Lecture 18 – Species-Tree Estimation

I. Causes of Incongruence - In our previous discussion of phylogeny estimation from multiple data sets (partitions and mixtures), we've made the assumption that all the data have the same history.

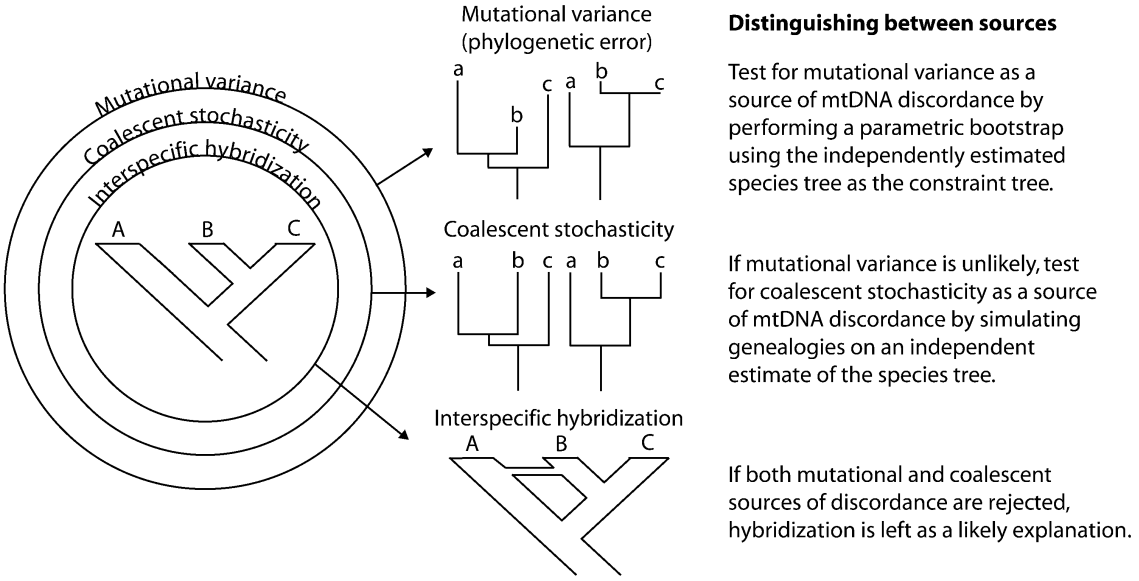
However, there are several important reasons that phylogeny estimates from separate genes might be incongruent.

1. Phylogenetic uncertainty – For much of the semester we've been dealing with methods for assessing and accommodating uncertainty. This may be simply the result of sampling or stochastic error, or it may be the result of one or more of the data sets being subject to systematic error. What's critical, though, is that these involve a common true history.

2. Coalescent stochasticity – Even if there has been only vertical transmission of genetic material, stochastic sorting of ancestral polymorphism (i.e., lineage sorting) may well lead to incongruence among gene trees. That is, there may be multiple true gene trees that have evolved within the same species tree.

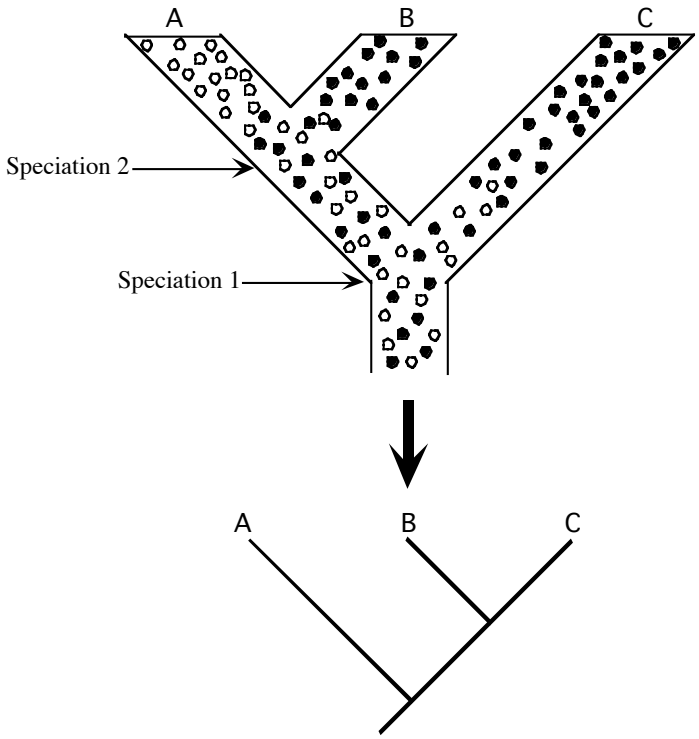
3. Hybridization (eukaryotes) and/or horizontal gene transfer (prokaryotes) – If there is a history of non-vertical transmission of genetic material (and evidence has accumulated that this may be pretty common), incongruence among gene trees may be reflecting different true histories.

These are actually listed in an order that suggests a sequence of testing to ascertain the cause of incongruence (from Reid et al. 2012. Syst. Biol.).



II. Lineage Sorting of Ancestral Polymorphism – Cause 2: Coalescent stochasticity.

Let's look at an ancestral population. Some sites will be polymorphic at the time of the first speciation event (that is, there may be more than a single allele).



