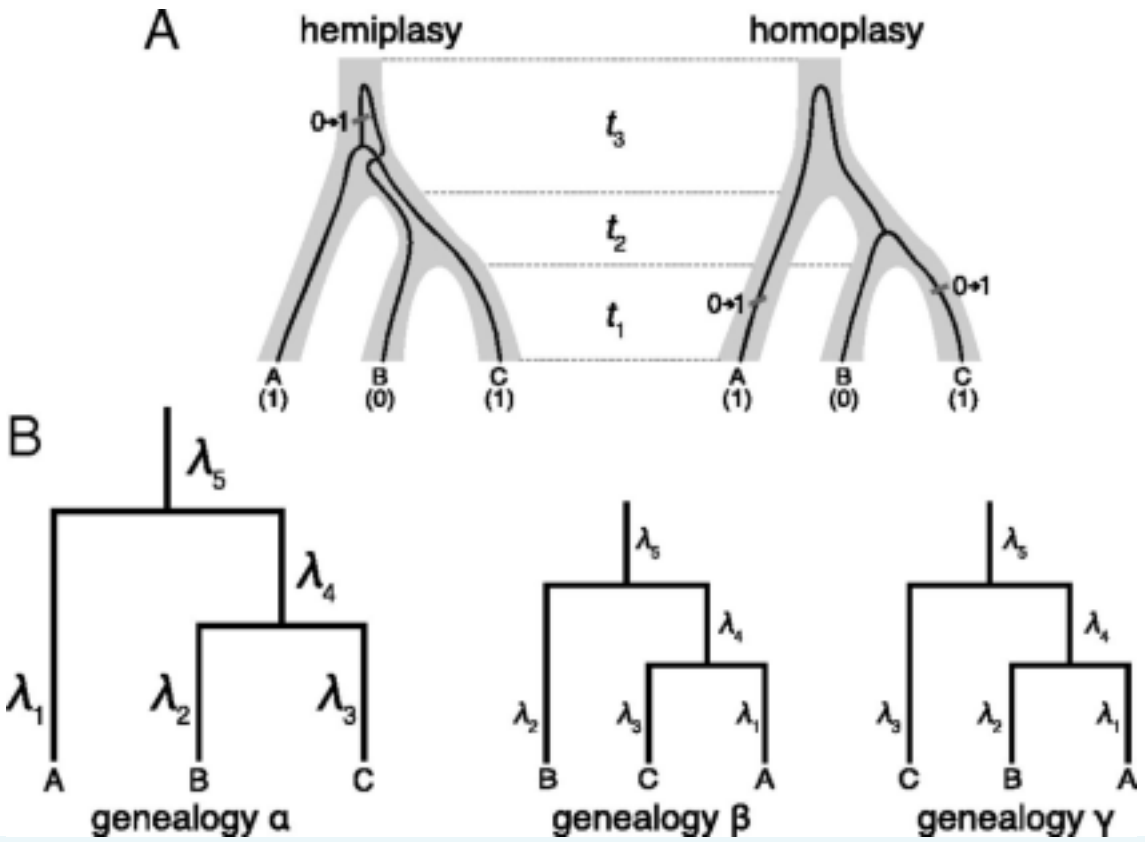
Complications to phylogenetic inference



EEB 5350

Lecture 13

Eric Gordon

Outline

- “Hard” problems in phylogenetics
  - Long branch attraction
  - Differential gene loss of paralogs
  - Incomplete lineage sorting (hemiplasy)
  - Horizontal gene transfer (Thursday lecture)
- Coalescent theory
- Statistical inconsistency of concatenation
- Coalescent phylogenetic programs
  - Those based on resolved gene trees
  - Those based on concurrent estimation of gene and species trees
  - Site-pattern methods
- Problems with coalescent methods and statistical binning
- Using coalescent theory for species delimitation

Phylogenies straight forward?

- If all sites evolved at the same rate within molecules and throughout the history of lineages, if all nucleotides were in equal proportion, if any nucleotide or amino acid evolved to any other with equal probability, if all taxa could be sampled, and if diversification happened at well-spaced intervals, then phylogenetic tree building would be easy--but it's not.

- Still would have at least four "hard" problems:
  - Long branch attraction (homoplasy overwriting true signal)

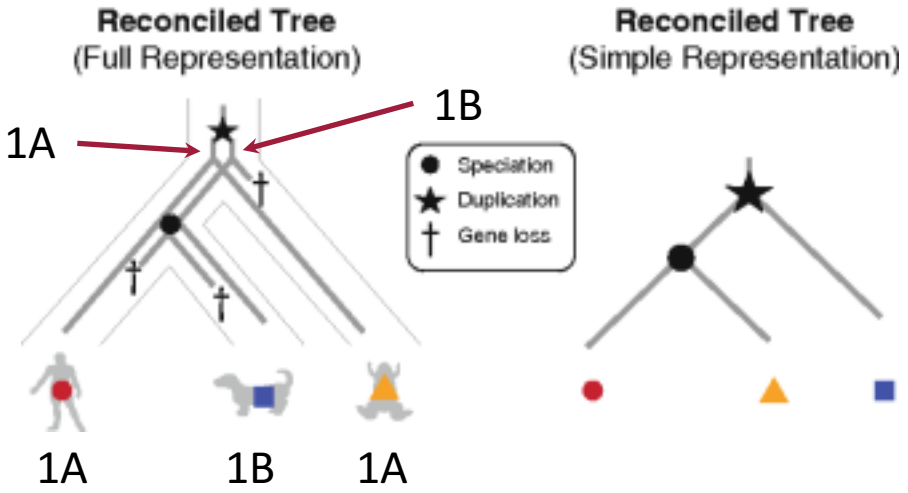
  Those related to gene trees having inconsistent histories with species trees
  - Gene duplication/extinction or paralogous sampling
  - Incomplete lineage sorting or deep coalescence (hemiplasy)
  - Reticulation (horizontal gene transfer, sex, recombination and hybridization)

