* changes located in the selected background of character A.
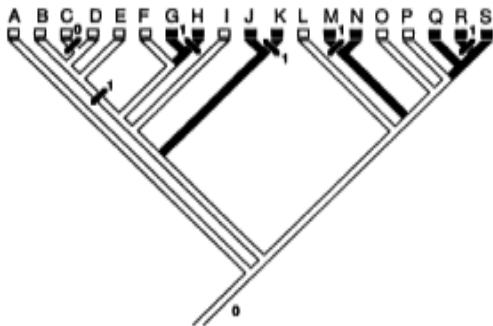

FIG. 1. Phylogenetic tree with the evolution of two characters traced on it. Changes in the first character are indicated by tick marks with the derived state listed ("0" or "1"). The states of the second character are indicated by the shading of the branch: white for one state and black for the other.

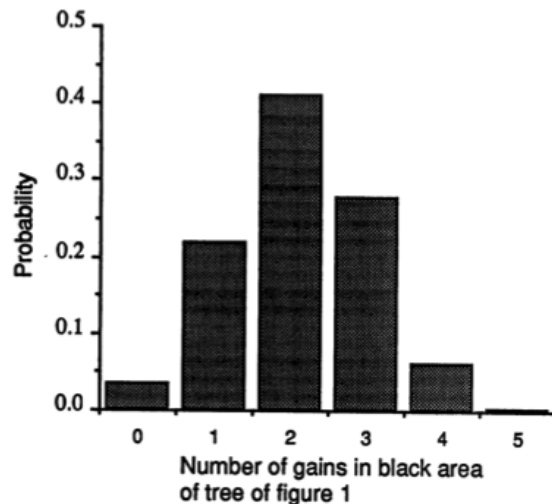≥4 gains (out of 5 gains, 1 loss) on black branches: p = 0.0638

FIG. 4. Probability distribution for numbers of gains in black area of phylogenetic tree in Figure 1 given five gains and one loss overall in the tree, under the null model that gains and losses are distributed randomly on the branches of the tree. Note that five gains in the black area would be judged a significant concentration.
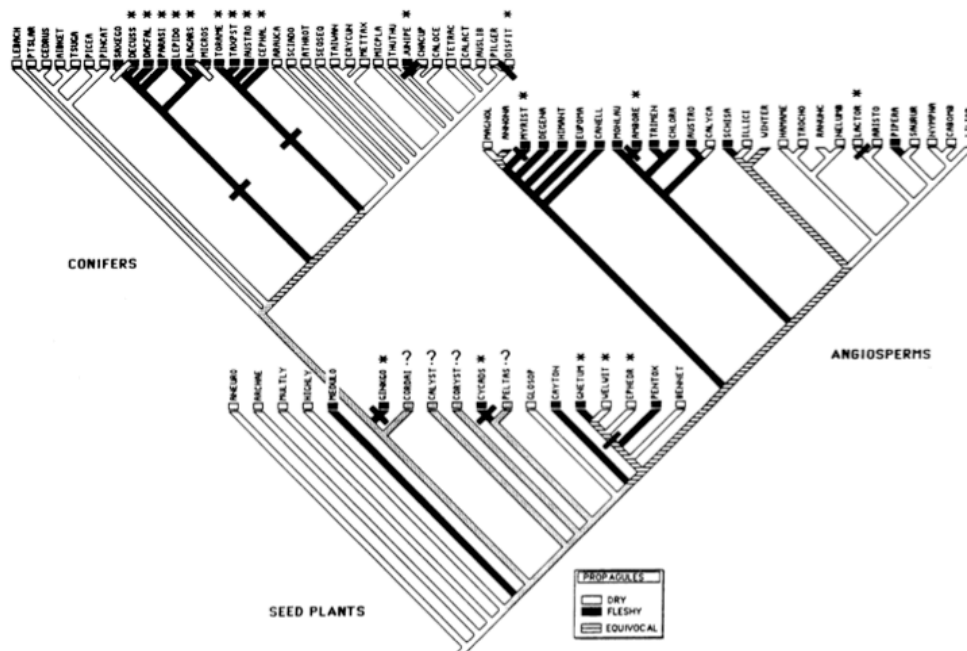
FIG. 2. A composite cladogram of seed plants based on Doyle and Donoghue (1986), Hart (1987), and Donoghue and Doyle (1989), showing parsimonious reconstructions of the evolution of fleshy propagules (animal dispersal) and dioecy (MacClade, version 2.1 [W. P. Maddison and D. R. Maddison, unpubl.]). Black branches indicate fleshy propagules, white branches indicate dry propagules, and striped branches represent equivocal conditions. Stippling reflects uncertainty concerning the mode of dispersal in taxa marked by "?." Taxa known to be dioecious are marked by asterisks; black crossbars indicate the origin of dioecy, open bars indicate reversal to monoecy. Maddison's (1990) statistical test was applied to two optimizations (see text): in optimization 1, stippled and striped branches are assigned dry propagules (colored white); in optimization 2, stippled and striped branches are assigned fleshy propagules (colored black). Additional information on the cladograms, terminal taxa, and characters is provided in the Appendix.
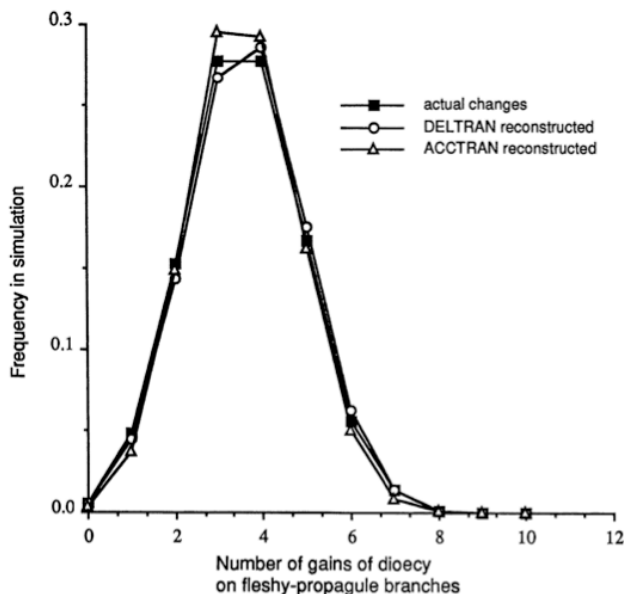