If the polymorphism persists through a second speciation event (which is called incomplete lineage sorting), there's a possibility that it will be sorted in a manner that is incongruent with the phylogeny, in this case ((B,C),A).

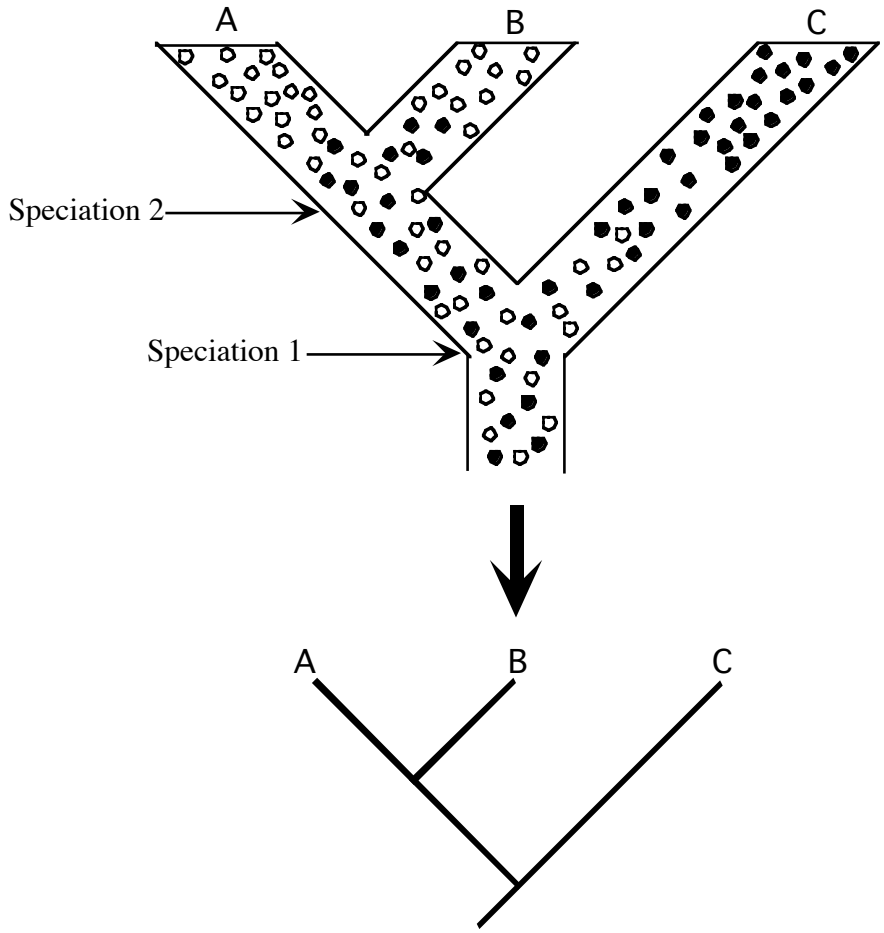
This has long been called incomplete lineage storing (ILS), but that's imprecise. More precise terms are anomalous lineage sorting or hemiplasy, the latter of which was coined by Avise & Robinson (2008. Syst. Biol., 57:503).

In these cases, we expect there to be more than a single true bifurcating history. One of these will represent the sequence of speciation (this is called the species phylogeny) others represent the history of lineage sorting for the particular gene. This is called the gene phylogeny.

In such a situation, the probability that there will be a conflict between the gene tree and the species tree is proportional to the length of time the polymorphism persists. This is, in turn, proportional to population size, so groups with large ancestral populations will be more susceptible to this type of conflict, especially for speciation events close in time.

Under simplifying assumptions, we expect ILS to lead to the other incorrect gene tree ((A,C),B) with the same frequency.

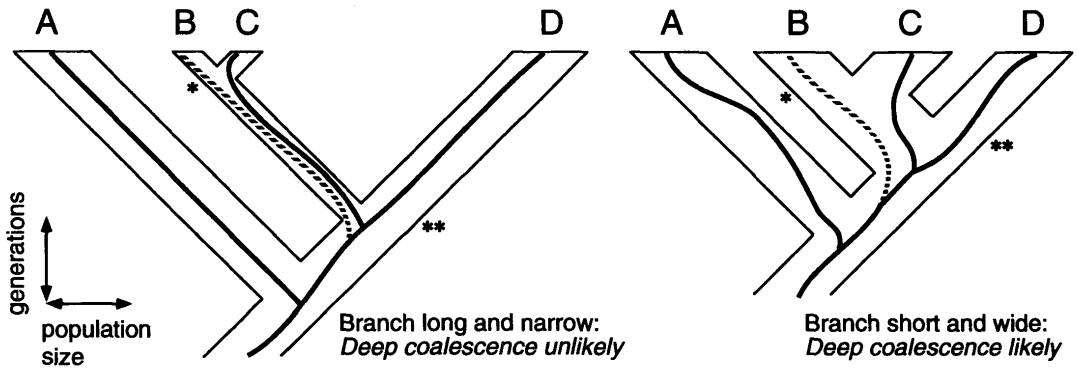
Again, in this situation, we expect some gene phylogenies to reflect the species phylogeny and others to conflict with it. This is because even with persistence of ancestral polymorphisms, there's a chance that the polymorphisms will sort in a manner congruent with the species phylogeny ((A,B),C).



