Species Tree Inference

Laura Kubatko

Evolution, Ecology and Organismal Biology
and Statistics
The Ohio State University

kubatko.2@osu.edu

Twitter: Laura_Kubatko

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

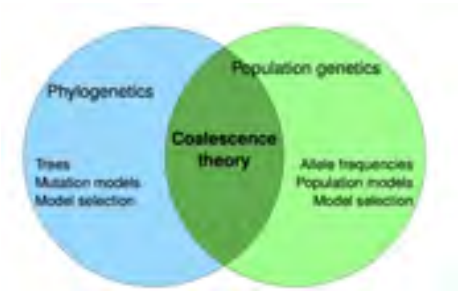
Relationship between population genetics and phylogenetics

- Need models at two levels:

1. Model what happens within each population

→ *coalescent model*

Peter's talk on Monday



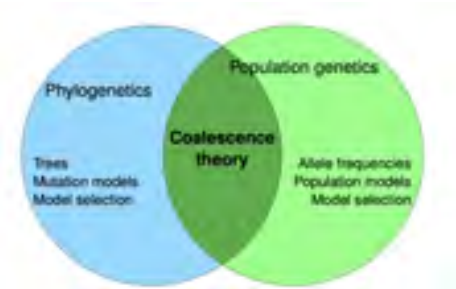
Relationship between population genetics and phylogenetics

- Need models at two levels:

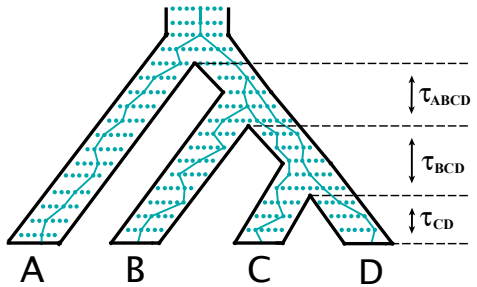
1. Model what happens within each population

→ *coalescent model*

Peter's talk on Monday



2. Link each within-population model on a phylogeny



Relationship between population genetics and phylogenetics

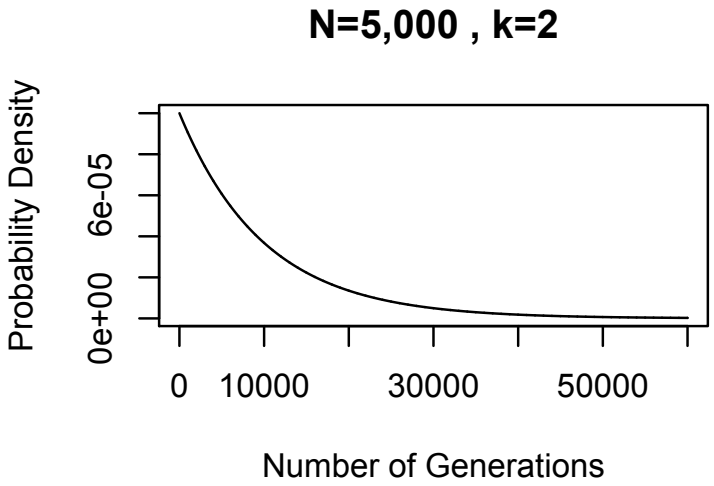
- Build up the species tree from many populations:

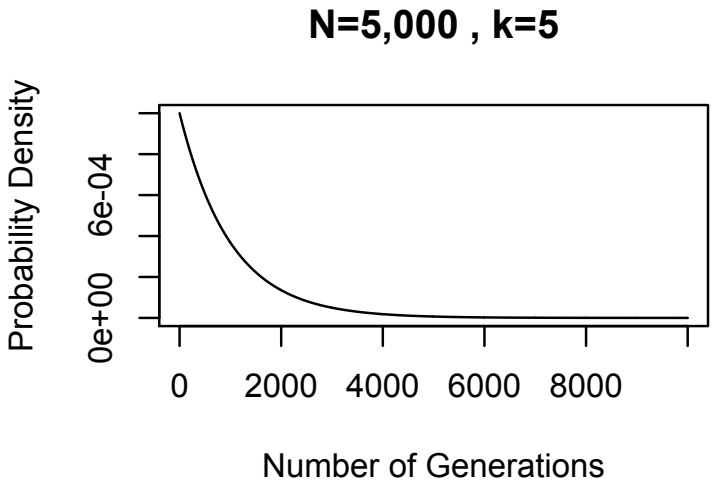
Coalescent review

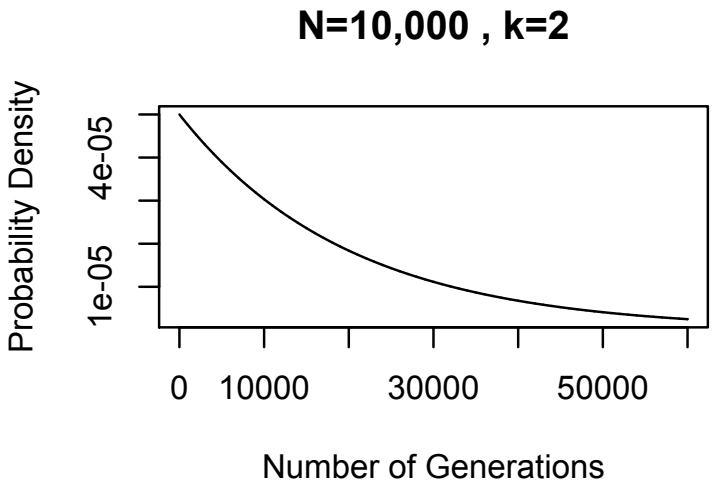
- Recall several important facts from Peter's lecture:
  - Kingman's coalescent: For a sample of k lineages, the distribution of the number of generations until two lineages coalesce is **exponential with rate** $\binom{k}{2}\frac{1}{2N}$
  - k=2: rate $= \frac{1}{2N}$ and mean time to coalescence is $2N$
  - k=5: rate $= \frac{10}{2N}$ and mean time to coalescence is $\frac{2N}{10}$
  - Larger N means that:
  - Larger k means that:
  - 

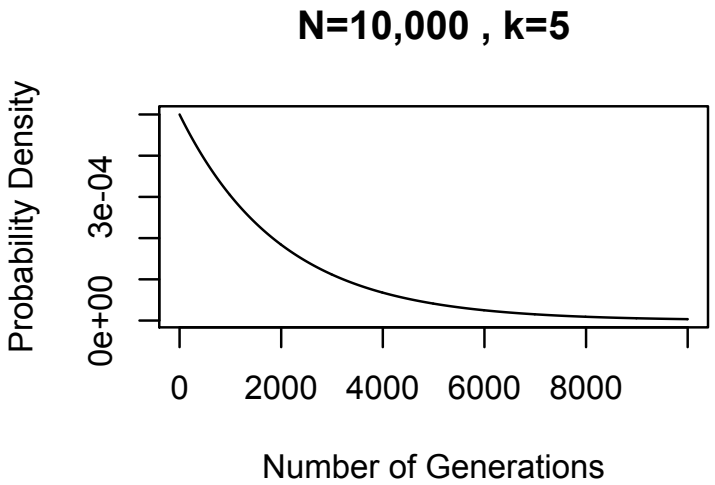
Coalescent review

- What does the exponential distribution look like?







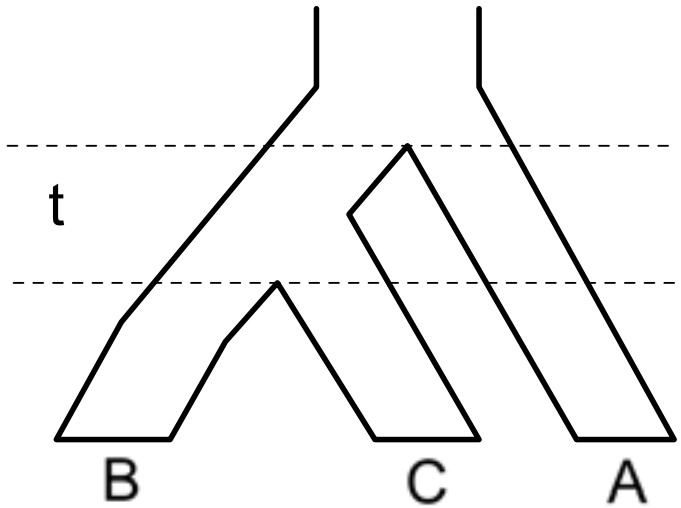


Coalescent units

- 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!

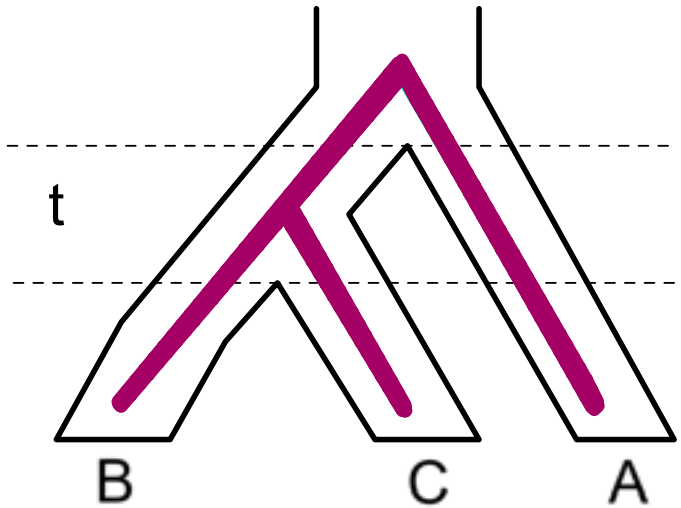
Phylogenetic coalescent model

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



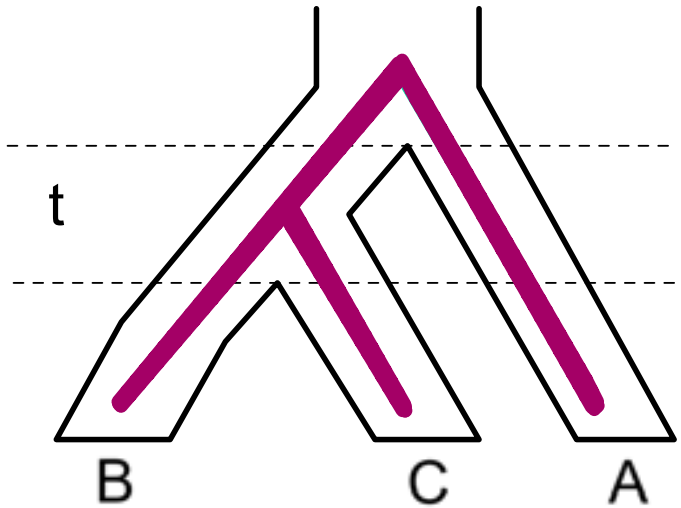
Phylogenetic coalescent model

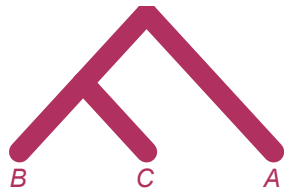
- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



Phylogenetic coalescent model

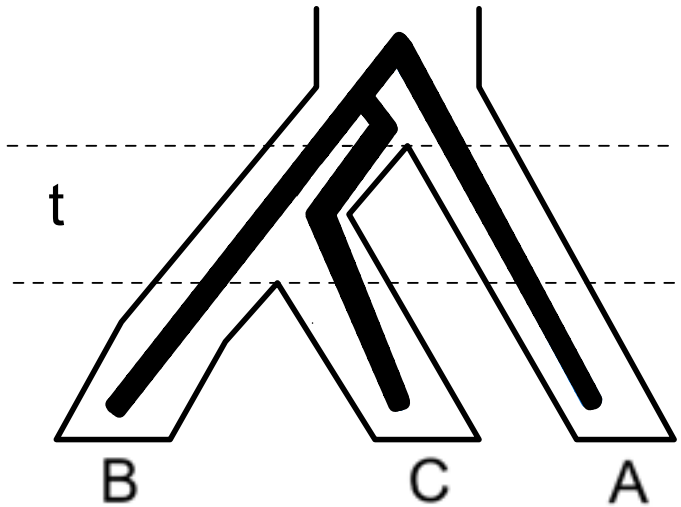
- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process