FIG. 6. Frequency distributions for numbers of gains of dioecy in animal-dispersed lineages on Donoghue's cladogram (1989, Fig. 2), given any number of losses in the black area, 10 gains and 2 losses overall in the tree, either state ancestral, under the null model of random distribution of changes (from simulations). Distribution labeled "Actual" is for the actual changes (sample size 12,987), "DELTRAN" for changes reconstructed by parsimony with ambiguity resolved using the DELTRAN algorithm (sample size 4,362), and "ACCTRAN" for changes reconstructed by parsimony with ambiguity resolved using the ACCTRAN algorithm (sample size 6,802).

Fitch parsimony — ancestral reconstruction of an unordered trait (binary or multistate); simplified algorithm for bifurcating tree (no polytomies) with no internally fixed states.

Set notation:
$\cap$ Intersection: the items shared by both sets
$\cup$ Union: the combined list of items from both sets
$\varnothing$ The null or empty set

Definitions:
downpass   an algorithm that starts at the terminal taxa and works
            downward through the ancestral nodes, so all daughter nodes
            must be visited before visiting a deeper ancestral node
uppass   an algorithm that starts at the root and progresses upwards
            to the terminals
MPR   Most Parsimonious Reconstruction
P,Q,M,N   nodes of the tree (see figure)
S   downpass state set
F   uppass state set
L   length of tree (number of changes)
$\leftarrow$ assign RHS to LHS

Start with L = 0
Algorithm 1: Fitch downpass algorithm
```
if (N is terminal) D_N ← observed trait value
else {
    D_N ← D_P ∩ D_Q
    if (D_N == ∅) {
        D_N ← D_P ∪ D_Q
        L ← L + 1
    }
}
```

Algorithm 2: Fitch uppass algorithm
```
if (N is root) U_N ← D_N
else {
    U_N ← D_S ∩ U_M
    if (U_N == ∅) U_N ← D_S ∪ U_M
}
```

Algorithm 3: Assign MPR at each node
$MPR_{root} \leftarrow D_N$
$MPR_N \leftarrow U_M \cap D_P \cap D_Q$
if (MPR_N == $\varnothing$) $MPR_N \leftarrow (U_M \cap D_P) \cup (D_P \cap D_Q) \cup (U_M \cap D_Q)$
if (MPR_N == $\varnothing$) $MPR_N \leftarrow U_M \cup D_P \cup D_Q$