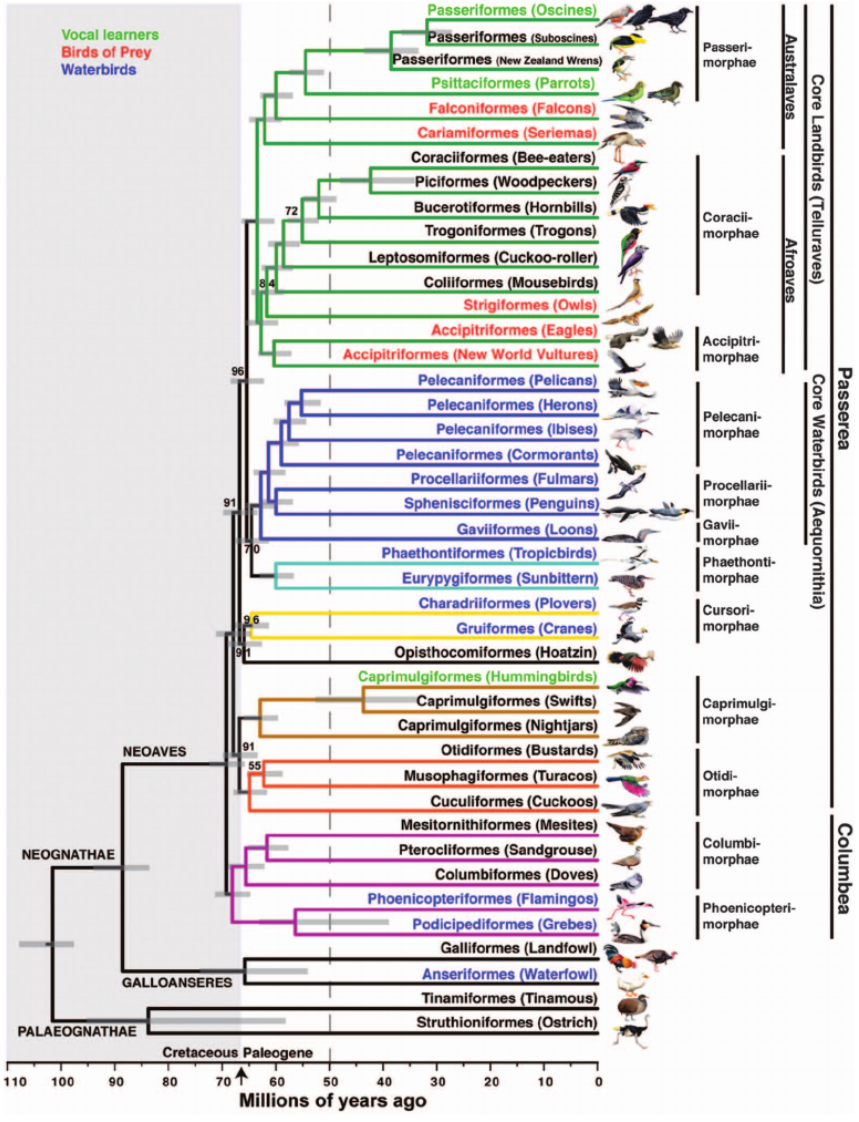
Handling discordance in phylogenetic inferences

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu

Fig. 1. Genome-scale phylogeny of birds. The dated TENT inferred with ExaML. Branch colors denote well-supported clades in this and other analyses. All BS values are 100% except where noted. Names on branches denote orders (-iformes) and English group terms (in parentheses); drawings are of the specific species sequenced (names in table S1 and fig. S1). Order names are according to (*36*, *37*) (SM6). To the right are superorder (-imorphae) and higher unranked names. In some groups, more than one species was sequenced, and these branches have been collapsed (noncollapsed version in fig. S1). Text color denotes groups of species with broadly shared traits, whether by homology or convergence. The arrow indicates the K-Pg boundary at 66 Ma, with the Cretaceous period shaded at left. The gray dashed line represents the approximate end time (50 Ma) by which nearly all neoavian orders diverged. Horizontal gray bars on each node indicate the 95% credible interval of divergence time in millions of years.



Jarvis et al. (2014)

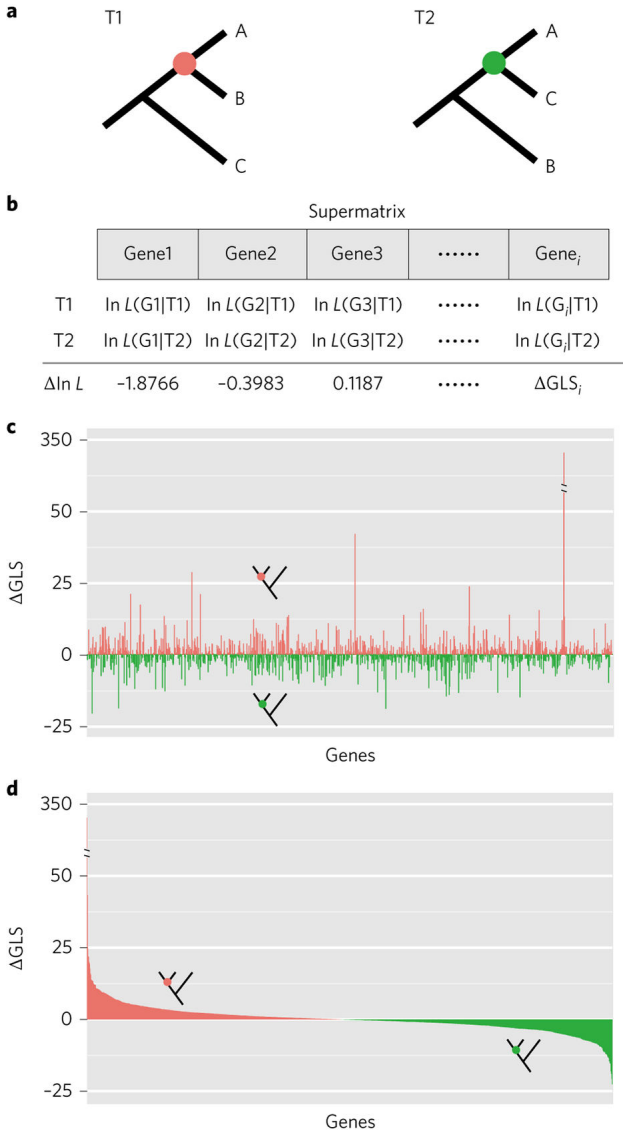
"Underlying this single topology was large-scale incongruence: none of the 14,536 trees from individual loci matched the inferred species tree, and many nodes with 100% bootstrap support appeared in <10% of the gene trees (Jarvis et al. 2014)."
Hahn and Nakhleh (2016)

How can you end up with 100% bootstrap support for relationships in only few of the gene trees?

