Coalescent Methods

Laura Kubatko
Departments of Statistics and
Evolution, Ecology, and Organismal Biology
The Ohio State University

kubatko.2@osu.edu
twitter: Laura_Kubatko

January 25, 2019

Relationship between population genetics and phylogenetics

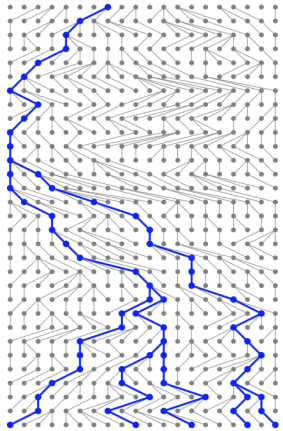
- Population genetics: Study of genetic variation within a population
- Phylogenetics: Use genetic variation between taxa (species, populations) to infer evolutionary relationships
- Previously:
  - Each taxon is represented by a single sequence – "exemplar sampling"
  - We have data for a single gene and wish to estimate the evolutionary history for that gene (the **gene tree** or **gene phylogeny**)
- Now:
  - Sample many individuals within each taxon (species, population, etc.)
  - Sequence many genes for all individuals

Relationship between population genetics and phylogenetics

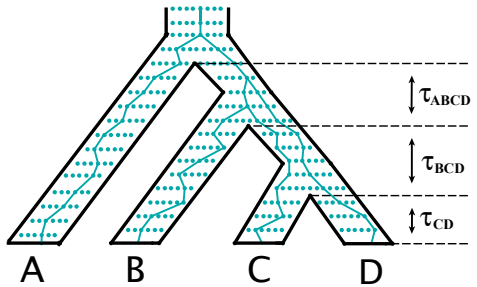
- Need models at two levels:

1. Model what happens within each population

→ *coalescent model*



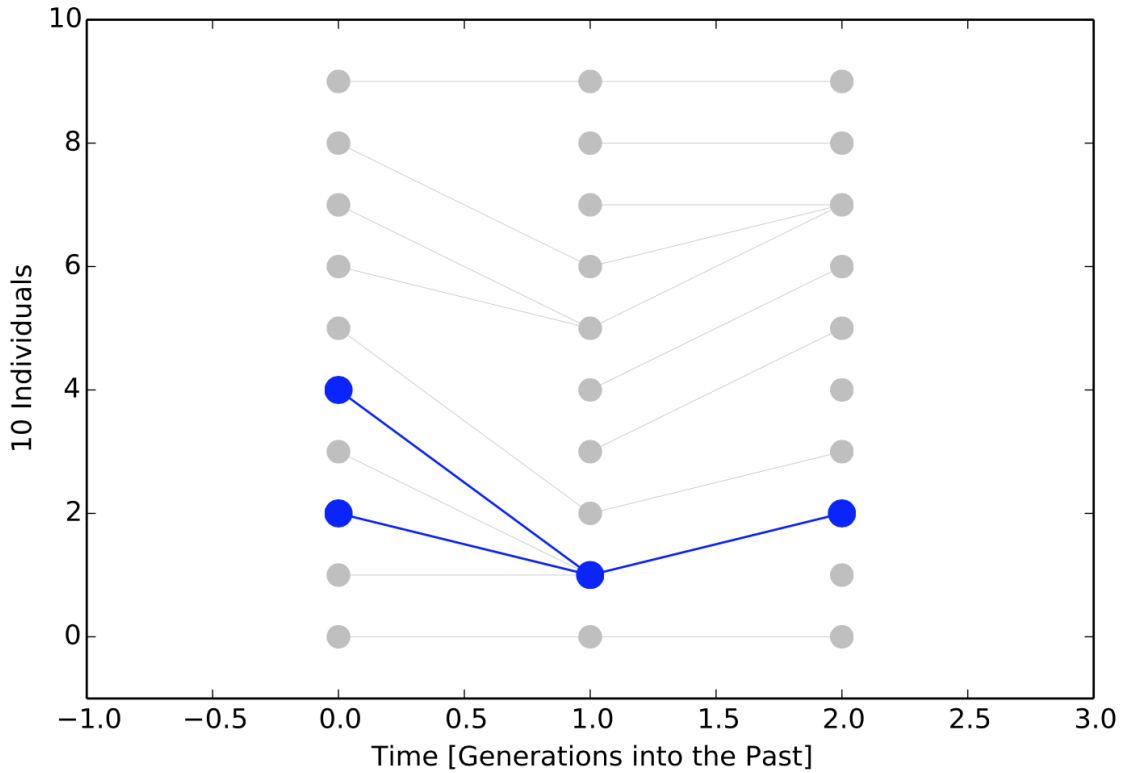
2. Link each within-population model on a phylogeny



- Assumptions:
  - Population of $2N$ gene copies
  - Discrete, non-overlapping generations of equal size
  - Parents of next generation of $2N$ genes are picked randomly with replacement from preceding generation (genetic differences have no fitness consequences)
  - Probability of a specific parent for a gene in the next generation is $\frac{1}{2N}$

Wright-Fisher model

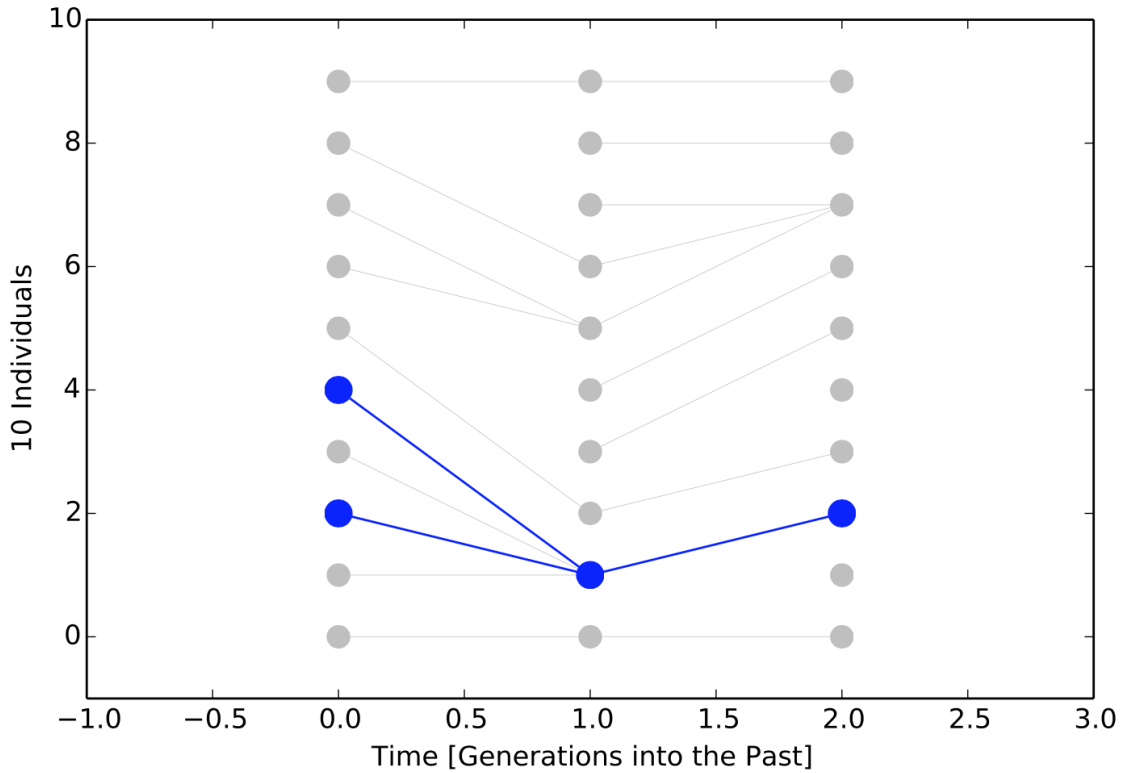
- Consider a population of $2N$ gene copies at time t
- **Question:** What is the probability that two randomly chosen individuals share a common ancestor in the previous generation?



Figures from PopVizard, by Peter Beerli

Wright-Fisher model

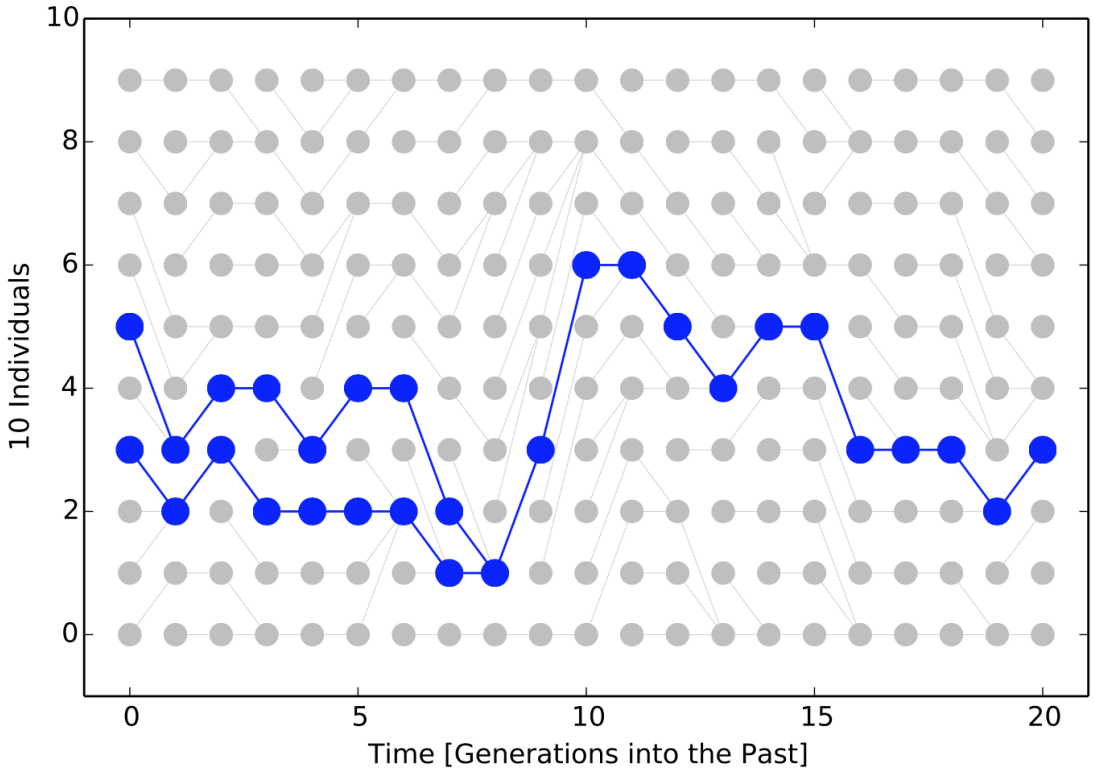
- Consider a population of $2N$ gene copies at time t
- **Question:** What is the probability that two randomly chosen individuals DO NOT share a common ancestor in the previous generation?



Figures from PopVizard, by Peter Beerli

Wright-Fisher model

- Consider a population of $2N$ gene copies at time t
- **Question:** How many generations do we need to wait until two randomly chosen individuals share a common ancestor?



Figures from PopVizard, by Peter Beerli

- Consider a population of $2N$ gene copies at time t
- **Question:** How many generations do we need to wait until two randomly chosen individuals share a common ancestor?
  - The number of generations, T, until two individuals share a common ancestor follows a geometric distribution with parameter $\frac{1}{2N}$, e.g.,

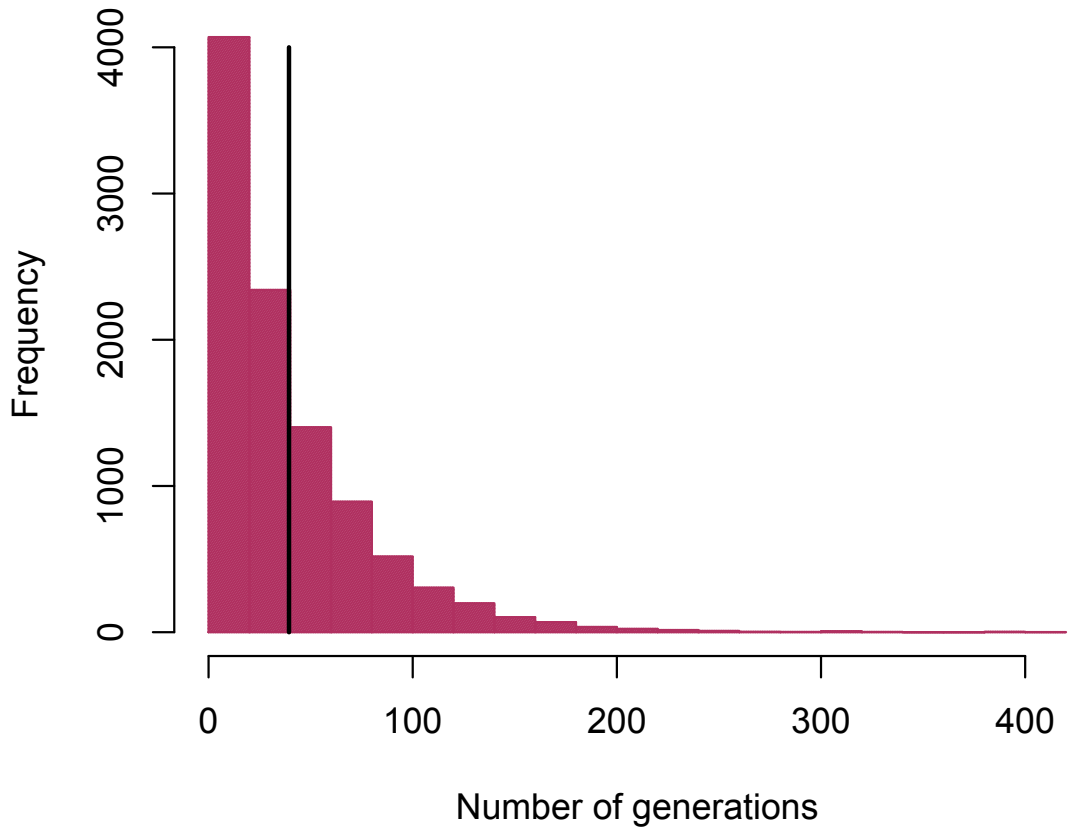
$$P(T = \tau) = \left(1 - \frac{1}{2N}\right)^{\tau - 1} \left(\frac{1}{2N}\right)$$

  - The expected number of generations until coalescence is

$$E(T) = 2N$$

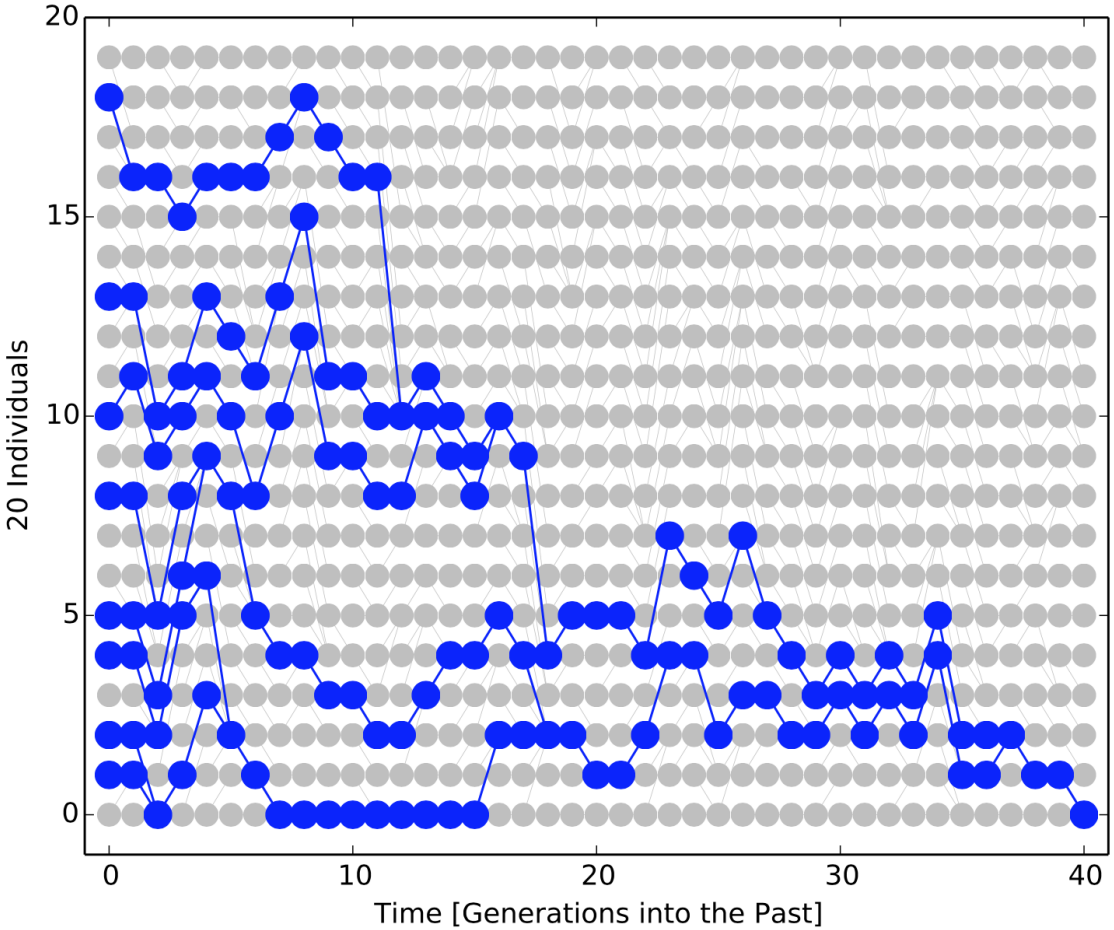
Wright-Fisher model

- Distribution of time to coalescence for two randomly sampled gene copies in a population of size $2N = 40$
- Observed mean (black line) = 39.40



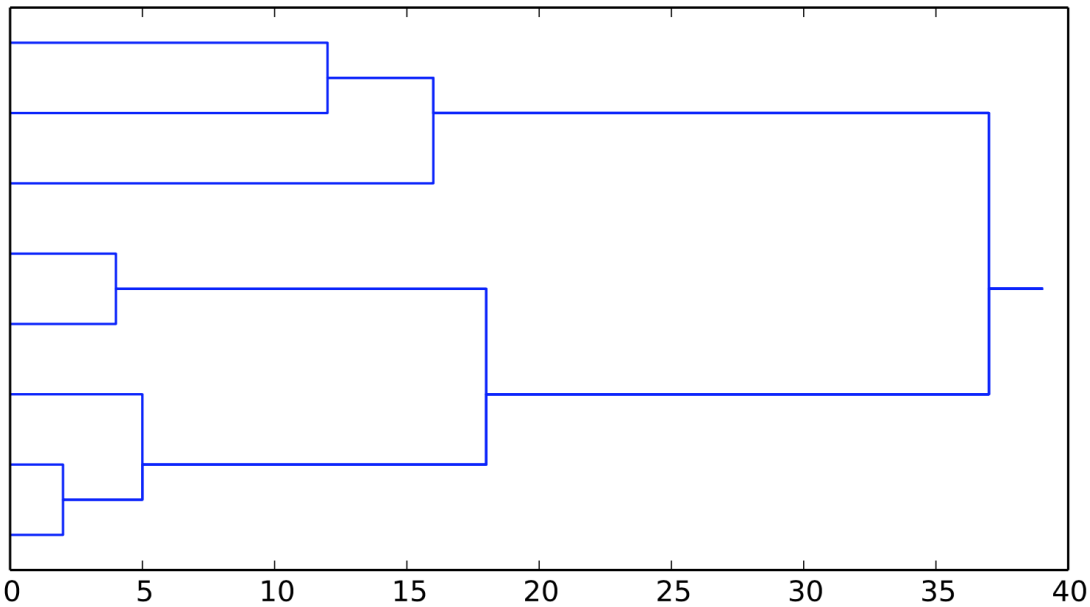
Wright-Fisher model

- Consider a population of $2N$ gene copies at time t
- **Question:** What about more than two gene copies?



Figures from PopVizard, by Peter Beerli

- Consider a population of $2N$ gene copies at time t
- **Question:** What about more than two gene copies?