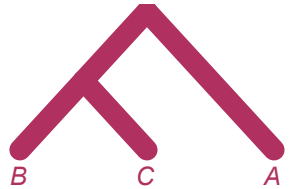


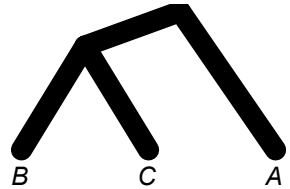


Phylogenetic coalescent model

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

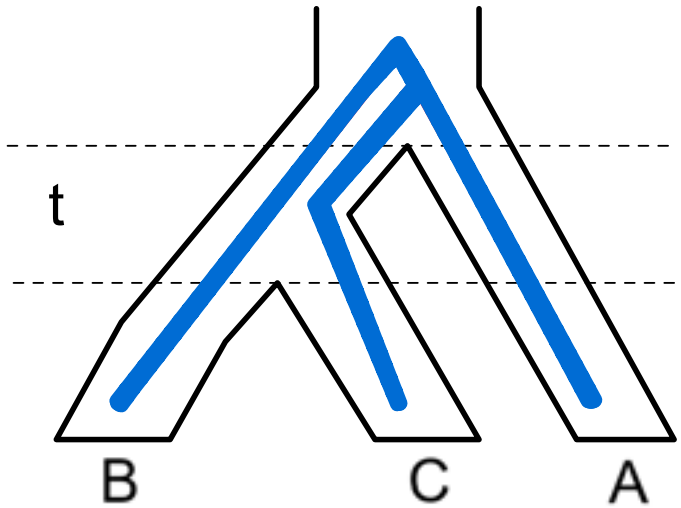


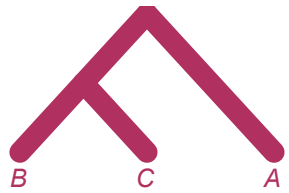


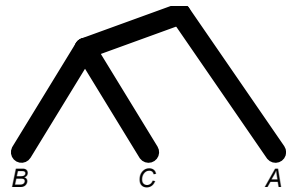


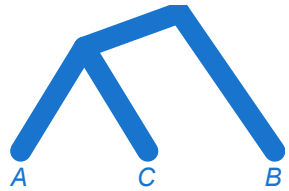
Phylogenetic coalescent model

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



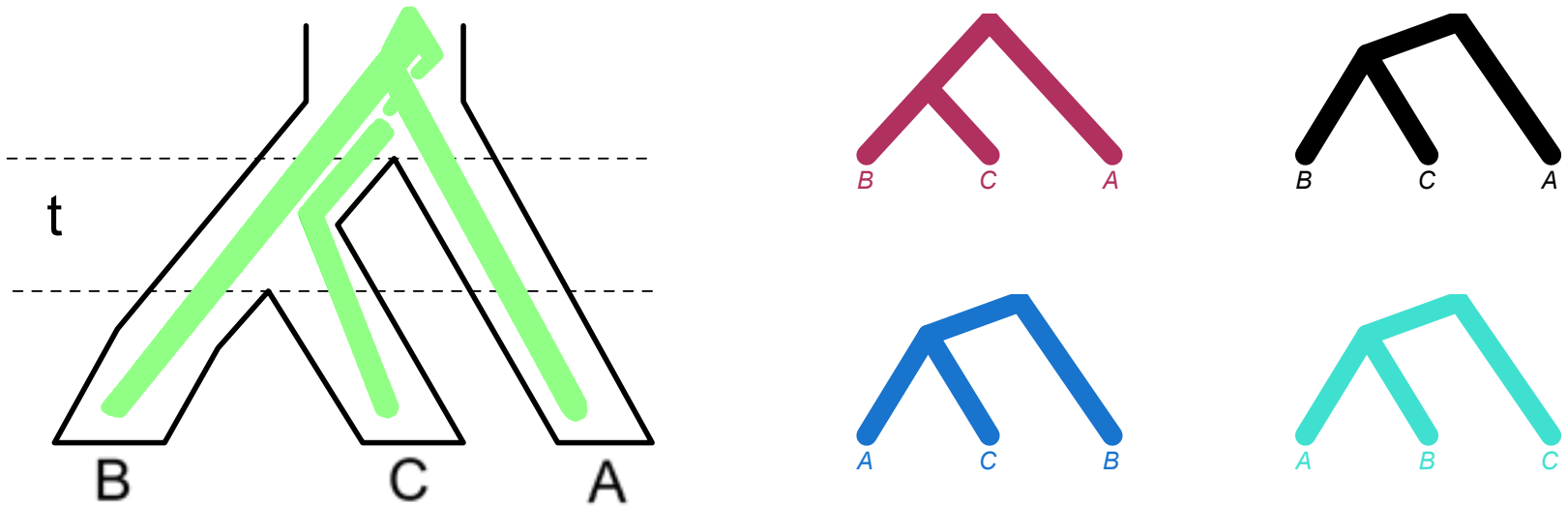






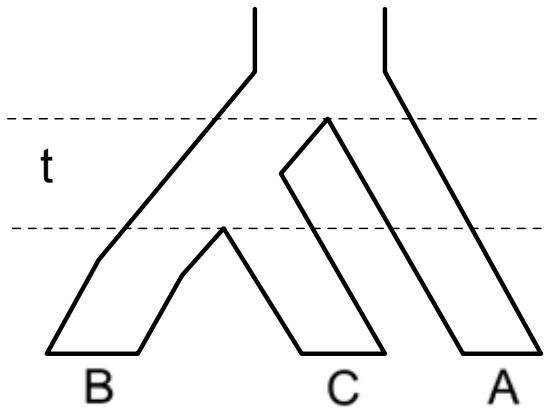
Phylogenetic coalescent model

- 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

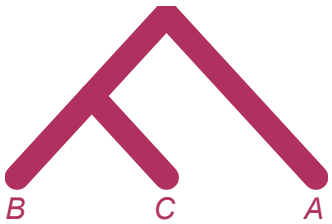
- t = length of interval between speciation events in **coalescent units**
  = number of $2N$ generations



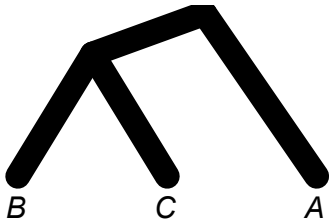
- **Example:** 1.2 coalescent units for an organism with population size $N = 10,000$ and a generation time of 3 years $= 1.2 \times 20,000 \times 3 = 72,000$ years

Phylogenetic coalescent model

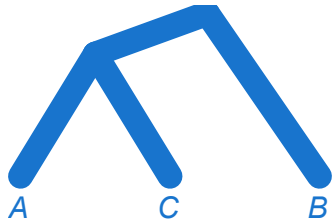
Probabilities of each gene tree history are shown below them
t = length of interval between speciation events



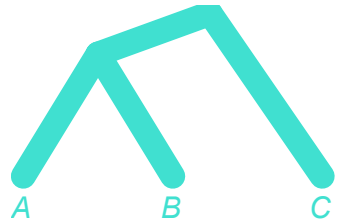
$1 - e^{-t}$



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

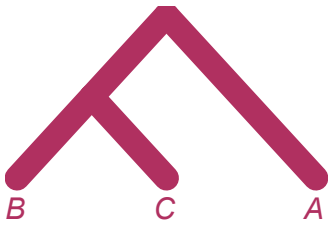


$\frac{1}{3}e^{-t}$



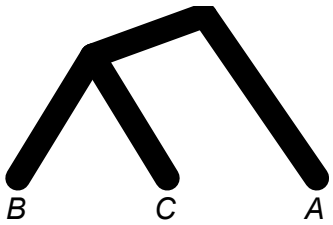
$\frac{1}{3}e^{-t}$

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



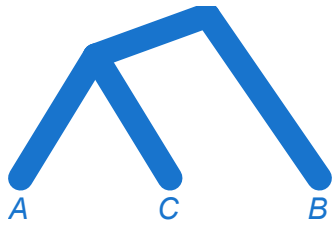
$1 - e^{-t}$

0.63



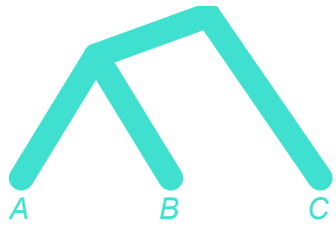
$\frac{1}{3}e^{-t}$

0.12



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

0.12

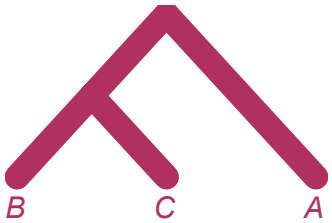


$\frac{1}{3}e^{-t}$

0.12

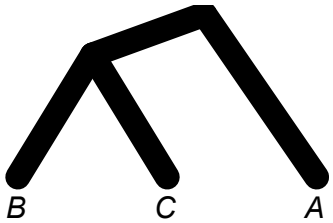
Phylogenetic coalescent model

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



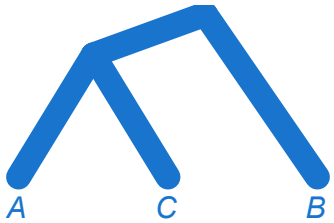
$1 - e^{-t}$

0.63
0.40



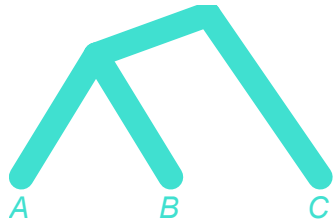
$\frac{1}{3}e^{-t}$

0.12
0.20



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

0.12
0.20

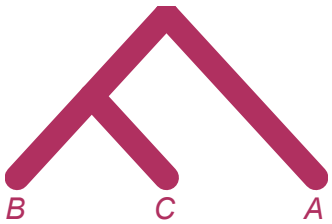


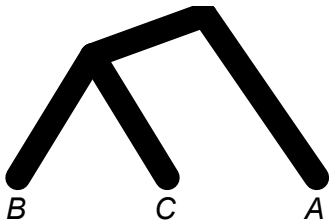
$\frac{1}{3}e^{-t}$

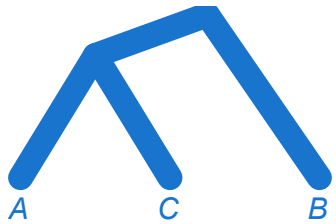
0.12
0.20

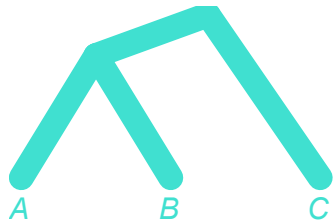
Phylogenetic coalescent model

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





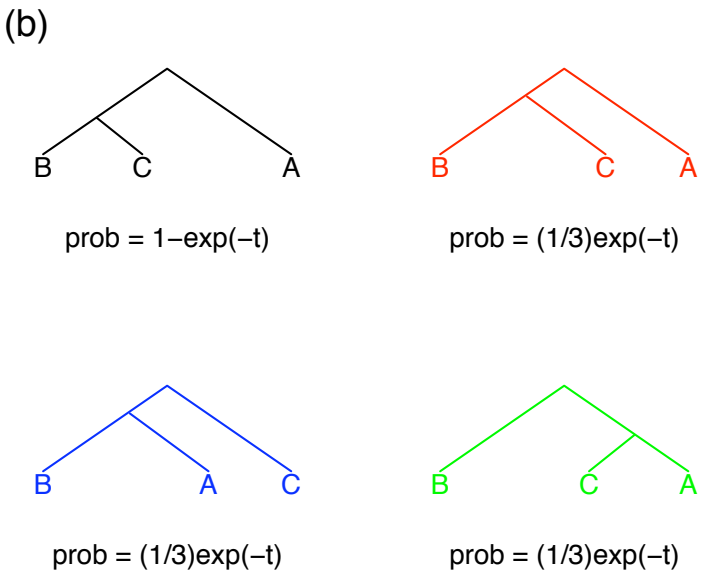


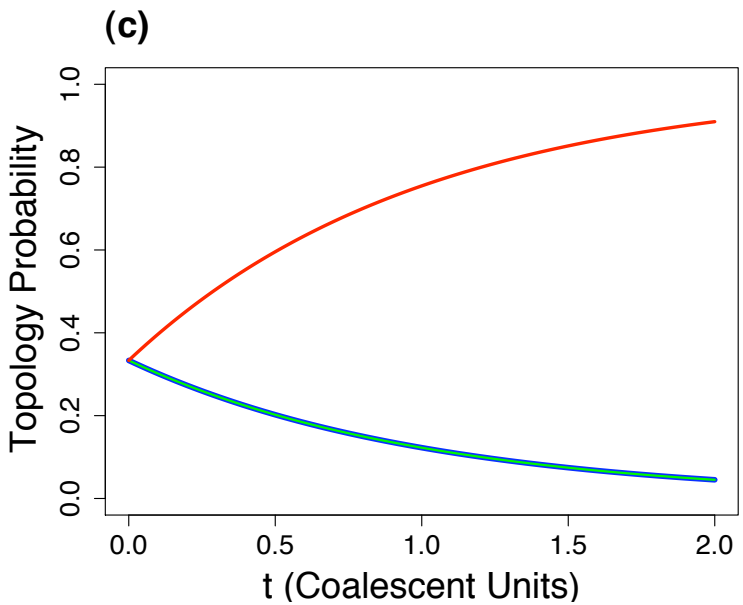


$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20
0.85	0.05	0.05	0.05

Effect of speciation time

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





- 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.