Long branch attraction

- Two lineages with high substitution rates are falsely reconstructed to be sisters to each other due to incorrect inference
- Because of only four states that rate of convergence is relatively high
- Maximum likelihood is more robust to reconstruction artifacts but is still susceptible in difficult cases
- How to fix?
  - Not a real way, but removing both potential taxa and rerunning the analysis to look for the same (highly supported) relationships is commonly used.
  - Can add in taxa to break up branches if possible

Differential gene loss of paralogs

- Gene 1 is duplicated at base of tree
  - 1A retained in circle and triangle (not sister groups) and 1B lost twice
  - 1B retained in square and 1A lost
- Accurate phylogeny of gene 1 is that square is sister to circle + triangle
- But speciation history is different

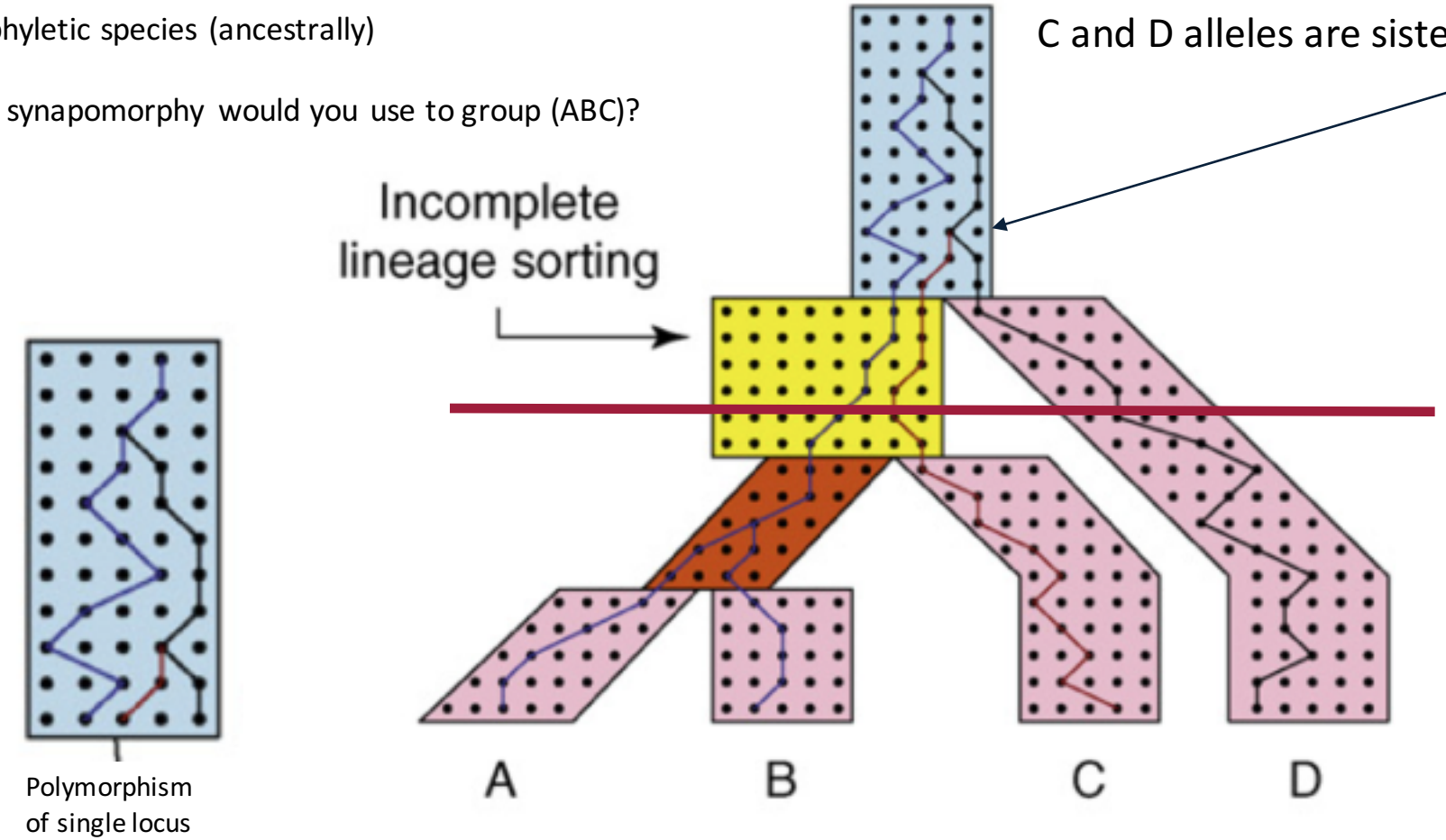


Altenhoff, A. M., & Dessimoz, C. (2012). Inferring orthology and paralogy. In *Evolutionary genomics* (pp. 259-279). Humana Press, Totowa, NJ.