Figures from PopVizard, by Peter Beerli

- Consider a population of $2N$ gene copies at time t
- **Question:** What about more than two gene copies?
  - Note from the previous slide that time to coalescence seems to vary with sample size
  - Specifically, the probability that two gene copies in a sample of k coalesce in the previous generation is

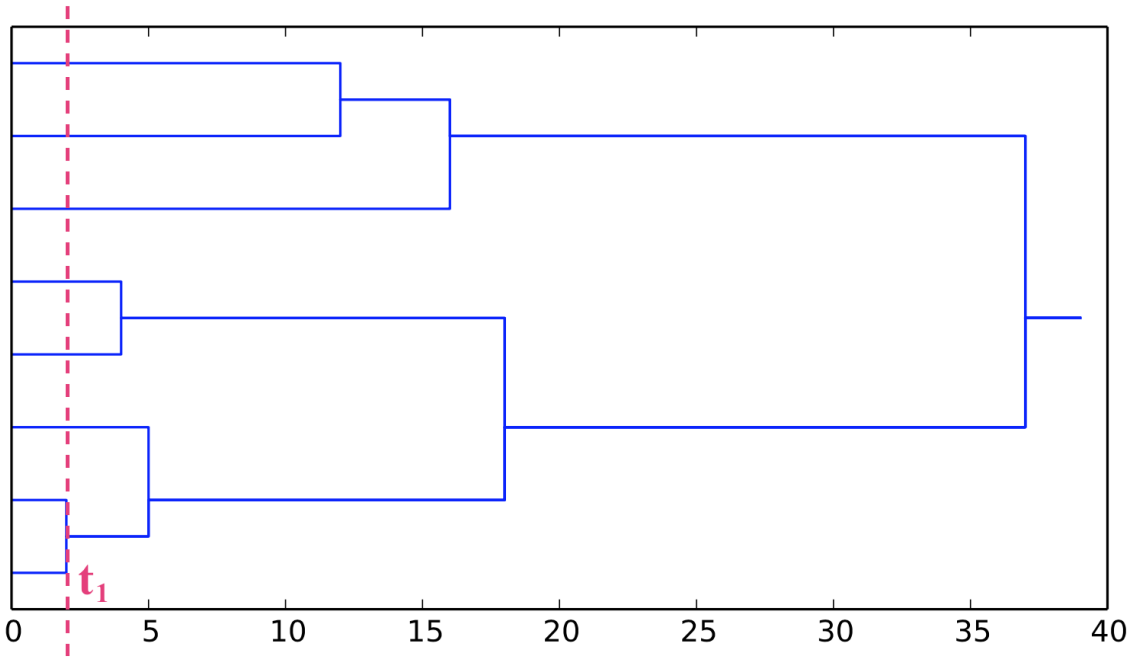
$$\binom{k}{2}\frac{1}{2N}$$

  - The distribution of the time to the first coalescent event is **geometric** with parameter $\binom{k}{2}\frac{1}{2N}$

- Cumbersome to work in this "discrete" setting where we think of things generation by generation

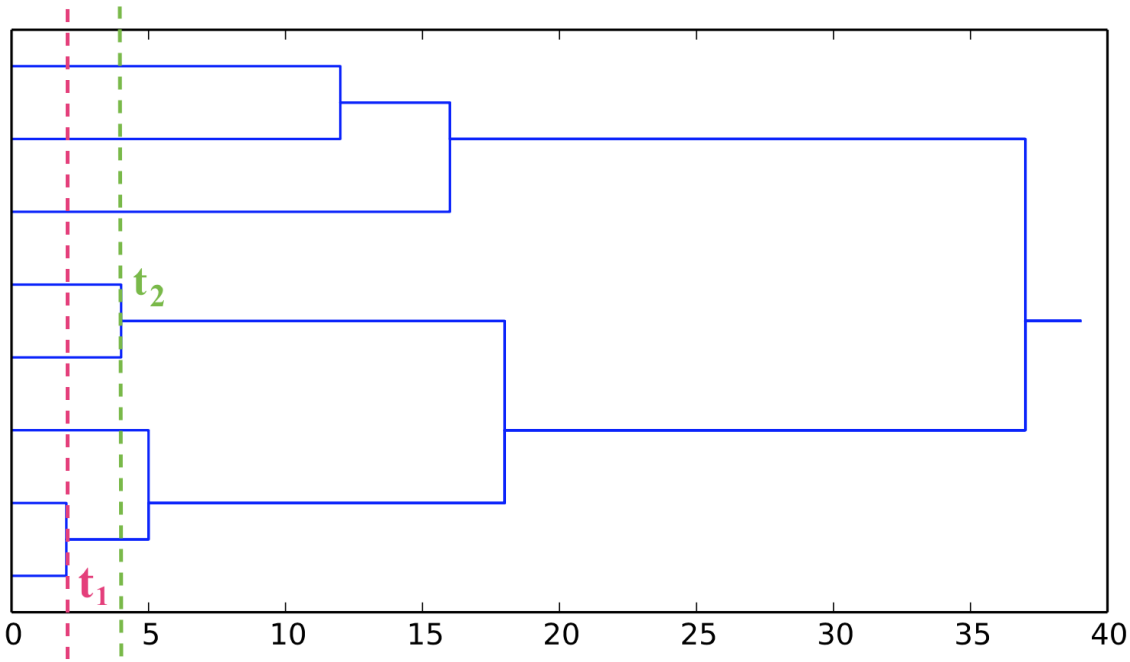
- Kingman's approximation: consider continuous time and a sample of k lineages. Then, the time back into the past until two lineages coalesce, U, is exponentially distributed with rate $\binom{k}{2}\frac{1}{2N}$
  - The probability density function is $g(u) = \binom{k}{2}\frac{1}{2N}e^{-\binom{k}{2}\frac{u}{2N}}$, for $u > 0$
  - The mean is $\frac{4N}{k(k-1)}$

Computing the probability density of a population tree under the coalescent



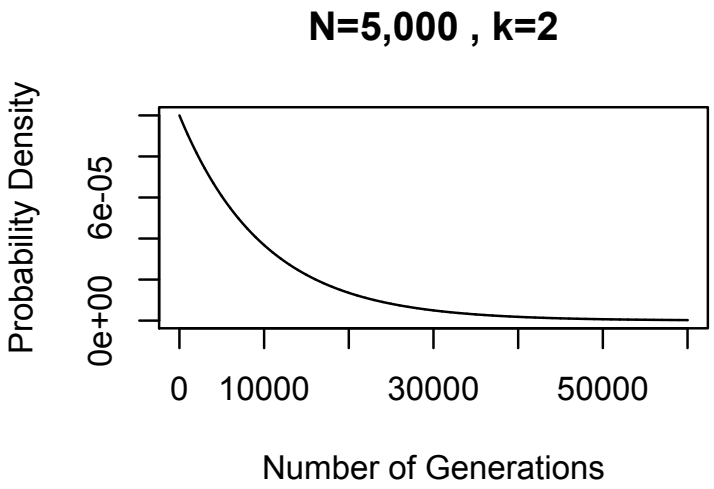
$$P(G) = \left(\frac{1}{2N} e^{-\frac{8(7)}{4N} t_1} \right) \times \ldots$$

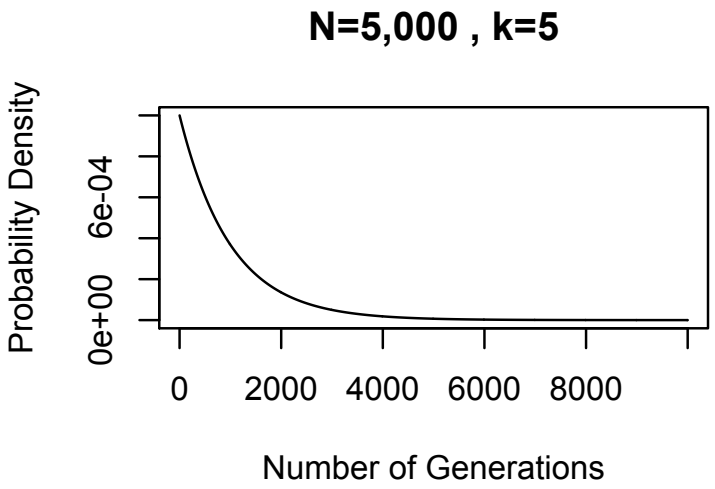
Computing the probability density of a population tree under the coalescent

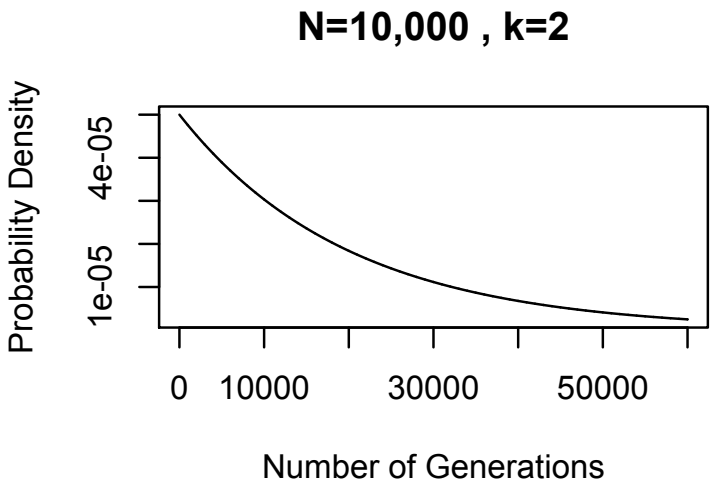


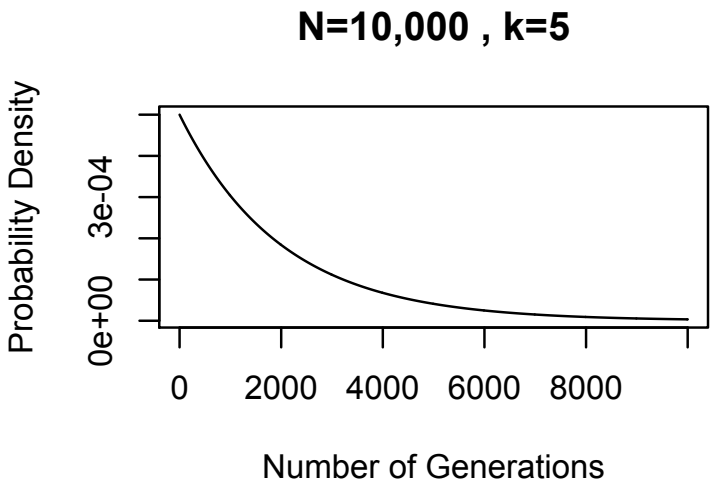
$$P(G) = \left(\frac{1}{2N} e^{-\frac{8(7)}{4N} t_1}\right)\left(\frac{1}{2N} e^{-\frac{7(6)}{4N} t_2}\right) \times \ldots$$

- What does the exponential distribution look like?







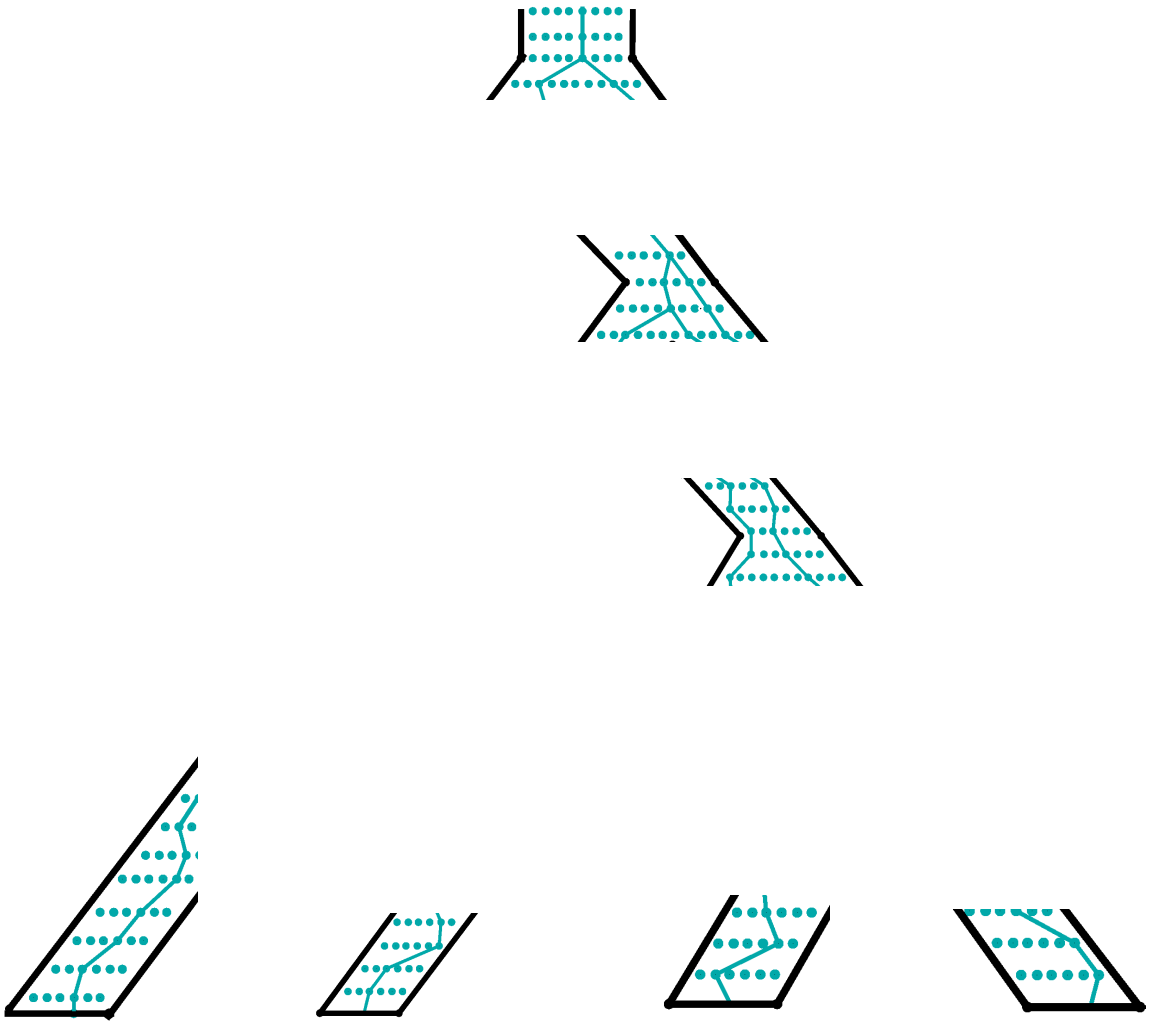


Coalescent units

- Define a common unit of time: coalescent unit, $t = \frac{u}{2N}$
- Examples:
  - $k = 2$ — exponential distribution with rate 1 and mean 1
  - $k = 5$ — exponential distribution with rate 10 and mean 0.1
- t "large" is now relative to population size, but the trends are the same:
  - Longer times lead to a higher probability of coalescence having occurred.
  - Coalescent events happen more quickly when the population size is smaller.
  - Coalescent events happen more quickly when the sample size is larger.
- Now we're ready to think about species trees!

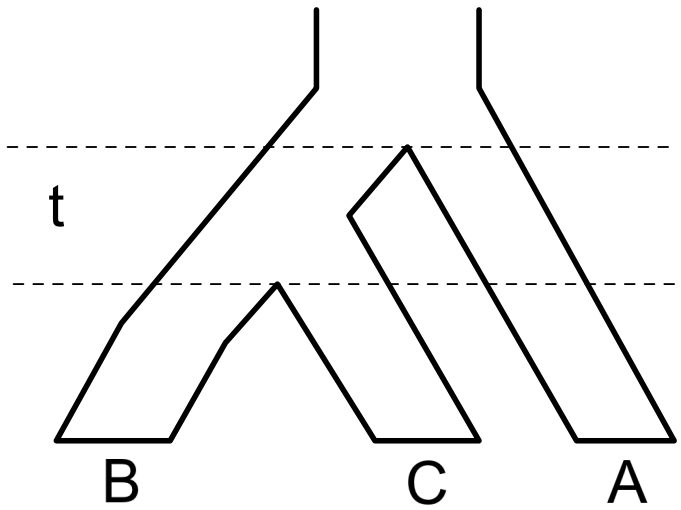
Putting it together . . . the coalescent model along a species tree

- Build up the species tree from many populations:



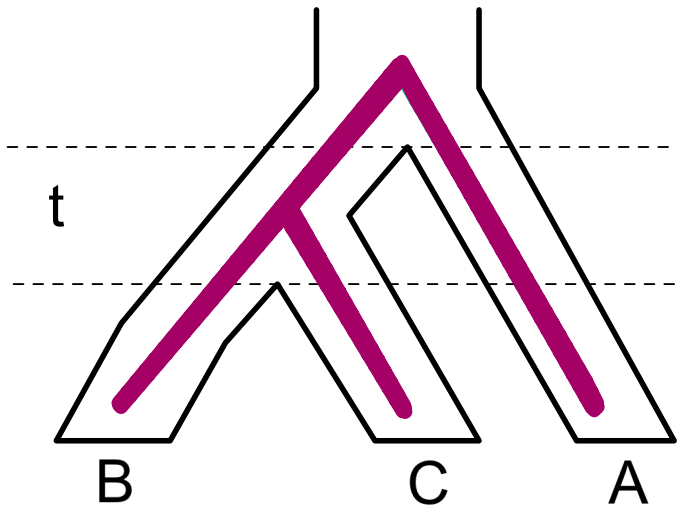
Phylogenetic coalescent model

- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



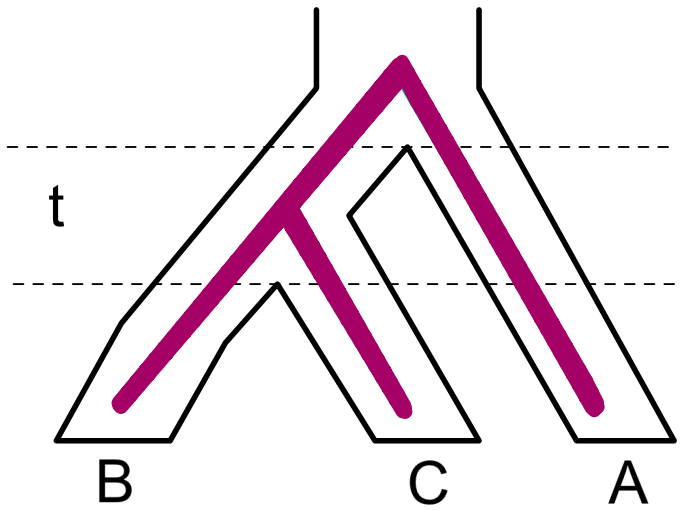
Phylogenetic coalescent model

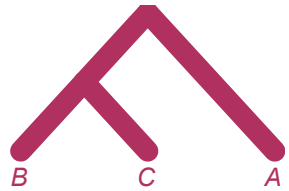
- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



Phylogenetic coalescent model

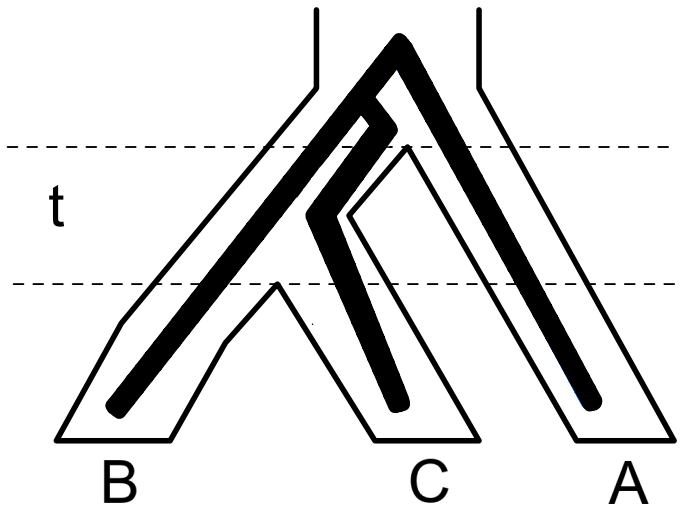
- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process

