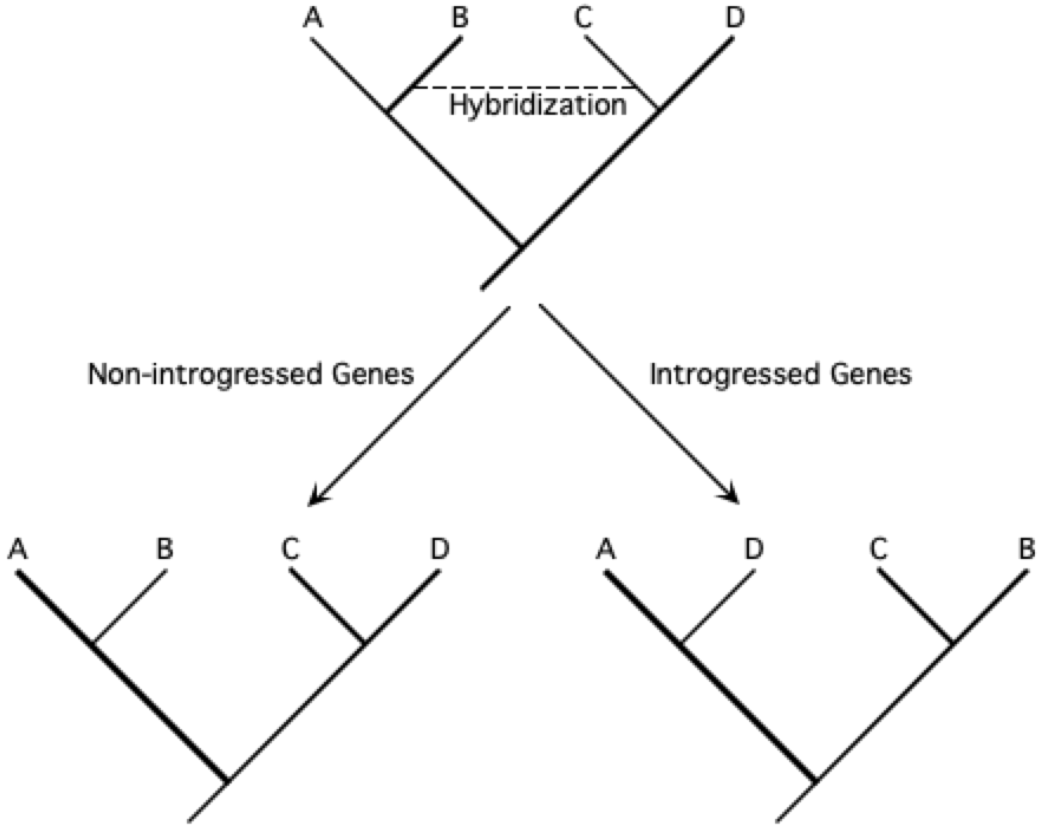
Lecture 19 – Network Based Methods

I. Introduction - Given that there are several **evolutionary processes** that may generate a complex, non-bifurcating history, increasing attention is focused on historical inferences that account for this. Specifically, non-vertical transmission can result in reticulating histories.

A. Hybridization – The early species concepts (at least for metazoans) focused on reproductive isolation as a criterion. Thus, animal phylogenies at or above the species level have classically been assumed to be strictly bifurcating. This is not true for plant phylogenies.

However, increasing molecular evidence indicates that hybridization is far more common in animals than has traditionally been recognized.

We expect there to be conflicting signal in cases of hybridization between non-sister taxa.



Hybridization between sister-taxa has less impact on expected tree topology, although it will certainly lead to conflicting signal with respect to branch lengths and reciprocal monophyly between species. It's a very cool and important topic but doesn't have the phylogenetic implications that hybridization between non-sister taxa has.

Hybridization may be current, or it may be historical. The more recent it is, the easier it is to differentiate from incomplete lineage sorting.

B. Horizontal Gene Transfer – There has long been evidence that microbial systems are subject to horizontal gene transfer.