How can you end up with a consensus tree that doesn't match any of your gene trees?

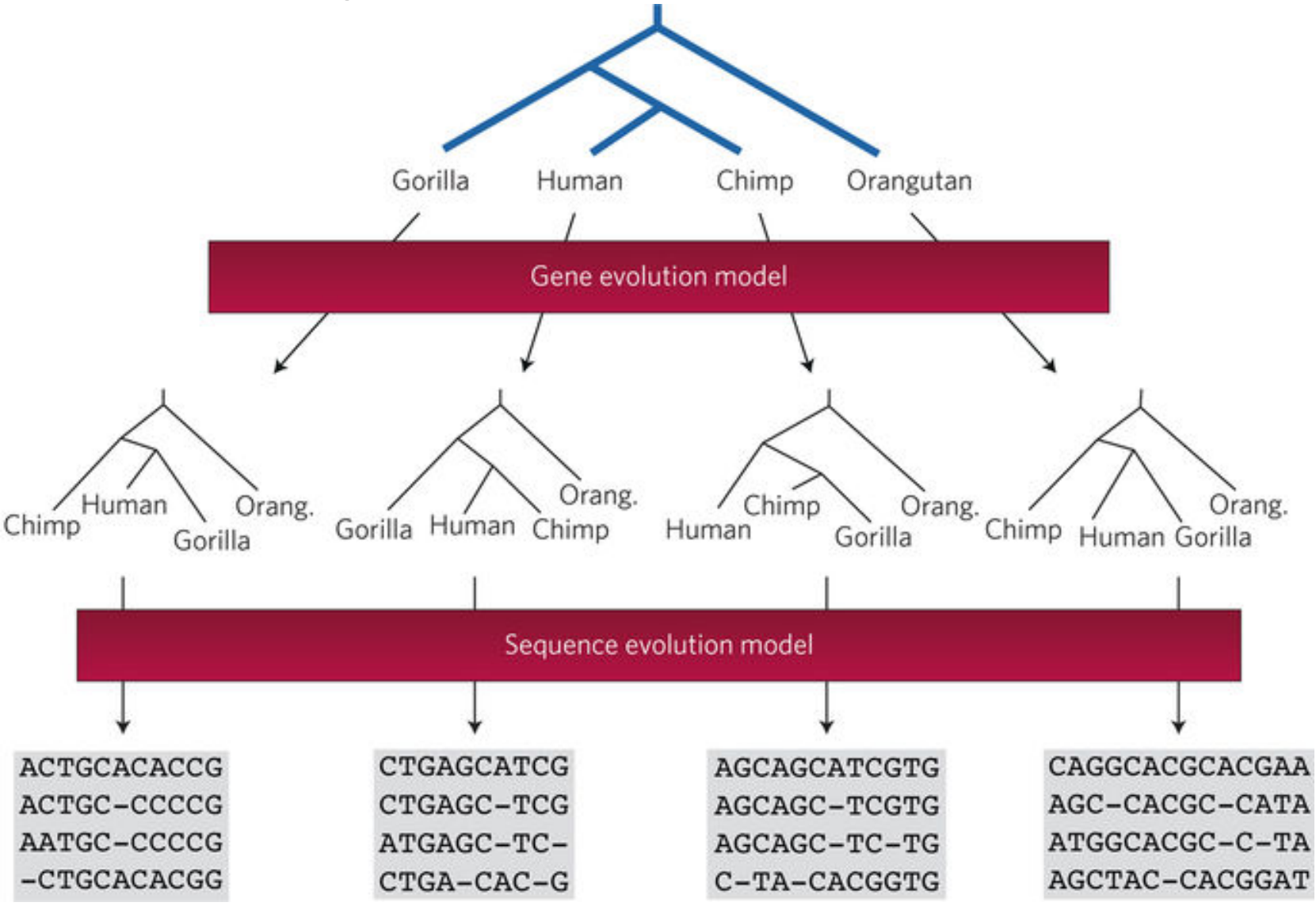
Taxa	Unrooted binary trees	Rooted binary trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	3×10^{7}
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

Genes may differ in support for topologies



Shen et al. (2017)

Should we exect gene trees to match species tree?



Mirarab (2017)

What are reasons that a 'gene tree' may not show the same relationships as the species tree?

- gene tree estimation error
- incomplete lineage sorting
- hybridization
- horizontal gene transfer

Gener tree estimation error: How can we get trees wrong?