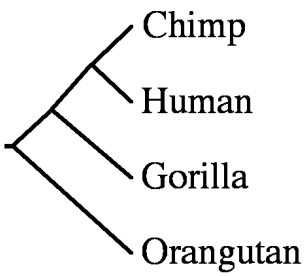
Summary of the three-taxon case

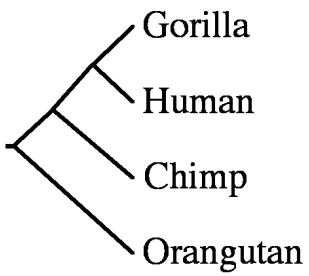
- What have we learned from considering 3 taxa?
  - Gene tree with topology that matches the species tree occurs with probability at least as large as the other two trees
  - The other two trees are expected to occur in equal frequency
  - Shorter intervals between speciation events lead to more disagreement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

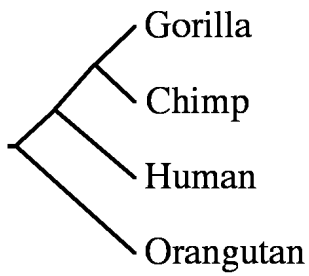
Application 1: Goodness of fit to empirical data

- 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.

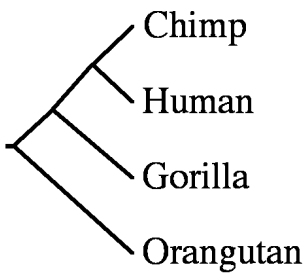
Application 1: Goodness of fit to empirical data



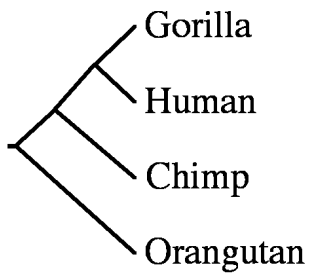




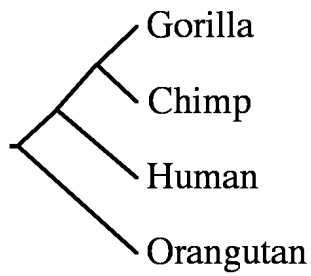
Application 1: Goodness of fit to empirical data



76.6%



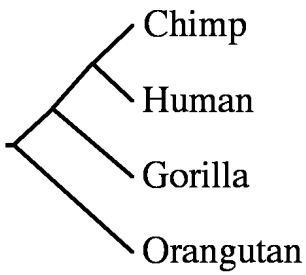
11.4%

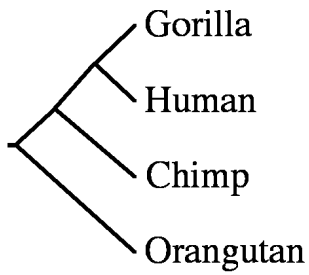


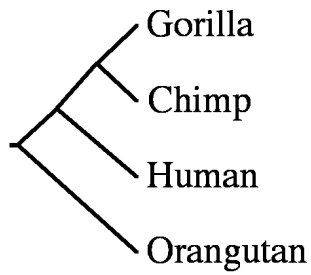
11.5%

Observed proportions of each gene tree among ML phylogenies

Application 1: Goodness of fit to empirical data







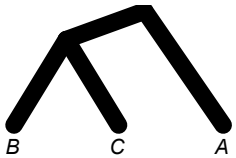
76.6%	11.4%	11.5%
79.1%	9.9%	9.9%

Observed proportions of each gene tree among ML phylogenies

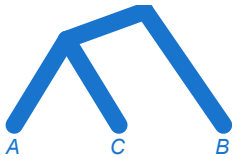
Predicted proportions using parameters from Rannala & Yang, 2003.

Application 2: Branch length estimation

- Suppose we are given a sample of gene trees, i.e.,



70 genes



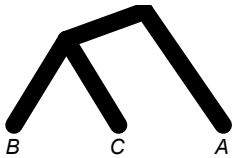
15 genes

15 genes

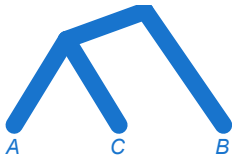
- What do the gene trees tell us?

Application 2: Branch length estimation

- Suppose we are given a sample of gene trees, i.e.,



70 genes

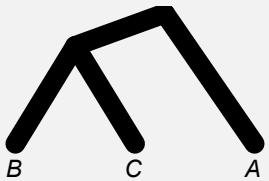


15 genes

15 genes

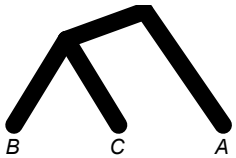
- What do the gene trees tell us?

The species tree

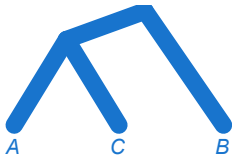


Application 2: Branch length estimation

- Suppose we are given a sample of gene trees, i.e.,



70 genes

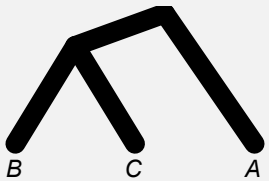


15 genes

15 genes

- What do the gene trees tell us?

The species tree



The branch length t:

Set $0.7 = 1 - \frac{2}{3}e^{-t}$
and solve for t

$$t = 0.7985$$

How general is this result?



J. Math. Biol. (2011) 62:833–862
DOI 10.1007/s00285-010-0355-7

Mathematical Biology

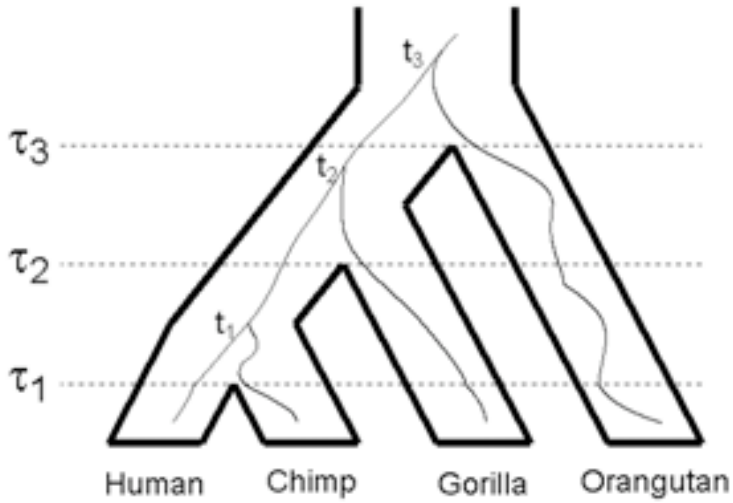
Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent

Elizabeth S. Allman · James H. Degnan · John A. Rhodes

- Four taxa: the distribution of unrooted gene trees determines the unrooted species tree and branch lengths
- Five or more taxa: the distribution of unrooted gene trees determines the rooted species tree and branch lengths.

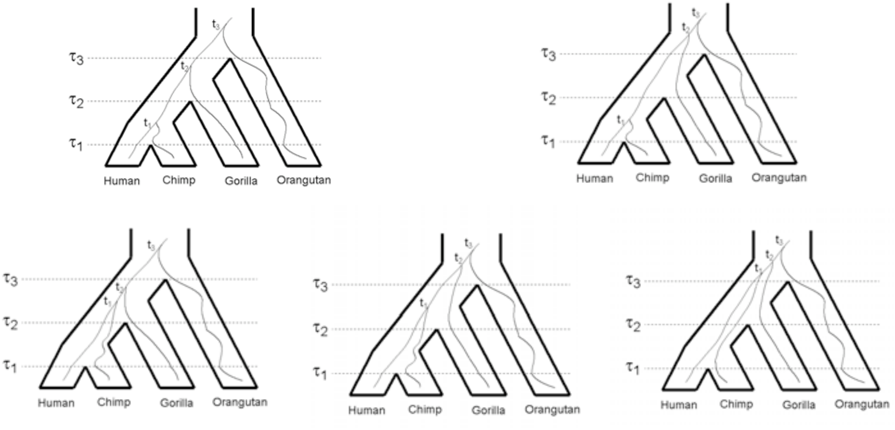
A slightly larger case

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



Coalescent histories for the 4-taxon example

- There are 5 possible histories for this example:



Enumerating 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}

Degnan and Salter, *Evolution*, 2005

- In the general case, we have the following:

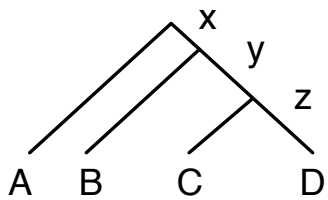
  The probability of a gene tree g given the species tree $\mathcal{S}$ is given by

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

- 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

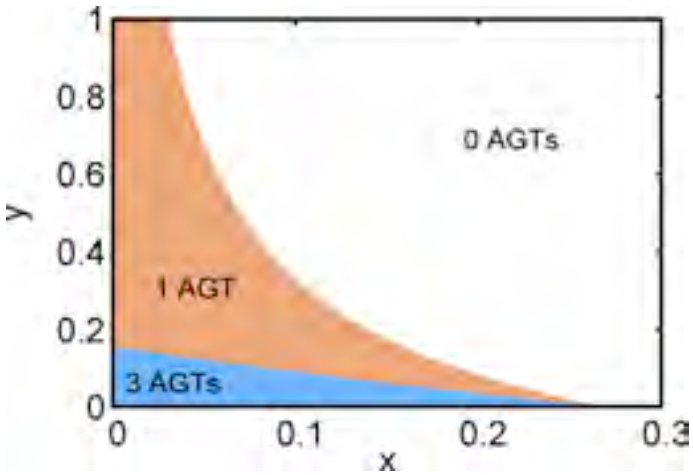
- In the three-taxon case, the gene tree with the highest probability has the same topology as the species tree
- Question: Must the distribution always look this way?
- Examine the entire distribution for four taxa – only 15 gene trees are possible
- For the species tree:



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

- https://lkubatko.shinyapps.io/GeneTreeProbs/

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

Can we use gene trees to estimate the species trees?