Incomplete lineage sorting

Paraphyletic species (ancestrally)

What synapomorphy would you use to group (ABC)?



James H. Degnan , Noah A. Rosenberg
Gene tree discordance, phylogenetic inference and the multispecies coalescent
Trends in Ecology & Evolution, Volume 24, Issue 6, 2009, 332 - 340
http://dx.doi.org/10.1016/j.tree.2009.01.009

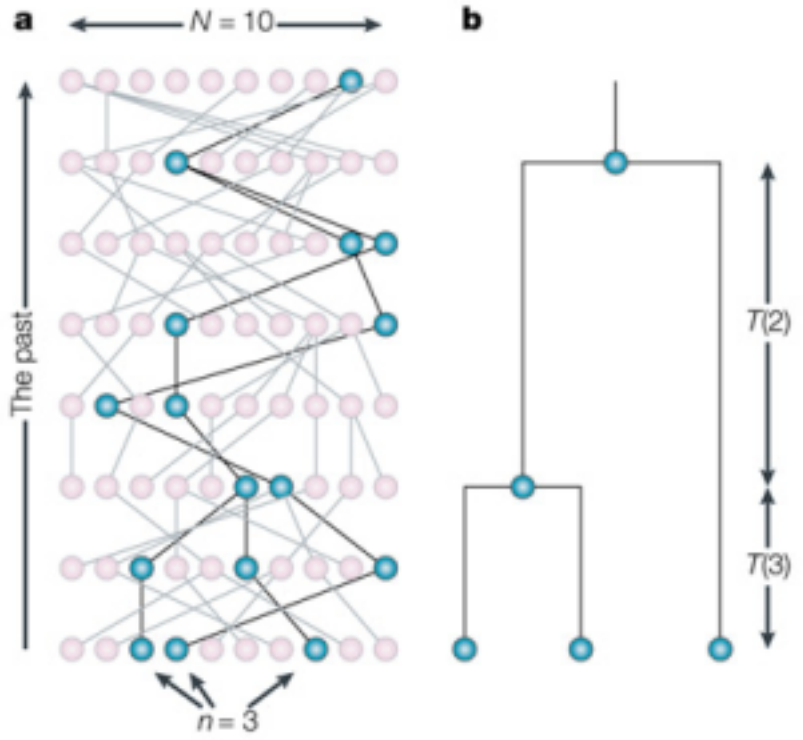
Can species be paraphyletic?

- Certain members of one species are more closely related to another group than to other members of its own species