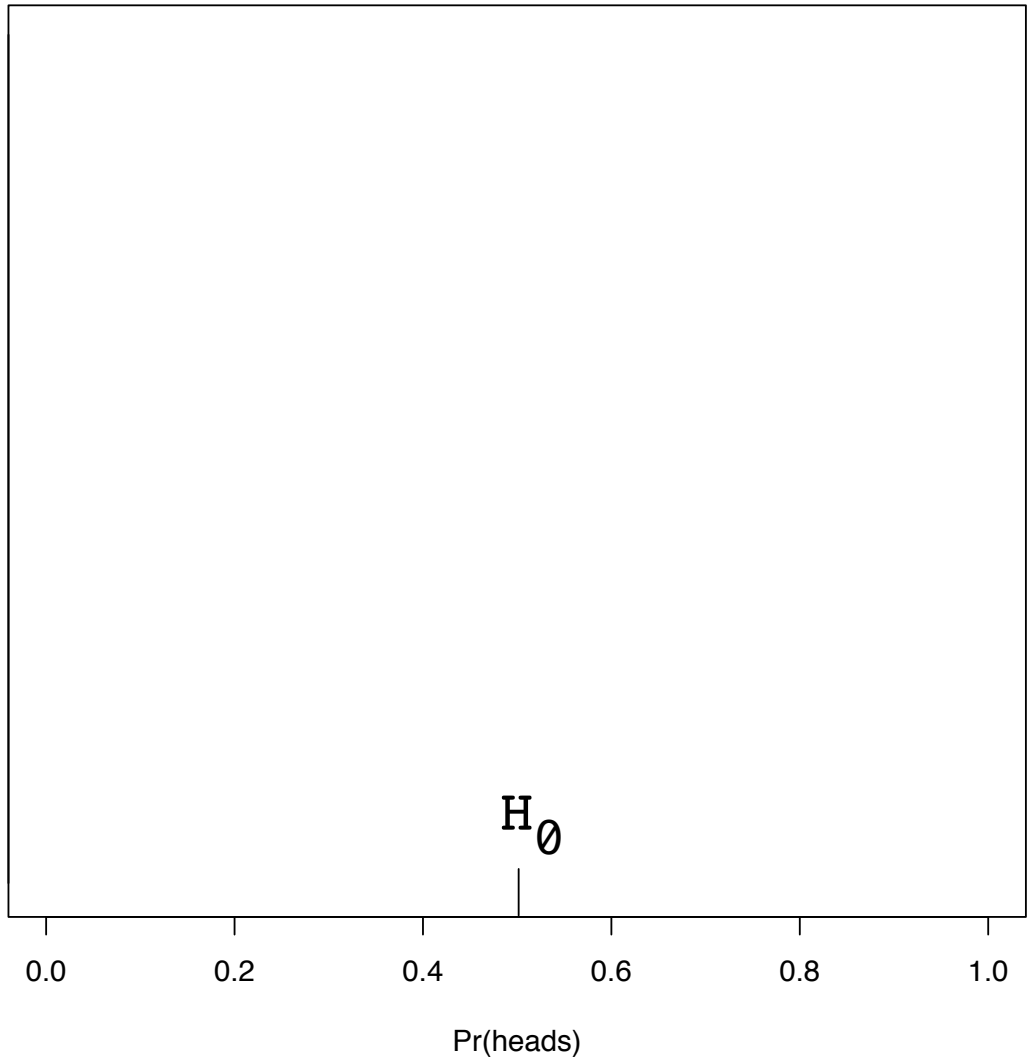
- Bad data - Garbage in Garbage out
- Incorrect model or incorrect inference methods
- Random error

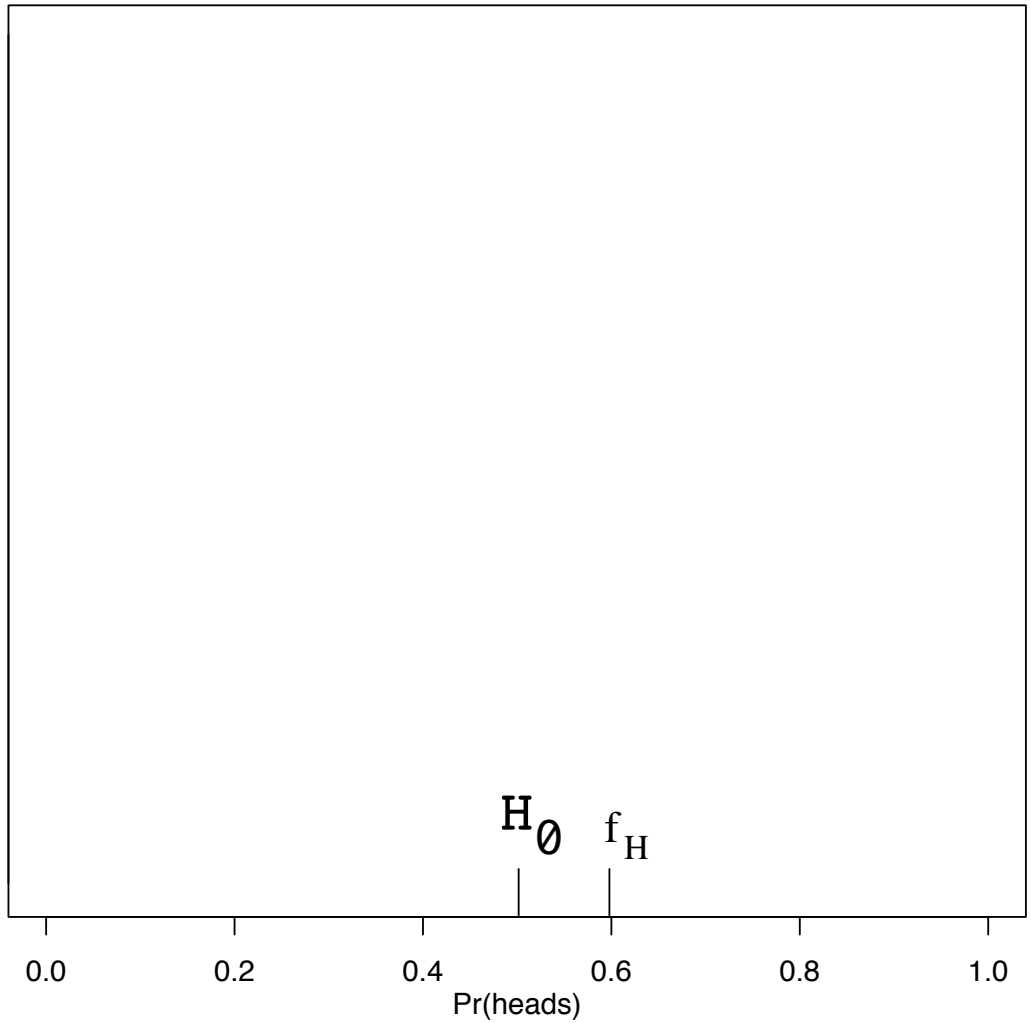
Frequentist hypothesis testing: coin flipping example

$N = 100$ and $h = 60$
Can we reject the fair coin hypothesis? $H_0 : \Pr(\text{heads}) = 0.5$