So, the probability of anomalous lineage sorting is dependent on the persistence of ancestral polymorphisms for at least two consecutive speciation events, as well as post-speciation fixation in a particular manner.

Obviously, the length of the internal branch of the species tree will impact this probability. If it's a long branch, the probability of anomalous lineage sorting will be lower than if the internal branch is short.

Furthermore, the persistence time of neutral polymorphisms is proportional to the (effective) size of the ancestral population. So, species with large population sizes will be more susceptible to anomalous lineage sorting than will species with relatively smaller population sizes.



It turns out we can model these probabilities with coalescent theory.

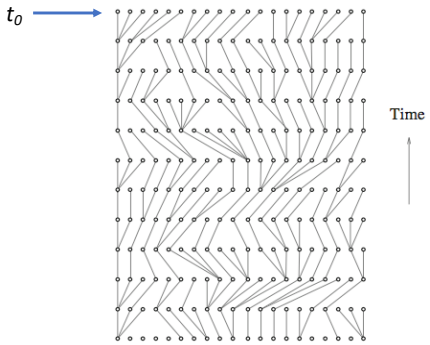
III. Brief Introduction to Coalescent Theory

A. Within a single population.

In 1982, J.F.C. Kingman provided an insight that flipped population genetics on its head.

Classical population genetics use recursion equations to describe allele frequency change over time, starting at the present, t_0, and going forward.

Coalescent theory starts at the present and looks back in time. Say we have a diploid population of 10 individuals, so 20 gene copies. Coalescent theory addresses ancestry of these copies and allows us to make inferences about the historical processes that have shape current genetic variation.



Assumptions (Ideal population).

- Constant N_e
- Discrete generations
- Random mating
- No selection
- Mutations occur regularly across time.

These make it simple to calculate the probabilities that two gene copies have descended from the same parent.

The probability that any two offspring have the same parent (*i.e.*, they coalesce) is $1/(2N)$.

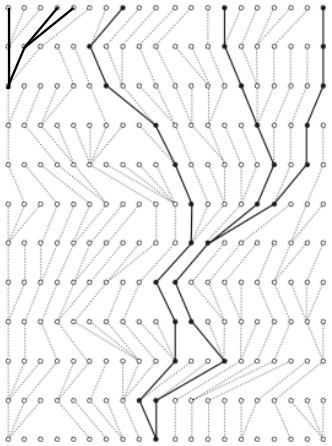
If we know N and have a sample of k copies, the *expected value* of the time to coalescence for the k copies is:

$$T = 4N\,(1 - (1/k))\text{ generations.}$$

Assume that we have a random sample of three gene copies from this population: $N = 10$, & $k = 3$.

Expected T (TMRCA) for $N = 10$ and $k = 3$ is:

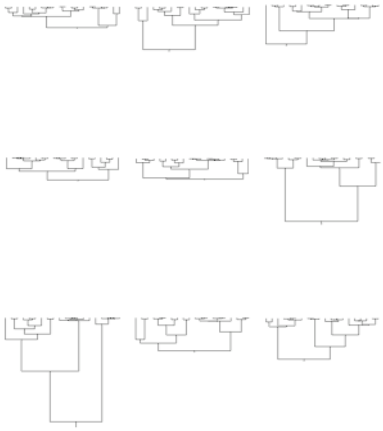
$$T = 4N\,(1 - (1/k)) = 40\,(1 - (1/3) = 40\,(2/3) = 26.67 \text{ generations.}$$



Some samples of three copies will have very short time to coalescence, and other samples will have much longer.

In fact, we actually expect there to be a lot of variation in the times to coalescence. This is what we mean by **coalescent stochasticity**.

The scale of the variability that this stochasticity generates is shown below. The figure illustrates nine different outcomes of the same coalescent process of 20 copies.