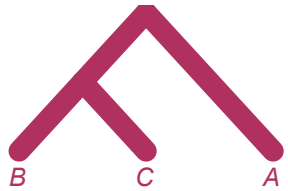


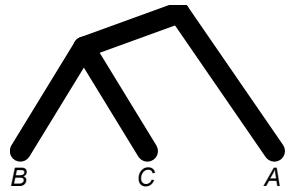


Phylogenetic coalescent model

- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process

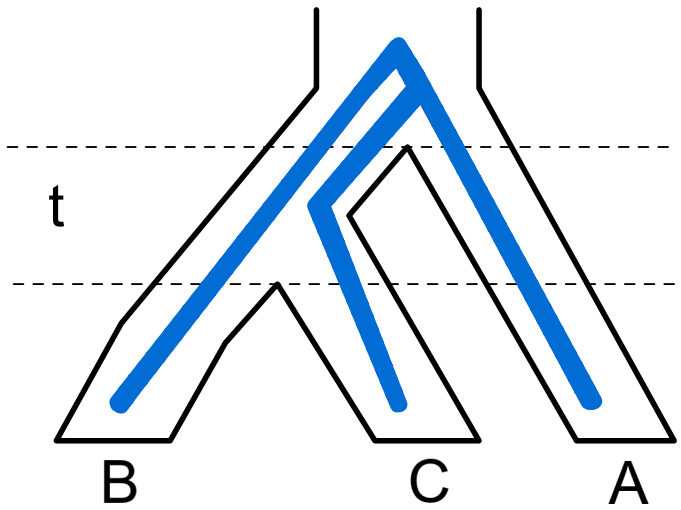


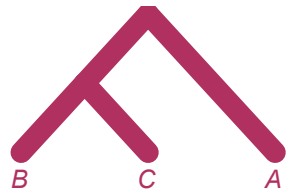


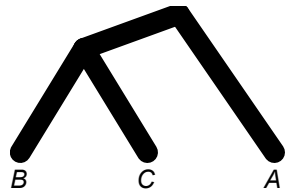


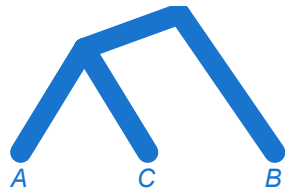
Phylogenetic coalescent model

- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



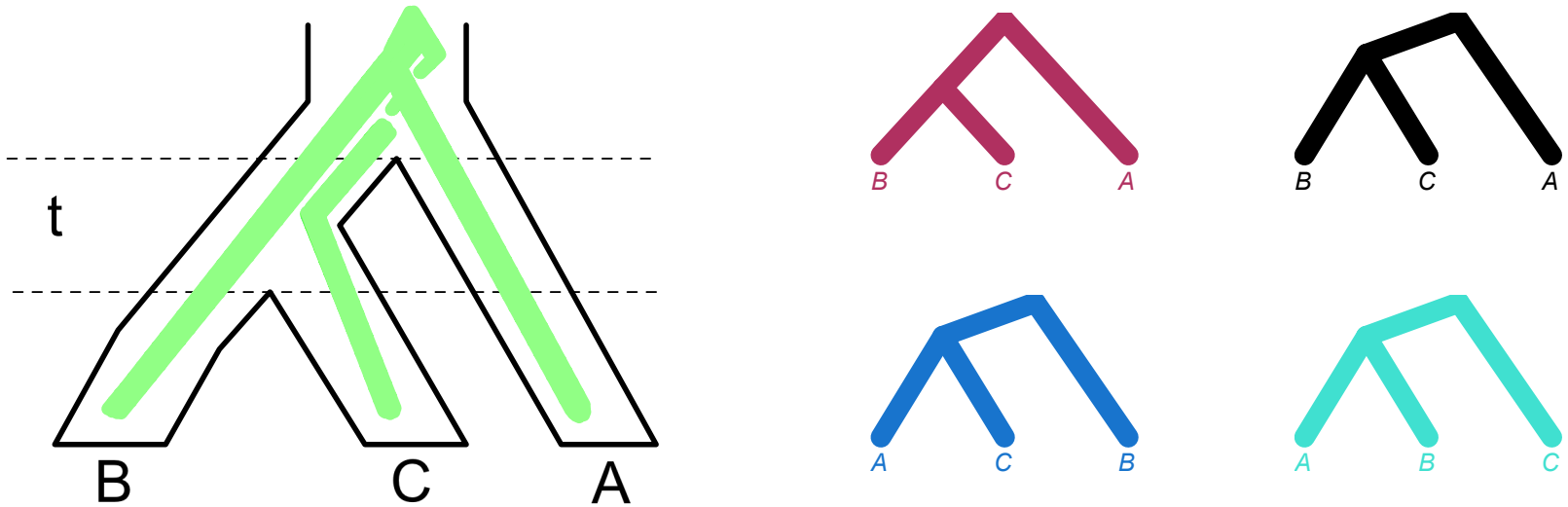






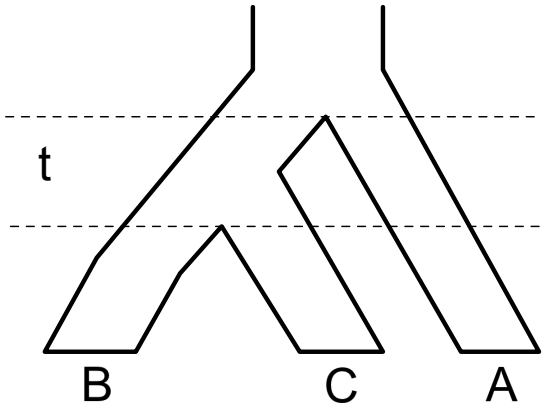
Phylogenetic coalescent model

- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



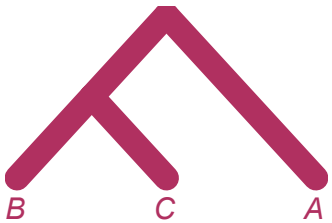
- Let's use what we've learned about the coalescent process to compute some probabilities

- t = length of interval between speciation events in **coalescent units**
  = number of $2N$ generations

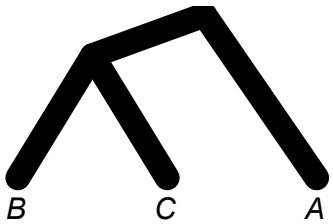


- **Example:** 1.2 coalescent units for an organism with population size $N = 10,000$ and a generation time of 3 years $= 1.2 \times 20,000 \times 3 = 72,000$ years

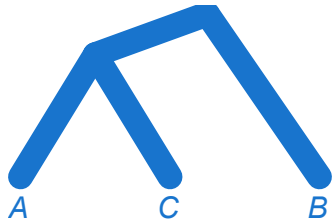
Probabilities of each gene tree history are shown below them
t = length of interval between speciation events



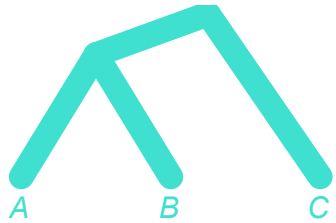
$1 - e^{-t}$



$\frac{1}{3}e^{-t}$



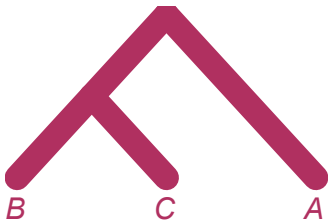
$\frac{1}{3}e^{-t}$



$\frac{1}{3}e^{-t}$

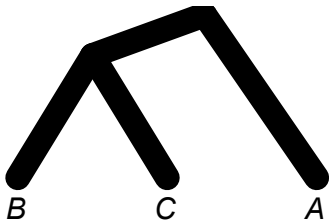
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0$



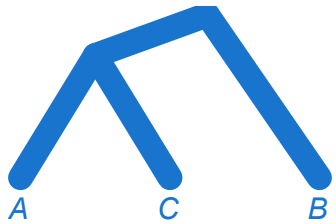
$1 - e^{-t}$

0.63



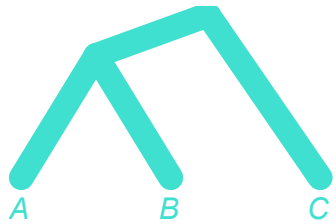
$\frac{1}{3}e^{-t}$

0.12



$\frac{1}{3}e^{-t}$

0.12

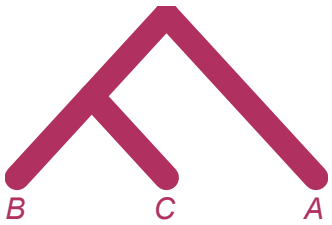


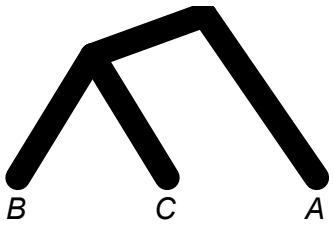
$\frac{1}{3}e^{-t}$

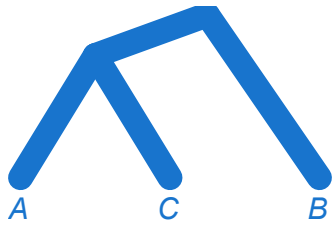
0.12

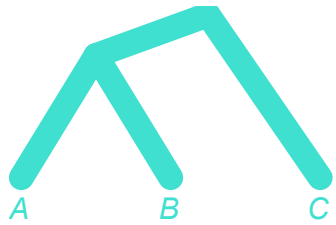
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0 = 0.5$





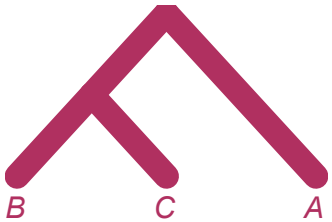


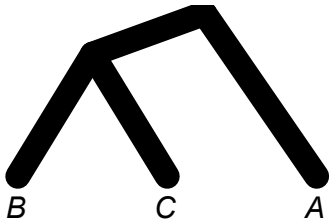


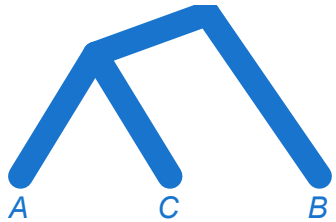
$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20

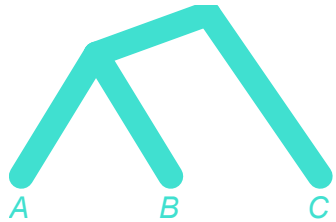
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0 = 0.5 = 2.0$





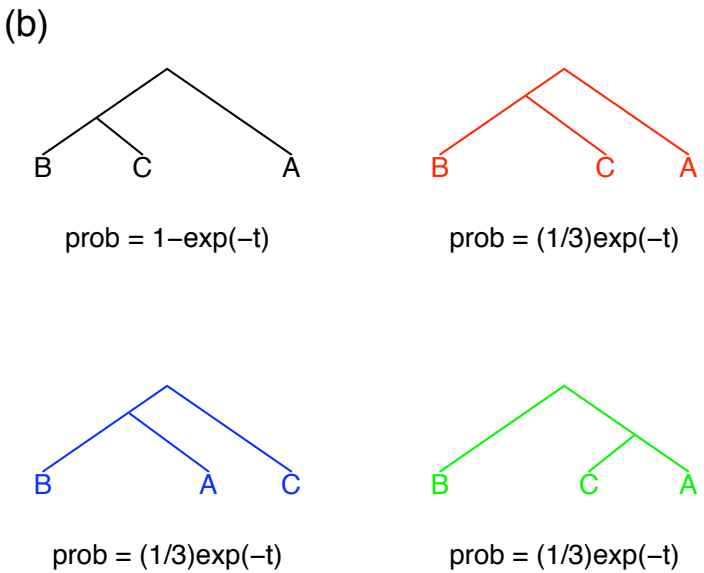


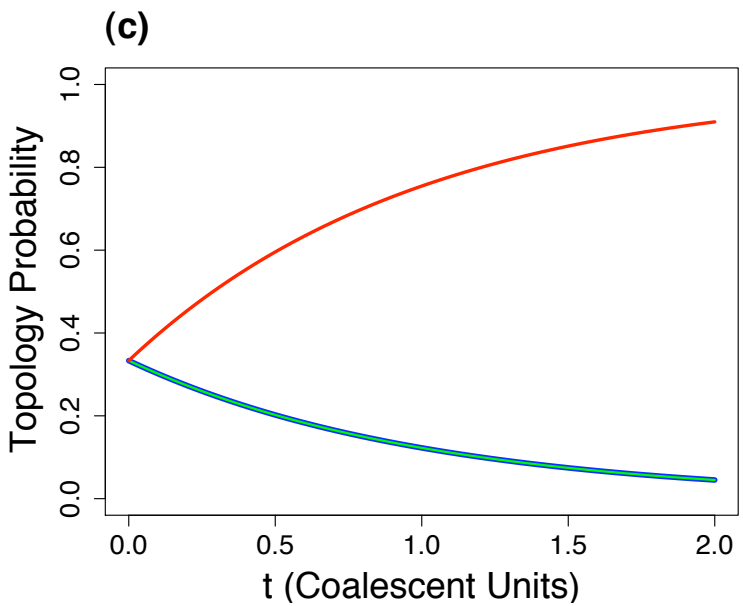


$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20
0.85	0.05	0.05	0.05

Effect of speciation time

- What are these probabilities like as a function of t, the length of time between speciation events?





Assumptions of the phylogenetic coalescent model

- What did we assume in carrying out these computations?
  - Events that occur in one population are independent of what happens in other populations within the phylogeny.
  - More specifically, given the number of lineages entering and leaving a population, coalescent events within populations are independent of other populations.
  - It is also important to recall an assumption we "inherit" from our population genetics model: all pairs of lineages are equally likely to coalesce within a population.
  - No gene flow occurs following speciation.
  - No other evolutionary processes (e.g., horizontal gene flow, duplication, . . .) have led to incongruence between gene trees and the species tree.

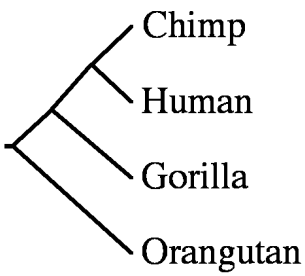
Summary of the three-taxon case

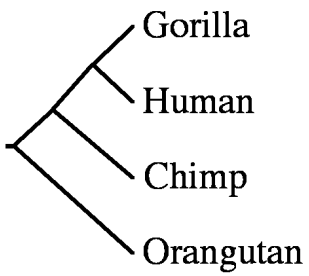
- What have we learned from considering 3 taxa?
  - Gene tree with topology that matches the species tree occurs with probability at least as large as the other two trees
  - The other two trees are expected to occur in equal frequency
  - Could model this with the multinomial distribution:
  - Shorter intervals between speciation events lead to more disagreement between gene trees and species trees

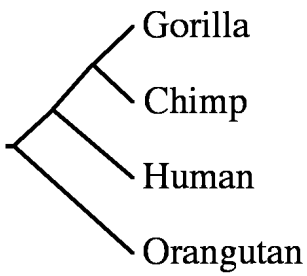
Application 1: Goodness of fit to empirical data

- Motivation: Paper by Ebersberger et al. 2007. *Mol. Biol. Evol.* 24:2266-2276
- Examined 23,210 distinct alignments for 5 primate taxa: Human, Chimp, Gorilla, Orangutan, Rhesus
- Looked at distribution of gene trees among these taxa - observed strongly supported incongruence only among the Human-Chimp-Gorilla clade.

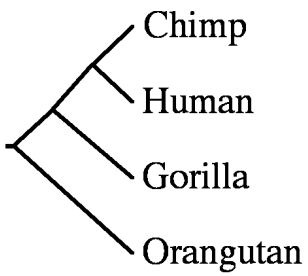
Application 1: Goodness of fit to empirical data



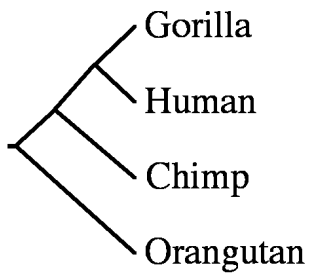




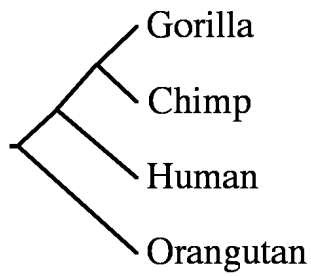
Application 1: Goodness of fit to empirical data



76.6%



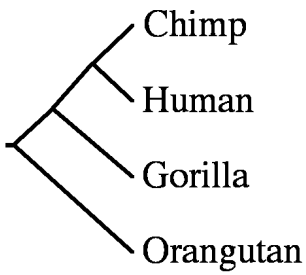
11.4%

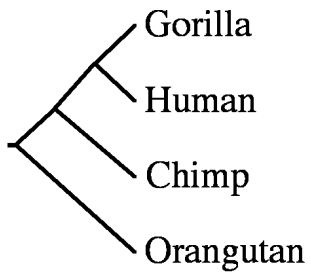


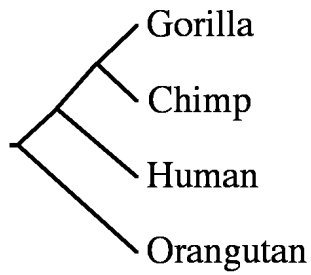
11.5%

Observed proportions of each gene tree among ML phylogenies

Application 1: Goodness of fit to empirical data







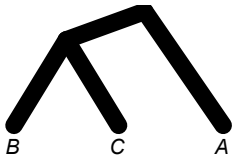
76.6%	11.4%	11.5%
79.1%	9.9%	9.9%

Observed proportions of each gene tree among ML phylogenies

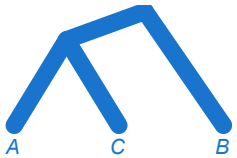
Predicted proportions using parameters from Rannala & Yang, 2003.

Application 2: Branch length estimation

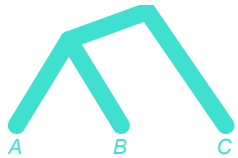
- Suppose you were given a sample of gene trees, i.e.,



70 genes



15 genes

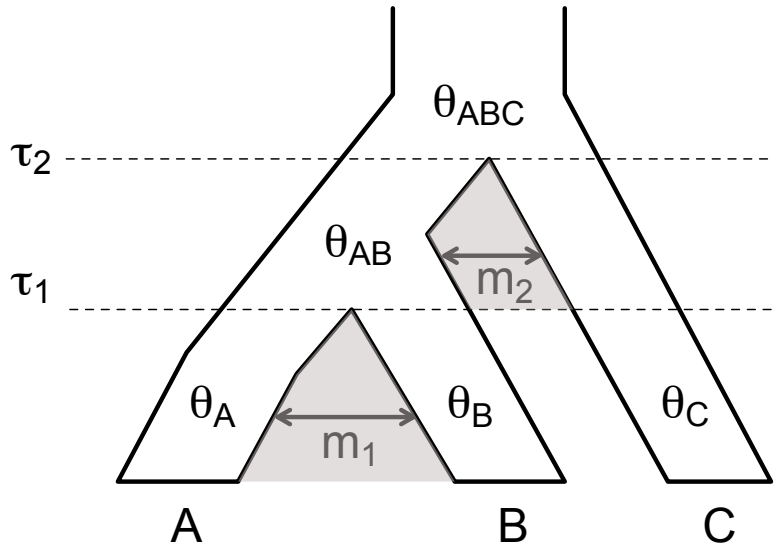


15 genes

- We also know from our earlier work that the probability that the gene tree matches the species tree, say p, is $\mathrm{p} = 1 - \frac{2}{3}e^{-t}$
- From the data, we estimate that $p = 0.7$ – use this to estimate t:

What about gene flow?

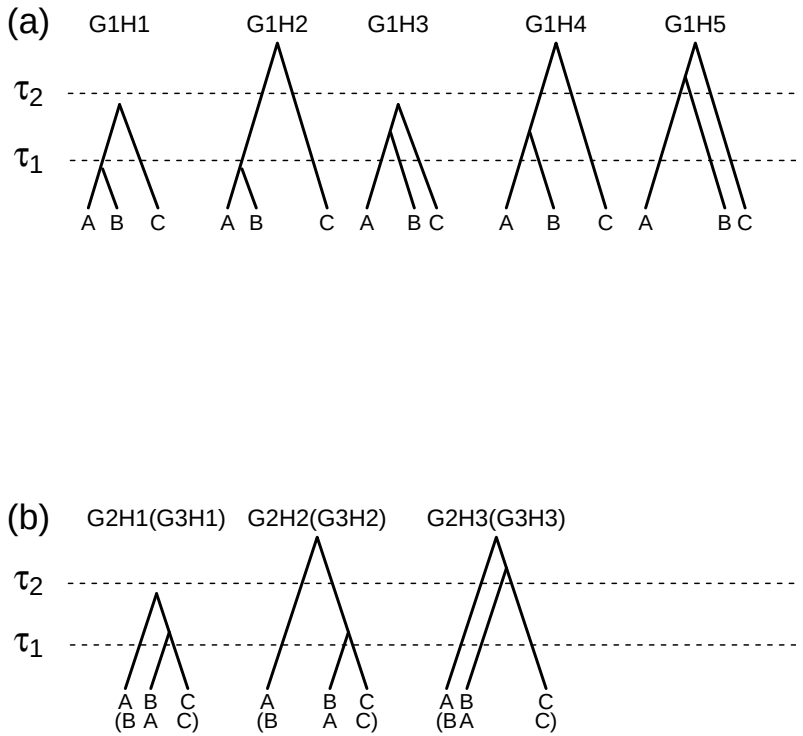
Question: What happens to gene tree topology probabilities under a model with gene flow?



Tian and Kubatko, MPE, 2017

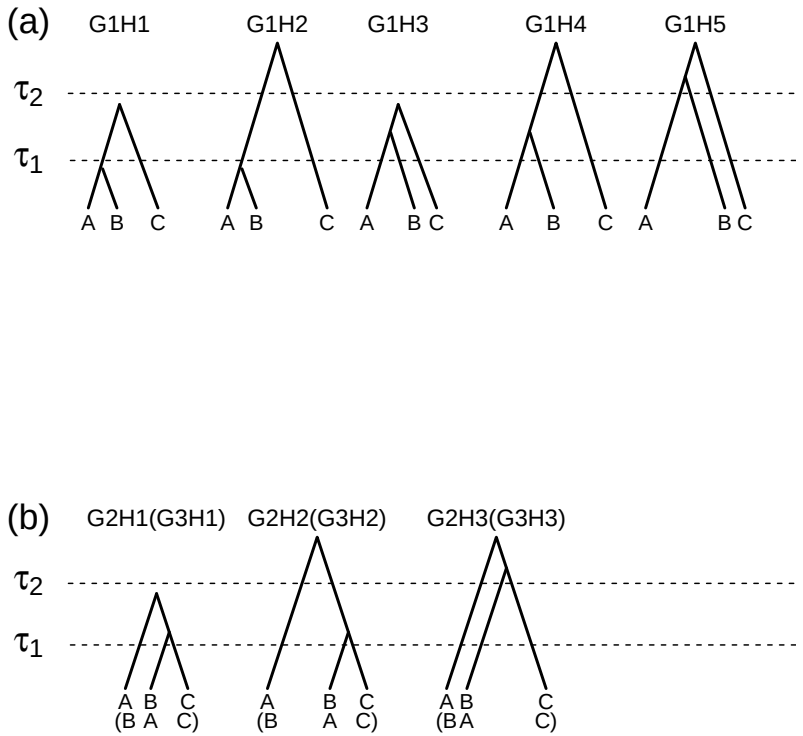
What about gene flow?

Complication: More **histories** are possible, because coalescent events can happen "before" speciation



What about gene flow?