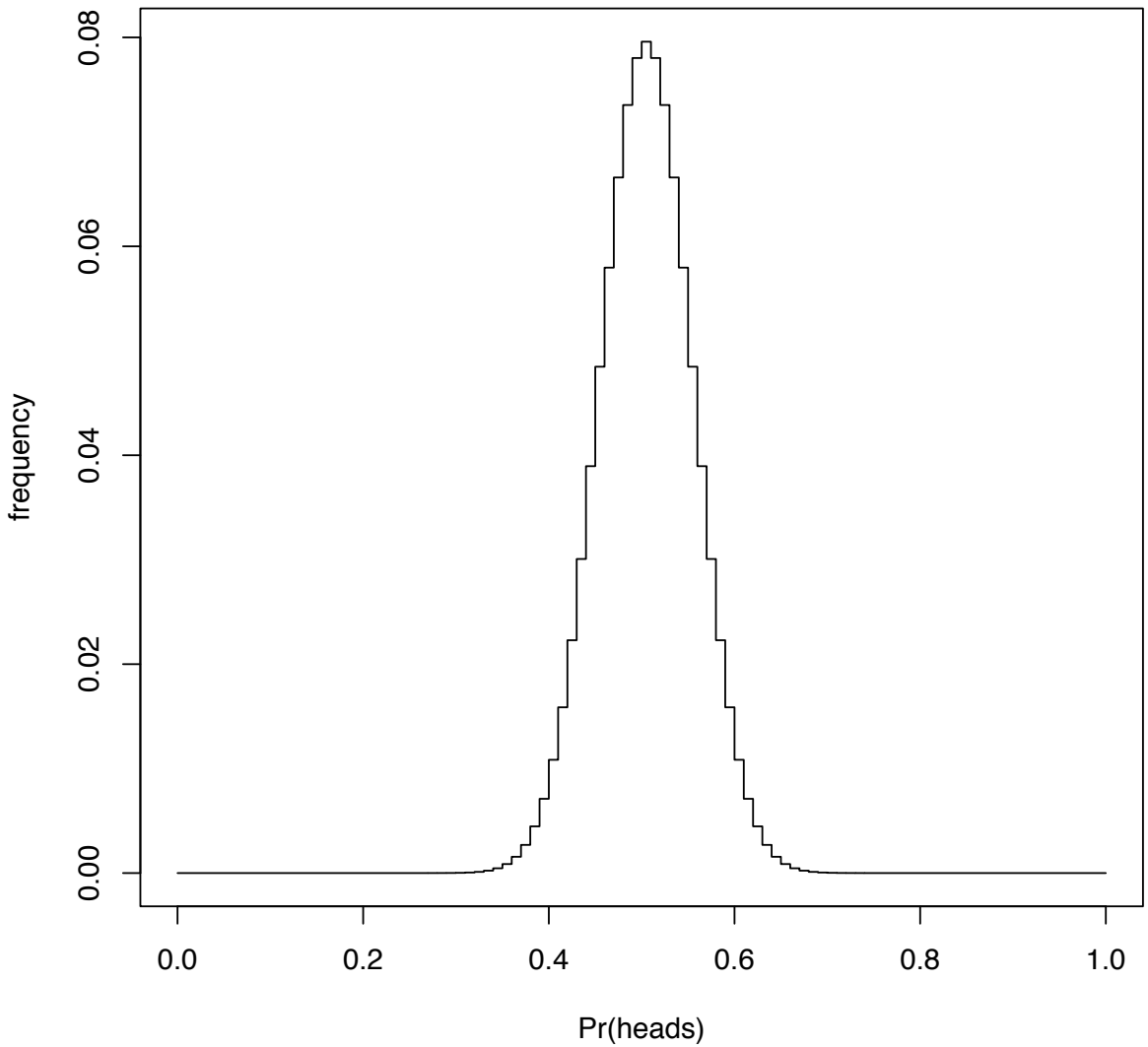
The "recipe" is:

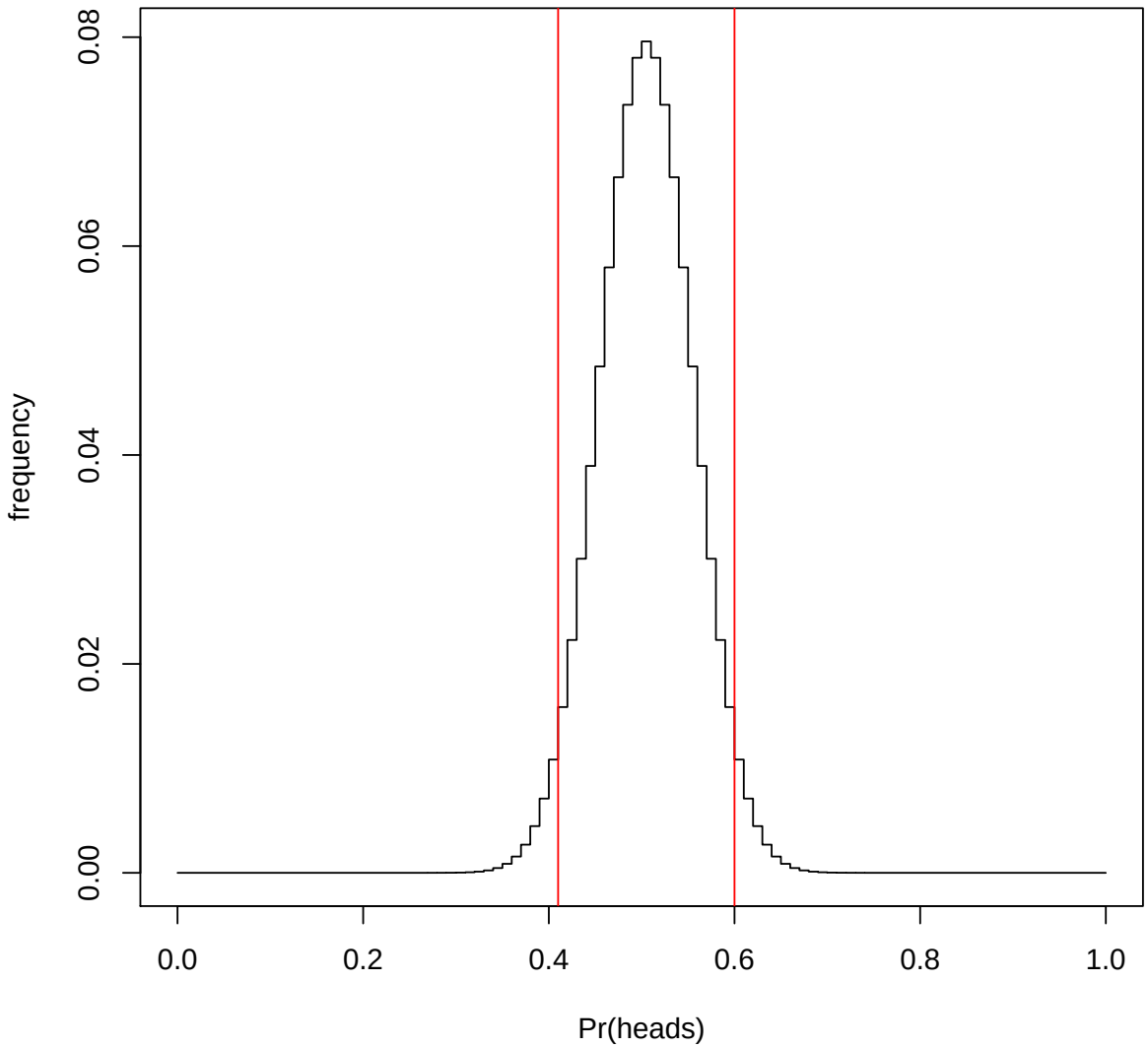
1. Formulate null (H_0) and alternative (H_A) hypotheses.
2. Choose an acceptable Type-I error rate (significance level)
3. Choose a test statistic: $f_H =$ fraction of heads in sample. $f_H = 0.6$
4. Characterize the null distribution of the test statistic
5. Calculate the P-value: The probability of a test statistic value more extreme than f_H arising *even if* H_0 *is true*.
6. Reject H_0 if P-value is $\leq$ your Type I error rate.