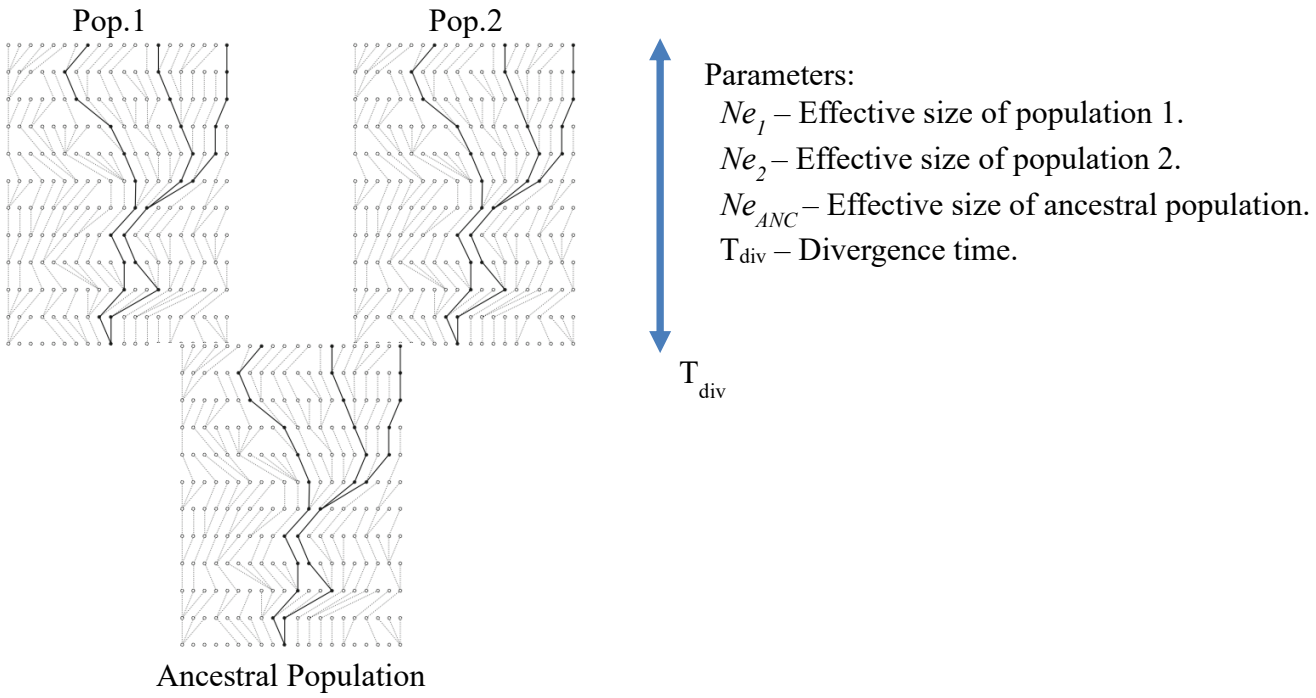


This is a simulation. However, these different coalescent histories could also apply to different genes from the same set of 10 individuals. We'll come back to this in species-tree estimation. For now, we'll note that this indicates that we should treat single-gene phylogenies with caution.

B. Mulitspecies Coalescent.

We can extend the coalescent to more than one species to model evolutionary divergence. We'll start with an ancestral species, in which the coalescent has been operating, and model a divergence event, at time T_{div}, that spawns two daughter species.

Simple multispecies coalescent model.



So, now we have a coalescent process occurring across speciation.

For genome-scale data, we can think of gene trees evolving within the species tree.

Here, we've omitted individuals, and trees of different colors represent different gene trees.

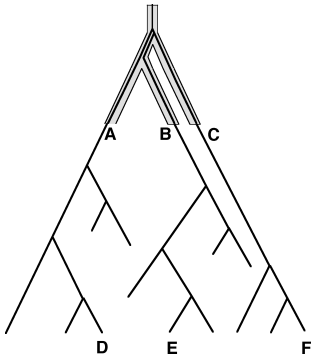
For each of gene, the independent coalescent process is occurring along each branch in the species tree.

III. Hemiplasy in deep time.

There's a somewhat widespread view that the effect of coalescent stochasticity on phylogeny estimation is only relevant to studies that examine relationships among closely related species.

While it may be true that its effects are most pronounced in such cases, coalescent stochasticity can cause incongruence (both between gene trees and species trees & among gene trees

themselves) even in cases where relatively ancient relationships are being examined. This is demonstrated below following Avise & Robinson (2008. Syst. Biol.).



It's therefore not the case that coalescent stochasticity will not affect deep relationships, the length of the internal branches affected will represent an increasingly short branch as the depth of the tree increases. However, it's not likely that clades B & C will be incorrectly inferred to be sister taxa with any *strong* support.

IV. Species Tree Estimation from Multiple Gene Trees

A. Parsimony Based Approach - MDC.

For any combination of estimated gene tree and putative species tree, we can use tree reconciliation approaches to assess how many deep coalescence events are required to resolve any incongruence.

This can be illustrated as shown below (From Tan & Nakhleh. 2009. PLoS Comp. Biol. 5:e1000501).