Complication: More **histories** are possible, because coalescent events can happen "before" speciation

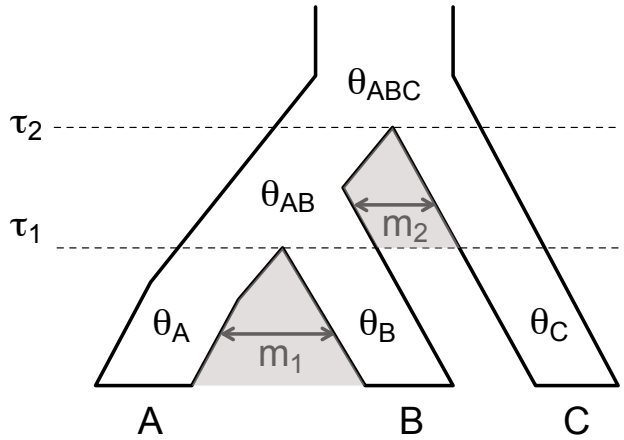


The gene tree that matches the species tree may not have the highest probability!

Recent results (Long and Kubatko, Systematic Biology, 2018):

- When $\theta_{AB} = \theta_C$, the gene tree that matches the species tree will have the highest probability (i.e., there are no anomalous gene trees)

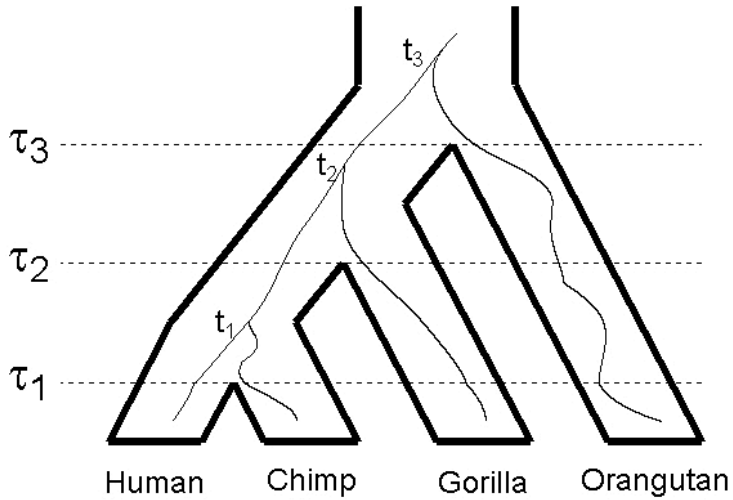
- When $\theta_{AB} \neq \theta_C$ and $m_2 > 0$, anomalous gene trees are possible – the probability of the gene tree matching the species tree could be as low as $\frac{1}{9}$ (leaving probability $\frac{4}{9}$ for each of the other two gene trees)



- When $\theta_{AB} \neq \theta_C$ and there is asymmetric gene flow between populations AB and C, anomalous gene trees are possible – the probability of the gene tree matching the species tree can go to 0 for highly asymmetric rates

A slightly larger case – no gene flow

- Consider 4 taxa – the human-chimp-gorilla problem



Coalescent histories for the 4-taxon example

- There are 5 possible histories for this example:

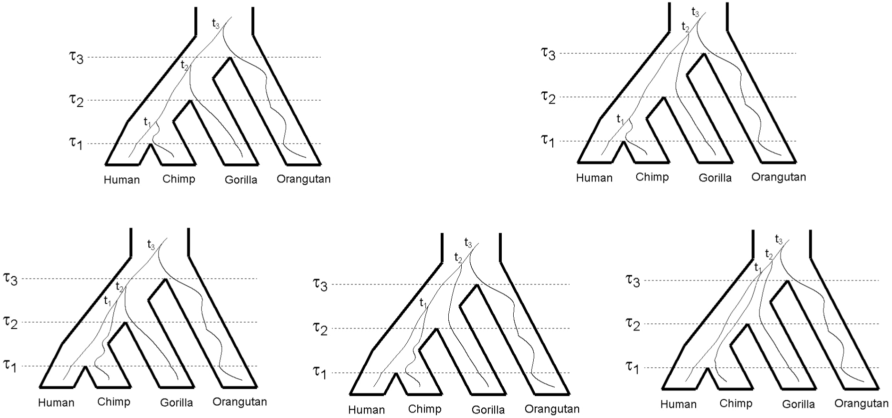


TABLE 3. The number of valid coalescent histories when the gene tree and species tree have the same topology. The number of histories is also the number of terms in the outer sum in equation (12).

	Number of histories		
Taxa	Asymmetric trees	Symmetric trees	Number of topologies
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1430	481	2,027,025
10	4862	1369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	6.190×10^{15}
20	1,767,263,190	100,360,324	8.201×10^{21}

Degnan and Salter, *Evolution*, 2005

- In the general case, we have the following:

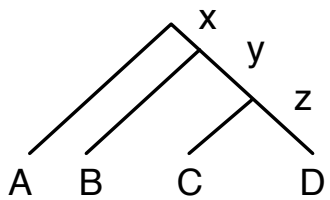
  The probability of a gene tree g gives the species tree $\mathcal{S}$ is given by

$$P\{G = g|\mathcal{S}\} = \sum_{histories} P\{G = g, history|\mathcal{S}\}$$

- Implemented in the software COAL (Degnan and Salter, *Evolution*, 2005)
- A more efficient method has been proposed (Wu, *Evolution*, 2012)

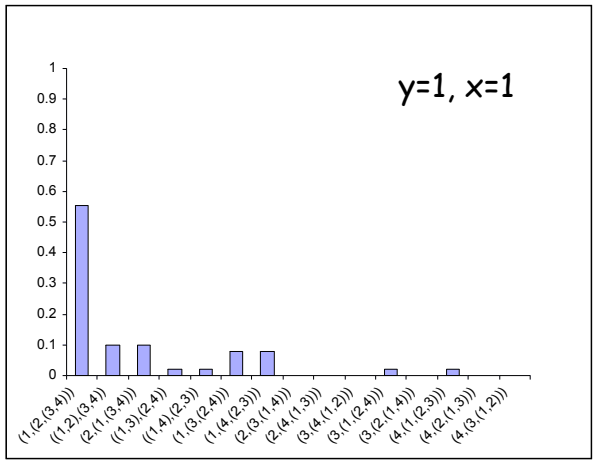
Gene tree distribution for four taxa

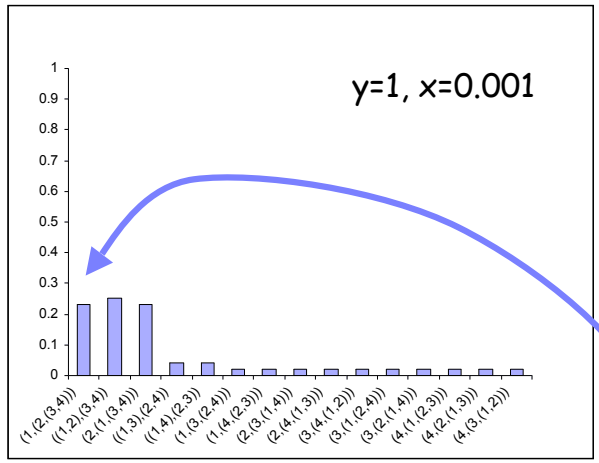
- In the three-taxon case in the absence of gene flow, there are no anomalous gene trees
- Question: Must the distribution always look this way?
- Examine the entire distribution for four taxa – only 15 gene trees are possible
- For the species tree:

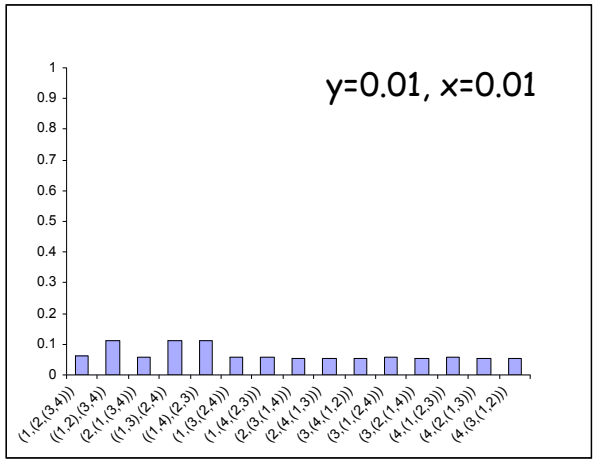


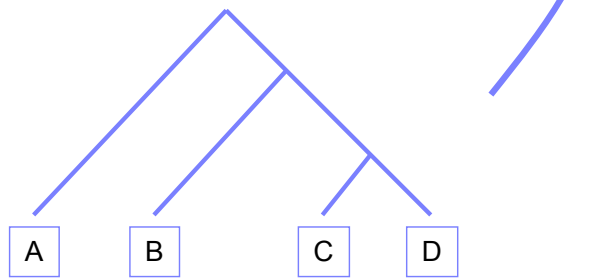
look at probabilities of all 15 gene tree topologies for values of x, y, and z

Gene tree distribution for four taxa

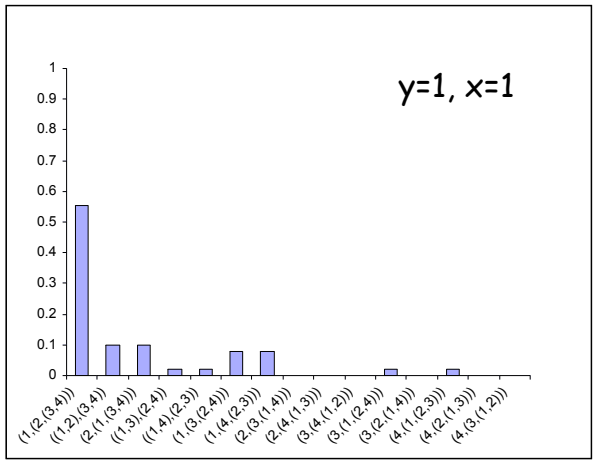


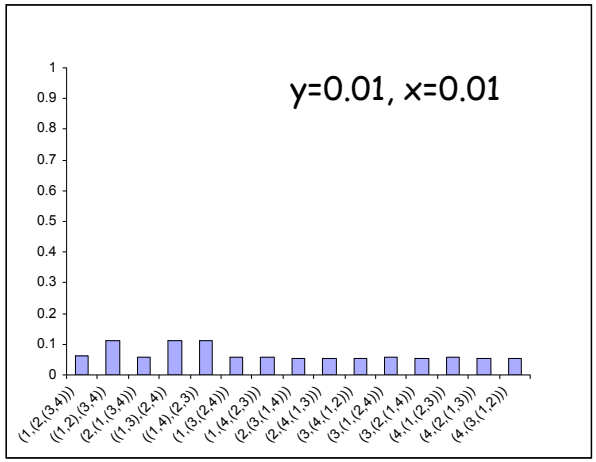


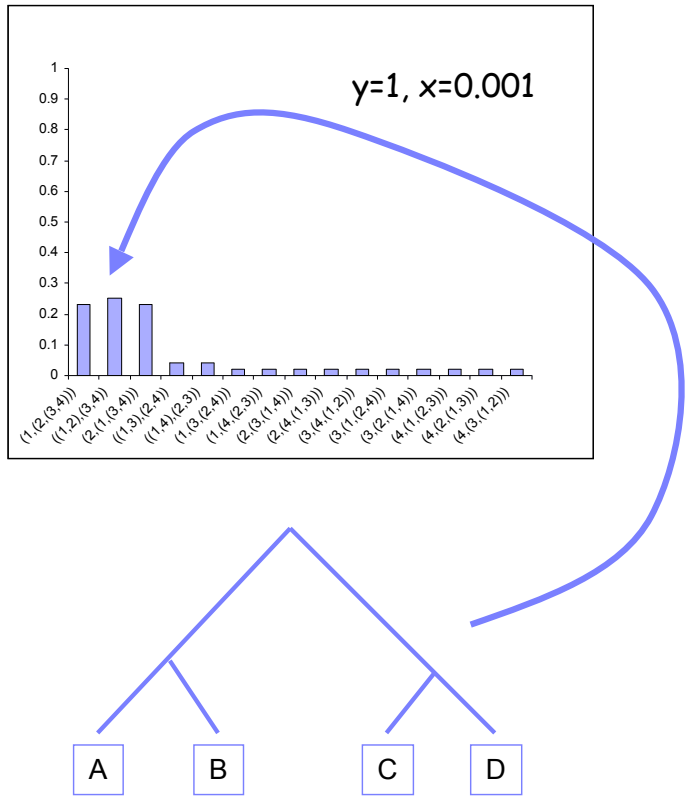




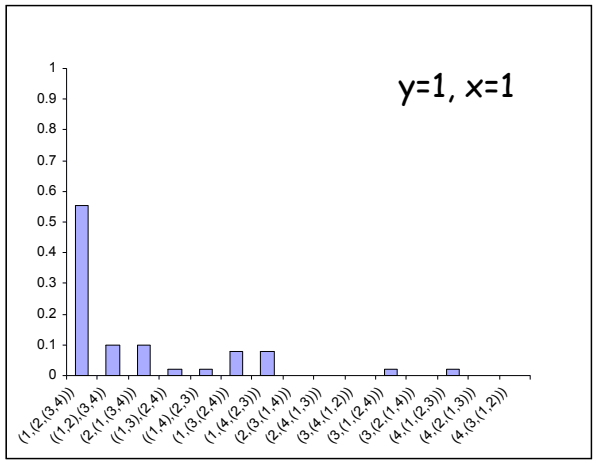
Gene tree distribution for four taxa

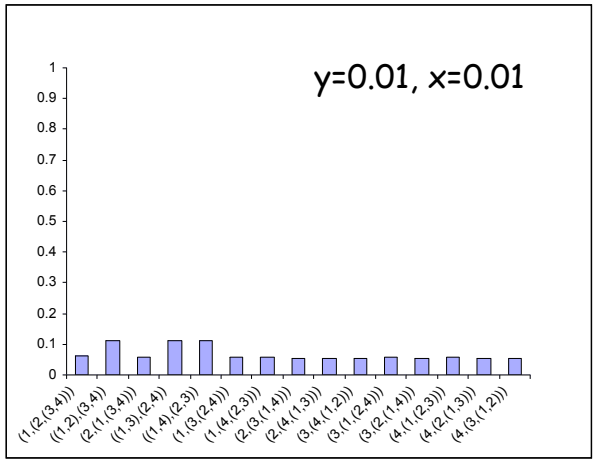


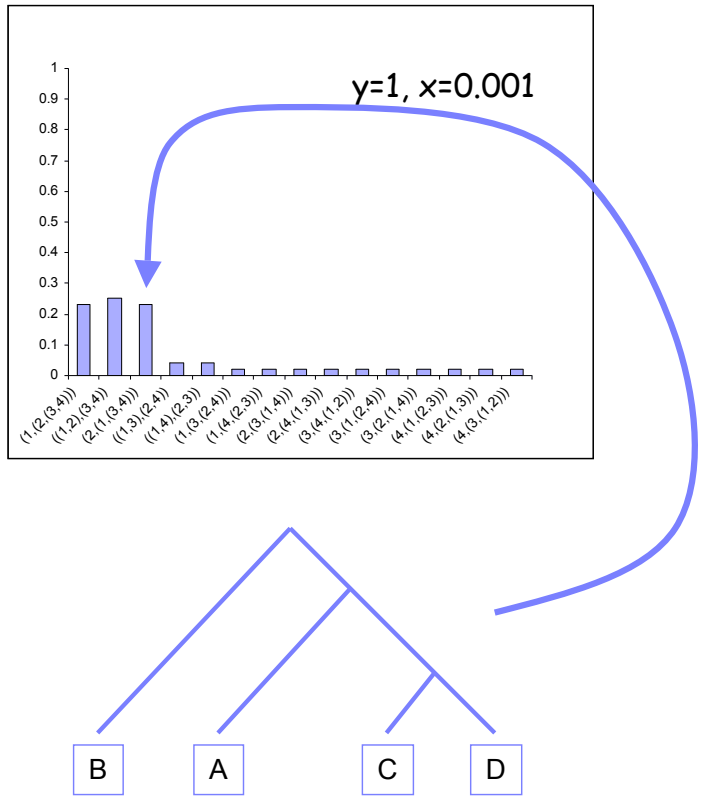




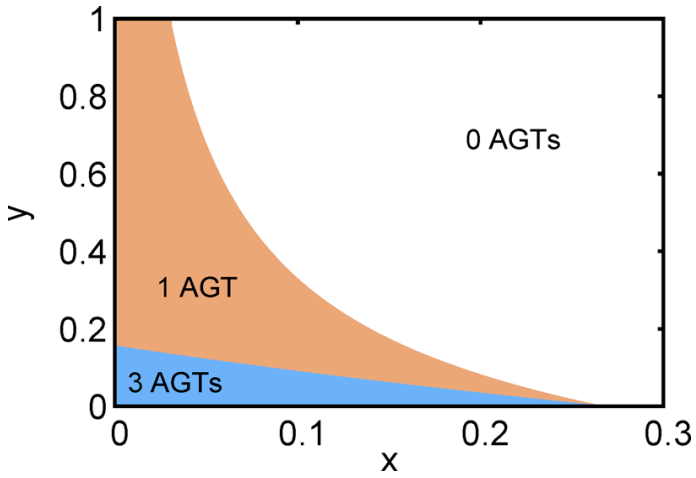
Gene tree distribution for four taxa







Gene tree distribution for four taxa



- The existence of anomalous gene trees has implications for the inference of species trees

Degnan and Rosenberg, *PLoS Genetics*, 2006

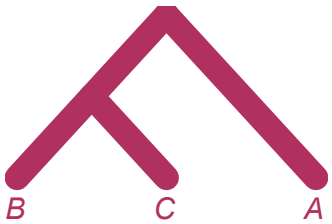
Rosenberg and Tao, *Systematic Biology*, 2008

What about mutation?

- What about mutation? How does this affect data analysis?
- The coalescent gives a model for determining gene tree probabilities for each gene.
- View DNA sequence data as the results of a two-stage process:
  - Coalescent process generates a gene tree topology.
  - Given this gene tree topology, DNA sequences evolve along the tree.
- Go back to our three-taxon example to get some intuition about the model

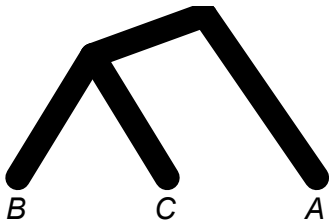
Phylogenetic coalescent model with mutation

t = length of interval between coalescent events $= 1.0$



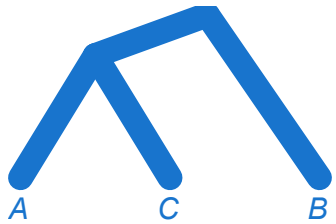
$1 - e^{-t}$

0.63



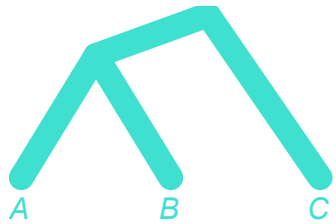
$\frac{1}{3}e^{-t}$

0.12



$\frac{1}{3}e^{-t}$

0.12

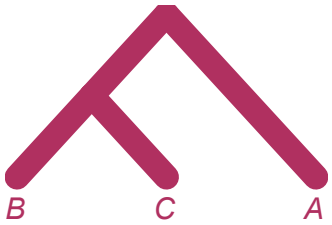


$\frac{1}{3}e^{-t}$

0.12

Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}

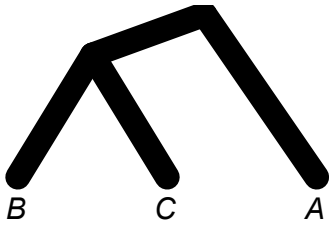
Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}



$1 - e^{-t}$

0.63

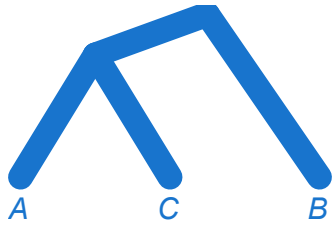
$p^{1a}_{TGT} = 0.1$



$\frac{1}{3}e^{-t}$

0.12

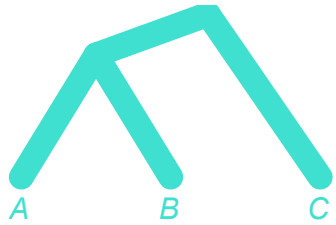
$p^{1b}_{TGT} = 0.025$



$\frac{1}{3}e^{-t}$

0.12

$p^{2}_{TGT} = 0.2$

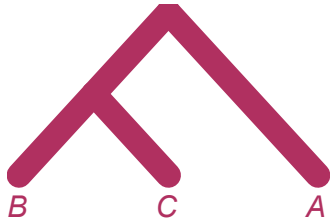


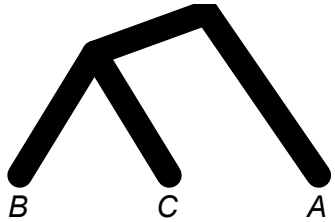
$\frac{1}{3}e^{-t}$

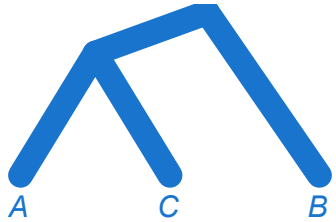
0.12

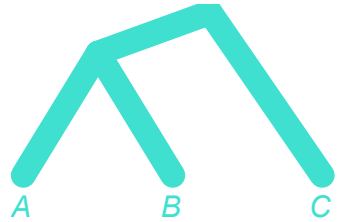
$p^{3}_{TGT} = 0.025$

Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}





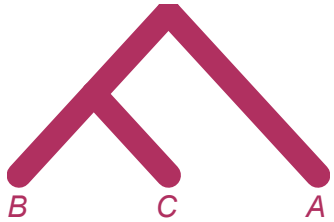


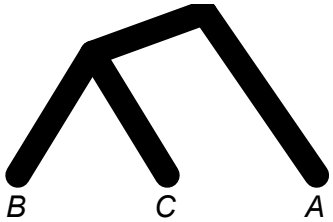


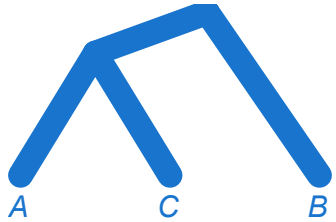
$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	s 0.12
$p_{TGT}^{1a} = 0.05$	$p_{TGT}^{1b} = 0.025$	$p_{TGT}^{2} = 0.2$	$p_{TGT}^{3} = 0.025$