H0
0.0
0.2
0.4
0.6
0.8
1.0
Pr(heads)


H0
fH
0.0
0.2
0.4
0.6
0.8
1.0
Pr(heads)

Null distribution





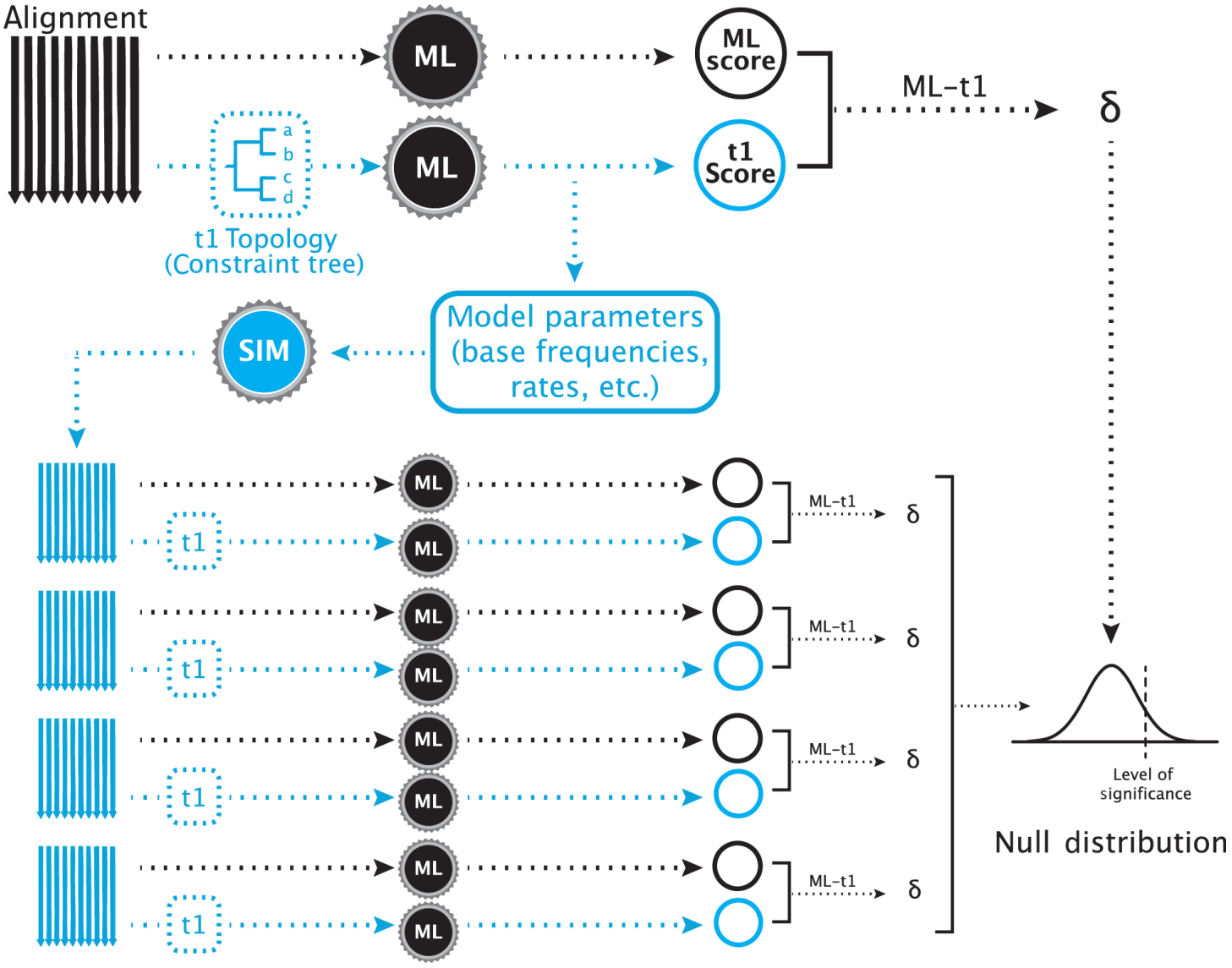
P-value ≈ 0.058

How to generate this variance distribution for trees?

Parametric bootstrapping, Swofford, Olson, Wadell, Hillis (SOWH) (Swofford et al., 1996)

- Simulate sequence data on alternative tree under estimated paramaters
- Test Statistic = Difference in Likelihood between ML tree and alternative tree
- Null Distribution = Expected distribution if the T1 hypothesis is assumed to be correct

script at `https://github.com/josephryan/sowhat`



(Church et al., 2015)

Problems with SOWH test:

- Many tree searches - computationally expensive
- Over simplified models result in high type 1 error