This is somewhat analogous to hybridization in eukaryotes, in that it results in some genes having a history of the fundamental (vertical) bifurcating relationships and some genes having a history of a hetero-specific (horizontal) exchange.

A major difference between these is that HGT is usually used to refer to reticulations among potentially strongly diverged organisms (i.e., large spans of evolutionary time). This is often plasmid-mediated, so its mechanisms are different than hybridization.

However, the effect is essentially the same in terms of phylogenetics, but it may be easier to detect HGT because the conflict should be stronger, at least in some instances.

Since it's becoming clear that circumstances arise which generate non-tree like histories, methods have been developed that try to assess historical signal without imposing a bifurcating tree.

II. Split Decomposition – Network-based methods incorporate reticulations into the network at points where the data have conflicting signals.

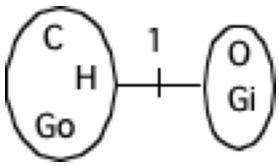
Split Decomposition was developed by Bandelt & Dress (1993. p. 123 in: Information and Classification, Opitz, Lausen, & Klar, ed., Springer-Verlag).

The idea is that the data are not forced into a single tree. Instead, each site is examined for the split that it supports.

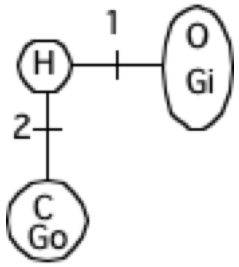
Let's take the following example:

```
Human       T C C T T A A A A
Chimp       T T C T A T A A A
Gorilla     T T A C A A T A A
Oranutan    C C A C A A A T A
Gibbon      C C A C A A A A T
```

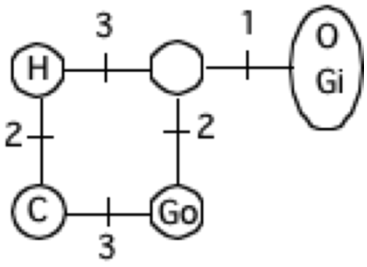
Site 1 splits {H, C, Go} {O,Gi}:



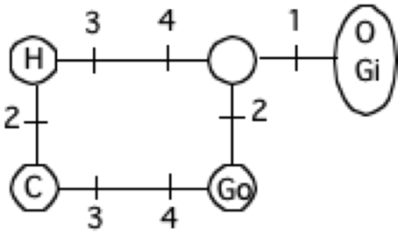
Site 2 splits {C,Go}{H,O,Gi}:



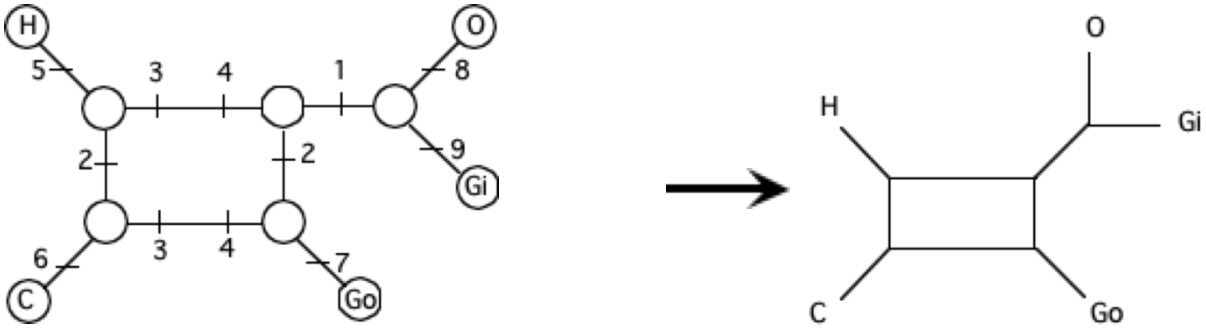
Site 3 conflicts with site 2 by splitting {H,C} from the rest, so we introduce a cycle (or reticulation) in the graph:



Site 4 does the same:

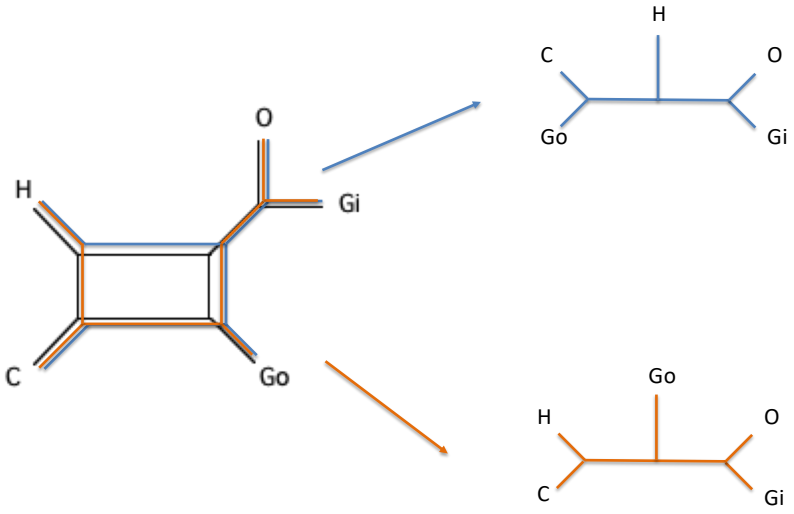


Sites 5 – 9 define terminal splits:



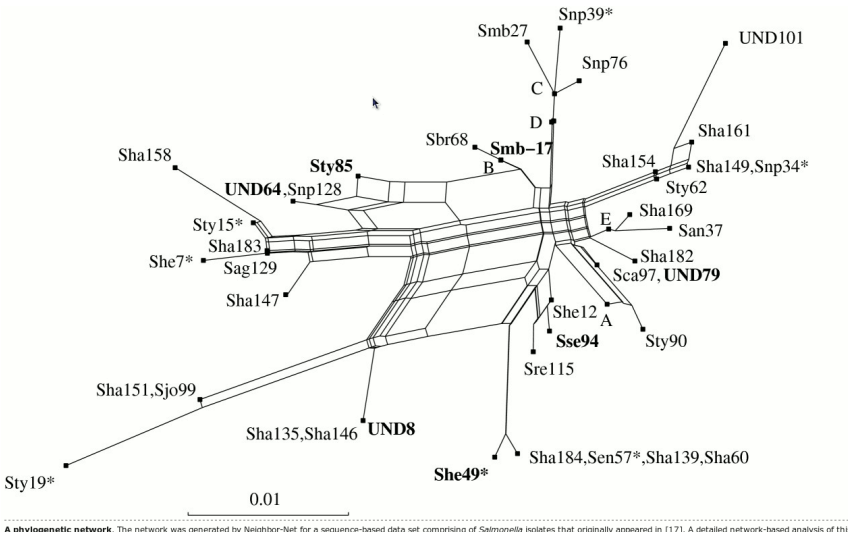
The partition that is not conflicted in the data, {H,C,Go} {O,Gi}, is represented by a bifurcation, but the conflict regarding resolution of {H,C,Go} is represented by a cycle.

So, there are two trees represented by the network, and the length of the branches represents the strength of the support for each (orange tree has more support).



In this case, this represents homoplasy, but it the method is very good as detecting conflict induced by any of the processes we discussed above.

These can get very complex: NeighborNet from *Salmonella*



They can be less complicated and locate where the deviation from a bifurcating tree is.