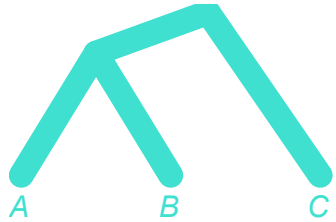
$$p_{TGT} = 0.63 \times 0.05 + 0.12 \times 0.025 + 0.12 \times 0.2 + 0.12 \times 0.025 = 0.0615$$

Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}







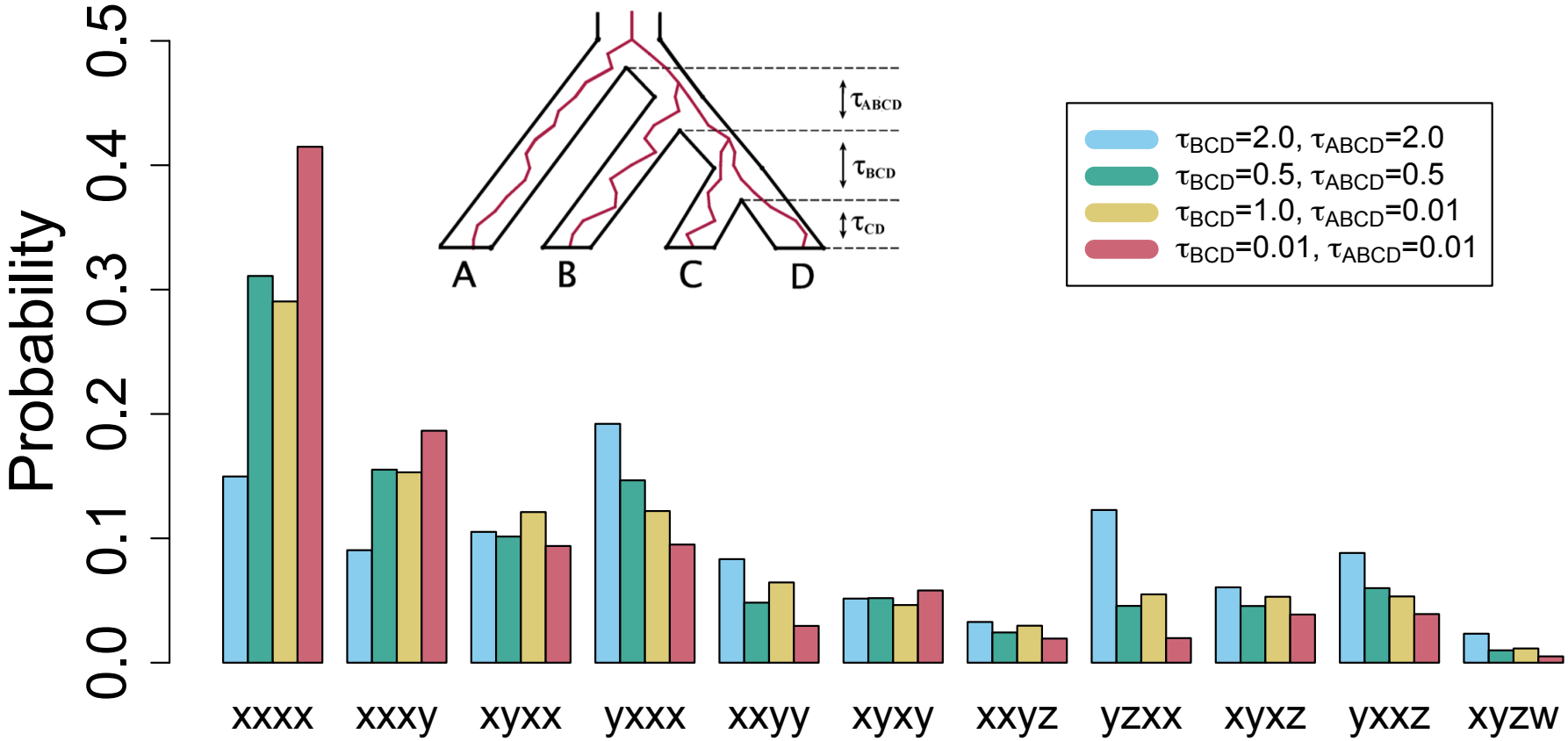


$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	s 0.12
$p_{TGT}^{1a} = 0.05$	$p_{TGT}^{1b} = 0.025$	$p_{TGT}^{2} = 0.2$	$p_{TGT}^{3} = 0.025$

$$p_{TGT} = 0.63 \times 0.05 + 0.12 \times 0.025 + 0.12 \times 0.2 + 0.12 \times 0.025 = 0.0615$$

$\Uparrow$ *For intuition only, not completely correct ...*

What does the site pattern probability distribution look like?



What about mutation?

Given this model, how should inference be carried out?

- As more data (genes) are added, the process of estimating species trees from concatenated data can be statistically inconsistent
- May fail to converge to any single tree topology if there are many equally likely trees.
- May converge to the wrong tree when a gene tree that is topologically incongruent with the species tree has the highest probability.
- The bootstrap may be positively misleading – show strong support for an incorrect clade

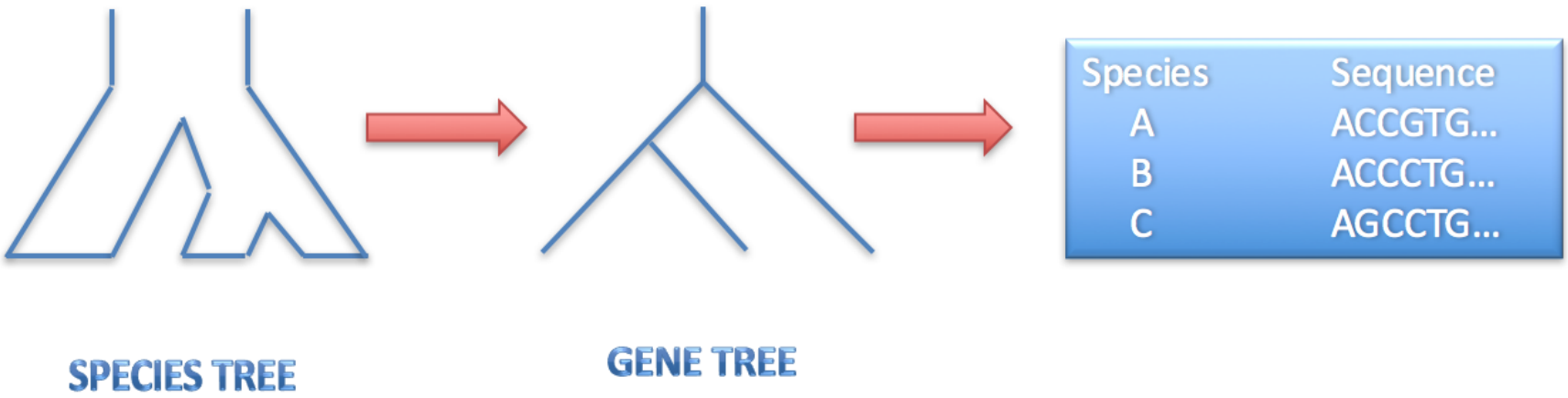
  Important note: This is NOT a failing of the bootstrap methodology; the observed "poor" performance is due to the use of an incorrect model (concatenation)

Kubatko and Degnan, 2007; Roch and Steel, 2015

Is there a better way to estimate species phylogenies?

Explicitly model the coalescent process!

Phylogenetic coalescent model with mutation



The likelihood function

- Suppose that we have available alignments for N genes, denoted by $D_1, D_2, \ldots, D_N$
- We would like to find the likelihood of the species phylogeny given these N alignments, assuming that
  - individual gene trees are randomly generated according to the coalescent
  - evolution of sequences along fixed gene trees occurs following a standard nucleotide-based Markov model
  - the data for the genes are independent given the species tree and associated parameters

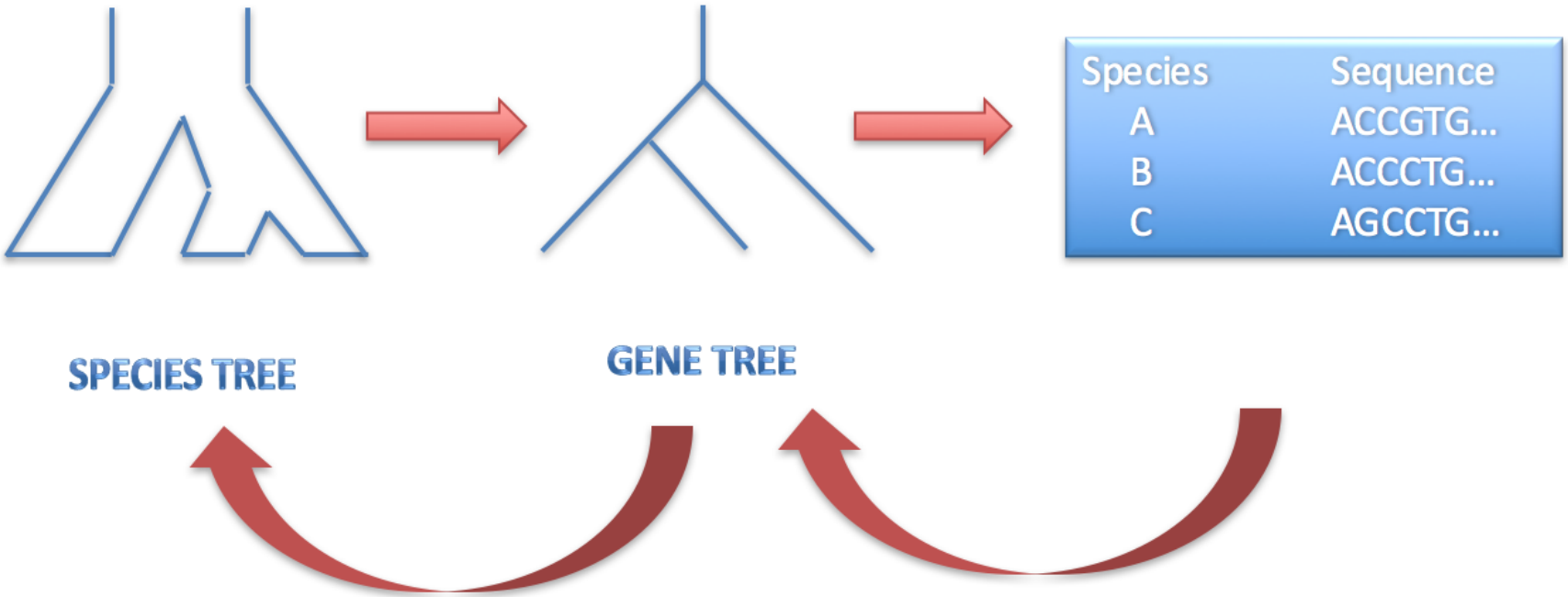
- Recall the Felsenstein equation from Peter's lecture, except that now we replace θ with S, the species tree. Use this to form the species tree likelihood for a multi-locus data set:

$$\begin{aligned} L(S|D_1, D_2, \ldots D_N) &= \prod_{i=1}^{N} P(D_i|S) \text{ [loci conditionally independent]} \\ &= \prod_{i=1}^{N} \sum_{j=1}^{G} P(D_i|g_j) f(g_j|S) \end{aligned}$$

  where S is the species tree (topology and branch lengths) and g_j represents a gene tree.

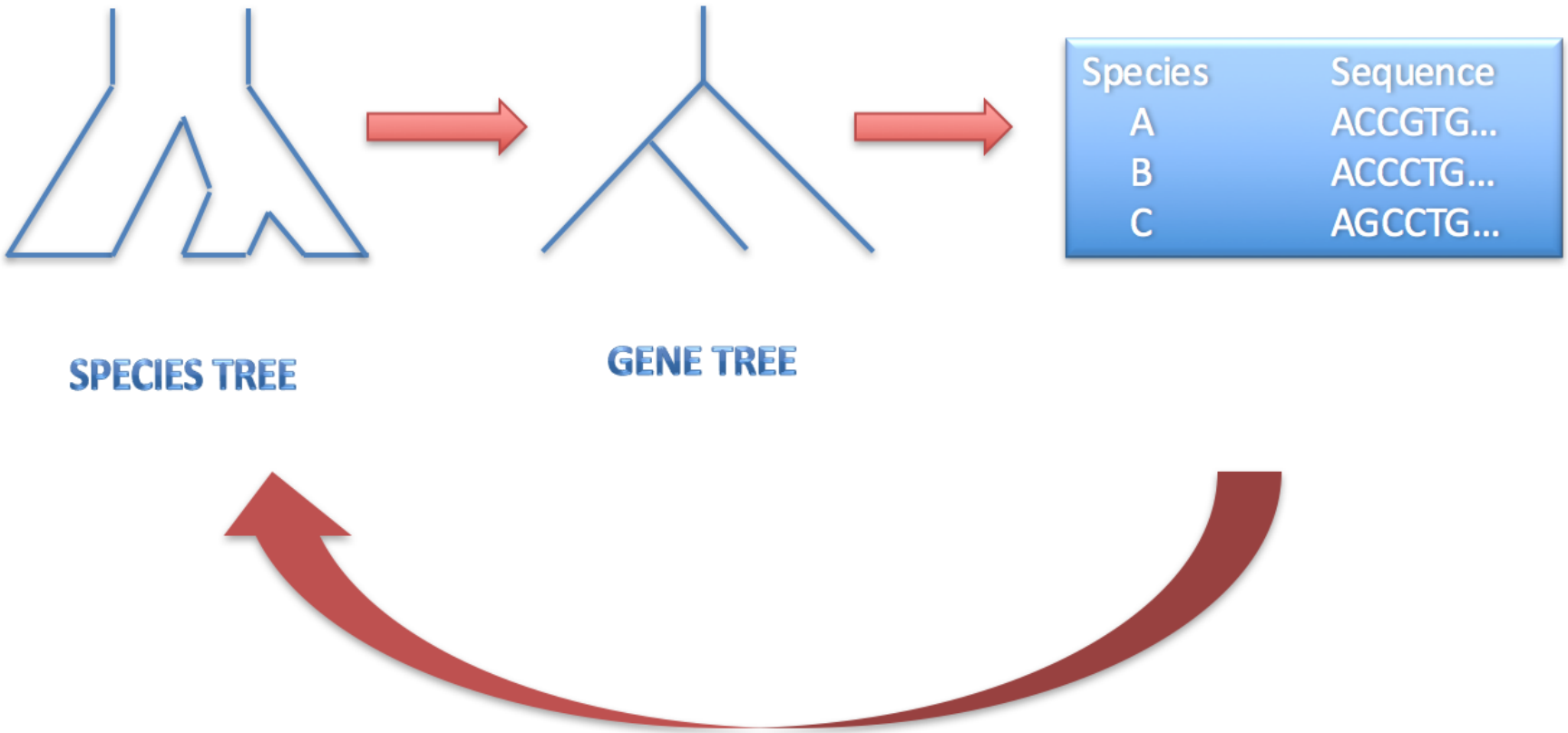
- This likelihood is difficult to evaluate directly, because of the dimension of the inner sum (which is really an integral)

Inference option 1: Summary statistics methods



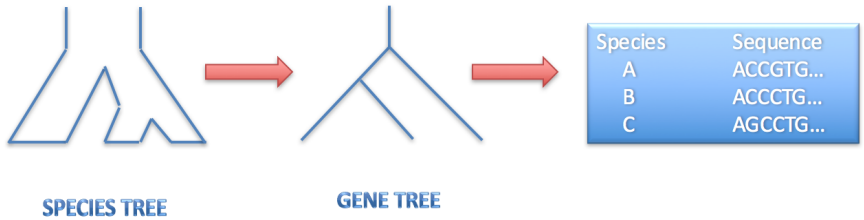
- Summary statistics methods: Start with estimated gene trees
  - Using estimated branch lengths:
    - STEM (Kubatko et al. 2009)
    - STEAC (Liu et al. 2009)
  - Using topology information only:
    - STAR (Liu et al. 2009)
    - Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 2009)
    - MP-EST (Liu et al. 2010)
    - ST-ABC (Fan and Kubatko 2011)
    - STELLS (Wu 2011)
    - ASTRAL (Mirarab et al. 2014)
    - Statistical binning (Bayzid et al. 2014)

Inference option 2: Full data methods



Full data methods I: BEST, *BEAST/STARBEAST2, BPP, SNAPP

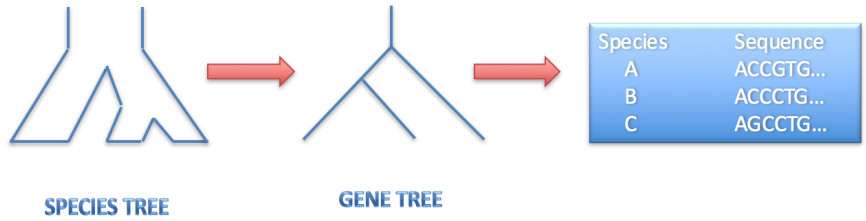
- Model the entire process of data generation
- Goal of these methods is to estimate the posterior distribution of the gene trees and species tree and associated model parameters



- BEST, *BEAST/STARBEAST2, and BPP use MCMC by considering both gene trees and the species tree, but their implementations are different
- SNAPP uses a clever two-step peeling algorithm to carry out the integration over gene trees, allowing it to consider a reduced space – but currently limited to biallelic data.

Full data methods II: SVDQuartets

- Model the entire process of data generation
- Avoid computing the likelihood by using algebraic structure in the distribution of site pattern probabilities under the model



- SVDQuartets is implemented in PAUP*
- SVDQuartets will be discussed in detail in this afternoon's lab

- Comparison of approaches:
  - Summary statistics methods
    - ⋆ Advantage: Quick
    - ⋆ Disadvantage: Ignore information in the data
    - ⋆ Most current implementations do not easily allow assessment of uncertainty (but bootstrap can be used, at the expense of computational efficiency)
  - Full data methods
    - ⋆ Advantage: Fully model-based framework
    - ⋆ Disadvantage: Computationally intensive, sometimes prohibitively so
    - ⋆ BEST, *BEAST/STARBEAST2, BPP, and SNAPP utilize a Bayesian framework and involve MCMC

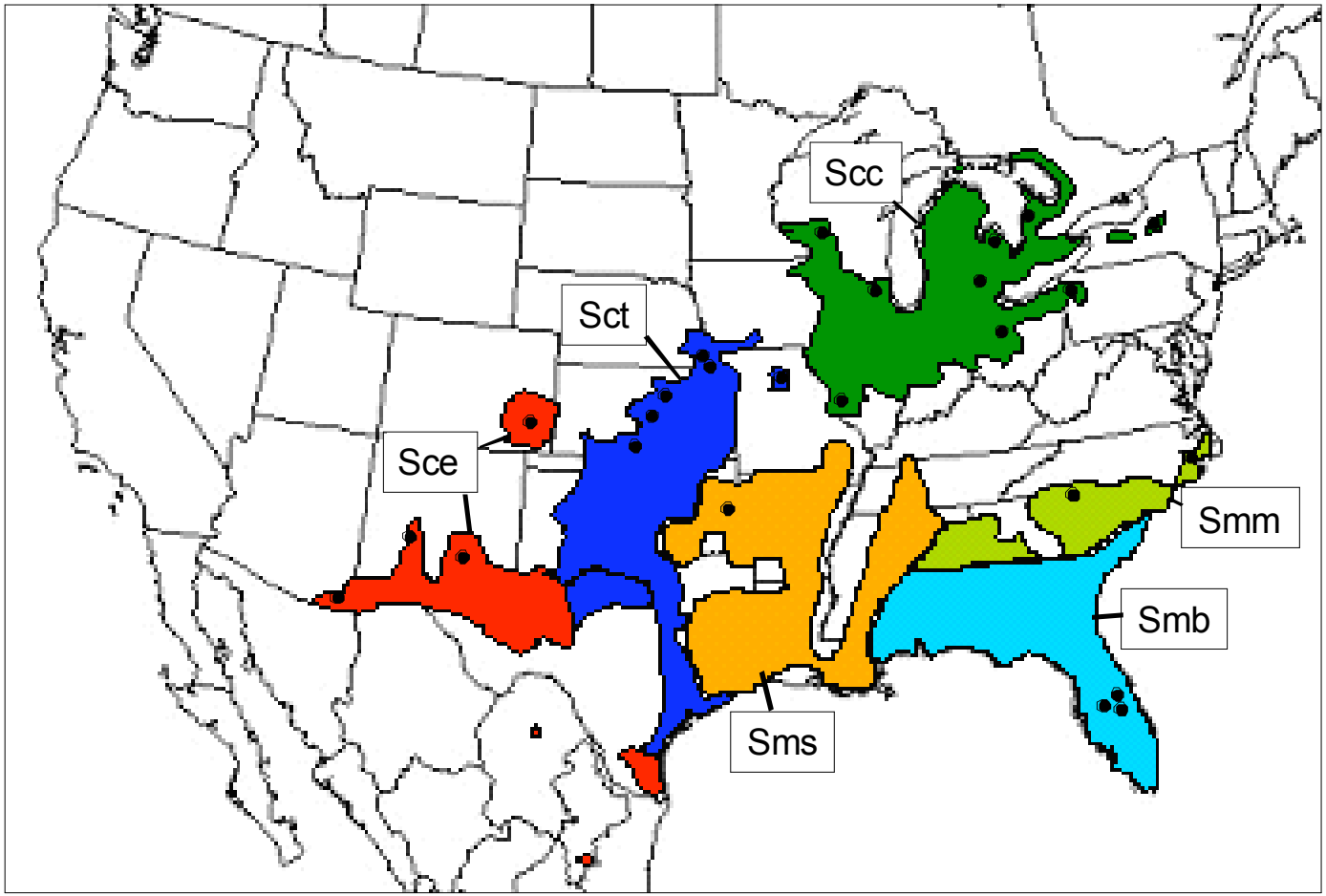
- Comparison of approaches:
  - Summary statistics methods
    - ⋆ Advantage: Quick
    - ⋆ Disadvantage: Ignore information in the data
    - ⋆ Most current implementations do not easily allow assessment of uncertainty (but bootstrap can be used, at the expense of computational efficiency)
  - Full data methods
    - ⋆ Advantage: Fully model-based framework
    - ⋆ Disadvantage: Computationally intensive, sometimes prohibitively so
    - ⋆ BEST, *BEAST/STARBEAST2, BPP, and SNAPP utilize a Bayesian framework and involve MCMC
- Ugh! Do we really need the coalescent? Why not just **concatenate**????
  - Well, the model is incorrect, and alternatives are available with a little effort
  - Also: the model matters for quantification of uncertainty and branch length estimation

Example 1: Sistrurus rattlesnakes

- North American Rattlesnakes - Joint work with Dr. Lisle Gibbs (EEOB at OSU)
- Of interest evolutionarily because of the diversity of venoms present in the various species and subspecies.
- Of conservation interest because population sizes in the eastern subspecies are very small.