- Two problems with using gene trees directly for inference:
- We don't observe gene trees directly

  Rather, we observe sequence data for each gene and need to estimate the gene trees

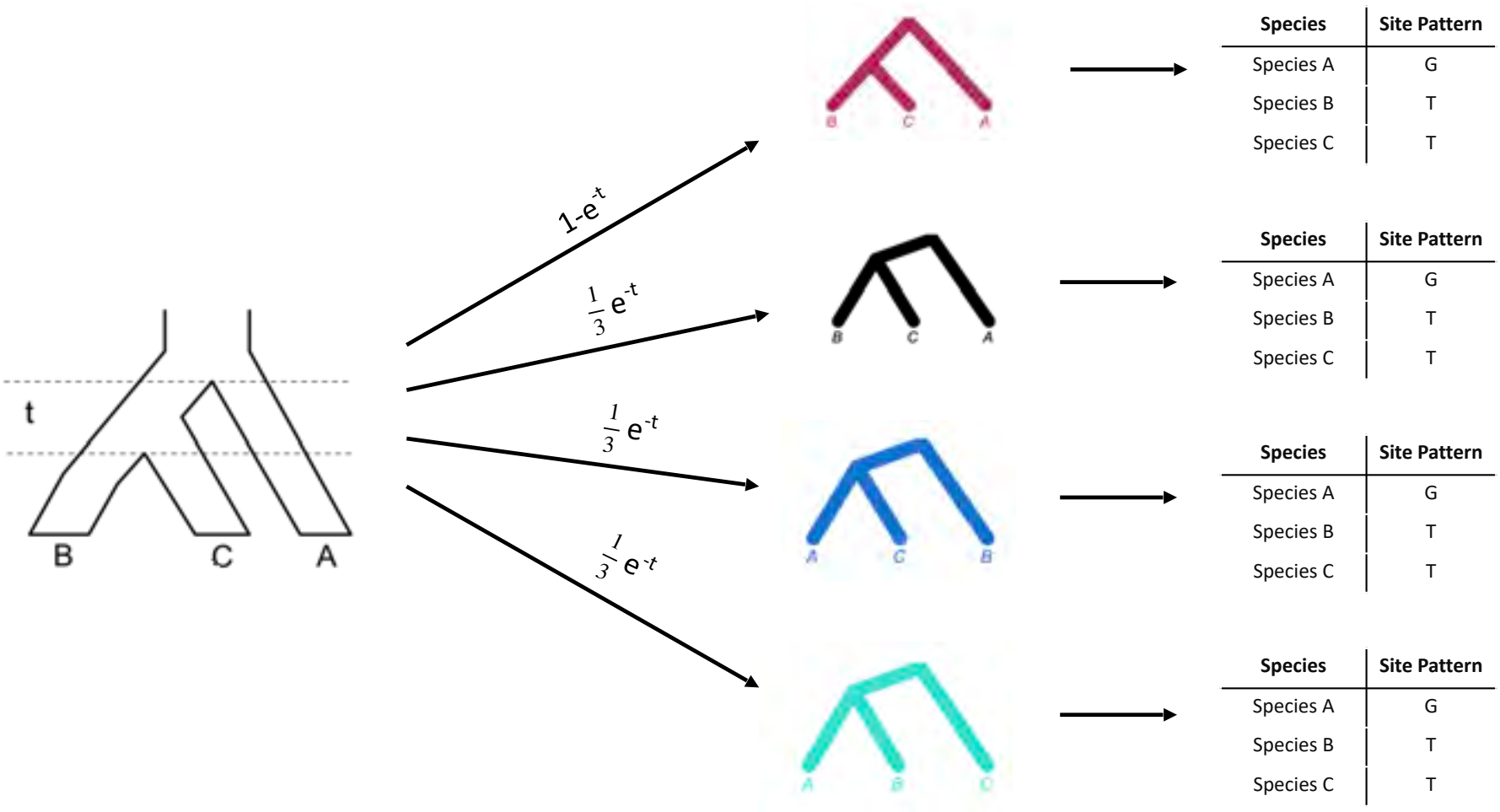
- Sampling error in the gene tree proportions would complicate inference

  For example, if the branch length t is long enough, we would only observe gene trees that matched the species tree ... and then how would we estimate t?

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

Sequence data



Species	Site Pattern
Species A	G
Species B	T
Species C	T

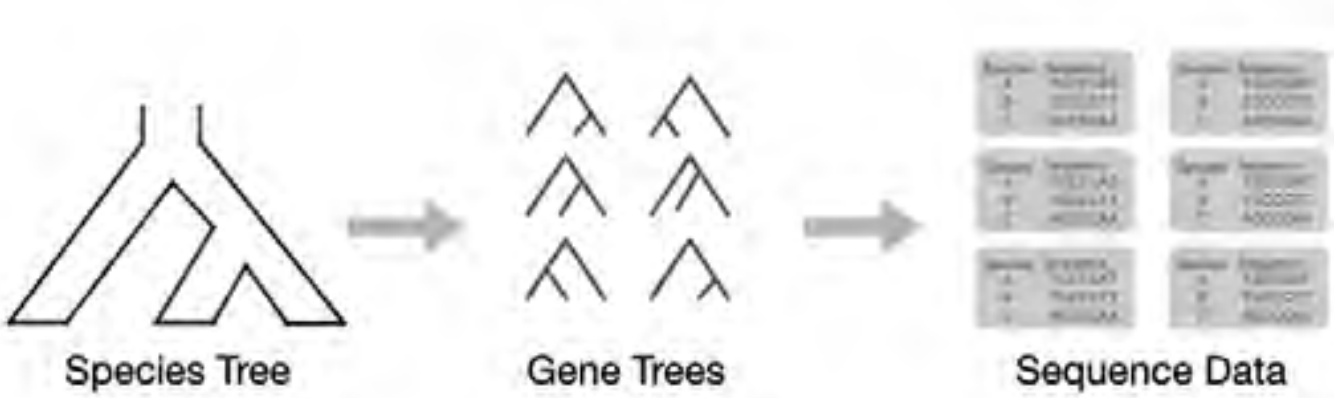
Species	Site Pattern
Species A	G
Species B	T
Species C	T

Species	Site Pattern
Species A	G
Species B	T
Species C	T

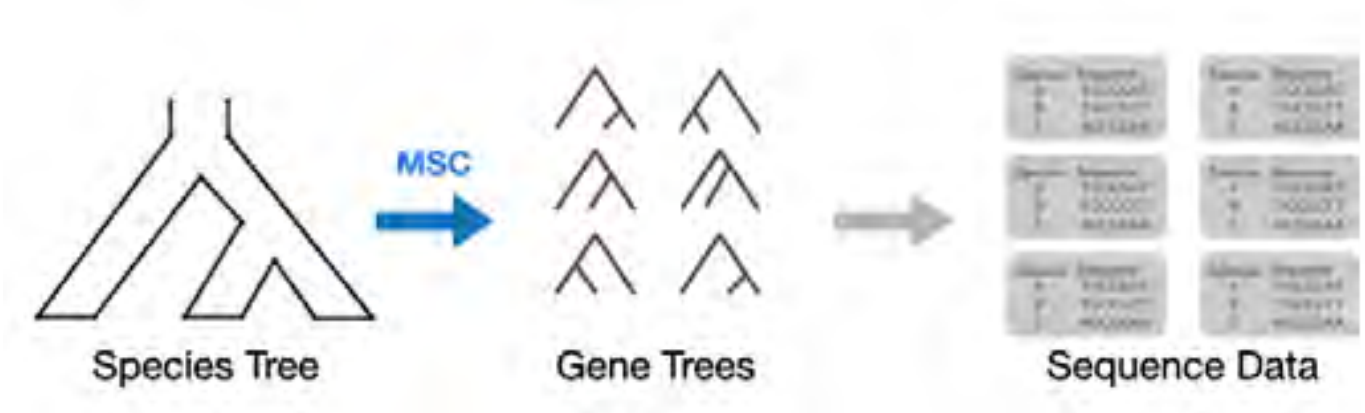
Species	Site Pattern
Species A	G
Species B	T
Species C	T

https://lkubatko.shinyapps.io/SitePatternsProbs/

Species Tree Inference under the Multispecies Coalescent



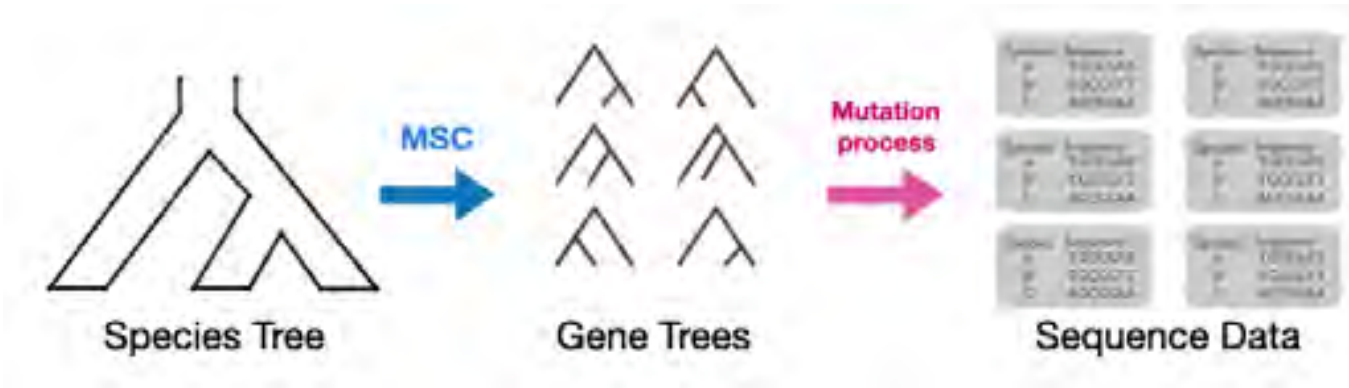
Species Tree Inference under the Multispecies Coalescent



- species tree $\rightarrow$ gene trees : : : multispecies coalescent model

  Times to coalescent events are exponentially distributed, with rate that varies with the number of potential lineages

Species Tree Inference under the Multispecies Coalescent



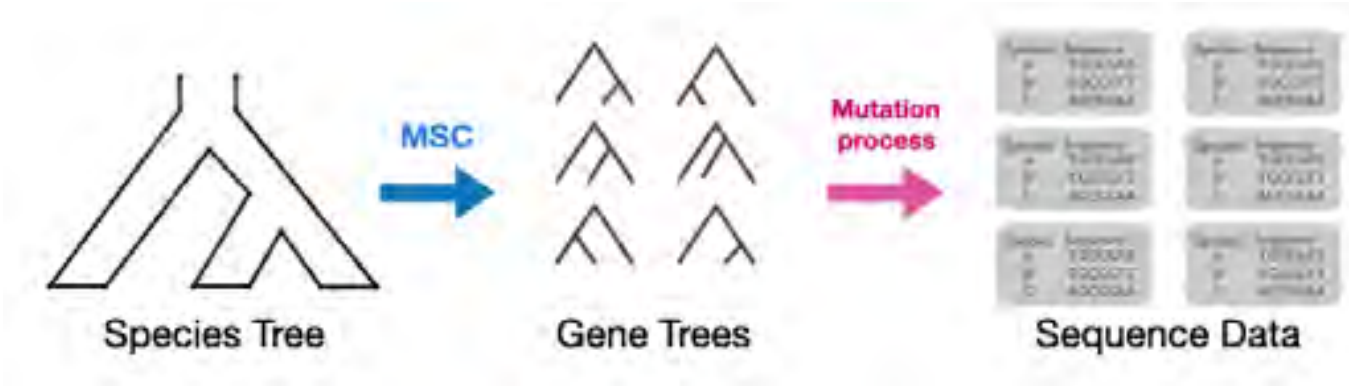
- species tree $\rightarrow$ gene trees : : : multispecies coalescent model

  Times to coalescent events are exponentially distributed, with rate that varies with the number of potential lineages

- gene trees $\rightarrow$ DNA sequences : : : standard nucleotide substitution models

  Continuous time Markov processes with states consisting of the four possible nucleotides (A, C, G, T) operate independently along each branch

Species Tree Inference under the Multispecies Coalescent



The likelihood of the species tree $(\mathcal{S}, \boldsymbol{\tau})$ for sequence data $\mathbf{D}$ is

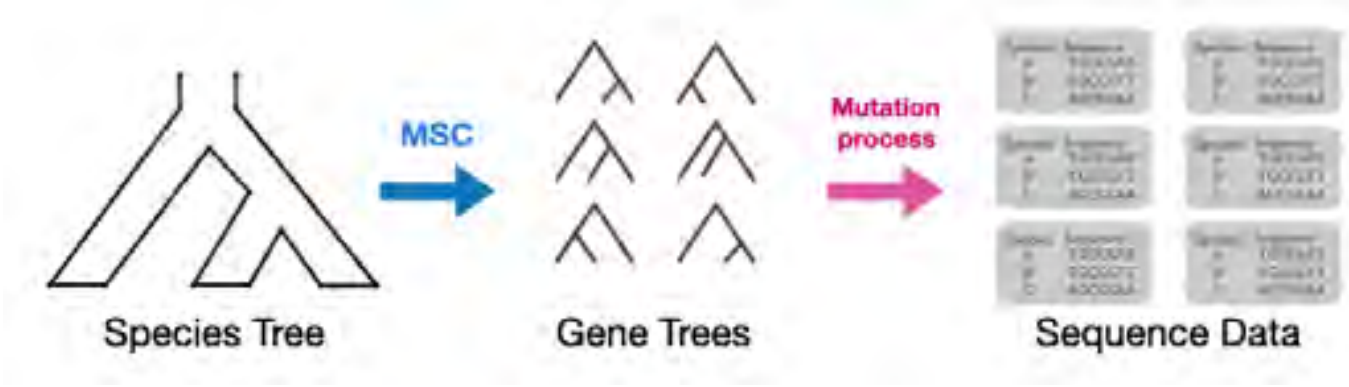
$$P(\mathbf{D}|(\mathcal{S},\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