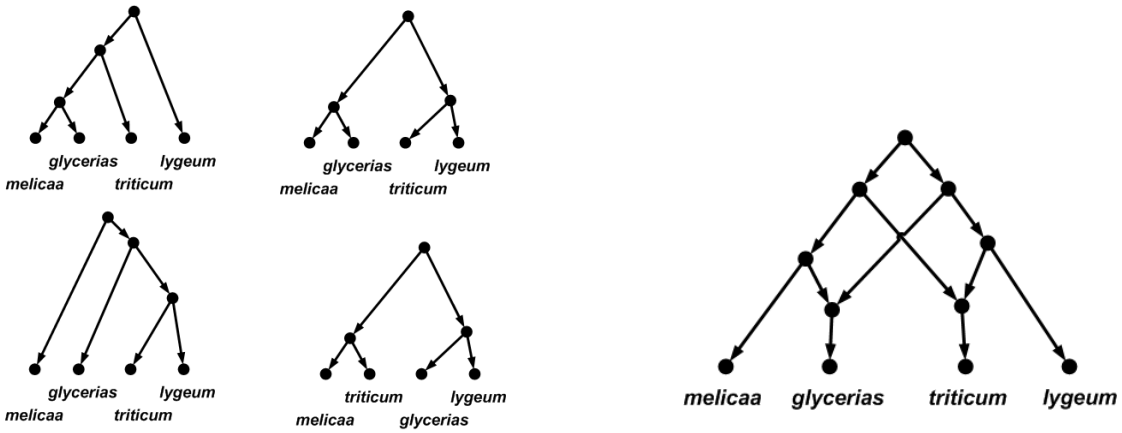
However, the method does not try to infer the processes (i.e., ILS or introgression) of to assess concerted homoplasy (systematic error).

III. Phylogenetic Networks

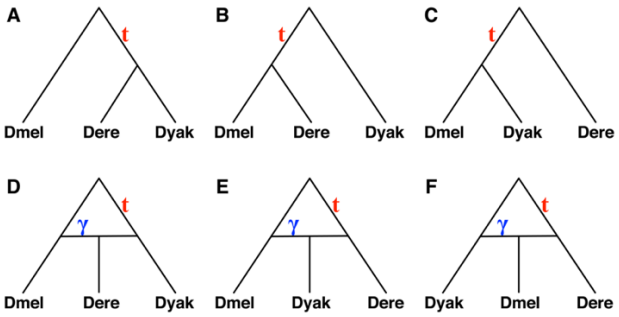
An increasing number of studies attempt to use reticulating graphs (networks) to infer hybridizations.

The idea is that different parts of the genome may be derived from different parent taxa and a network can "contain" several gene trees.

In a parsimony framework, we can search for the network that requires the fewest reticulation events but still contains all the gene trees.



Yu et al. (2012. PLoS Genetics) provided the likelihoods for gene trees evolving within a phylogenetic network.



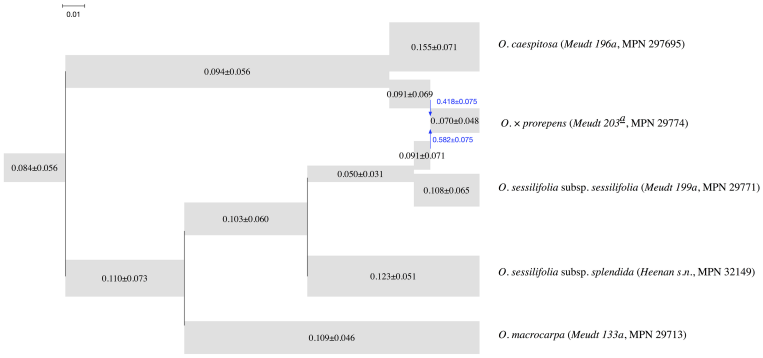
Of course, adding reticulations will always improve the likelihood score, so some type of model selection is required.

Species phylogeny	$-\ln L$	t	γ	AIC	AICc	BIC
Figure 3A	9070	0.46	N/A	18143	18143	18150
Figure 3B	10233	$1E-10$	N/A	20469	20469	20476
Figure 3C	10233	$1E-10$	N/A	20469	20469	20476
Figure 3D	9045	0.58	0.11	18095	18095	18109
Figure 3E	9070	0.46	0.0	18145	18145	18159
Figure 3F	10233	$1E-10$	0.0	20471	20471	20485

doi:10.1371/journal.pgen.1002660.t003

Gamma represents proportion of the genome of hybrid taxa that derive from the minor parent, and its estimation requires computation of partial likelihoods for all possible pathways on the network.