[Pictures by Jimmy Chiucchi and Brian Fedorko]

Geographic Distribution of Snake Populations

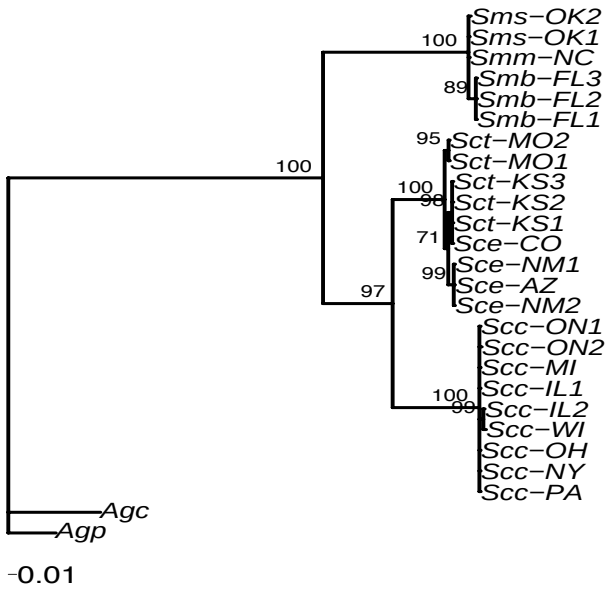


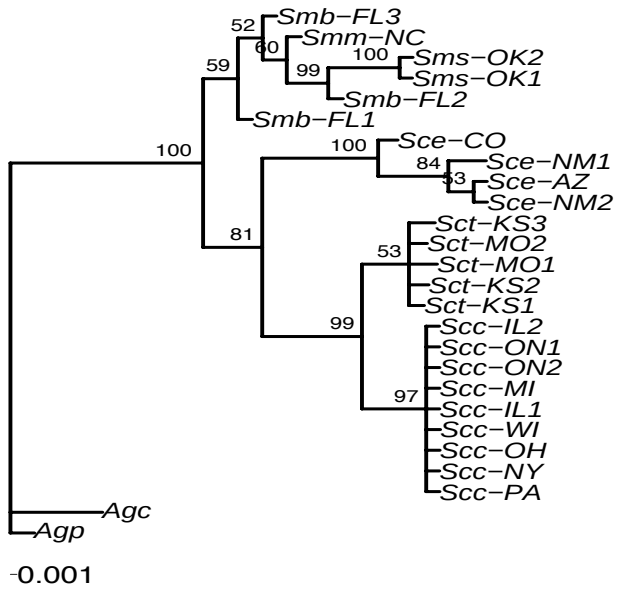
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

Species	Location	No. of individuals per gene
S. catenatus catenatus	Eastern U.S. and Canada	9
S. c. edwardsii	Western U.S.	4
S. c. tergeminus	Western and Central U.S.	5
S. miliarius miliarius	Southeastern U.S.	1
S. m. barbouri	Southeastern U.S.	3
S. m. streckerii	Southeastern U.S.	2
Agkistrodon sp. (outgroup)	U.S.	2

Individual Gene Tree Estimates

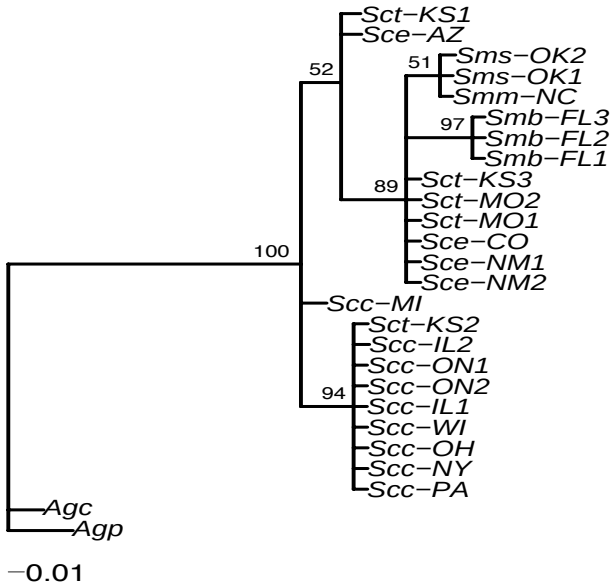
Some are very informative:

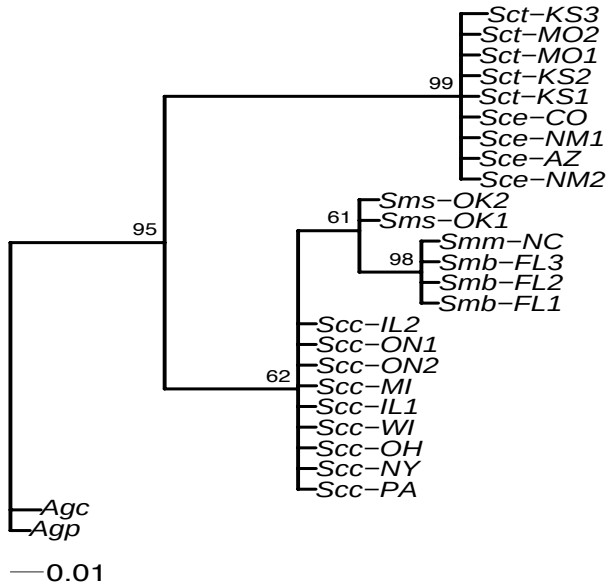




Individual Gene Tree Estimates

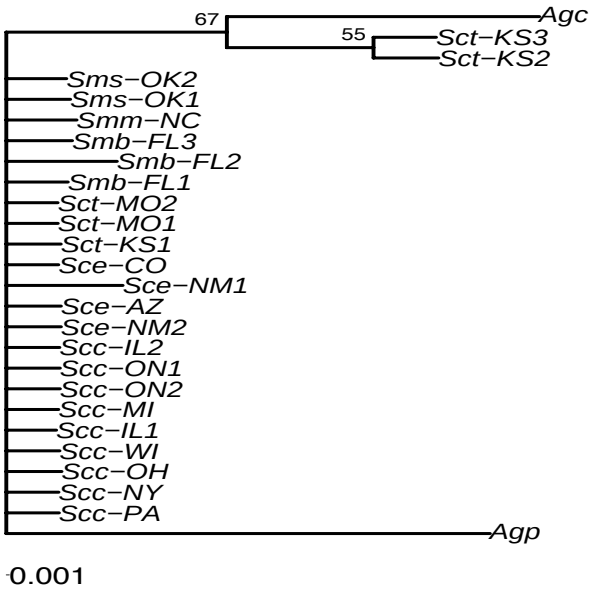
Some are a little informative:

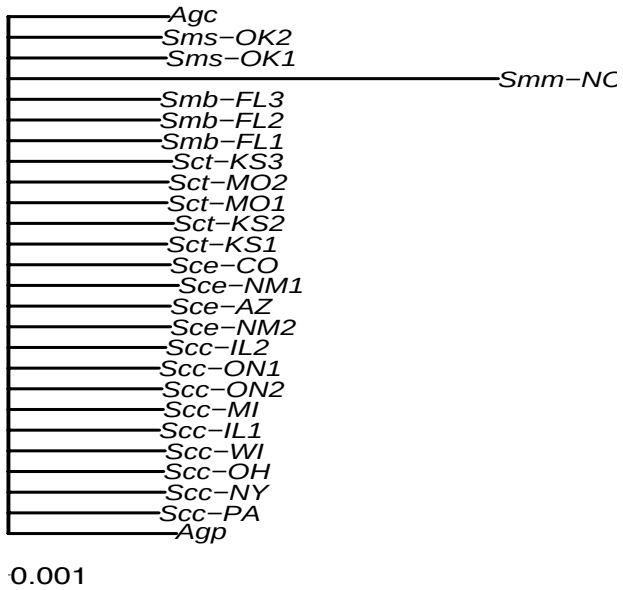




Individual Gene Tree Estimates

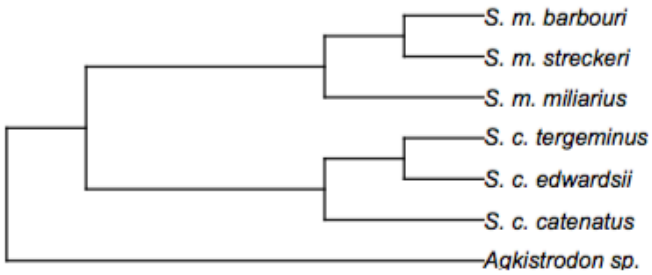
And then there are others .....



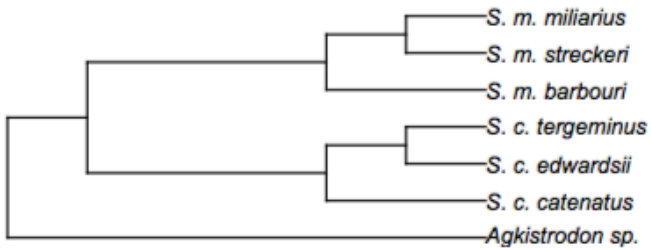


Example 1: Sistrurus rattlesnakes

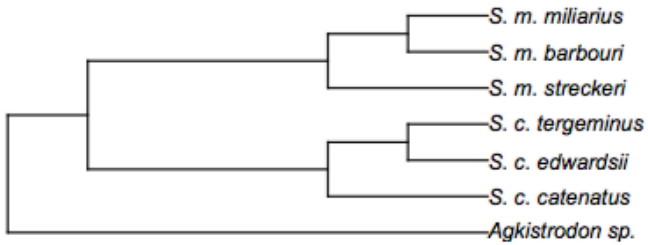
STEM, STEAC



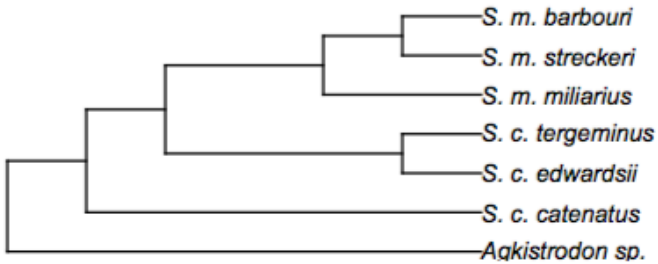
BEAST (concatenated data), *BEAST



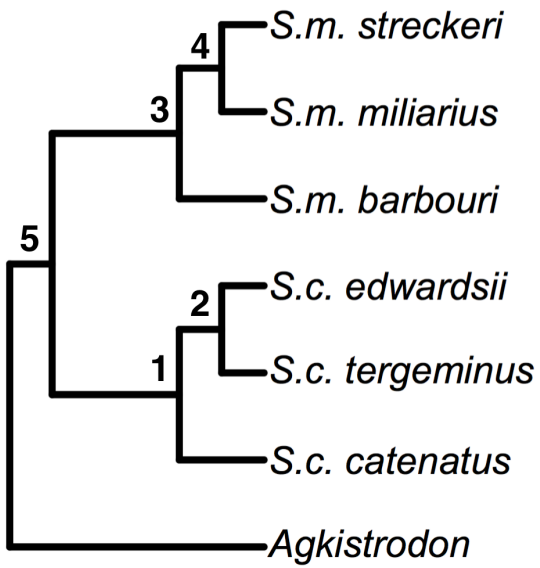
BEST, Parsimony & MrBayes (concatenated data), Astral



PhyloNet, STAR



Example 1: Sistrurus rattlesnakes

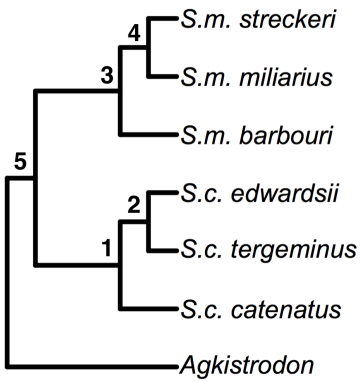


Node	1	2	3	4	5
*BEAST	100	100	100	46*	100
BPP	100	99	100	33*	100
SVDQ	93	100	100	46	100

* = This clade was not in the maximum clade credibility (*S. m. miliarius* and *S. m. barbouri* received 48.78% posterior probability with *BEAST and 59% posterior probability with BPP)

Example 1: Sistrurus rattlesnakes

Very rough ideas of computational time ...



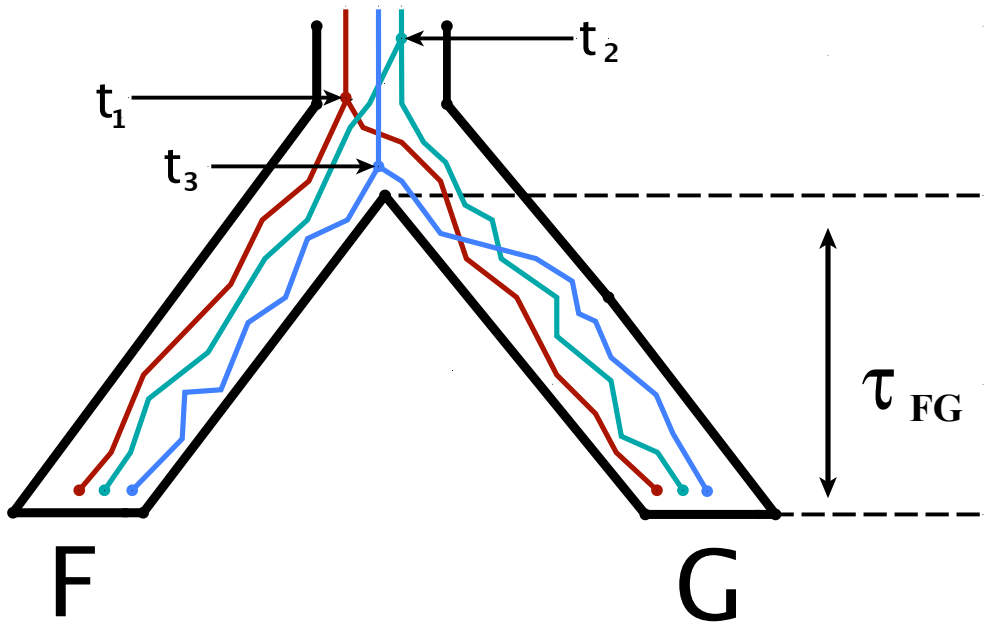
Program	Time	Details
BEST	∼3 days	11,770,000 iterations (not converged)
*BEAST	16.8 hours	100,000,000 iterations all ESS $>$ 200 except 1 ($>$100)
BPP	4 days	500,000 iterations
SVDQ	11 minutes	all quartets sampled 100 bootstrap reps
ASTRAL	2.215 sec	given gene trees! also need bootstrap

- How does concatenation do?
  - Tree agrees with estimated species tree (both with BEAST and with ML in PAUP*)
    - ★ BEAST: posterior probability on *miliarius* clade: 73%
  - Speciation time estimates are severely biased:

Dated node	Divergence estimates from concatenated gene tree (Ma)[a]	Divergence estimates from species tree (Ma)[a]	Percent difference[b] (%)
(Scc (Sce,Sct)) vs. (Sms(Smb, Smm))	9.45 (9.14, 10.24)	10.04 (9.25, 12.97)	+6
Scc vs. (Sce, Sct)	6.06 (5.22, 7.02)	2.92 (1.58,4.90)	−52
Sce vs. Sct	2.41 (2.01, 2.88)	0.47 (0.24, 0.86)	−79
Smb vs. (Smb, Sms)	1.98 (1.60, 2.47)	0.77 (0.44,1.31)	−62
Sms vs. Smm	1.60 (1.23, 2.06)	0.49 (0.25, 0.92)	−69

Example 1: Sistrurus rattlesnakes

- Why are speciation times biased?
  - We estimate different quantities when using a gene tree vs. species tree analysis!