$\mathcal{H}$ = set of all gene tree histories

$h \in \mathcal{H}$ = a gene tree history with branch lengths $\mathbf{t}_h$

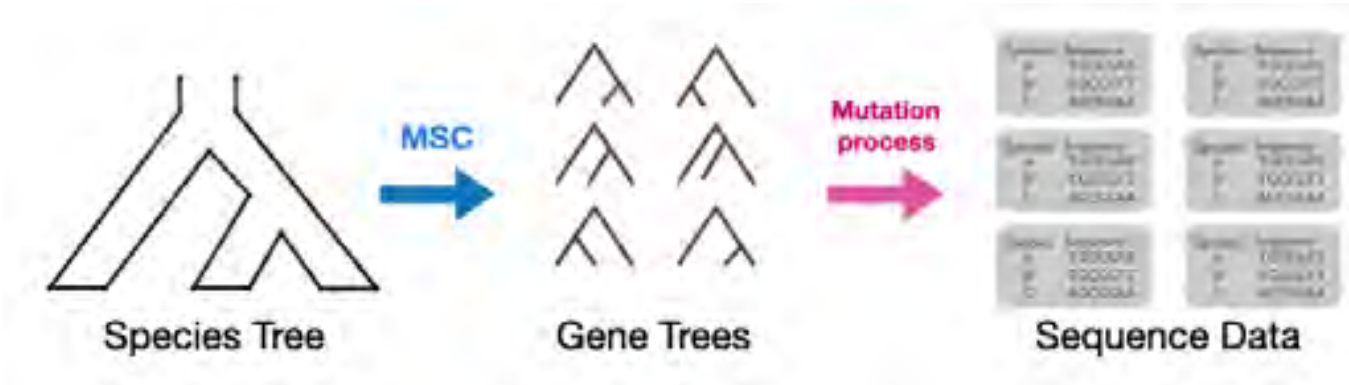
$(\mathcal{S}, \tau)$ = species tree with topology $\mathcal{S}$ and speciation times τ

Species Tree Inference under the Multispecies Coalescent



$$P(\mathbf{D}|(S,\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

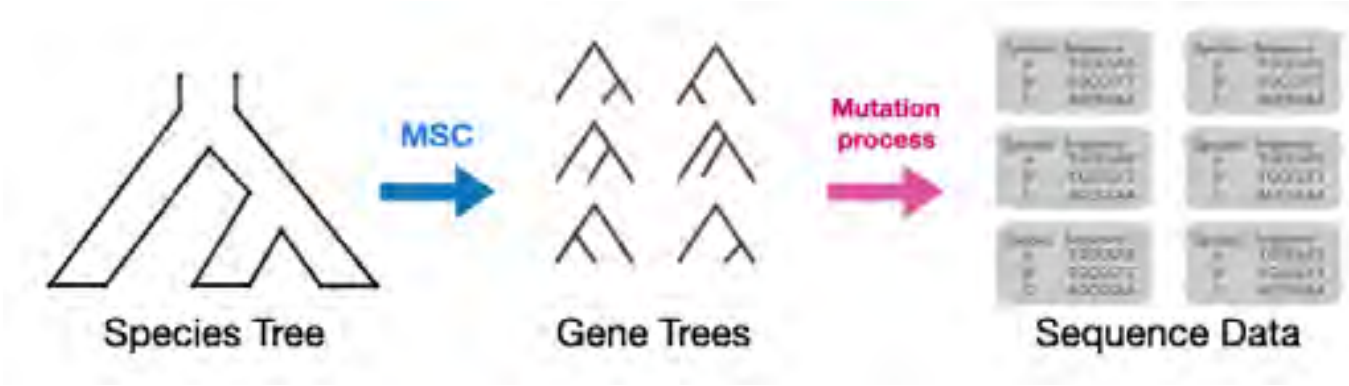
Species Tree Inference under the Multispecies Coalescent



$$P(\mathbf{D}|(S,\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

$|\mathcal{H}|$ is greater than the number of trees for a fixed number of species, n

Species Tree Inference under the Multispecies Coalescent

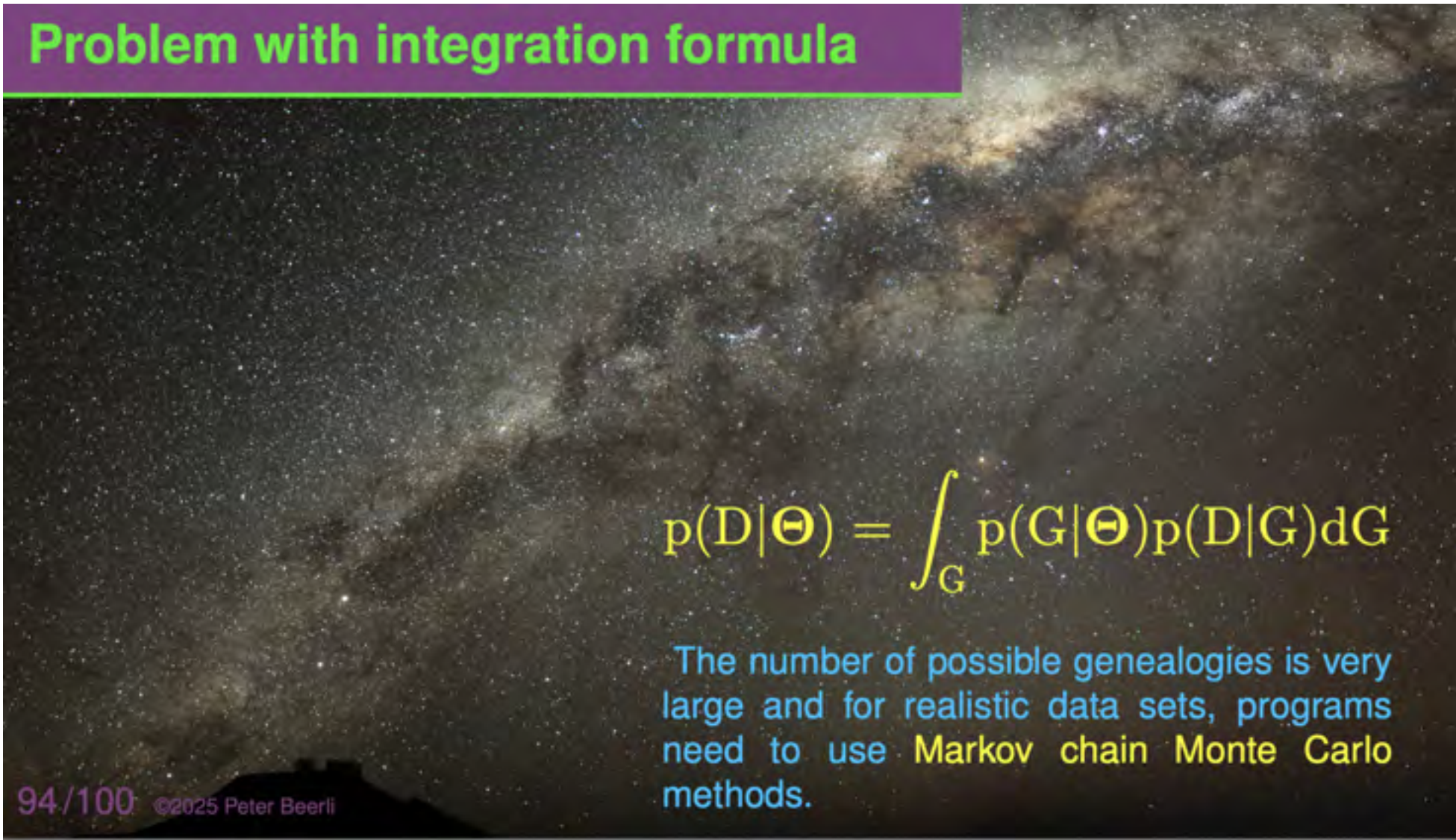


$$P(\mathbf{D}|(S,\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

The dimension of $\mathcal{H}$ is greater than the number of trees for a fixed number of species, n

For each $h \in \mathcal{H}$, we need to compute an $(n-1)$-dimensional integral

Problem with integration formula



Inference of population size

Why do species tree inference?

- This seems really hard – do I **really** need to do species tree inference???? Why can't I just use a gene tree method on the concatenated data????

Why do species tree inference?

- This seems really hard – do I **really** need to do species tree inference???? Why can't I just use a gene tree method on the concatenated data????
- **Three reasons** to use a method designed for species tree inference:

Why do species tree inference?

- This seems really hard – do I **really** need to do species tree inference???? Why can't I just use a gene tree method on the concatenated data????
- **Three reasons** to use a method designed for species tree inference:

1. Concatenation can be statistically inconsistent
   [Roch and Steel, 2015; Kubatko and Degnan, 2007]

Why do species tree inference?

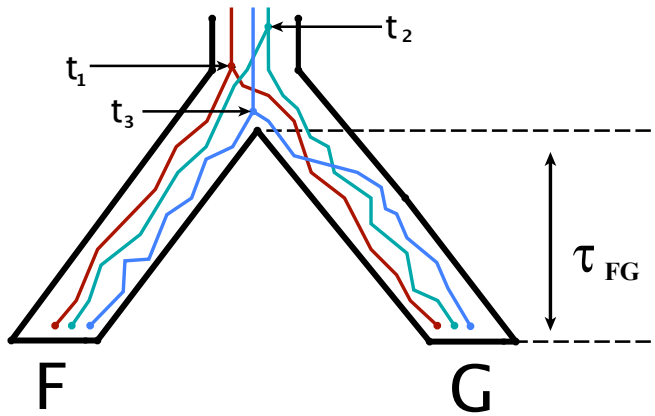
- This seems really hard – do I **really** need to do species tree inference????
  Why can't I just use a gene tree method on the concatenated data????
- **Three reasons** to use a method designed for species tree inference:

1. Concatenation can be statistically inconsistent
   [Roch and Steel, 2015; Kubatko and Degnan, 2007]
2. Bootstrap values / posterior probabilities will be too large
   Consider data from two gene trees: 510,000bp and 490,000bp
   Probability of a bootstrap sample with more sites from tree 2 ≈ 0

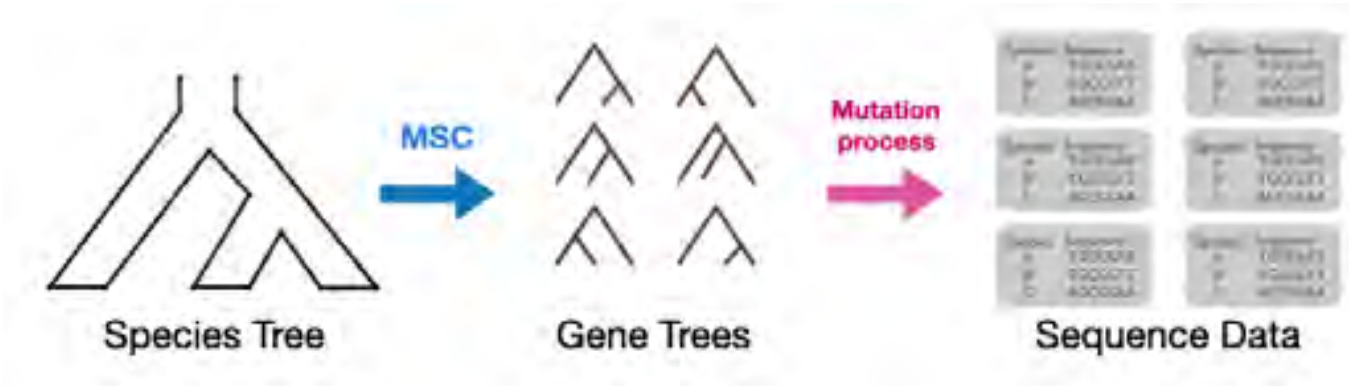
Why do species tree inference?