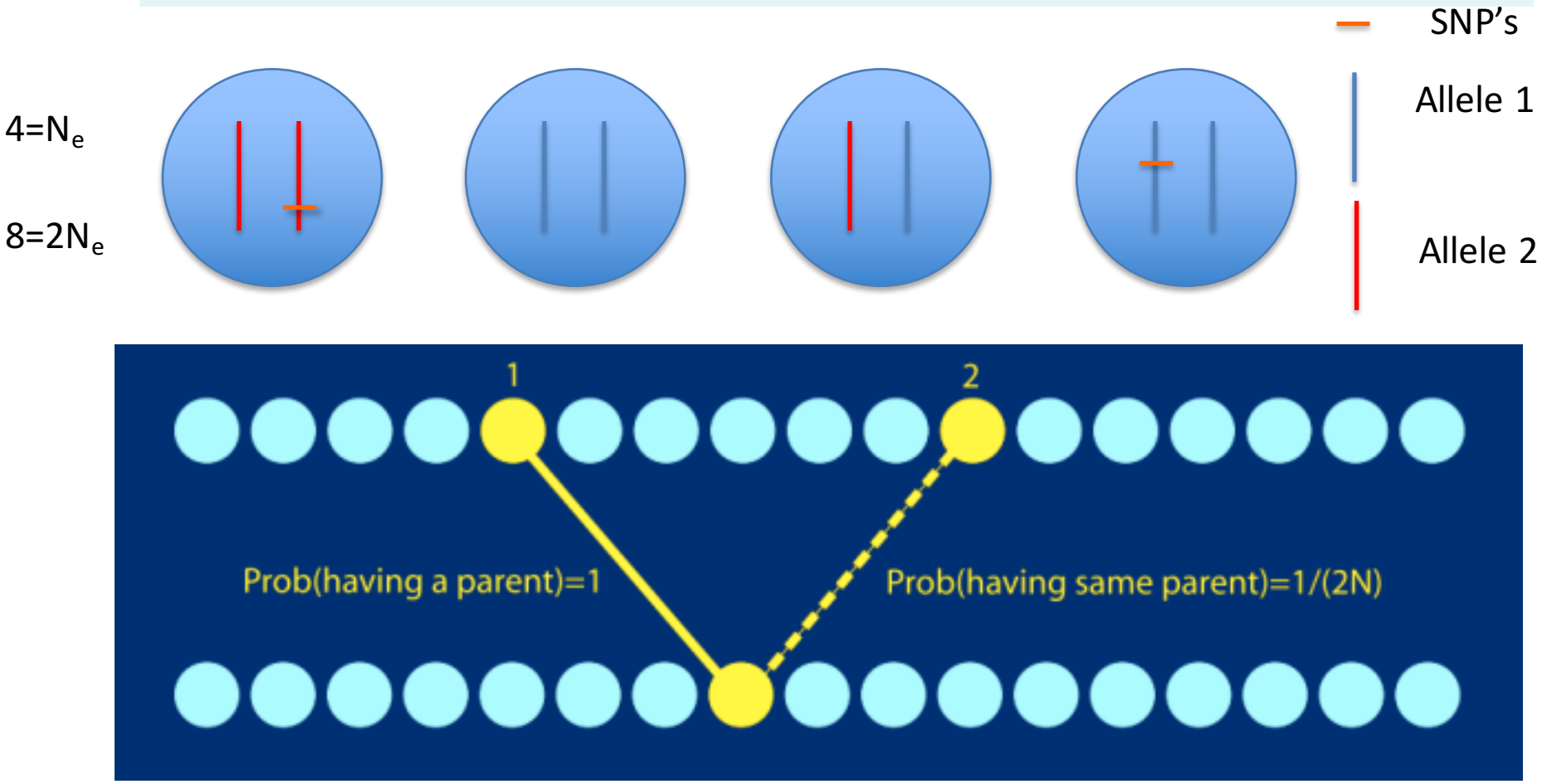
- One hypothetical example
  - Almost worldwide species colonizes a new island and rapidly adapts and is reproductively isolated from founder population
  - Founder population is paraphyletic with respect to this new species until panmictic mating causes coalescence within this nearly global population (dependent on effective population size which is very large)
  - If additional speciation events occur within this time period before reciprocal monophyly, then called incomplete lineage sorting/deep coalescence

Coalescent theory in population genetics

- John Kingman in 1982
- Natural extension to Fisher-Wright ecological null model (neutral evolution)
- Assumes
  - constant effective population size (N_e)
  - no recombination among alleles
  - random mating
  - no selection
  - constant mutation rate (normal genetic drift)
- Used to reconstruct "antecedents" (ancestors) for different alleles (orthologs) within a population



Rosenburg and Nordborg 2002



Probability that two alleles "coalesce" in the previous generation is $1/2N_e$
Probability that they don't is $1-1/2N_e$

$$P_c(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \left(\frac{1}{2N_e}\right).$$

$$P_c(t) = \frac{1}{2N_e} e^{-\frac{t-1}{2N_e}}.$$

Population size is important

- For two alleles, the time it takes to coalesce is $2N_e$ generations (coalescent unit)
- Chances are that all alleles coalesce in $4N_e$ generations (2 coalescent units)

Violation of assumptions:

If population size varies across generations (likely), then bottleneck effects erase previous history

If selection and linkage vary expect larger amount of certain gene trees in contrast to the null model

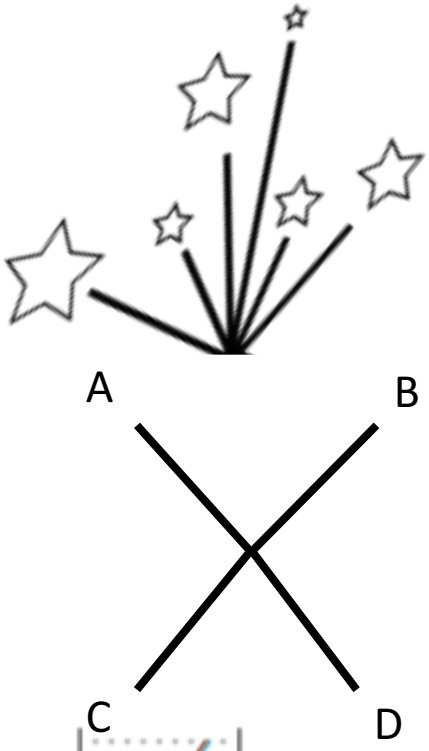
If mutation rates vary, ???