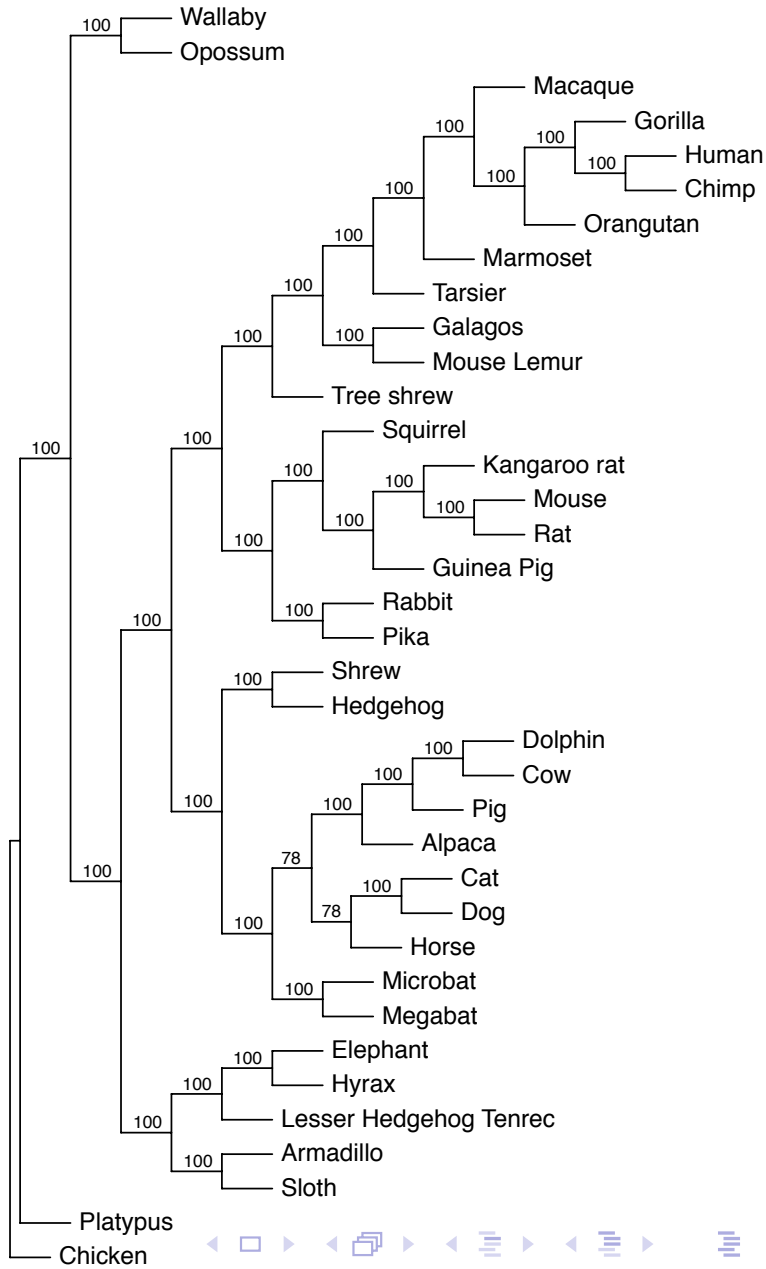
Multilocus data example 2: Mammals

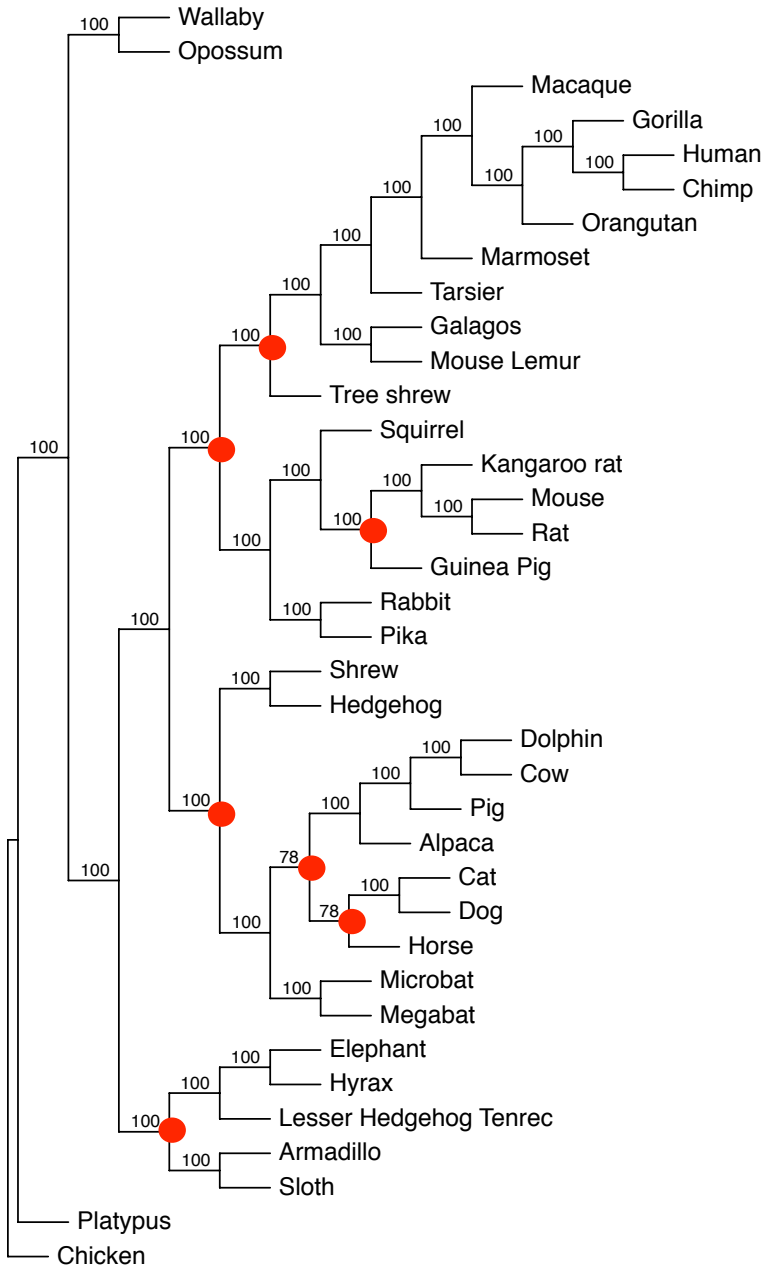
- Series of papers in the literature debating proper phylogenetic relationships among a group of mammals
  - Meredith RW, et al. (Science, 2011) criticized by Song et al. (PNAS, 2012):
    - ⋆ Amount of data "insufficient" (26 genes, 35,603 bp, 164 mammals)
    - ⋆ Concatenation not appropriate
  - Response by Gatesy and Springer (PNAS, 2013) criticizing Song et al.:
    - ⋆ Loci chosen not representative ("concatalescence" – exons 'pasted' together)
    - ⋆ Many nodes still not well supported
    - ⋆ Subset of 36 species
  - Wu et al. (PNAS, 2013) criticize Gatesy and Springer's response:
    - ⋆ Concatenation of all genes is worse than within a few genes
    - ⋆ The approach of treating exons from a single gene with introns stripped has worked well in other cases
  - etc. . . .

Example 2: Mammals

- Dataset: obtained from Liang Liu, 36 mammal species + outgroup, $\sim$ 1.4 million bp from 447 genes

- SVDQ run on 8-year old dual-core linux machine – 27 hours required to estimate the tree and obtain bootstrap support from 100 replicates





- "Historically problematic nodes" identified by McCormack et al. (Genome Research, 2012) are identified with a red circle
- Overall, the SVDQ analysis agrees with the analysis of Song et al. (2012), who used the coalescence-based method MP-EST
- The SVDQ analysis differs from analyses based on concatenation for some of the difficult nodes, but agrees with concatenation for the two nodes with lower bootstrap support

Species Tree Inference Summary – Comparison of Methods

Software	Data Type	Measure of Uncertainty	Computation Time	Models Included
BEST	multilocus	posterior probability	long; can be run in parallel	coalescent; all reversible substitution models
*BEAST/ STARBEAST2	multilocus	posterior probability	intermediate; can be run in parallel	coalesent; all reversible substitution models; relaxed clock; variable population sizes
BPP	multilocus	posterior probability	long	coalescent; JC69 model on molecular clock; species delimitation
SVDQ	multilocus; SNP	bootstrap	short	coalescent; all reversible substitution models; non-cl gene flow; parameter estimation ?
SNAPP	biallelic SNP; AFLP	posterior probability	long; can be run in parallel	coalescent; two-state substitution model; Bayes factor delimitation
ASTRAL	unrooted gene trees	local posterior probability	short given gene trees	no specific model assumed
MP-EST	rooted gene trees	bootstrap	short given gene trees	coalescent model

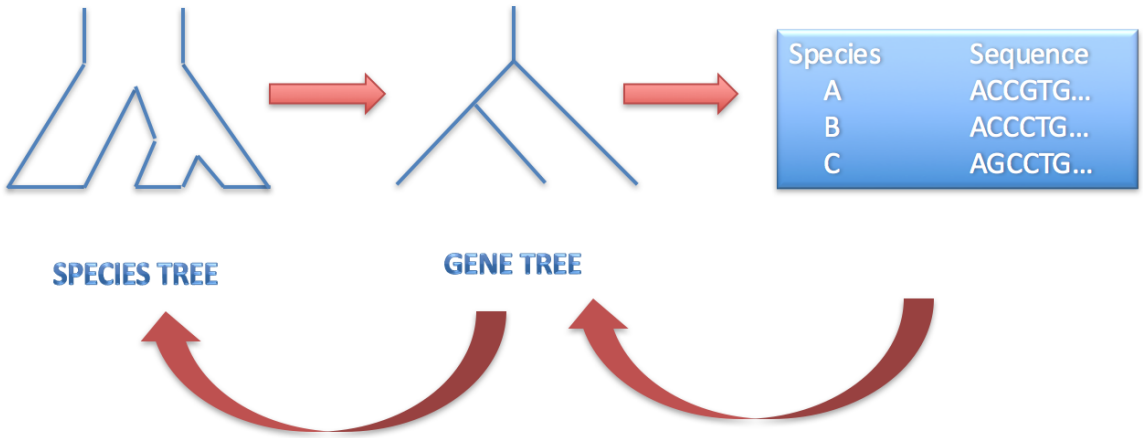
Species tree inference summary

- Failure to incorporate the coalescent model in estimation of the species tree can lead to statistical inconsistency, even when a method that is statistically consistent is applied.
- Many new methods for inferring species trees are being developed – each has its advantages and disadvantages.
- In addition, we should continue to think about other ways of using multi-locus data to its full advantage .... and we should be thinking beyond estimation of the species tree.
- Lots of areas emerging: species delimitation, incorporating horizontal events along the phylogeny, etc.

Key points to take away ....

- Gene trees and species trees are different – both conceptually and physically
- The coalescent model predicts a distribution of gene trees for a given species tree
  - 3 taxa:
  - Empirical data often fit this predicted distribution
- Three reasons a species tree analysis is preferred over concatenation:

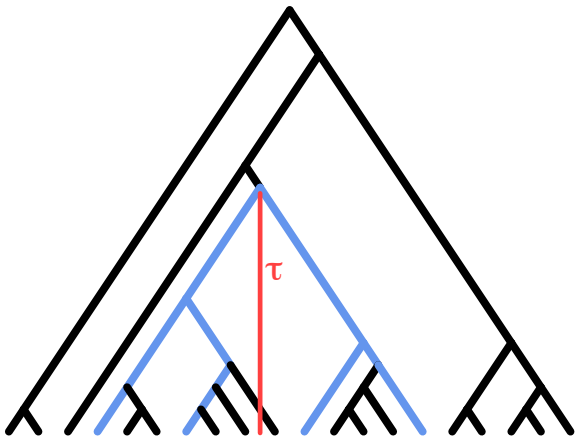
- So far: understand why we need a species tree and how difficult it can be to accurately estimate it!
- We have specified a model:

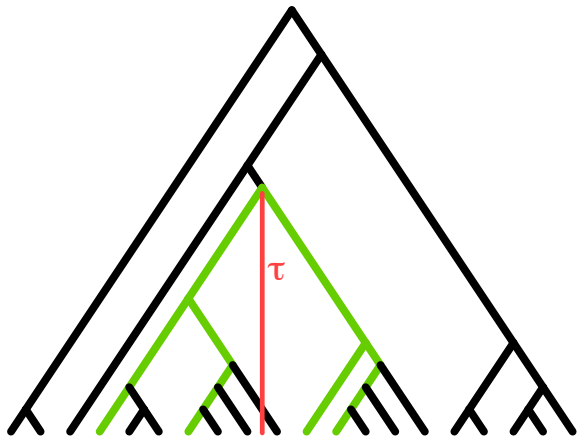


- Now consider the details of two methods for species tree inference under this model: ASTRAL and SVDQuartets

- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus
  - **Step 2.** Extract all quartet relationships from the estimated gene trees
  - **Step 3.** Find the species tree that "agrees" with as many quartets as possible

- **Step 2.** Extract all quartet relationships from the estimated gene trees





- **Step 3.** Find the species tree that "agrees" with as many quartets as possible
  - This is a non-trivial problem .... recall that we expect substantial incongruence among trees
  - However, *unrooted* gene trees cannot be anomalous for four taxa in the absence of gene flow, so *if the gene trees are correct*, then this is easy
  - ASTRAL uses the Weighted Quartet Score of a candidate species tree – defined to be the number of quartets from the set of input gene trees that agree with the candidate species tree
  - Optimization problem – need to search for the species tree that maximizes the Weighted Quartet Score

ASTRAL

- Example:

ASTRAL

- Example:

Additional features of ASTRAL

- ASTRAL → ASTRAL-II
  - Expand the set of input quartets beyond those found in the input gene trees
  - Improve search for the species tree that optimizes the Weighted Quartet score
  - Allow polytomies in input trees
- ASTRAL-II→ ASTRAL-III
  - Better options for handling multiple individuals per species
  - Improved algorithms (faster)
- ASTRAL can also estimate branch lengths (in coalescent units) and can provides a measure of uncertainty (local posterior probability)

ASTRAL performance

- ASTRAL is claimed to be statistically consistent
  - This is true when the gene trees are known without error

ASTRAL performance

- ASTRAL is claimed to be statistically consistent
  - This is true when the gene trees are known without error
- ASTRAL will perform well when the gene trees can be estimated well

ASTRAL performance

- ASTRAL is claimed to be statistically consistent
  - This is true when the gene trees are known without error
- ASTRAL will perform well when the gene trees can be estimated well
- Computational efficiency: the estimation of gene trees is the time-consuming step, but can be parallelized

ASTRAL performance

- ASTRAL is claimed to be statistically consistent
  - This is true when the gene trees are known without error
- ASTRAL will perform well when the gene trees can be estimated well
- Computational efficiency: the estimation of gene trees is the time-consuming step, but can be parallelized
- Crucial assumption: true unrooted quartets have higher probability than other quartet relationships

ASTRAL performance

- ASTRAL is claimed to be statistically consistent
  - This is true when the gene trees are known without error
- ASTRAL will perform well when the gene trees can be estimated well
- Computational efficiency: the estimation of gene trees is the time-consuming step, but can be parallelized
- Crucial assumption: true unrooted quartets have higher probability than other quartet relationships
- Assessment of uncertainty: use the local posterior probability (now recommended over the bootstrap)

Goal of this work:

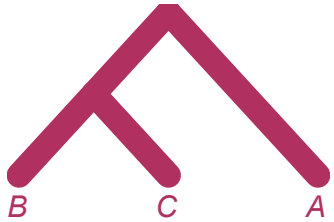
Develop a full data approach that is computationally feasible for large-scale data

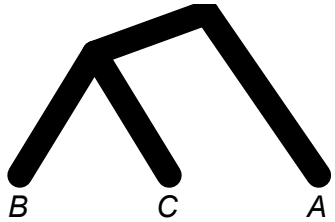
How?

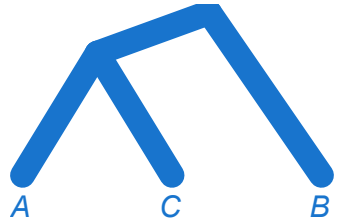
- Summarize data differently, so that model requires less computation
- Develop theory to infer relationships among quartets of taxa very accurately
- Use a quartet assembly method to build a large tree

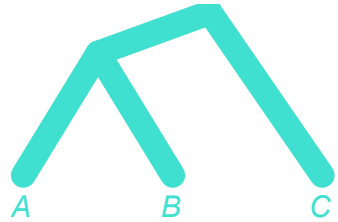
Recall the phylogenetic coalescent model with mutation

Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}









$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
$p_{TGT}^{1a} = 0.05$	$p_{TGT}^{1b} = 0.025$	$p_{TGT}^{2} = 0.2$	$p_{TGT}^{3} = 0.025$

$$p_{TGT} = 0.63 \times 0.05 + 0.12 \times 0.025 + 0.12 \times 0.2 + 0.12 \times 0.025 = 0.0615$$

$\Uparrow$ *For intuition only, not completely correct ...*

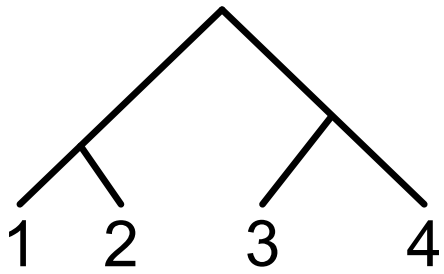
But .... there are a lot of histories!

TABLE 3. The number of valid coalescent histories when the gene tree and species tree have the same topology. The number of histories is also the number of terms in the outer sum in equation (12).

	Number of histories		
Taxa	Asymmetric trees	Symmetric trees	Number of topologies
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1430	481	2,027,025
10	4862	1369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	6.190×10^{15}
20	1,767,263,190	100,360,324	8.201×10^{21}

- This means that calculating the likelihood – and thus using likelihood-based methods for inference – will be difficult, especially for large-scale data
- Alternative approach: compute explicitly (i.e., write formulas for) the site pattern probabilities for 4-taxon trees, and look for "structure"

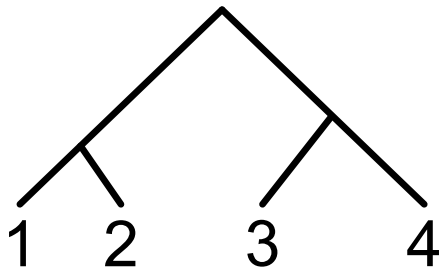
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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

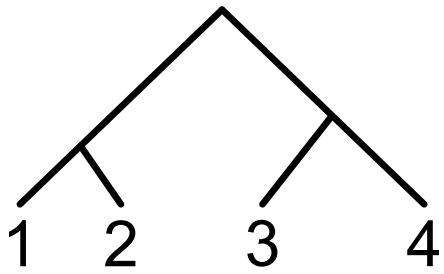
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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

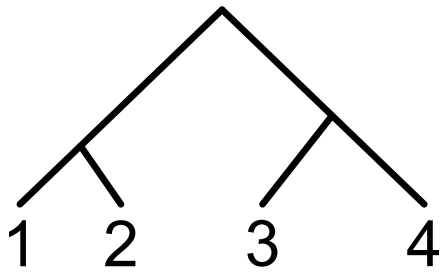
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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

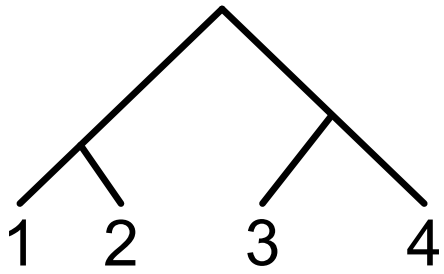
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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

Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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

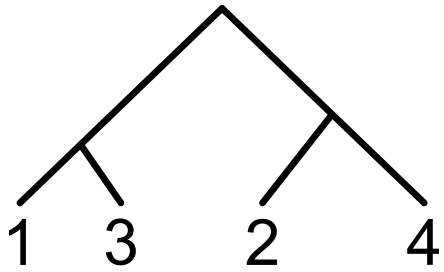
These two columns are identical – matrix rank is reduced by one

Main Result:

- Species tree inference: For a flattening matrix constructed on the true four-taxon tree, **the matrix rank is 10** under the following model
  - species tree $\rightarrow$ gene tree ::: coalescent process
  - gene tree $\rightarrow$ data ::: nucleotide substitution models: GTR+I+Γ and submodels

- **This result still holds** when the species tree violates the molecular clock and/or when there is variation in effective population size across the branches and/or when there is gene flow between sister taxa

What about the incorrect tree?



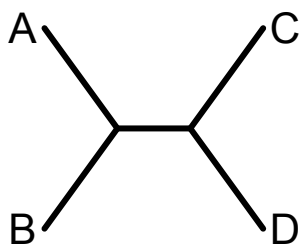
Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

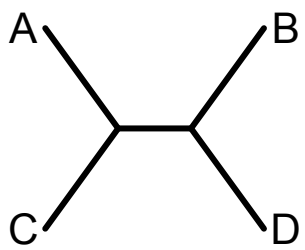
$$\mathbf{Flat}_{12|34}(\mathbf{P}) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\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

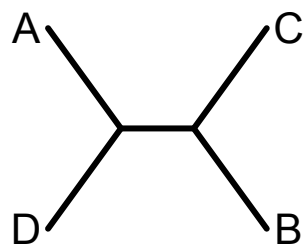
These two columns are no longer identical – full rank matrix in both cases (rank = 16)

How can we use these facts to estimate the species tree?

- **Basic idea:**
  - Data: aligned DNA sequences for multiple loci or for a collection of SNPs
  - Estimate the flattening matrix for each of the following trees:





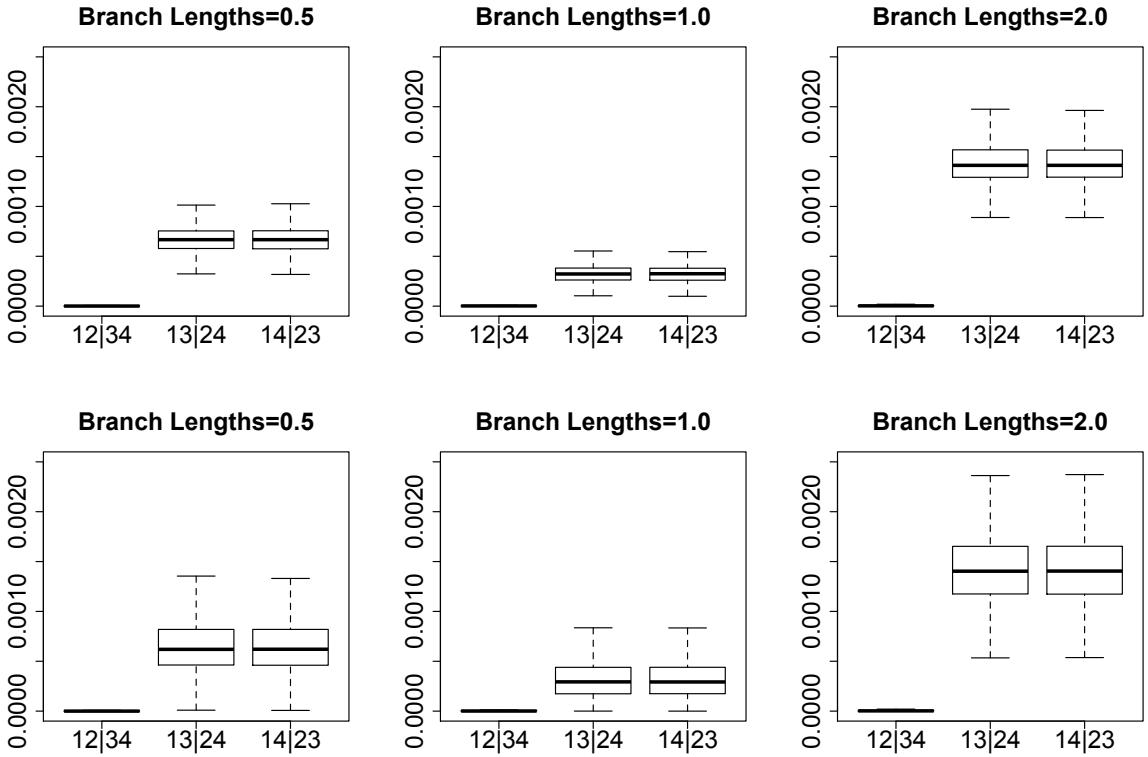


  - Compute a measure of how close each of the three observed flattening matrices is to a matrix with rank 10 – we use the SVDScore
  - Pick the tree relationship that gives the smallest SVDScore

Simulation study 1 – can we detect the correct split?