- This seems really hard – do I **really** need to do species tree inference???? Why can't I just use a gene tree method on the concatenated data????
- **Three reasons** to use a method designed for species tree inference:

1. Concatenation can be statistically inconsistent
   [Roch and Steel, 2015; Kubatko and Degnan, 2007]
2. Bootstrap values / posterior probabilities will be too large
   Consider data from two gene trees: 510,000bp and 490,000bp
   Probability of a bootstrap sample with more sites from tree 2 ≈ 0
3. Speciation times are overestimated
   (often significantly)



Species tree inference



- Outline for the rest of the talk:
  - How can we estimate a species tree under the MSC?
  - Empirical examples

How do we estimate a species tree?

- Classes of methods for species tree estimation:
  - Summary methods / two-step methods
    Estimate gene trees from sequences, estimate the species tree from the gene trees

How do we estimate a species tree?

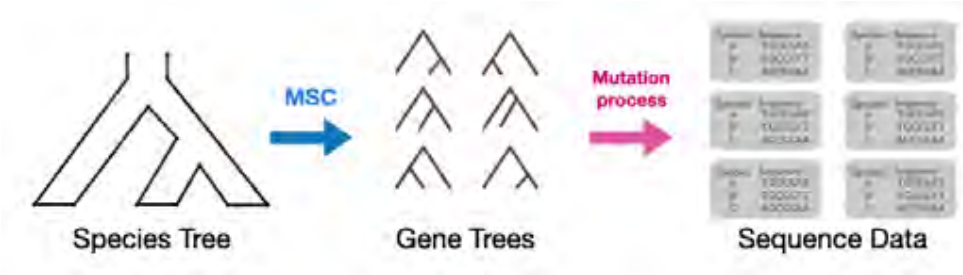
- Classes of methods for species tree estimation:
  - Summary methods / two-step methods
    Estimate gene trees from sequences, estimate the species tree from the gene trees
  - Bayesian co-estimation of gene trees and species trees
    Use MCMC to explore the joint space of gene trees and the species tree

How do we estimate a species tree?

- Classes of methods for species tree estimation:
  - Summary methods / two-step methods
    Estimate gene trees from sequences, estimate the species tree from the gene trees
  - Bayesian co-estimation of gene trees and species trees
    Use MCMC to explore the joint space of gene trees and the species tree
  - Site-based methods
    Ignore grouping of sites into loci and treat sites as independent observations from the MSC

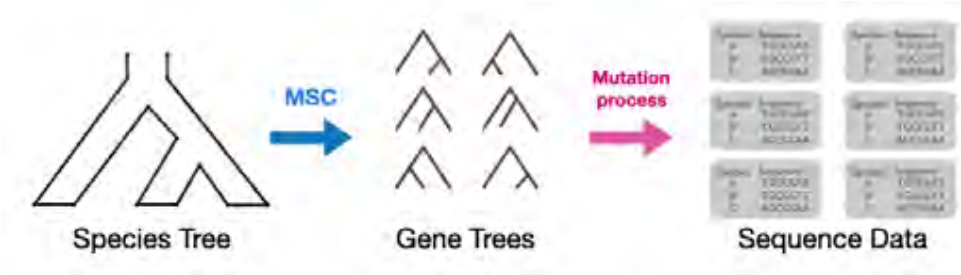
- Start with estimated gene trees
  - Using estimated branch lengths:
    - ⋆ STEM (Kubatko et al. 2009)
    - ⋆ STEAC (Liu et al. 2009)
  - Using topology information only:
    - ⋆ STAR (Liu et al. 2009)
    - ⋆ Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 2009)
    - ⋆ MP-EST (Liu et al. 2010)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

- Recall our ideas about inference under the phylogenetic coalescent model



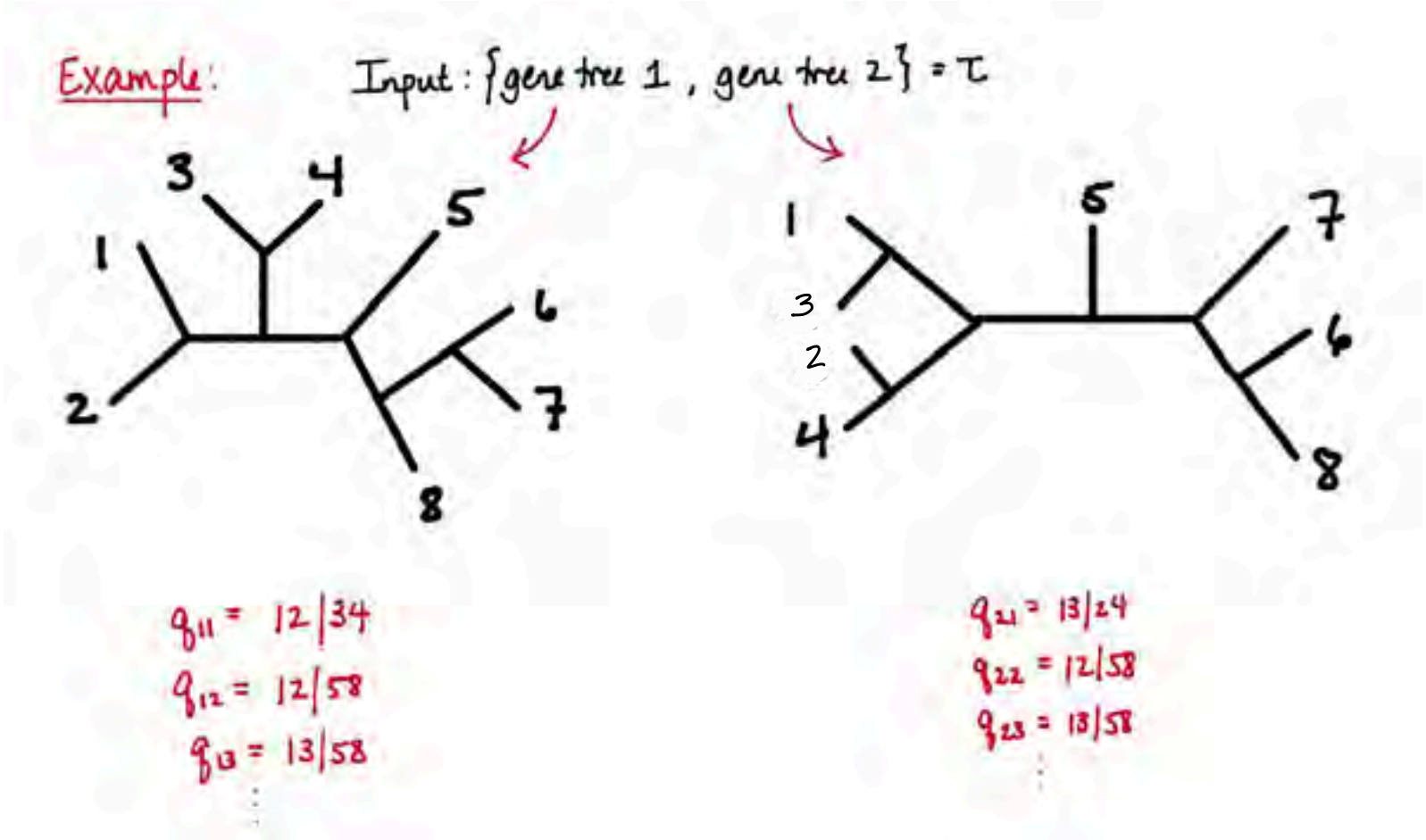
- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus
  - **Step 2.** Extract all quartet relationships from the estimated gene trees
  - **Step 3.** Find the species tree that "agrees" with as many quartets as possible

- Recall our ideas about inference under the phylogenetic coalescent model

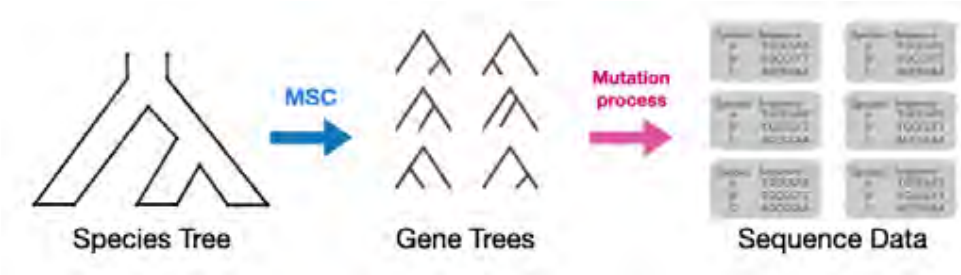


- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus ✓
  - **Step 2.** Extract all quartet relationships from the estimated gene trees
  - **Step 3.** Find the species tree that "agrees" with as many quartets as possible

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



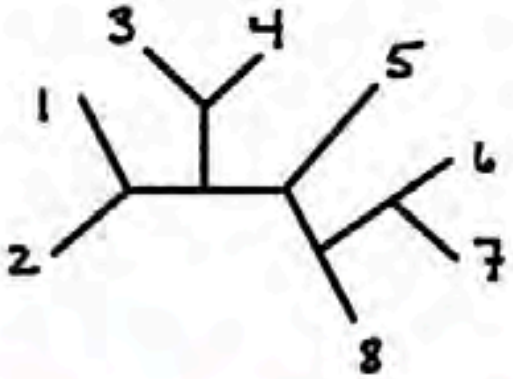
- Recall our ideas about inference under the phylogenetic coalescent model

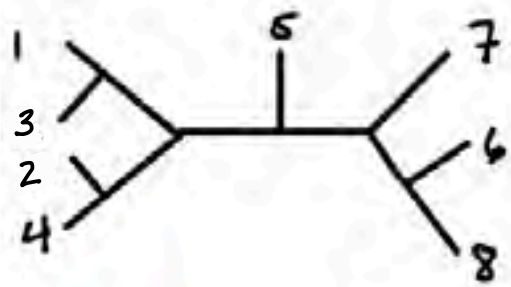


- **ASTRAL** is a summary statistic method for species tree estimation:
  - **Step 1.** Estimate gene trees for each locus ✓
  - **Step 2.** Extract all quartet relationships from the estimated gene trees ✓
  - **Step 3.** Find the species tree that "agrees" with as many quartets as possible

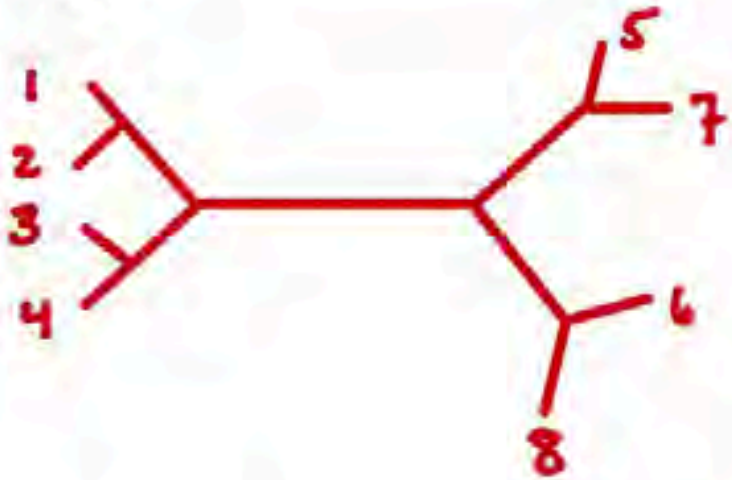
- **Step 3.** Find the species tree that "agrees" with as many quartets as possible
  - This is a non-trivial problem .... recall that we expect substantial incongruence among trees
  - However, *unrooted* gene trees cannot be anomalous for four taxa in the absence of gene flow, so *if the gene trees 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

- Example:





Consider the species tree T_1 =



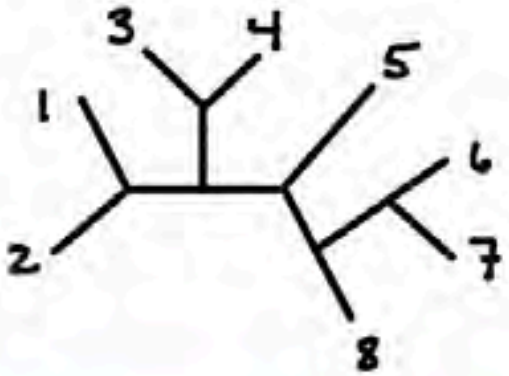
$$\text{Score}(T_1) = \sum_{\substack{\text{quartets in} \\ \text{tree } T_1, q}} w(q, T)$$

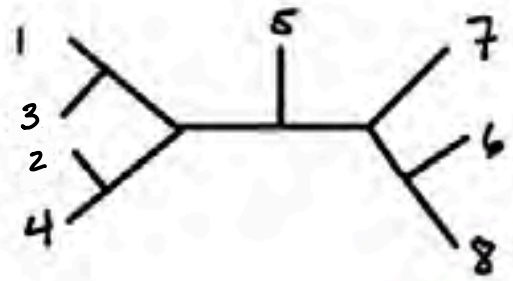
$q_1 = 12|34 \longrightarrow w(q_1, T) = 1$ (appears in gene tree 1)

$q_2 = 12|58 \longrightarrow w(q_2, T) = 2$ (appears in both input gene trees)

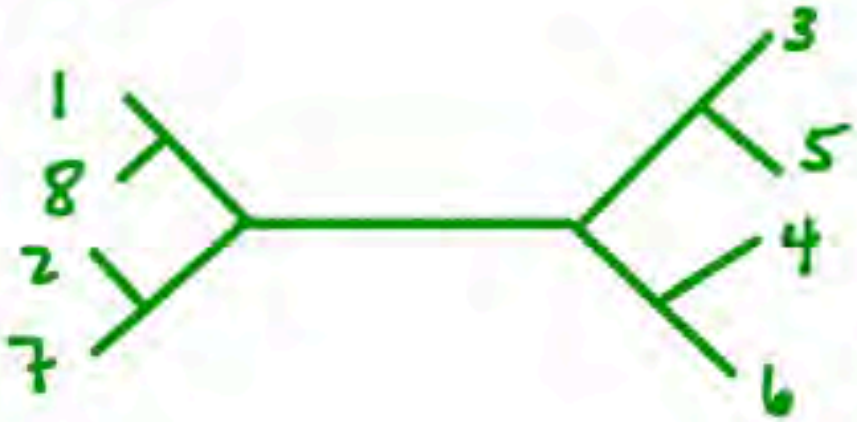
+ all quartets in T_1

- Example:





Consider the species tree T_2 =



$$Score(T_2) = \sum_{\substack{\text{quartets in} \\ \text{tree } T_2, q}} w(q, T)$$

$q_1 = 18|27 \longrightarrow w(q_1, T) = 0$ (doesn't appear in either input gene tree)

$q_2 = 12|34 \longrightarrow w(q_2, T) = 1$ (appears in gene tree 1)

$q_3 = 18|23 \longrightarrow w(q_3, T) = 0$ (doesn't appear in either input gene tree)

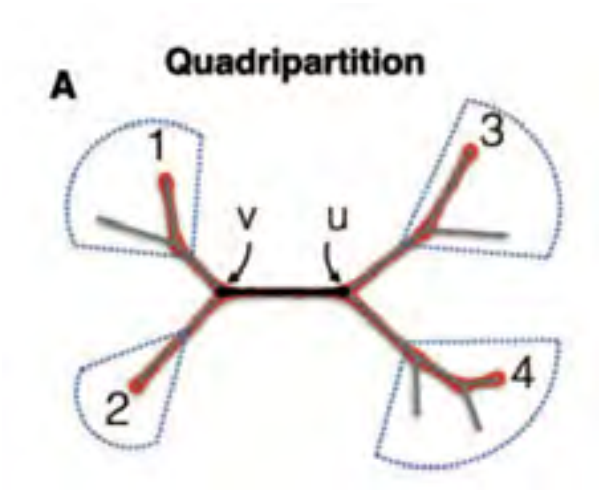
+ all quartets in T_2

- Recall our ideas about inference under the phylogenetic coalescent model

- **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 ✓

Additional features of ASTRAL

- ASTRAL can also estimate branch lengths (in coalescent units)
- ASTRAL also provides a measure of uncertainty: *local posterior probability*



Sayyari and Mirarab, 2016

- Assume that the "clusters" on each edge of the branch under consideration are correct
- Use the gene trees to obtain quartet frequencies for the three possible arrangements of clusters
- Assume a prior distribution on the quartet trees (Yule prior with parameter λ)
- Compute the posterior probability that this branch appears in the true species tree, given the observed quartet frequencies

ASTRAL performance

- ASTRAL is statistically consistent *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)

- **Recall** the difficulty with model-based species tree estimation:

$$P(\mathbf{D}|(S,\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

- **Recall** the difficulty with model-based species tree estimation:

$$P(\mathbf{D}|(S,\tau)) = \sum_{h \in \mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

- If we knew the gene trees for each gene, then the calculation is feasible

- **Recall** the difficulty with model-based species tree estimation:

$$P(\mathbf{D}|(S,\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

- If we knew the gene trees for each gene, then the calculation is feasible

- Bayesian species tree inference methods propose gene trees AND the species tree together – thus making calculation of **a** likelihood possible

- Current software for Bayesian co-estimation:
  - StarBEAST/StarBEAST2 – Ogilvie et al. (2017)
    Estimate the species tree, speciation times, model parameters, posterior probabilities
  - BPP – Flouri et al. (2015)
    Estimate the species tree, speciation times, model parameters, posterior probabilities; also handles species delimitation and species networks
  - SNAPP – Leache et al. (2014)
    Method for SNP and AFLP data

Performance of Bayesian co-estimation methods

- Strengths:
  - Fully model-based
  - Estimates of all model parameters
  - Built-in method for uncertainty quantification via posterior probabilities

- Challenges:
  - Need to specify prior distributions
  - Convergence (and assessing convergence) can be a significant challenge
  - Currently limited to dozens of species and hundreds of genes – doesn't scale well to truly genome-scale data

Site-based methods

- The skull equation one more time:

 $P(\mathbf{D}|(S,\tau)) = \sum_{h \in \mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$ 

Site-based methods

- The skull equation one more time:

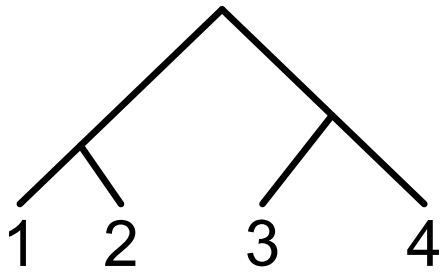
$$P(\mathbf{D}|(S,\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

- Simplify the likelihood by making two assumptions:
  - Suppose that each locus only has 1 bp – sites are unlinked
  - Consider only trees with four taxa – the sum then has either 25 or 31 terms, and there are only 3 integrals for each term

- With these assumptions, we can compute the probabilities!

$$P(\mathbf{D}|(S,\tau)) = \sum_{h\in\mathcal{H}} \int_{\mathbf{t}_h} P(\mathbf{D}|(h,\mathbf{t}_h)) f((h,\mathbf{t}_h)|(\mathcal{S},\tau)) d\mathbf{t}_h$$

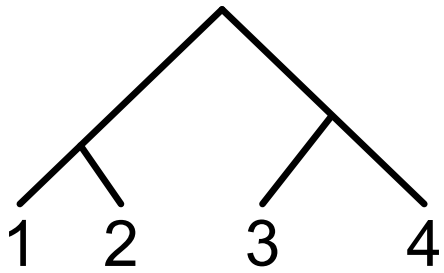
Site-based methods: SVDQuartets



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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

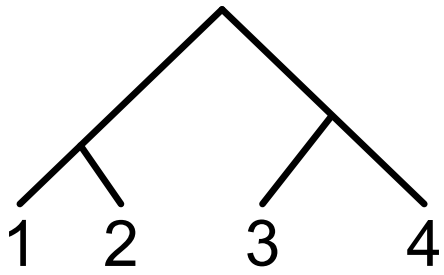
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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

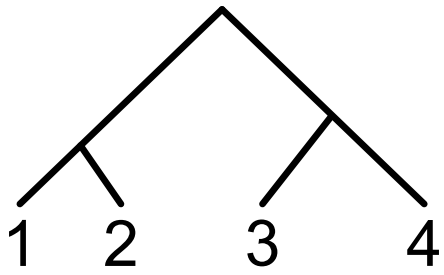
Looking for structure in site pattern probabilities ....



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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

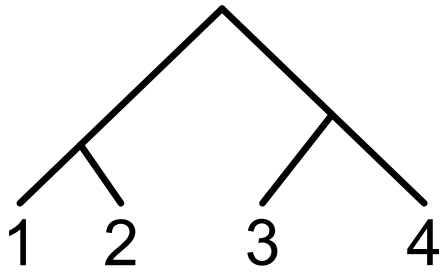
Looking for structure in site pattern probabilities ....



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

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

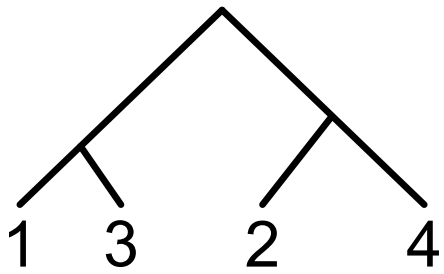
These two columns are identical – matrix rank is reduced by one

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?



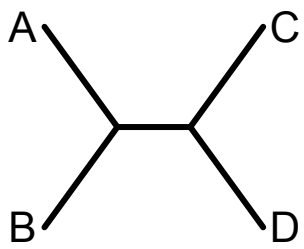
Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

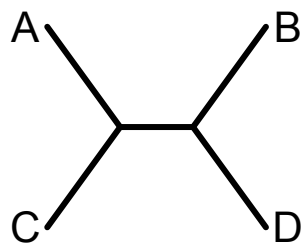
$$\mathbf{Flat}_{12|34}(\mathbf{P}) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\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

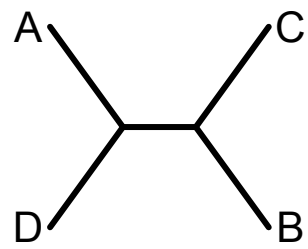
These two columns are no longer identical – full rank matrix in both cases (rank = 16)

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:





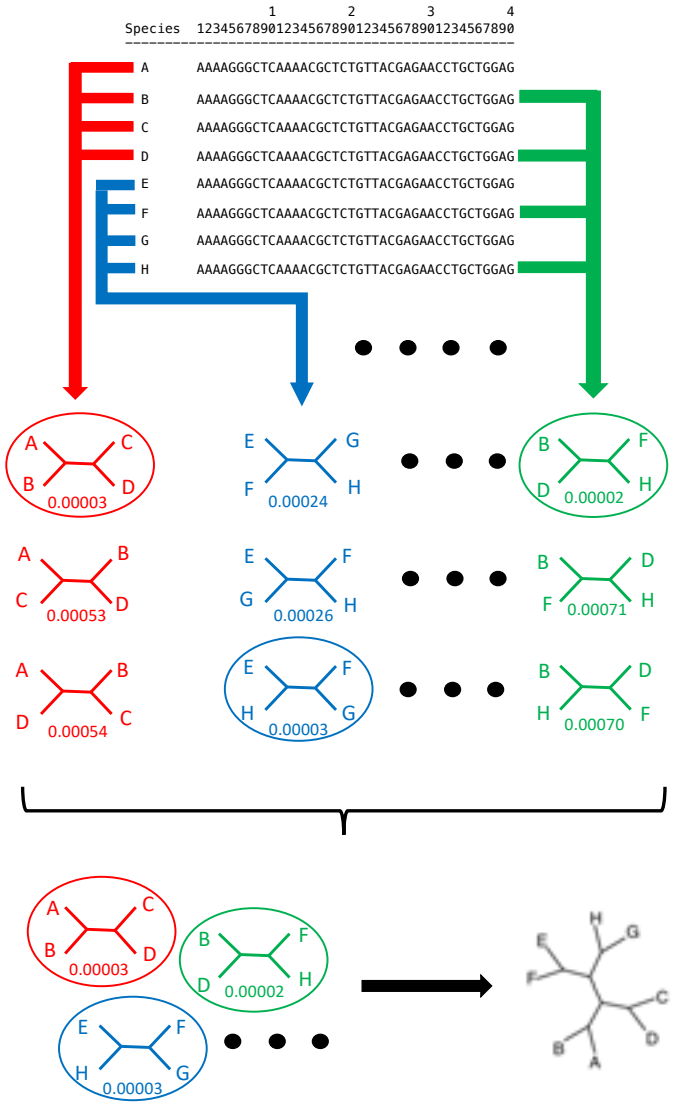


  - Compute a measure of how close each of the three observed flattening matrices is to a matrix with rank 10 – we use the SVDScore
  - Pick the tree relationship that gives the smallest SVDScore

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

Extension to larger trees



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.

