9/26 Biogeography

I. Understand how we can use CTMC models to fit and make inferences about biogeographic history. Understand the basic model of Ree and Smith and be able to follow the logic of the Q matrix. Understand the DEC model.
    A. How would you test asymmetric dispersal between two islands? How would you test if there is no dispersal? Be able to come up with and formulate Q-matrices to test these hypotheses. How does the size of the Q matrix change with the number of areas?
    B. What happens at cladogenesis?

II. How can biogeographic data be used to date trees? Why can this work?
    A. Understand the basic elements of how Landis et al. fit their model. Obviously don't need to know every detail or how to do it, but understand that every part of the model (morphological evolution, molecular evolution, biogeography, uncertainty of island age) all inform each other, and are jointly inferred. In other words, molecular data could inform the estimate of geologic age!

10/1 Revbayes lab

I. Understand how to diagnose a Bayesian analysis. What is convergence? What is mixing? What should a parameter trace look like? What should acceptance probabilities be for maximal efficiency?

II. Be able to explain, in general terms, the total evidence analysis we ran in class with its 3 components.

10/3 Dating phylogenetic trees

I. Why can't you get dates from molecular data alone? Understand relationship between Rate X Time = Change; what's estimated in a molecular analysis, and what we want.
    A. What is a strict clock? How could you test the hypothesis of a strict clock?
    B. Relaxed clock models, understand what each of these terms implies
        1. Local clocks (fixed changes based on a priori assignment)
        2. Random local clocks (let data tell you where clock model changes)
        3. Uncorrelated clocks (Branch rates come from a distribution, such as a lognormal distribution. Each branch randomly drawn from distribution)
        4. Autocorrelated clocks (Branch rates come from a distribution, such as a lognormal distribution, but the mean of that distribution is set to the previous branches' value. In other words, there is phylogenetic signal in the clock rate and it evolves down the tree.)

II. How are fossils used to time-calibrate a phylogeny?
    A. Use fossils as minimum ages and set priors on node ages
        1. What is the drawback of using priors in this way? (How many fossils can you use)
    B. Compare to "tip-dating" and the "fossilized birth-death model"
        1. FBDP - Birth death model that incorporates diversification dynamics, all fossils, places them as tips, and simultaneously dates the tree.

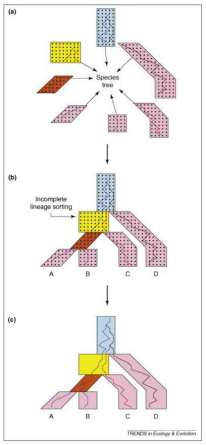
2. Includes some probability that some fossils are *observed ancestors*. Understand how this inference can be made by thinking about why we think the probability of observed ancestors increases toward the present, and how this is related to the birth-death process.
3. Morphological model, molecular model, birth-death model, stratigraphic/sampling model all combined in "total evidence approach". What are the benefits?

10/8 The Multispecies Coalescent

I. Why is the risk of concatenating gene sequences and analyzing? (Anomaly zone, what is it?)
   A. Incomplete lineage sorting means not all gene trees (or even most, in some cases) match the true species tree.
   B. Compare and contrast the 3 sources of "error" in species tree estimation: Mutational variance, Coalescent stochasticity, and hybridization. What do we do about dealing with mutational variance? What about coalescent stochasticity?

II. Coalescent theory
   A. What are the important parameters? What makes ILS more likely? What makes it less likely?
   B. What is the role of ancestral polymorphism?
   C. Coalescent model tells us how likely discordance from ILS will be given a particular species tree and demographic history...under what assumptions?
   D. Example, we expect mitochondrial genes to sort how many times faster than nuclear genes? Why?
   E. How can you do things like estimate past population sizes using the coalescent using gene trees? What would gene trees look like if a species had a bottleneck vs. if the species was expanding in population size?

III. Understand how the multispecies coalescent model roughly works according to this slide



$P(D_i \mid Q_i, \pi_i, \boldsymbol{\psi_i})$ = standard likelihood of gene tree

$P(\boldsymbol{\psi_i} \mid S)$ = Likelihood of gene tree given the species tree

"AND" rule:
$P(D_1, D_2...D_n \mid S) =$
$P(D_1 \mid Q_1, \pi_1, \boldsymbol{\psi_1}) * P(\boldsymbol{\psi_1} \mid S)$ x ...
x $P(D_n \mid Q_n, \pi_n, \boldsymbol{\psi_n}) * P(\boldsymbol{\psi_n} \mid S)$

. You don't need to use the right parameter symbols or memorize all the details, but be able to write down likelihood equations like depicted here and distinguish between what a concatenated analysis assumes, and what a multispecies coalescent model assumes. (i.e. that a coalescent

model doesn't assume gene tree =species tree, but calculates the likelihood of the gene tree given the species tree under the coalescent).