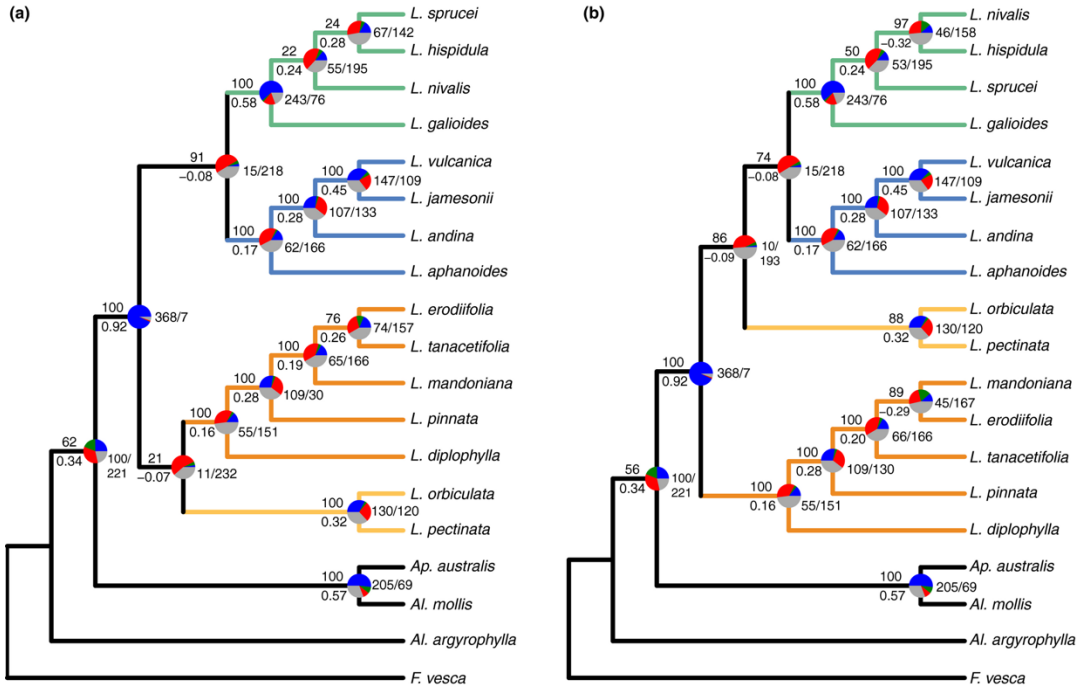
PhyloNET includes a Bayesian implementation (Zhu et al. 2018. PLoS Comp. Biol.) that applies the multispecies coalescent that accounts for ILS and therefore estimates a species network.