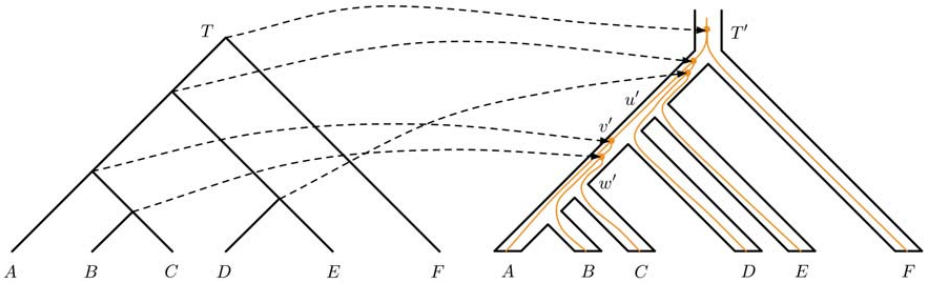


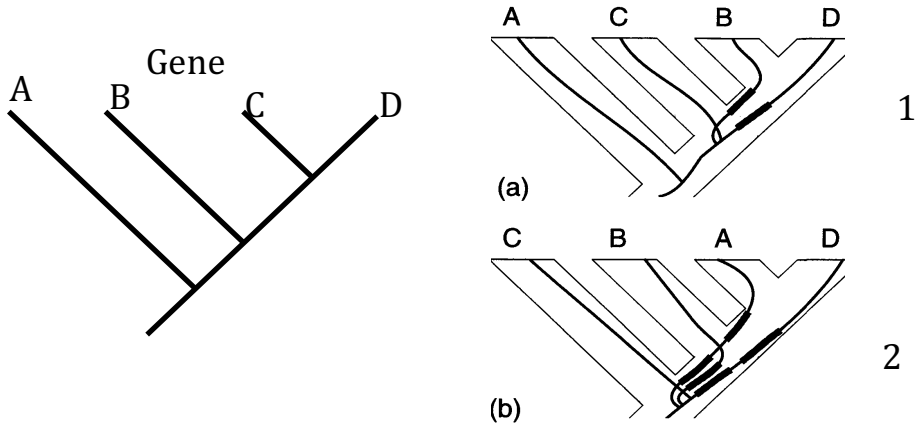
Figure 15. Fitting a gene tree *T* into a species tree *T′*. Here, only mappings of internal nodes of *T* are shown.
doi:10.1371/journal.pcbi.1000501.g015

So, in the reconciliation, two incongruent deep coalescent events are required.

The gene-tree reconciliation for each gene in a data set is evaluated on a putative species tree and the number of deep coalescences required is summed across all genes.

The species tree that requires the fewest incongruent, deep coalescences is the MDC estimate

of the species tree (Maddison 1997. Syst. Biol., 46:523).



Tree space is searched using approaches we've already discussed, and the tree that minimizes incongruence is taken as the species tree. Note that this idea, maximizing gene-tree congruence to estimate a species tree, has been established for ~25 years.

B. Maximum-Likelihood Estimation of Species Trees - STEM

Kubatko and Degnan (2007. Syst. Biol. 56:17) pointed out that there are conditions in species-tree space under which the most likely gene trees conflict with the species tree (which they called *the anomaly zone*). That is, there are combinations of internal branch lengths (of the species tree) and θ (which is $4N_e\mu$, and indexes population size), where concatenation will lead to inconsistent estimation (in the statistical sense) of the species tree.

Furthermore, if we expect there to be a lot of incomplete lineage sorting, the tree that minimizes deep coalescences (i.e., the MDC tree) may not be a good estimate of the species tree.

This suggests that we need to model the coalescent process in species-tree estimation.

Recall that in our discussion of lineage sorting, we pointed out that we can calculate the probabilities of gene tree/species tree discord using coalescence theory.

Recent approaches to species-tree estimation use this property to infer the most likely species tree from a collection of gene trees.

$$P(D \mid \tau_s) = \prod_{i=1}^{l} \sum_{j=1}^{g} P(D_l \mid \tau_G) * P(\tau_G \mid \tau_s),$$

where the product is across all loci and the sum is across all possible gene trees.

$P(D_l \mid \tau_G)$ is simply the regular likelihood function (for a gene tree) and the quantity $P(\tau_G \mid \tau_S)$ is the probability of a particular gene tree given a species tree. This is derived from coalescence theory in Degnan and Salter (2005) and Rannala and Yang (2003. Genetics, 1644:1645). It's based in Pamilo and Nei (1987) and, further, assumes that there is no recombination within genes and free recombination among genes.

So, given a set of gene trees, we can calculate the likelihood of a species tree, and STEM (Kubatko et al. 2009. Bioinformatics, 25:971) uses simulate annealing (remember that?) to search the space of species trees.

This approach requires that gene trees are estimated well and that they are pretty clock-like.

C. Bayesian Estimation of Species Trees (BEST, *BEAST, & BPP)

Each of the above methods (MDC & STEM) estimates the species tree from a collection of gene trees that have been estimated previously.

A few implementations (BEST, Liu & Perl 2007. Syst. Biol. 56:504; *BEAST, Heled and Drummond 2010. Mol. Biol. Evol., 27:570; BPP, Flouri et al. 2018. Mol. Biol. Evol., 35:2585) of a Bayesian approach treat gene trees as nuisance parameters and estimate the species tree directly from multi-locus sequence data.

$$P(S \mid D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i \mid g_i) P(g_i \mid S)\right) P(S) dG,$$

where $D = d_1, d_2, \ldots d_n$ is the set of aligned sequences, $G = (G_1 * G_2 * \ldots * G_n)$ is the space of gene trees and g_i is one of the possible gene trees in G_i.

As above $P(d_i \mid g_i)$ is the regular likelihood function (i.e., the probability of the data for gene i given the tree for gene i).

$P(g_i \mid S)$ is the probability of gene tree i given the species tree and again is derived from Rannala and Yang (2003).

$P(S)$ is the prior of on species trees (usually, a Yule process prior).

These approaches are perhaps the theoretically best justified, because they accommodate uncertainty in estimating gene trees directly into species-tree estimation. They are extremely expensive computationally.