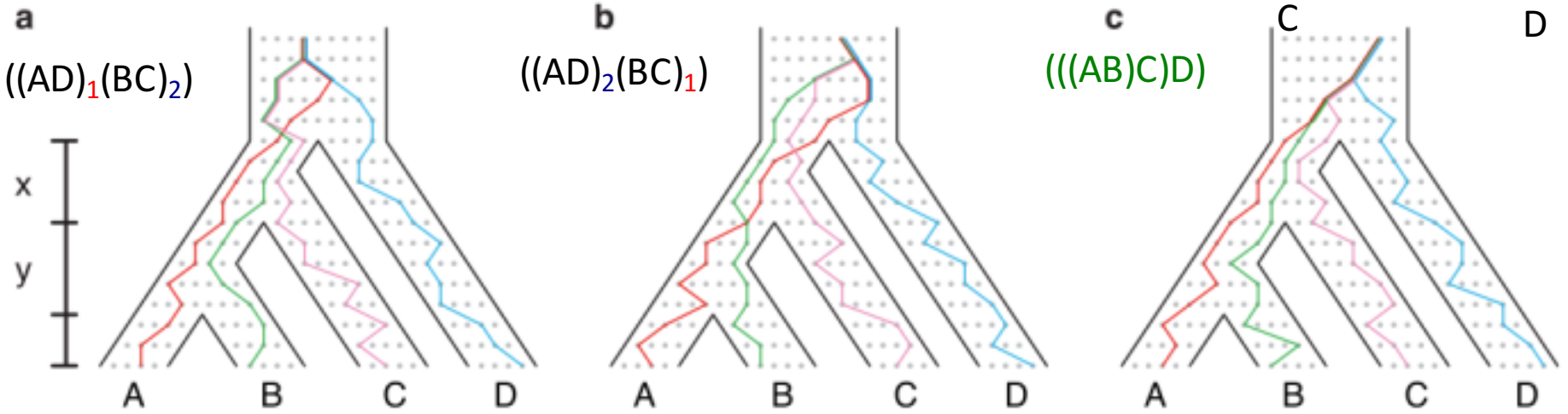
So why incorporate coalescence into a traditional phylogeny?

- Concatenation does not allow modeling of different gene histories.
- Since recombination certainly does occur between different concatenated loci. Inherent assumption of method is wrong.
- Certain genes may have entirely different histories (esp. mitochondrial)? Even homoplasy-free regions give different phylogenies.

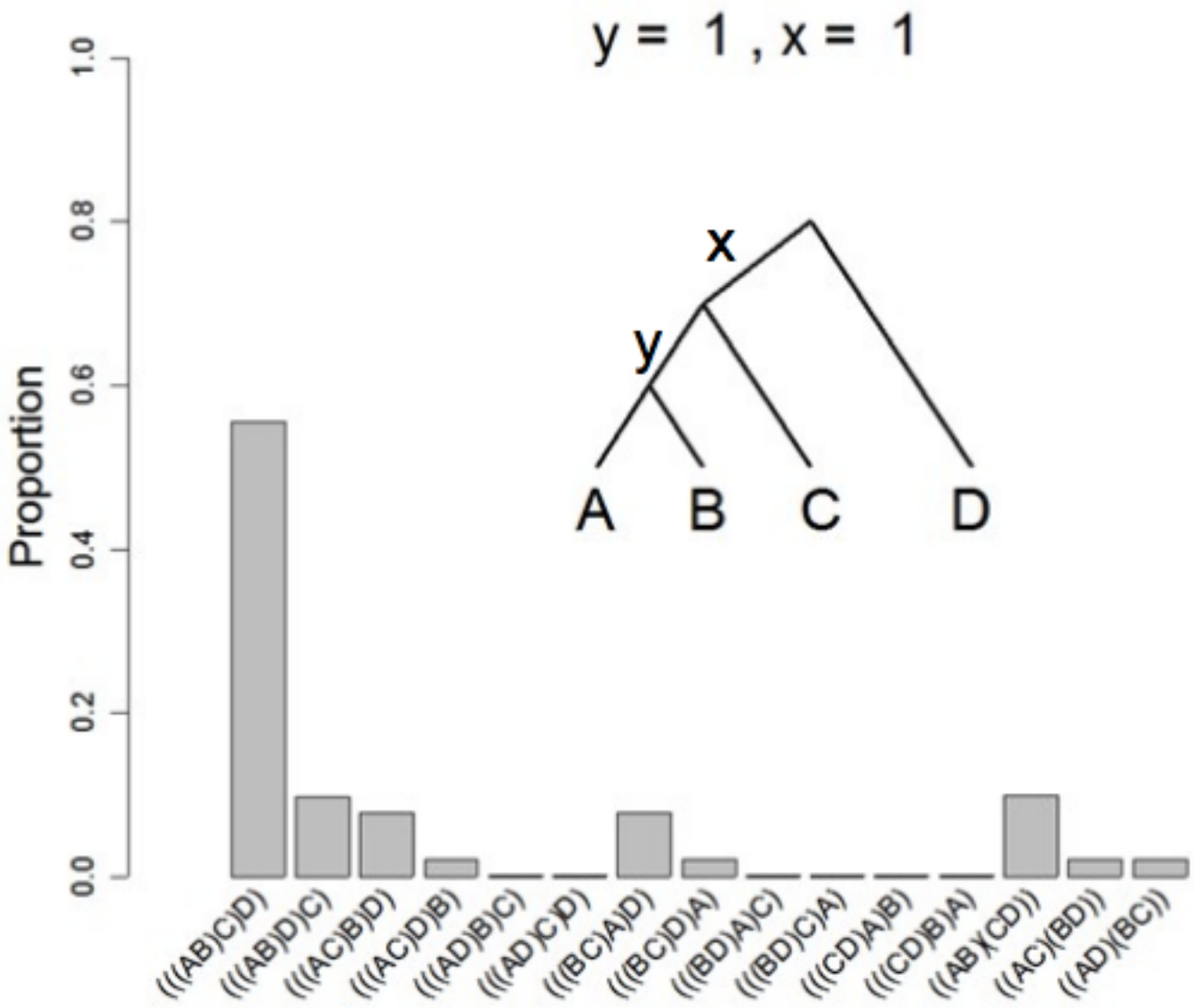
Anomaly zone

- Concatenation will give high support for wrong relationships in cases when there is a true polytomy (or very close to one)
- In some cases (anomaly zone), the most common gene tree is **wrong** (even without homoplasy!)
- Anomalous gene trees (AGTs)