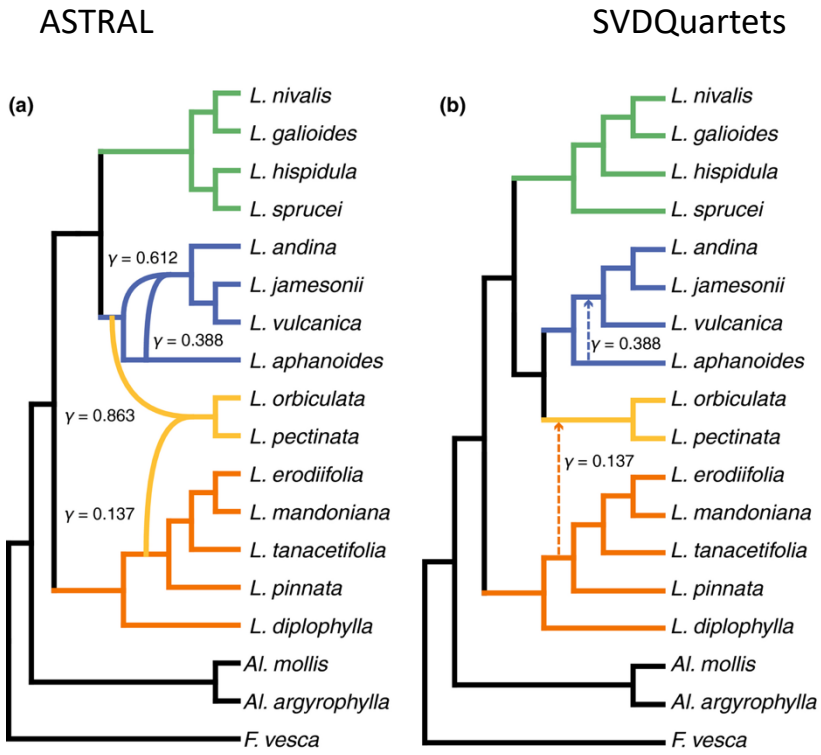


If we set the number of reticulations to one, we can estimate variable effective population sizes and inheritance coefficients.

The two examples shown are really easy examples, in that there's just a single reticulation an it's between tip taxa.

It's possible, but difficult, to estimate more complex histories. One of the best I've seen is from Diego's dissertation in *Lechemilla* (Morales-Biones et al. 2018. New Phytologist).