D. Semi-parametric and Summary-statistic Approaches

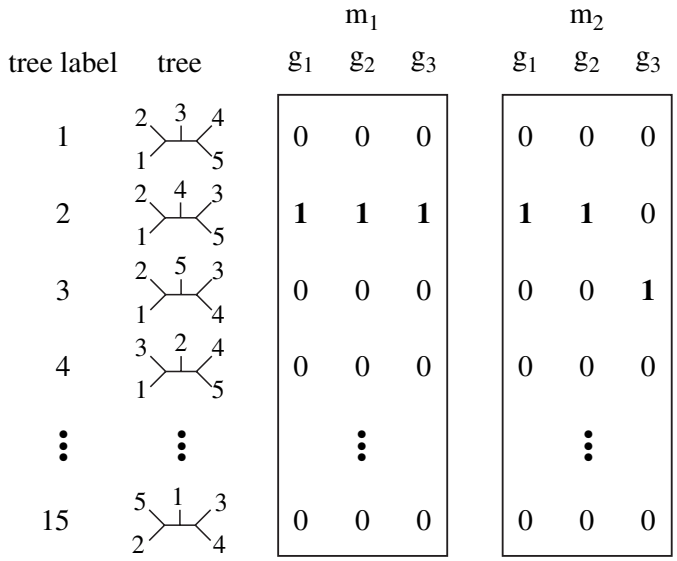
The advances above come at an obvious cost because we're increasing the complexity of our solution space enormously and convergence is a huge issue. The MDC/reconciliation

approach described above is an example of a non-parametric approach, as is use of supertrees.

BCA/BUCKy:

One semi-parametric approach has been developed by Cecile Ané, and is called Bayesian Concordance Analysis (BCA). This approach can be used to assess species trees like MDC, STEM and Bayesian approaches, but it makes no assumptions about the causes of incongruence. Therefore, it's suitable for assessing introgression as well.

It operates using the concept of a Gene-Tree Map (Ané et al., 2007. Mol. Biol. Evol., 24:412).



. 1.—Examples of GTMs. Left box (m_1): GTM showing c
ance among gene trees because all genes are mapped t
ox (m_2): GTM showing some discordance. The first and
are the same tree (tree 2), but the third gene has a diffe

Here, the cells of the gene by tree matrix indicate which tree each gene supports.

In mapping m_1, all three genes support tree 2 and the gene-tree map (2,2,2) is entirely concordant. In the mapping m_2, two genes support tree 2 and the third gene supports tree 3 (2,2,3).

Of course, there's phylogenetic uncertainty and BCA uses Bayesian estimation to fill the cells with posterior probabilities.

They then introduce a 'concordance factor' (α) to model the probability that two randomly chosen genes will have the same gene tree and use this to modify the gene-tree maps during a second-stage MCMC. It's also treated as a random variable that is estimated from the data.

$\alpha = 0$, there's no correlation among gene trees (each gene has unique gene tree). $\alpha = \infty$, the

approach converges to concatenation (there a single gene tree for all genes).

(a)

	single-gene		
	gene		
tree	1	2	3
2	1	0	0
3	0	0.9	0.2
4	0	0.1	0.2
15	0	0	0.6

(b)

GTM	Likelihood*	Prior Prob.	Post. Prob.
(2,3,3)	0.18	$8.38\ 10^{-4}$	0.6600
(2,3,4)	0.18	$7.62\ 10^{-5}$	0.0600
(2,3,15)	0.54	$7.62\ 10^{-5}$	0.1800
(2,4,3)	0.02	$7.62\ 10^{-5}$	0.0067
(2,4,4)	0.02	$8.38\ 10^{-4}$	0.0733
(2,4,15)	0.06	$7.62\ 10^{-5}$	0.0200

(c)

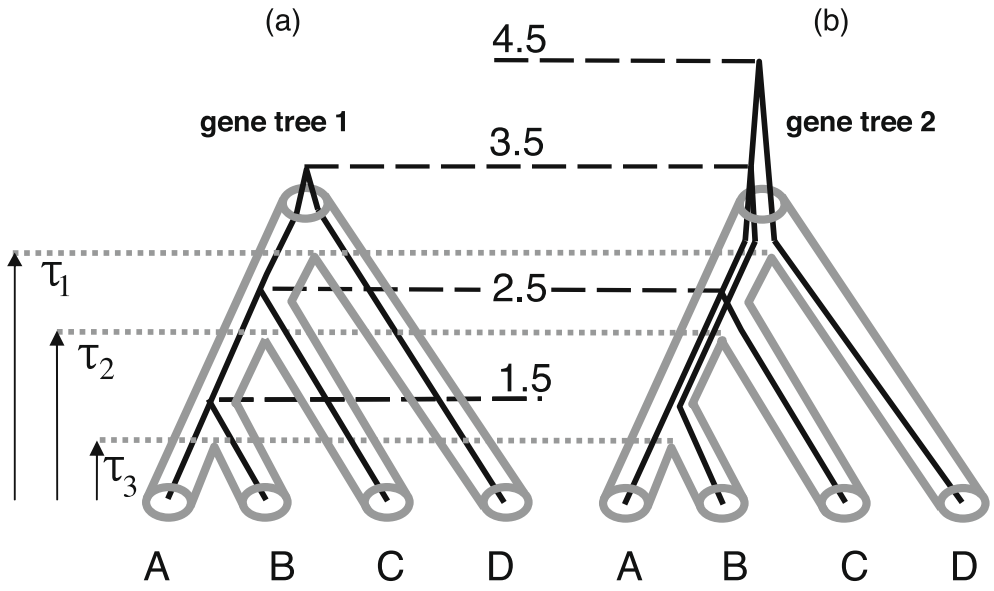
	multi-gene		
	gene		
tree	1	2	3
2	1	0	0
3	0	0.9	0.67
4	0	0.1	0.13
15	0	0	0.2

FIG. 2.—(*a*) Example of the single-gene posterior distribution for 3 genes, with one distribution in each column. (*b*) Posterior probabilities of GTMs after concordance analysis of the single-gene distributions shown in (*a*). All posterior probability is concentrated on 6 of the $15^3 = 3375$ GTMs. Likelihood is proportional to the product of single-gene posterior probabilities as in equation (2). This product is reported here. The posterior probabilities are proportional to the product of the prior probability and likelihood and sum to 1. (*c*) Posterior distribution for each gene conditional on the data from all 3 genes assuming $\alpha = 1.5$ to account for the expected concordance, derived from (*b*).