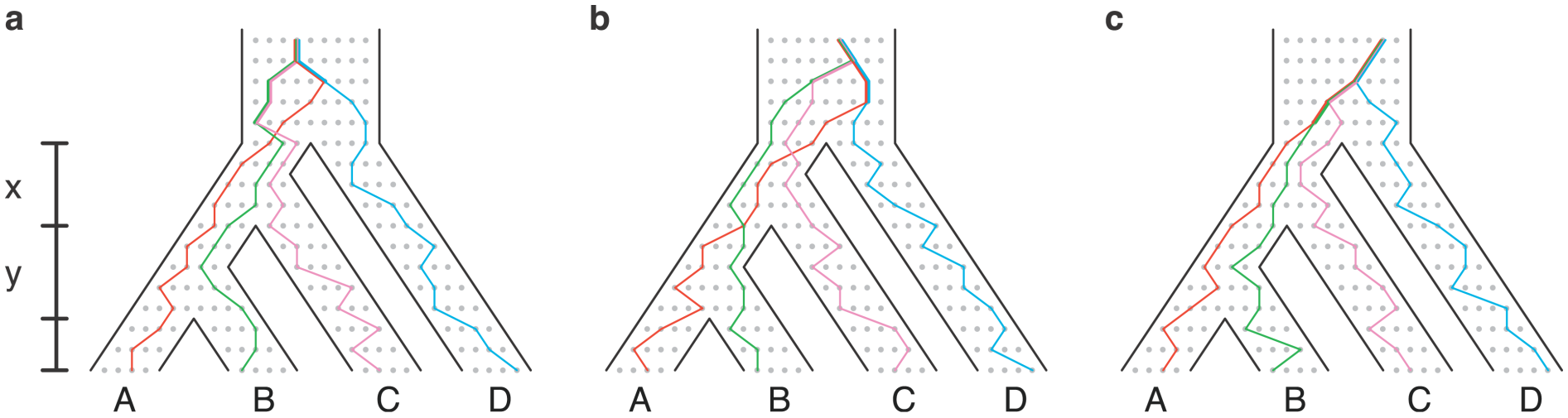
If you reject a null hypothesis based on parametric bootstrapping, the appropriate conclusion is that the degree of support observed is too large to be easily explained by chance assuming that the simulated model of sequence evolution adequately mimics the level of conflicting signal in the true generating process.
McTavish and Holder, Encyclopedia of Evolution, 2016

There are mamy non-parametric approaches to topology comparison as well - KH test, SH test RELL bootstrap.
If you are interested I can send you McTavish and Holder, Encyclopedia of Evolution, 2016

What are reasons that a 'gene tree' may not show the same relationships as the species tree?

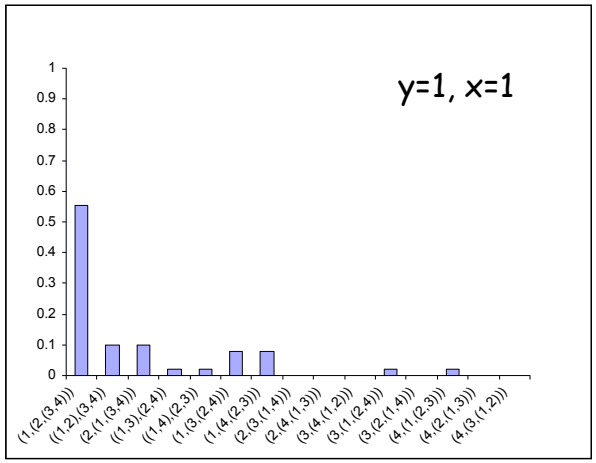
- ~~gene tree estimation error~~
- incomplete lineage sorting
- hybridization/horizontal gene transfer

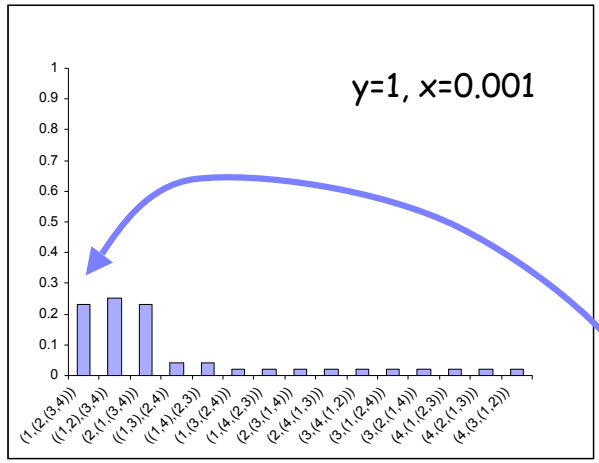
Is the gene tree that matches the species tree always the most common gene tree?

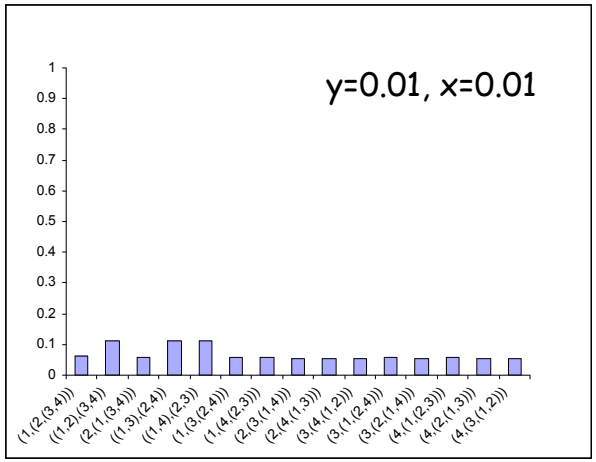


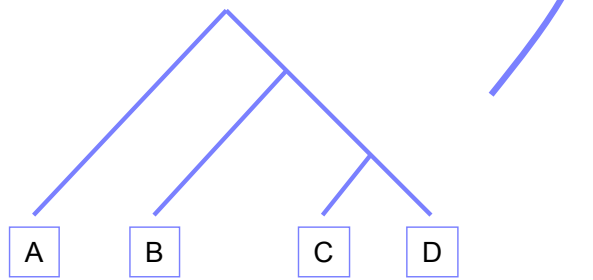
If the internal branches of the species tree—x and y—are short so that coalescences occur deep in the tree, the two sequences of coalescences that produce a given symmetric gene tree topology together have higher probability than the single sequence that produces the topology that matches the species tree. (a) and (b) Two coalescence sequences leading to gene tree topology ((AD)(BC)). In (a), the lineages from B and C coalesce more recently than those from A and D, and in (b), the reverse is true. (c) The single sequence of coalescences leading to gene tree topology (((AB)C)D). (Degnan and Rosenberg, 2006)