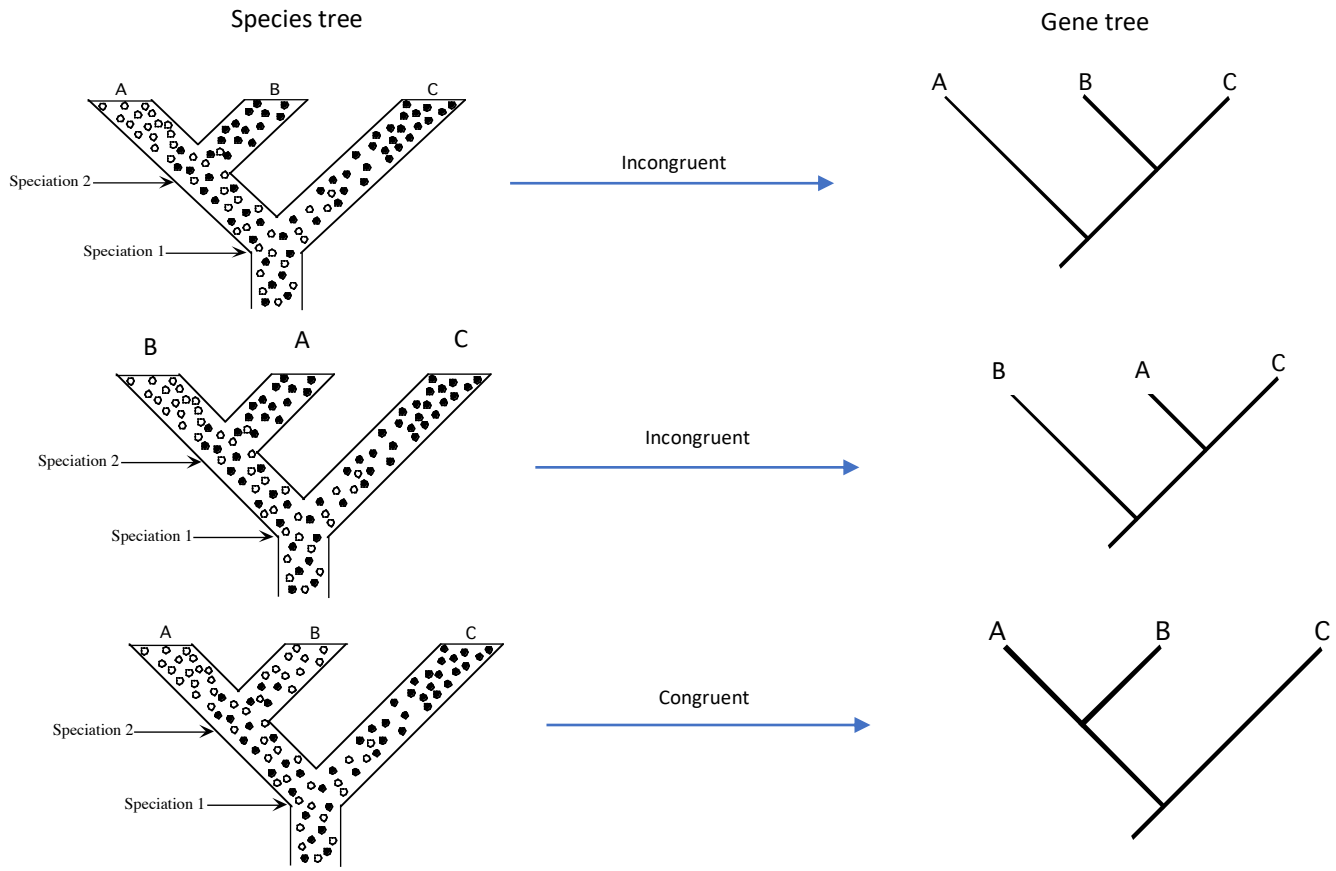




Species network – two ancient hybridization events.

IV. Differentiating ILS from Hybridization using quartets.

Remember (from Lecture 18) that, if ancestral polymorphisms persist across two successive speciation events, coalescent stochasticity results in three potential gene trees:



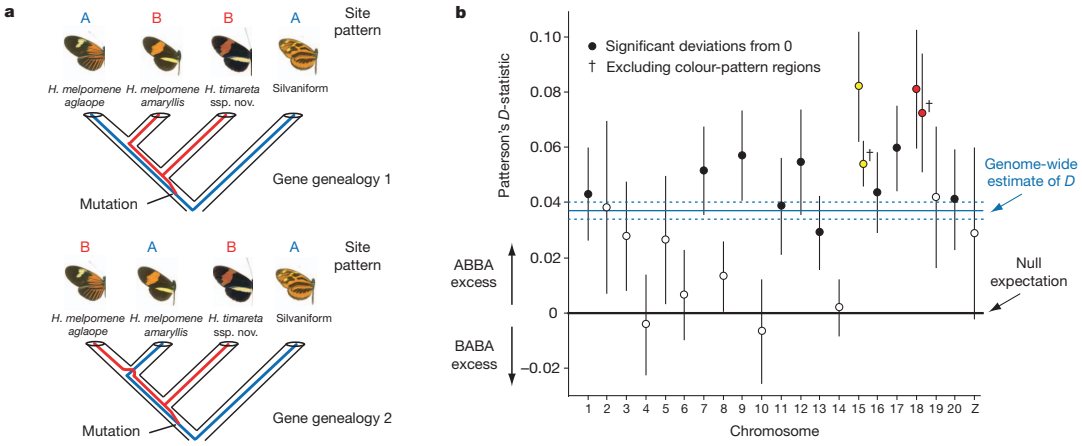
Furthermore, if ILS is the only source of gene-tree/species-tree incongruence, these are all expected to occur with equal frequency.

A. ABBA/BABA tests

Thus, if we have an outgroup to provide for inference of the states in the ancestor, we can make inferences about the frequencies of these gene trees from the site patterns in our data.

Let's assume that the polymorphism is due to a single mutation (i.e., there have been no multiple hits at the polymorphic site).

Under ILS, there should be no asymmetry in frequency of sites with pattern ABBA versus those with the pattern BABA.



Patterson's D measures the deviation from equal frequencies:

$$D = \frac{C_{ABBA} - C_{BABA}}{C_{ABBA} + C_{BABA}}$$

This has become quite widely used, but it carries the requirement that the species tree has been estimated, in addition to its inherent assumption of not multiple hits.

B. HyDe & SVDquartets

This quartet framework aligns very well with that of the last species-tree estimation approach we discussed.

In fact, an intermediate step in species-tree estimation under this approach is to calculate the scores for all three resolutions of each quartet, which obviously includes the resolutions that are incongruent with the species tree.

We can illustrate this with the chipmunk data you used in the last lab.

HyDe detects no introgression.

But BPP (in the hands of its developers) does.