IV. Have at least a cursory knowledge of the general way that methods of species tree using coalescent methods vary. Axis ranging for full models that take a long time and have trouble converging to methods that use summary statistics, do not use a full model, analyze only parts of the data at a time, assume perfect gene trees etc. (shortcuts to make it computationally tractable).

V. Using all data may not be best. Many genes noisy. May want to break down your gene trees into components and *interrogate your data.* Example:

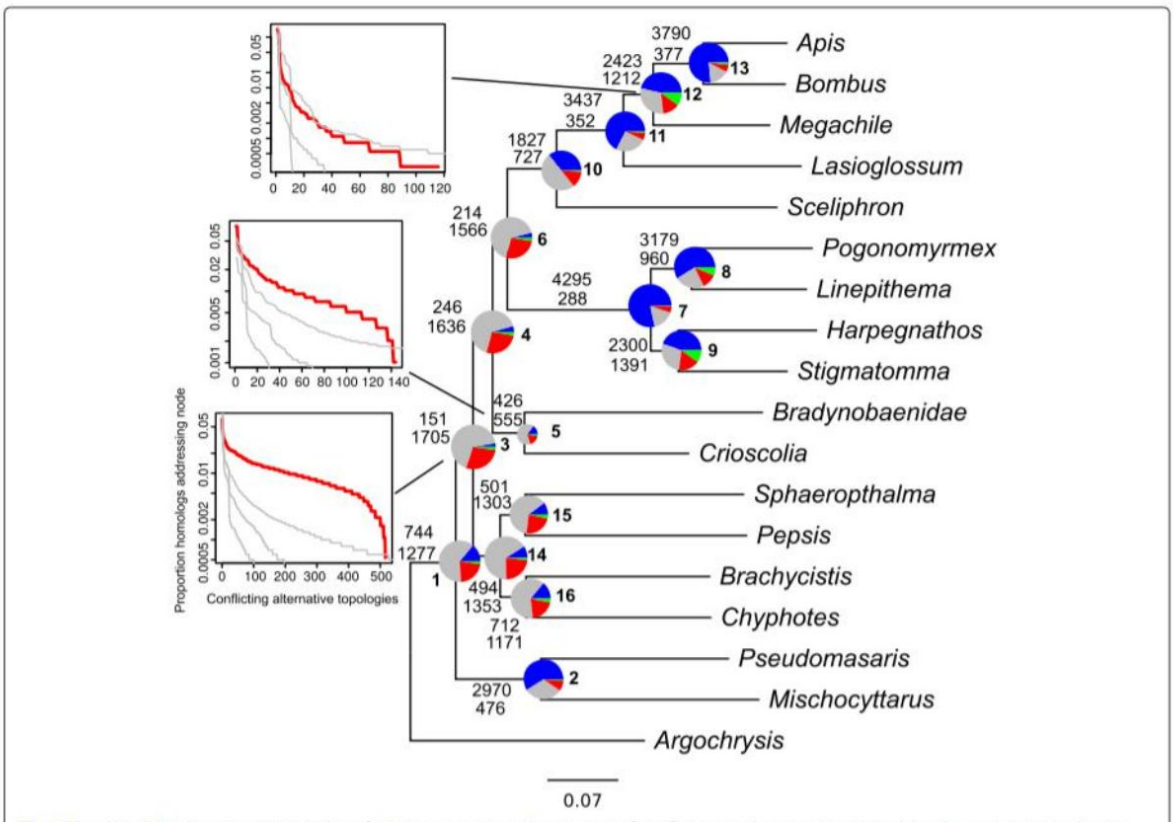


Fig. 2 Combined ML (species tree) topology for Hymenoptera, with summary of conflicting and concordant homologs. For each branch, the top number indicates the number of homologs concordant with the species tree at that node, and the bottom number indicates the number of homologs in conflict with that clade in the species tree. The pie charts at each node present the proportion of homologs that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion that inform (conflict or support) this clade that have less than 50 % bootstrap support (grey). The histograms show, for three nodes, the proportion of the total homologs that support each conflicting alternative resolution for the clade in question, sorted from largest to smallest. Grey lines represent distributions of conflicting alternative resolutions based on coalescent simulations generated with three tree heights. The histograms for other nodes are presented in Additional file 2: Figure S5

VI.

10/10 Tara Pelletier

I. Questions about Tara's paper, especially her goals, questions, general methodological approach, findings and research philosophy are fair game!