Simulate data from the Jukes-Cantor model for a 4-taxon tree and examine split scores

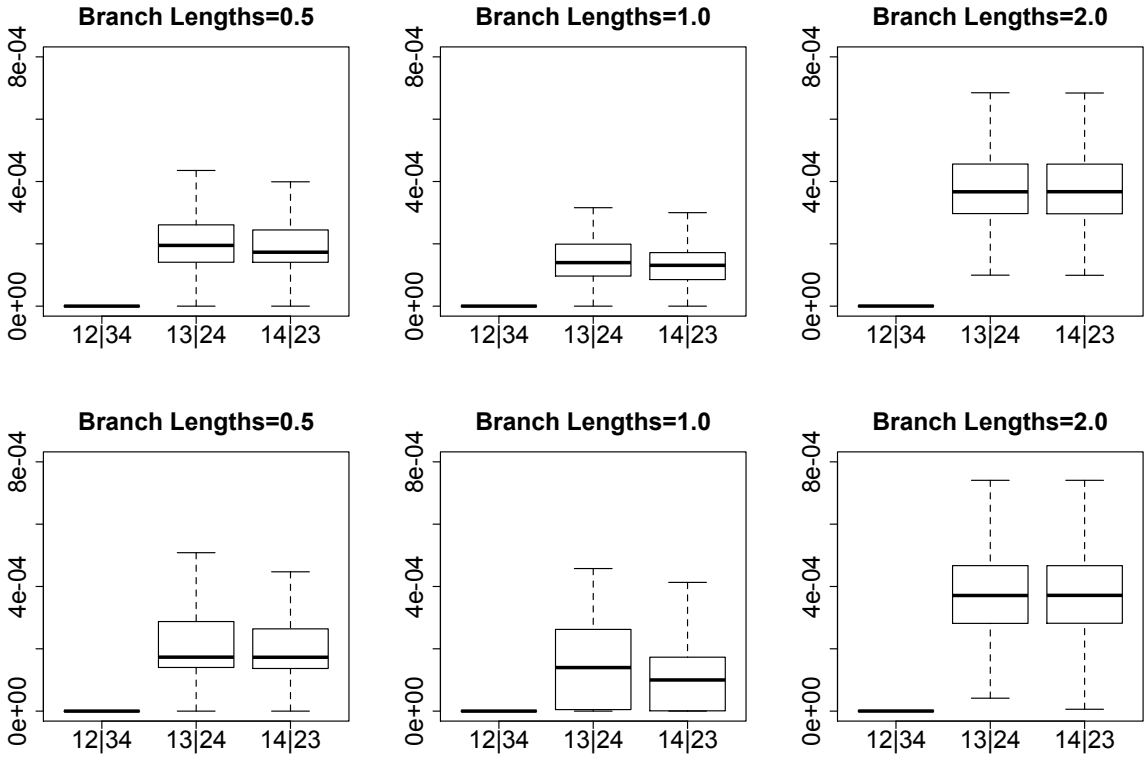
First row: 5,000 SNP sites; Second row: 10 genes of 500bp



Simulation study 1 – can we detect the correct split?

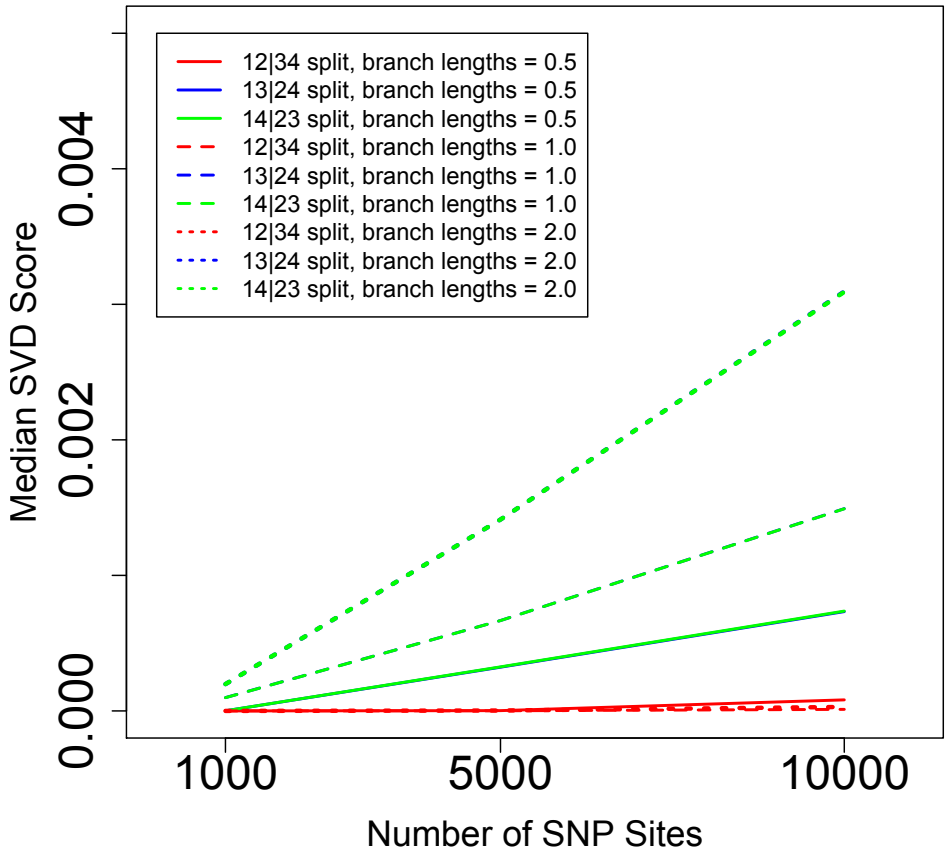
Simulate data from the GTR+I+Γ model for a 4-taxon tree and examine split scores

First row: 5,000 SNP sites; Second row: 10 genes of 500bp



Simulation study 1 – can we detect the correct split?

Change in scores as amount of data increases

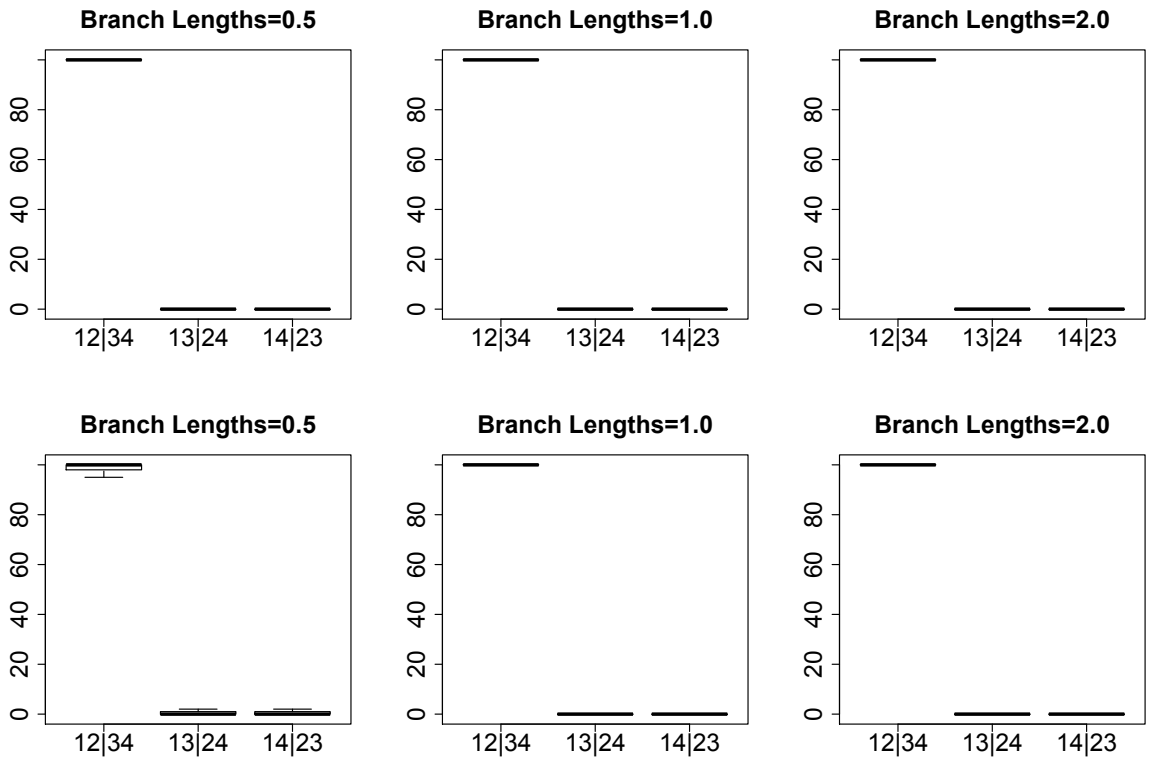


How do we assess variability?

- How can we measure confidence in the inferred split?
- Use a nonparametric bootstrap procedure
  - Generate bootstrap data sets from the original data matrix
  - Compute split scores on all three splits for each bootstrap data matrix
  - Record the number of bootstrap data sets for which each split is inferred, and use the proportion of these as a bootstrap support measure
- Evaluate performance of the bootstrap procedure using the same simulated data

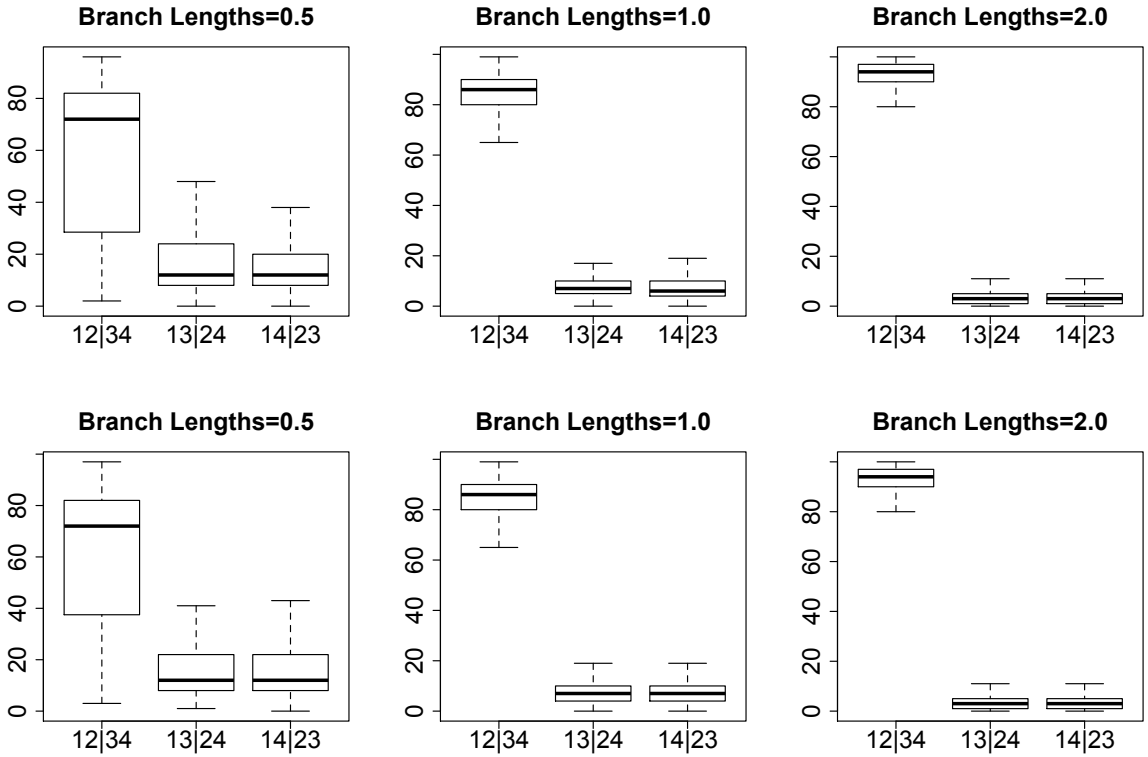
Assessing support using the bootstrap

Simulate data from the Jukes-Cantor model for a 4-taxon tree and examine bootstrap support scores



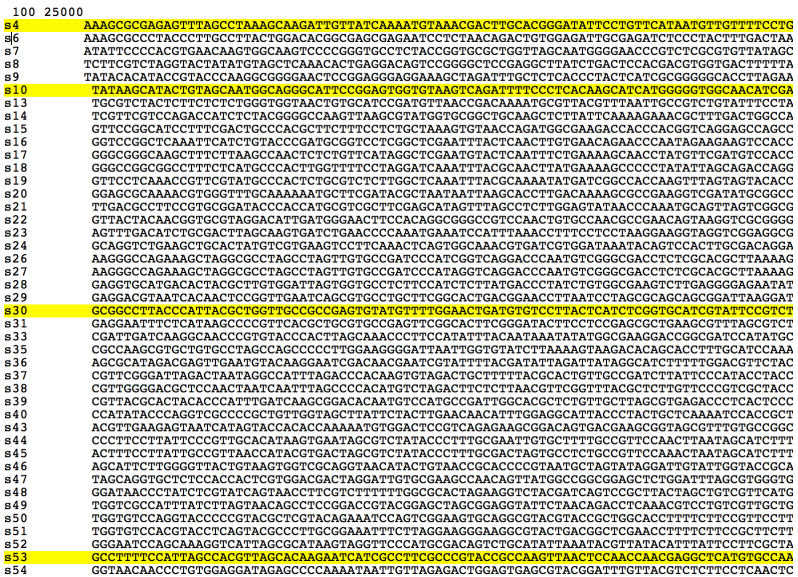
Assessing support using the bootstrap

Simulate data from the GTR+I+Γ model for a 4-taxon tree and examine bootstrap support scores



Algorithm

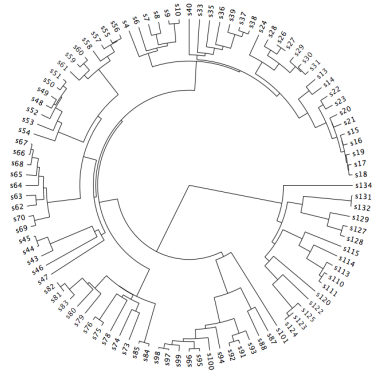
1. Generate all quartets (small problems) or sample quartets (large problems)
2. Estimate the correct quartet relationship for each sampled quartet
3. Use a quartet assembly method to build the tree - PAUP* uses the method of Reaz-Bayzid-Rahman (2014), called QFM, to build the tree.



→

1 2 | 3 4
3 5 | 2 17
19 6 | 16 1
5 22 | 3 7
. . . .

→



Extension to larger trees

- Multiple lineages are handled as follows:
  1. Sample four species
  2. Select one lineage at random from each species
  3. Estimate the quartet relationships among the four sampled lineages
  4. Restore the species labels (but lineage quartets are saved, too)
- Quantify uncertainty using the bootstrap

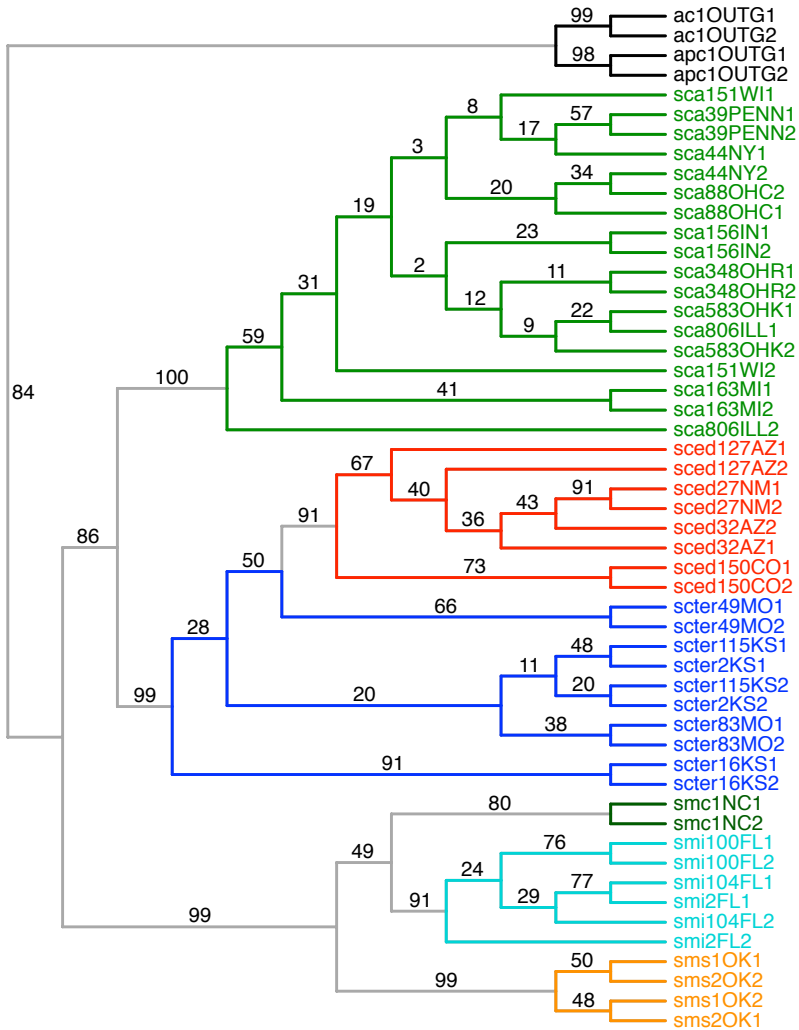
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

Species	Location	No. of individuals per gene
S. catenatus catenatus	Eastern U.S. and Canada	9
S. c. edwardsii	Western U.S.	4
S. c. tergeminus	Western and Central U.S.	5
S. miliarius miliarius	Southeastern U.S.	1
S. m. barbouri	Southeastern U.S.	3
S. m. streckerii	Southeastern U.S.	2
Agkistrodon sp. (outgroup)	U.S.	2

Empirical example: Sistrurus rattlesnakes

All quartets and 100 bootstrap replicates

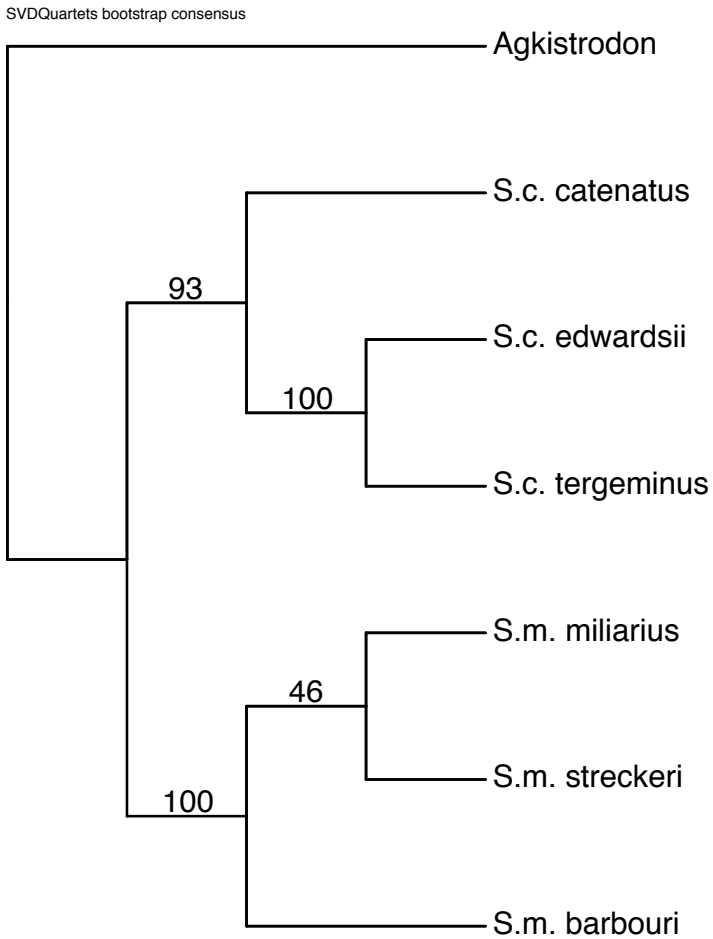
$\sim$ 11 minutes



Empirical example: Sistrurus rattlesnakes

All quartets and 100 bootstrap replicates

$\sim$ 11 minutes



- SVDQuartets
  - Will perform well when there are a lot of data (multilocus or SNP) available
  - More complex model $\Longrightarrow$ more data needed
  - Valid when the molecular clock is violated
  - Valid when there is gene flow between sister taxa
  - Computationally efficient, including bootstrapping
  - Will soon include estimates of branch lengths
- ASTRAL
  - Will perform well when gene trees can be estimated well
  - Gene flow can cause the method to fail (because then quartets can be anomalous)
  - Computationally efficient after individual gene trees have been estimated
  - Can provide estimates of branch lengths

- How do these compare to Bayesian methods, such as STARBEAST2 and BPP?
  - STARBEAST2 and BPP carry out estimation under the model, including all model components
  - Estimation of the posterior distribution provides a natural way to quantify uncertainty
  - ASTRAL and SVDQ use features of the model to assess fit of the data to the model
    - ASTRAL: gene trees
    - SVDQ: site pattern probabilities
  - Trade-offs involved in choosing among methods: computational efficiency, robustness to the model, etc.