Applications of the topology distribution - example 2

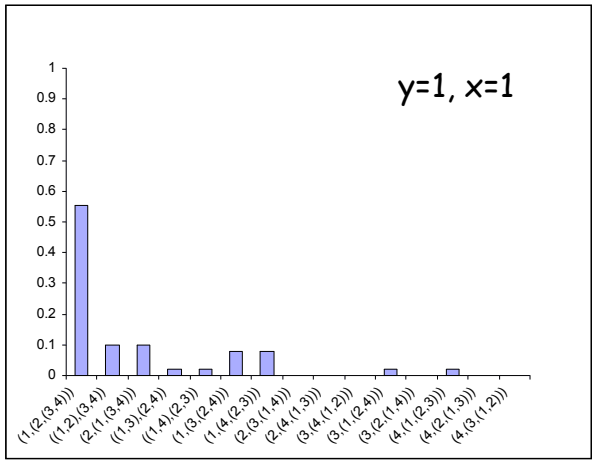


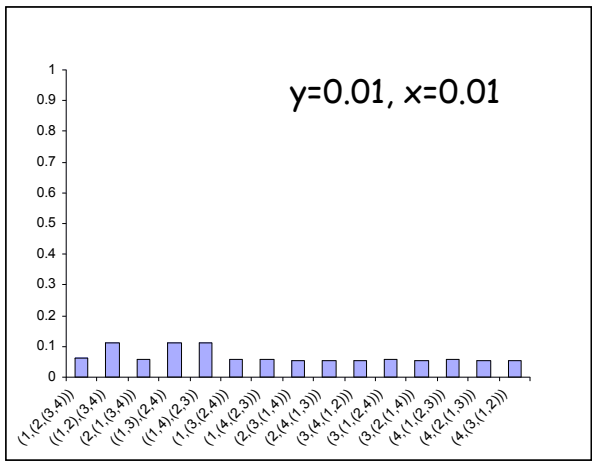


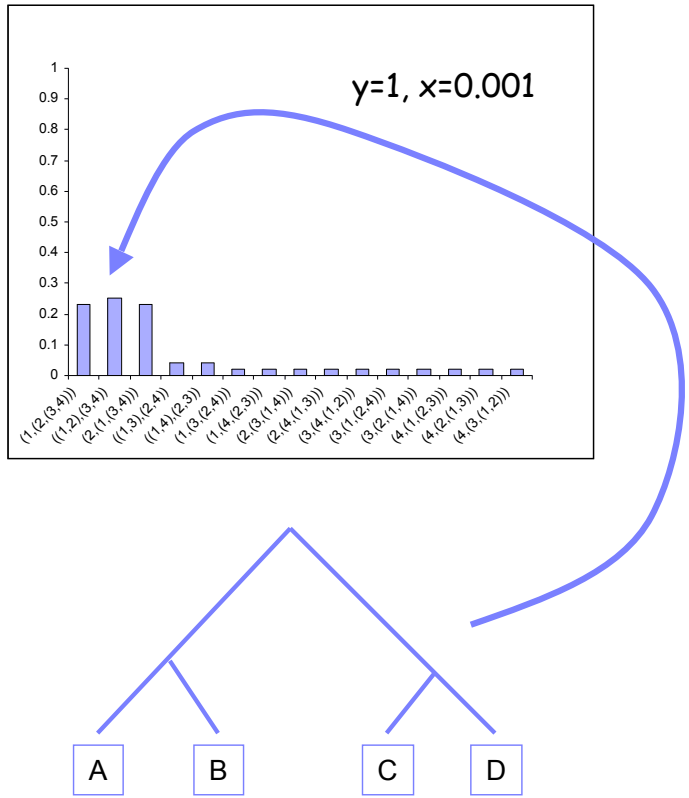




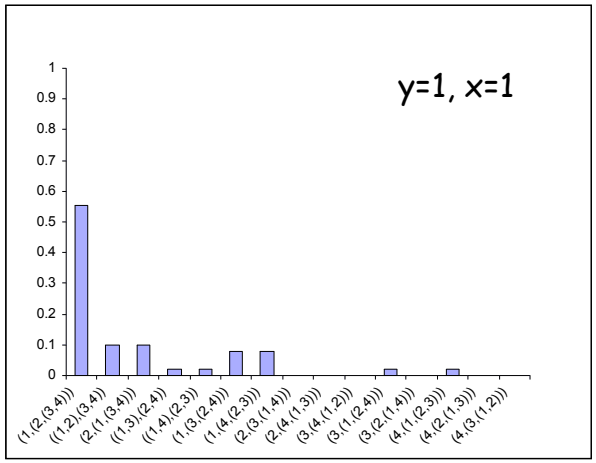
Applications of the topology distribution - example 2