The inference is that tree 3 in the concordance tree, and the support for tree 2 in gene 1 is due to some process that hasn't been assessed. Thus, the approach does not employ a coalescent model and does not assume that coalescent stochasticity is the only source of incongruence among gene tees.

A number of **summary approaches** (reviewed by Liu et al. 2009. MP&E, 53:320) operate under the following prediction from coalescent theory:

Gene coalescence times always predate species divergence time. For example:

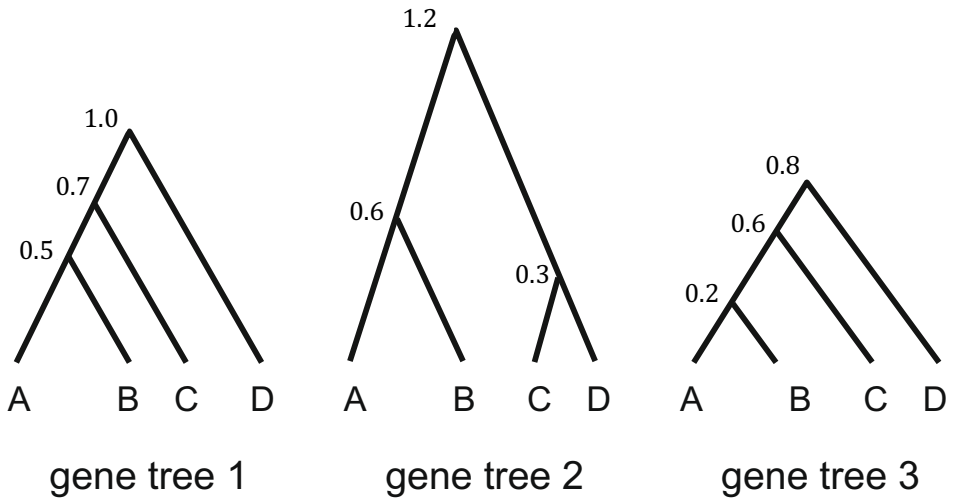


Even in the tree on the left (congruence between GT & ST), the coalescences are earlier than the divergence times (τ_x).

If we can summarize coalescence times for all pairs of taxa and across all sampled loci, we can estimate the timing of speciation events and therefore the species tree.

GLASS (Global Latest Split: Mossel, E., Roch, S., 2007. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. Available from: http://arxiv.org/abs/0710.0262.) takes the approach that the minimum coalescences will follow the specie tree.

(a)

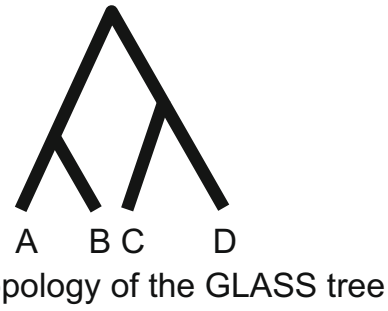


So we have three gene trees with times to coalescence estiamted for each.

We erect a piarwise matrix that fills each cell with the minimum time to coalescence across all of the gene trees.

(b)

	A	B	C	D
A	--	0.2	0.6	0.8
B	0.2	--	0.6	0.8
C	0.6	0.6	--	0.3
D	0.8	0.8	0.3	--



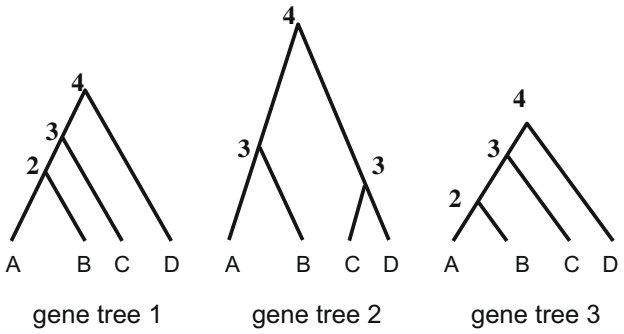
This matrix is subject to clustering (actually, UPGMA) to derive the GLASS estimate of the species tree.

Liu et al. (2009. Syst. Biol. 58:468) proposed estimating the species tree using the average ranks of coalescence times across genes (**STAR**).

We have the same three gene trees, but now we rank the coalescence times, beginning by

assigning the root a rank of *n* (with *n* terminals).

(a)



Again, a pairwise matrix is erected, here with the cells containing twice the average rank.

(c)

	A	B	C	D
A	--	4.66	6.66	8
B	4.66	--	6.66	8
C	6.66	6.66	--	7.34
D	8	8	7.34	--

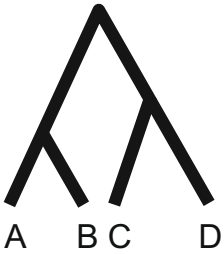
topology of the STAR tree

The species tree is estimated by subjecting the matrix to NJ.