“Positively misleading”



Degnan 2008

Other benefits of coalescent methods

- Natural genetic drift within populations. Species are more than one individual! Allows estimation of this parameter
- Genes have different phylogenies (not due to simple artifacts in reconstructions or homoplasy)
- Deep coalescence is especially common with large population sizes, short branches (i.e, rapid radiations or those under severe selective pressure)

Coalescent methods for phylogenetics

- Attempt to model possible different coalescent histories for genes given a species trees
  - Multispecies coalescent
  - Coalescent histories are independent and random
  - Coalescent events have to occur on species branches but can go back further in time than species divergences
- Often break tree into quartets topologies and estimate most consistent species tree
- Several kinds of methods:
  - Summary methods
  - Coestimation
  - Site patterns

Summary or “shortcut” coalescence methods

- Relies on individual gene tree topologies
- More accurate when based on large loci since homoplasy is less suppressive of phylogenetic signal in individual gene trees
- Can be used (with caveats) when genes are missing from certain taxa randomly
- Uses computational power to assess species tree fit to reconstructed gene trees
- Some estimate branch lengths
- STAR, STEAC, STEM, MP-EST, ASTRAL and ASTRID

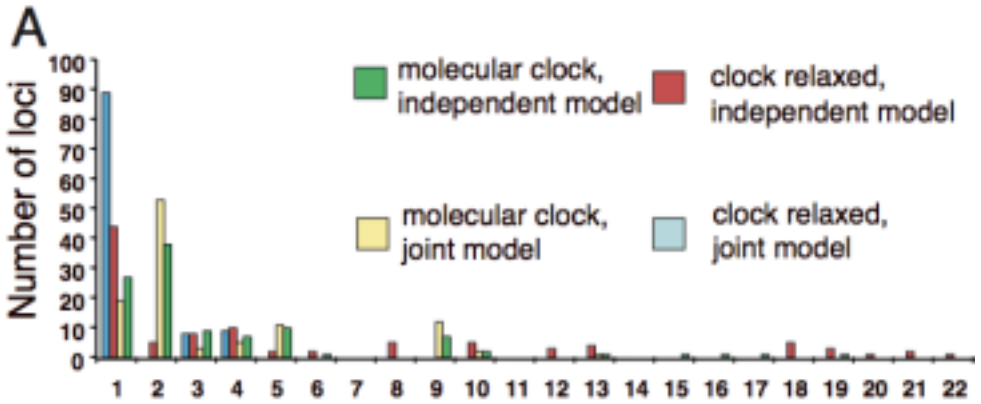
Liu, L., Wu, S., & Yu, L. (2015). Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution*, *53*(5), 380-390.

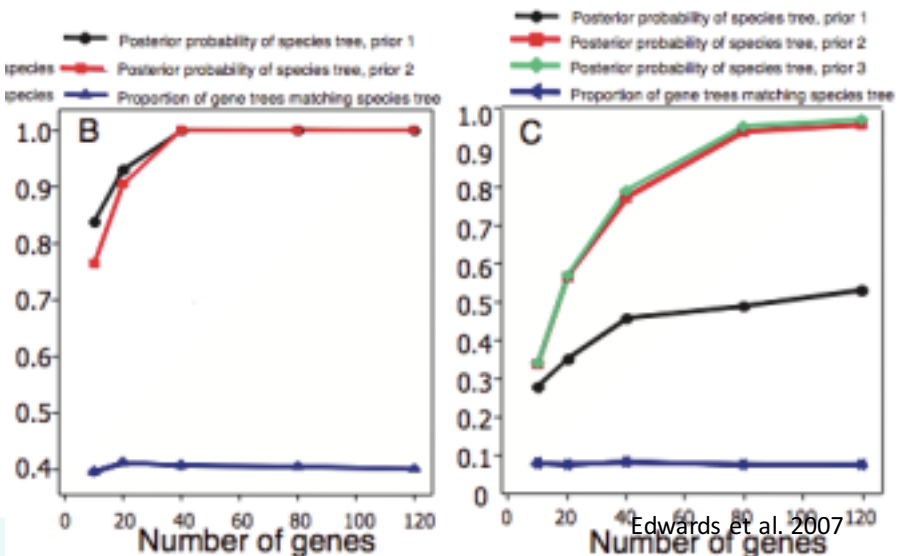
Bayesian Estimation of Species Trees and "star"BEAST