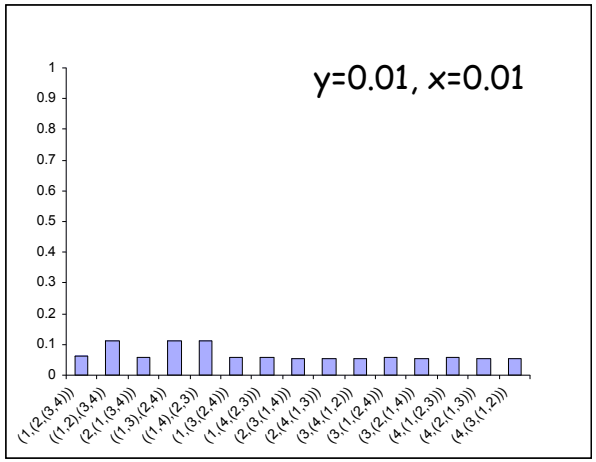


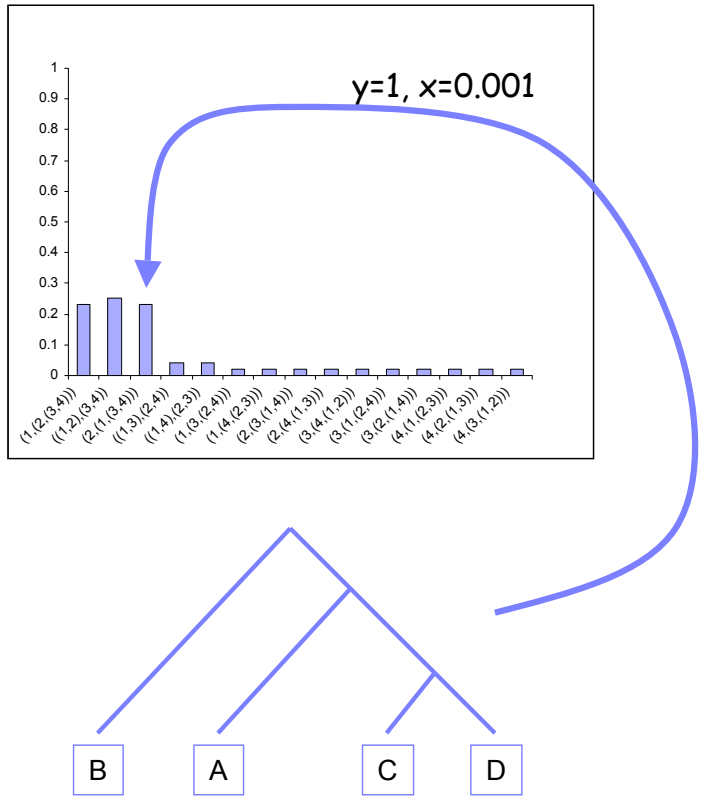




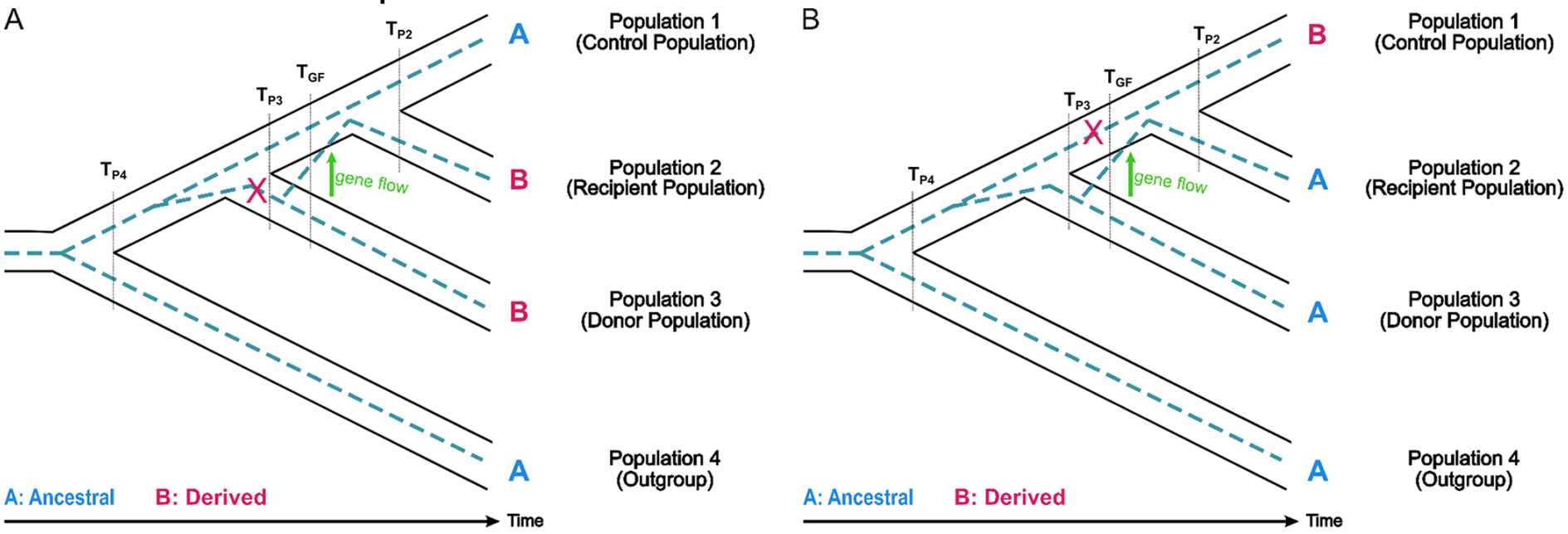
Applications of the topology distribution - example 2







Introgression/horizontal gene transfer can also result in gene trees that conflict with the species tree



(Lopez-Fang et al. 2023)

How do you estimate relationships in the face of these processes?

How do you estimate relationships in the face of these processes?
It depends what question you are trying to answer!

Concatenation

Combine all genes or regions for each sample or taxon.

Advantages:

- Relatively fast
- Provides single answer (interpretable!)
- May result in similar or same inferences are more complex methods, especially when gene tree incongruence is rare. Tonini et al. (2015)

Disadvantages:

- Final tree often doesn't match any of the individual gene trees, and can be the wrong spp tree Kubatko and Degnan (2007)
- Incorporating coalescent models can improve accuracy Edwards et al. (2016)
- Potentially interesting conflicting signals are lost Hahn and Nakhleh (2016)

Gene tree methods

Infer individual gene trees, and combine.

Advantages:

- Captures gene tree variation
- Can focus on loci of interest (e.g. Hahn and Nakhleh (2016))
- Can model variation across gene trees in multiple ways - ILS, HGT, hybridization, and assess model fit.

Disadvantages:

- Shorter sequences result in higher error
- Loci from many approaches for generating genomic data are too short to estimate individual trees (SNPs or short loci)

Gene tree methods

Infer individual gene trees, and combine.

Coalescent analyses

- Astral: `https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md`
- MP-EST: `https://github.com/lliu1871/mp-est`
- BUCKY: `http://pages.stat.wisc.edu/~ane/bucky/index.html`

Network/hybridization inference

- PhyloNet: `https://wiki.rice.edu/confluence/pages/viewpage.action?pageId=39500205`
- SNaQ: `https://github.com/crsl4/PhyloNetworks.jl/wiki`

Full data methods

Joint inference of gene trees, model and species tree.

Advantages:

- Model describes the processes generating the data
- Full joint likelihood calculation

Disadvantages:

- Complex models, often very slow to infer for large numbers of taxa (months!)

Full data methods

Gene sequences

- *BEAST, starBEAST2: `https://taming-the-beast.org/tutorials/StarBeast-Tutorial/`
- BPP: *Can jointly estimate coalescence and introgression* https://hal.archives-ouvertes.fr/hal-02536475/document

SNPs or short loci from across the genome

- SVDQuartets: *fast, quartet based, so handles missing data well* `http://www.phylosolutions.com/tutorials/ssb2018/svdquartets-tutorial.html`
- SNAPP: `http://evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2018/01/BFD-tutorial-1.pdf`

Conclusions:

- The importance of gene tree discordance, as well as how to address it, depends on both your data and your question!

Church, S. H., Ryan, J. F., and Dunn, C. W. (2015). Automation and Evaluation of the SOWH Test with SOWHAT. *Systematic Biology*, 64(6):1048–1058. Number: 6.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462.

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17. Number: 1.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., Fonseca, R. R. d., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N.,

Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., and Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331. Number: 6215.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology*, 56(1):17–24. Number: 1 Publisher: Oxford Academic.

Mirarab, S. (2017). Phylogenomics: Constrained gene tree inference. *Nature Ecology & Evolution*, 1(2):0056. Number: 2.

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5):0126. Number: 5.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference.

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., and Ortí, G. (2015). Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions. *PLOS Currents Tree of Life*. Publisher: Public Library of Science.