In the **STEAC** approach, the matrix is filled by the 2X the average coalescence time for each pair of taxa across genes.

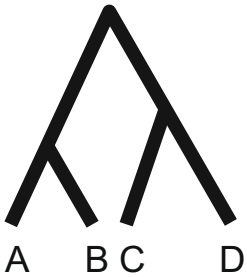
(d)

	A	B	C	D
A	--	0.86	1.86	2.2
B	0.86	--	1.86	2.2
C	1.86	1.86	--	1.4
D	2.2	2.2	1.4	--



topology of the STEAC tree

NJST uses the average node-distances.

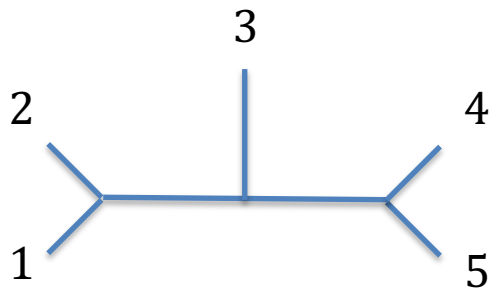


Topology of the NJST

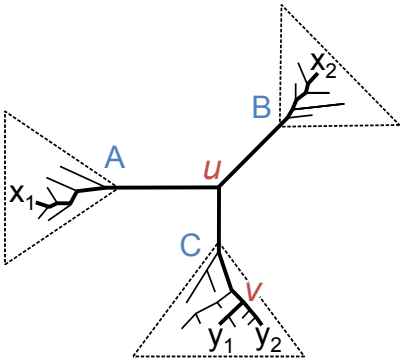
E. Quartets Approaches.

A couple of novel approaches go back to the quartets approach we discussed earlier this semester, and these have become widely used.

{1,2,3,4}	{1,2,3,5}	{1,2,4,5}	{1,3,4,5}	{2,3,4,5}
((1,2)3,4)	**((1,2)3,5)**	**((1,2)4,5)**	**((1,3)4,5)**	**((2,3)4,5)**
((1,3)2,4)	((1,3)2,5)	((1,4)2,5)	((1,4)3,5)	((2,4)3,5)
((1,4)2,3)	((1,5)2,3)	((1,5)2,4)	((1,5)3,4)	((2,5)3,4)



These first of these (ASTRAL; Mirarab et al. 2014. Bioinformatics) leverages proofs that collections of unrooted gene trees will permit consistent estimation of the species tree (e.g., Allman et al. 2011. J. Math. Biol. 62:333; Degnan. 2014. Syst. Biol. 62:574).

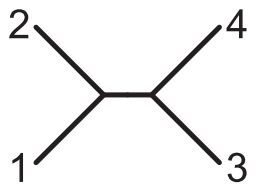


So, for the unrooted gene tree, the quartet tree (in bold) maps to internal nodes *u* and *v*. ASTRAL estimates the species tree by finding the internal nodes in the gene trees to which most quartets map.

Chipman and Kubatko (2014. Bioinf.) have developed a quartet-based approach to inferring the species tree from SNP data: SVDQuartets.

This is important as high throughput approaches such as ddRAD-Seq may be the most cost-effective method for generating genome-wide genetic data, in that data are in the form of SNPs.

The idea here is that for any quartet tree, there is a valid bipartition and two invalid bipartitions:



- Valid: 12|34
- Not valid: 13|24
  14|23

Infer the species tree in the 4-taxon case from SNP data using coalescent theory and a GTR model of sequence evolution.

For each resolution of each quartet, we can use the frequencies of each site pattern that is consistent with it as a measure of its support.

$$Flat L_1|L_2(\widehat{P})$$

$$\begin{pmatrix} \hat{p}_{AAAA} & \hat{p}_{AAAC} & \hat{p}_{AAAG} & \hat{p}_{AAAT} & \hat{p}_{AACA} & \cdots & \hat{p}_{AATT} \\ \hat{p}_{ACAA} & \hat{p}_{ACAC} & \hat{p}_{ACAG} & \hat{p}_{ACAT} & \hat{p}_{ACCA} & \cdots & \hat{p}_{ACTT} \\ \hat{p}_{AGAA} & \hat{p}_{AGAC} & \hat{p}_{AGAG} & \hat{p}_{AGAT} & \hat{p}_{AGCA} & \cdots & \hat{p}_{AGTT} \\ \hat{p}_{ATAA} & \hat{p}_{ATAC} & \hat{p}_{ATAG} & \hat{p}_{ATAT} & \hat{p}_{ATCA} & \cdots & \hat{p}_{ATTT} \\ \hat{p}_{CAAA} & \hat{p}_{CAAC} & \hat{p}_{CAAG} & \hat{p}_{CAAT} & \hat{p}_{CACA} & \cdots & \hat{p}_{CATT} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{p}_{TTAA} & \hat{p}_{TTAC} & \hat{p}_{TTAG} & \hat{p}_{TTAT} & \hat{p}_{TTCA} & \cdots & \hat{p}_{TTTT} \end{pmatrix},$$

We can represent this with a Singular Value Decomposition, *SVD (L_1|L_2)*. The true resolution of the quartet is the one with the lowest SVD score.

For very large data sets a random sample of (say 100,000) quartets can be used to estimate the species tree.