- BEST and *BEAST
- Estimate posterior distribution of gene trees given the data. The prior on gene trees is determined for distributions of all gene trees given species trees (considering all possible species trees under coalescence)
- Separate mutation rate per locus. Better results when assuming no molecular clock at all (for highly divergent taxa)
- Unfortunately, currently computationally limited to equivalent of ~20 taxa and 100 genes
- Multiple individuals per species

Edwards et al. 2007

- Used BEST on famous Rokas 2003 yeast dataset
- Found that coalescence methods were able to reconstruct phylogeny correctly with much fewer genes. (8 with high confidence as opposed to 20)
- Simulations found that coalescence will reconstruct correct tree when concatenation will construct wrong tree with increasingly high support





Edwards SV, Liu L, Pearl DK. High-resolution species trees without concatenation, Proc. Natl. Acad. Sci. USA , 2007, vol. 104 (pg. 5936-5941).

Edwards et al. 2007

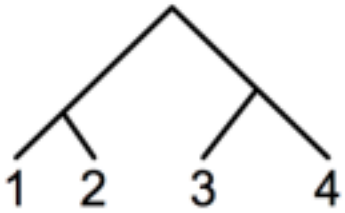
SVDQuartets and SVDQuest

- Different from most other approaches in only using shared site patterns
- Relies on tree symmetry generating same patterns of base pairs in sister taxa in quartets
- Matrix decomposition
- Works on data simulated under coalescent processes and robust to gene tree estimation error
- Particularly useful for RADseq data but need lots of data for an accurate reconstruction

Chifman, J. and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent, Bioinformatics, 30(23): 3317-3324.
Vachaspati, P., & Warnow, T. (2018). SVDquest: Improving SVDquartets species tree estimation using exact optimization within a constrained search space. *Molecular phylogenetics and evolution*, *124*, 122-136.

Summarize site patterns through matrices and quartets

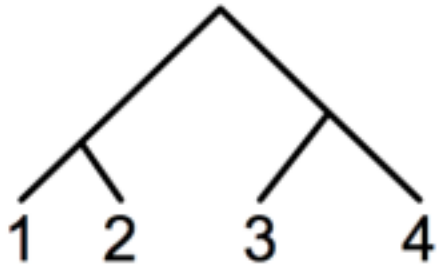
Methods – data representation



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & 5 & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & 2 & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

Methods – data representation



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & 5 & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & 2 & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

These two columns are identical – matrix rank is reduced by one

Process

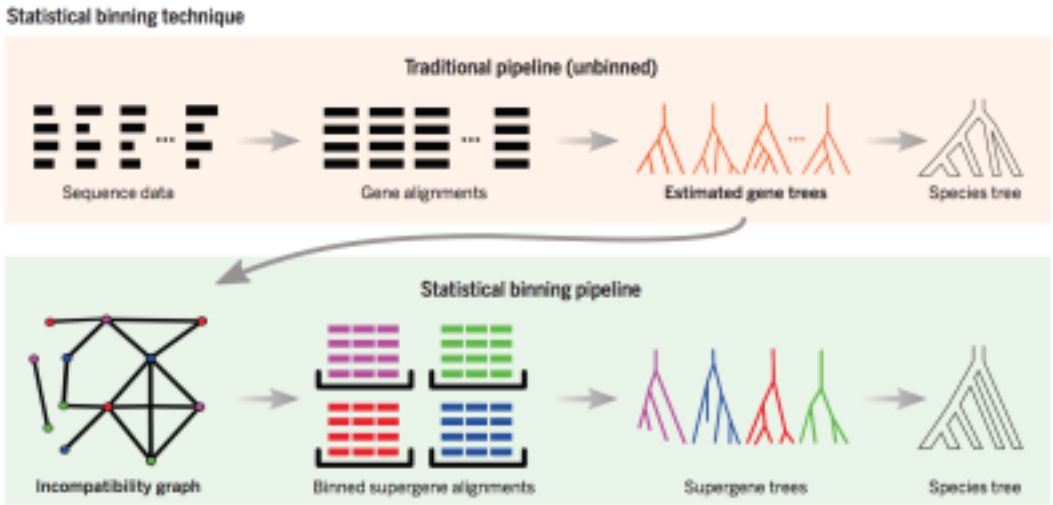
- Make matrix for each of three possible trees for four taxa
- Matrix “rank” of true quartet relationship should approach 10 (6 symmetrical columns can be flattened)
- Matrix of non sister taxa should be full rank of 16 independent columns
- Rank of matrices are estimated with “Singular Value Decomposition” score
- Pick the best quartet tree for all quartets or sample quartet sets from taxa
- Finally, use a quartet assembly method to build the species tree

Problems with coalescent methods

- Incomplete lineage sorting is inherently assumed to be responsible for different gene trees thus may be overestimated
- Gene trees reflect incorrect topologies due to homoplasy or other processes
- "Gene trees" may be based on data that is spatially very far apart in genome and thus may combine data that has separate evolutionary history due to recombination (sometimes called c-genes)
    - Simulations showed that unrecognized recombination has only a minor effect (Lanier and Knowles 2012)
    - Recombination within loci is okay in long branches since the history of recombining loci are the same once alleles have coalesced in a long branch (chances are by $5N_e$ generations)
    - Recombination during retention of ancestral polymorphism is definitely a problem but how likely?
- A single resolved species tree fails to visualize complex history

Is “concatalescence” a problem or a solution?

- Frequently reconstruct a different gene tree for each partition (this is not reflective of true gene trees but rather homoplasy).
- Simulated data with naïve lumping of loci together gave better results and allowed current version of *BEAST to work on larger datasets (Bayzid and Warnow 2013)
- Should gene trees reconstructed as having the same or very similar histories be lumped together to swamp out potential homoplasy within individual genes?
- “Statistical binning”…combine gene trees with similar history..shown to improve estimation of phylogeny



The statistical binning pipeline for estimating species trees from gene trees. Loci are grouped into bins based on a statistical test for combinability, before estimating gene trees.

Bayzid, M. S., & Warnow, T. (2013). Naive binning improves phylogenomic analyses. *Bioinformatics*, *29*(18), 2277-2284.

Mirarab, S., Bayzid, M. S., Boussau, B., & Warnow, T. (2014). Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science, 346(6215), 1250463.

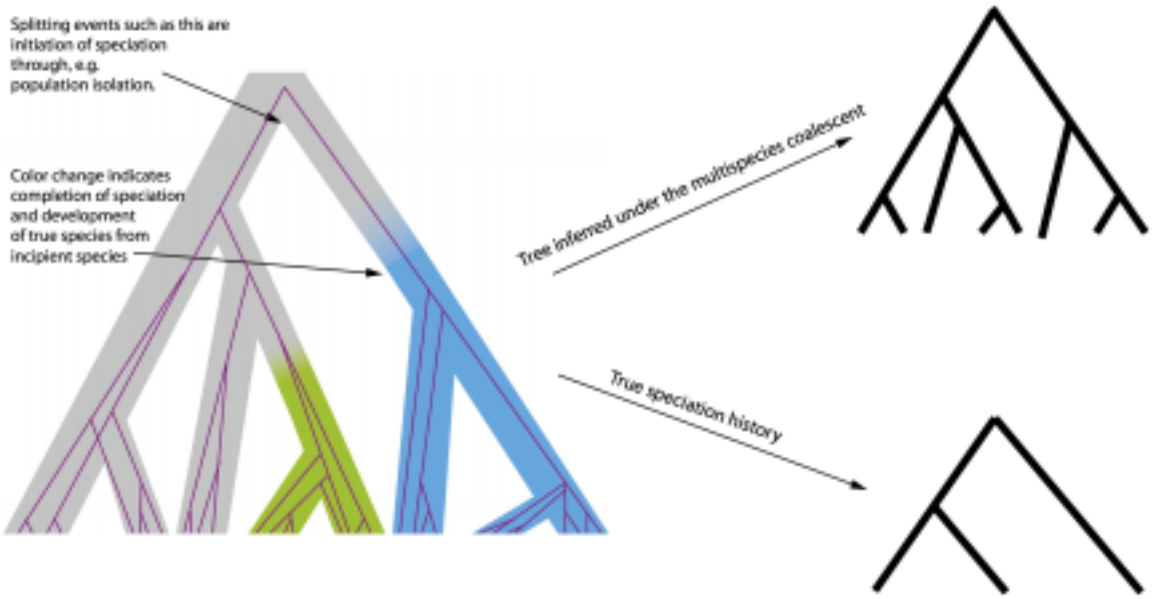
Coalescent models for species delimitation

- GMYC-Generalized mixed Yule-coalescent model (implemented in Bayesian programs like BPP)
- DNA-based taxonomy (e.g., cryptic species or morphologically impoverished semaphoronts)
- Models probability of tokogenetic "net-like" coalescence versus cladogenetic branching at each node and finds threshold of speciation across ultrametric tree
- Better than arbitrary limits e.g, 97% difference between COI species

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. Current Zoology, 61(5), 854-865.

Problems with GMYC

- Incorrect phylogenetic inference will bias results
- Requires several genes, many individuals sampled per population and does not work well on rapid radiations
- Can incorporate different threshold for different branches of tree
- Can not delimit incipient speciation from any reproductively-isolated population e.g, initiation of speciation events do not always result in new species thus always overestimates species numbers



Reid, N. M., & Carstens, B. C. (2012). Phylogenetic estimation error can decrease the accuracy of species delimitation: a Bayesian implementation of the general mixed Yule-coalescent model. BMC evolutionary biology, 12(1), 196.
Sukumaran, J., & Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Sciences, 114(7), 1607-1612.

Summary

- Phylogenetic estimation is hard for various reasons and does not always follow a single tree
- Coalescent methods for phylogenetic reconstruction are a promising method of overcoming incomplete lineage sorting
  - Allow for gene trees to vary in topology
  - Allows estimation of population size
  - More accurate reconstruction of species' histories
- But they may overestimate rate of incomplete lineage sorting
- SVDQuartets method is robust to gene tree error thus solving one of the problems with coalescent methods and becoming more commonly used
- Can use modeling of coalescent process to delimit species based on molecular data