Extension to larger trees

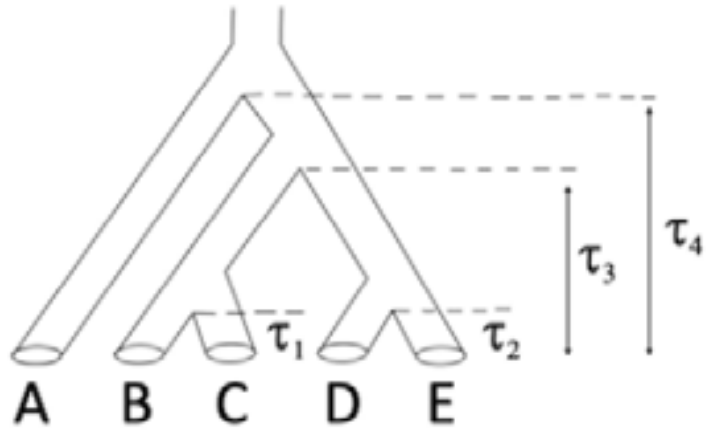
- Multiple lineages are handled as follows:
  1. Sample four species
  2. Select one lineage at random from each species
  3. Estimate the quartet relationships among the four sampled lineages
  4. Restore the species labels (but lineage quartets are saved, too)
- Quantify uncertainty using the bootstrap

Performance of SVDQuartets

- Statistically consistent (Wascher and Kubatko 2021)
- Robust to underlying substitution model
- Scales well to large numbers of species
- Scales well to large numbers of sites
- Perhaps less powerful than a method that more directly uses the likelihood
- Provide an estimate of topology only – but the qAge method can provide estimates of speciation times

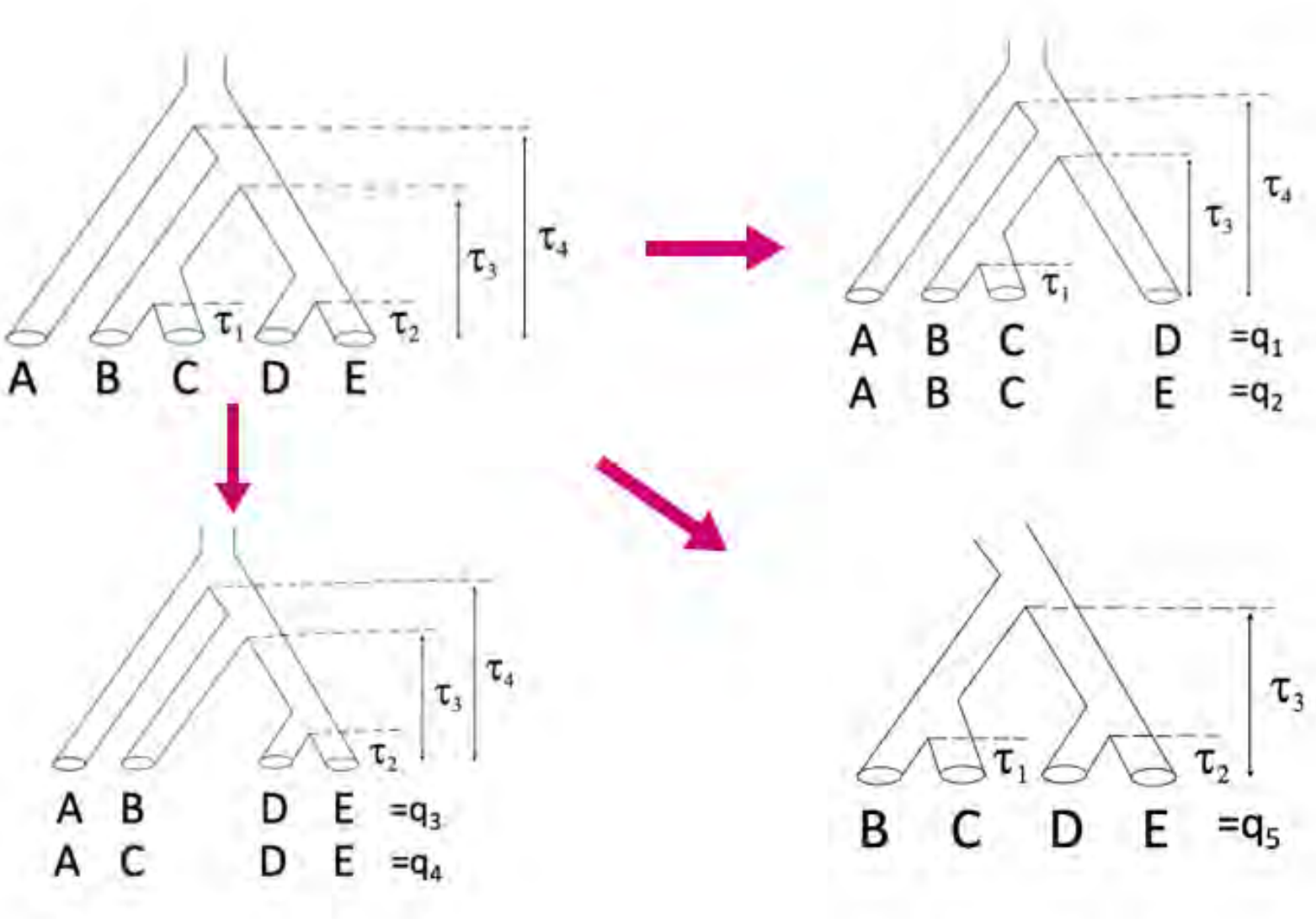
Site-based methods: Composite likelihood

- Consider the following species tree:

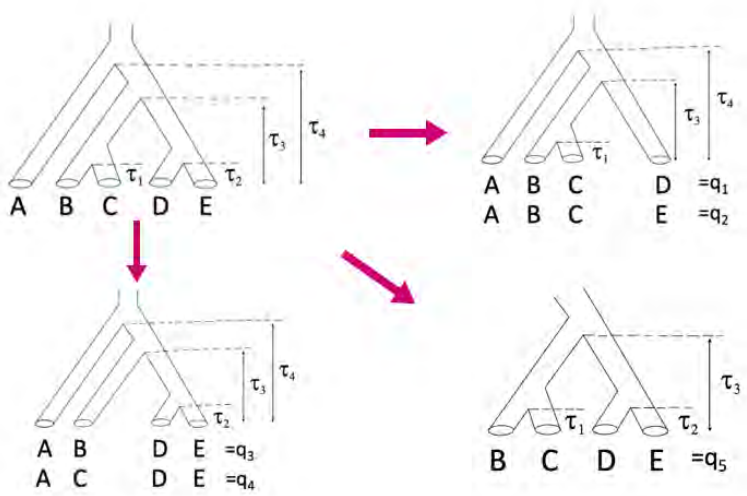


- Idea:
  - Decompose the tree into all 4-taxon subsets
  - Compute the likelihood for each of these
  - Multiply the likelihoods to form the composite likelihood

Composite likelihood



Composite likelihood



The composite likelihood can then be computed as

$$\mathcal{L}_C((\mathcal{S}, \boldsymbol{\tau})|D) = \prod_{i=1}^{5} \mathcal{L}_i((\mathcal{S}_i, \boldsymbol{\tau})|D)$$

where

- $\boldsymbol{\tau}$ is the vector of species tree branch lengths
- $\mathcal{L}_i$ is the likelihood for quartet tree $\mathcal{S}_i$

Why composite likelihood?

- CL methods have a long history in statistics, with much theoretical development:
  - CL estimators are consistent and asymptotically normal
    e.g., Lindsay (1988); Arnold and Strauss (1991); Cox and Reid (2004); Molenberghs and Verbeke (2005)

Why composite likelihood?

- CL methods have a long history in statistics, with much theoretical development:
  - CL estimators are consistent and asymptotically normal
    e.g., Lindsay (1988); Arnold and Strauss (1991); Cox and Reid (2004); Molenberghs and Verbeke (2005)
  - CL can be used for model selection via AIC and BIC
    e.g., Varin and Vidoni (2005); Gao and Song (2010); Ng and Joe (2014)

Why composite likelihood?

- CL methods have a long history in statistics, with much theoretical development:
  - CL estimators are consistent and asymptotically normal
    e.g., Lindsay (1988); Arnold and Strauss (1991); Cox and Reid (2004); Molenberghs and Verbeke (2005)
  - CL can be used for model selection via AIC and BIC
    e.g., Varin and Vidoni (2005); Gao and Song (2010); Ng and Joe (2014)
  - CL can be used to conduct likelihood ratio tests
    e.g., Molenberghs and Verbeke (2005); Chandler and Bate (2007); Pace et al. (2011); Chen et al. (2018)

Why composite likelihood?

- CL methods have a long history in statistics, with much theoretical development:
  - CL estimators are consistent and asymptotically normal
    e.g., Lindsay (1988); Arnold and Strauss (1991); Cox and Reid (2004); Molenberghs and Verbeke (2005)
  - CL can be used for model selection via AIC and BIC
    e.g., Varin and Vidoni (2005); Gao and Song (2010); Ng and Joe (2014)
  - CL can be used to conduct likelihood ratio tests
    e.g., Molenberghs and Verbeke (2005); Chandler and Bate (2007); Pace et al. (2011); Chen et al. (2018)
  - CL can be used in Bayesian settings, including in Markov chain Monte Carlo (MCMC)
    e.g., Pauli et al. (2011); Ribatet et al. (2012); Miller (2021)

Why composite likelihood?

- CL methods also have a long history in population genetics

  Reviewed by Larribe and Fernhead (2011)

Why composite likelihood?

- CL methods also have a long history in population genetics

  Reviewed by Larribe and Fernhead (2011)

- CL (pseudolikelihood) methods have also been used in species tree inference, for example:
  - MP-EST – Liu et al. (2010)
  - PhyloNet (e.g., MPL) – Yu and Nakhleh (2015)
  - SNaQ – Solís-Lemus and Ané (2016)

  but mostly applied to inferring species trees from estimated gene trees

Why composite likelihood?

- CL methods also have a long history in population genetics

  Reviewed by Larribe and Fernhead (2011)

- CL (pseudolikelihood) methods have also been used in species tree inference, for example:
  - MP-EST – Liu et al. (2010)
  - PhyloNet (e.g., MPL) – Yu and Nakhleh (2015)
  - SNaQ – Solís-Lemus and Ané (2016)

  but mostly applied to inferring species trees from estimated gene trees

Maximum likelihood phylogenetics has always been a composite likelihood method

Site-based methods: composite likelihood

- Some composite likelihood species tree methods:
  - qAge – implemented in PAUP* (along with SVDQuartets) – estimates speciation times on a fixed species tree
  - PhyNEST – estimation of species networks using composite likelihood
  - PICL – set of tools for phylogenetic inference (right now, species trees only) using composite likelihood; includes both multilocus data and SNPs
  - Bayesian speciation time estimation using composite likelihood – dissertation work of Shawn Chen (see talk in Thursday's virtual Evolution meeting)
- I'm really excited about these approaches!

Pros and cons of composite likelihood

- A computationally-tractable approach that directly uses the model-based likelihood and has firm theoretical foundations
- Requires search over tree space if the tree that maximizes the composite likelihood is to be used
- Uncertainty quantification uses the bootstrap

- Please note!
  - This is NOT a comprehensive list of methods
  - The methods discussed here largely deal ONLY with the phenomenon of incomplete lineage sorting that is modeled by the multispecies coalescent
  - Other processes – e.g., horizontal gene transfer, gene duplication and loss – are often important, too, and can be modeled
  - The methods discussed here apply to sexually-reproducing organisms for which variation in gene history along a chromosome arises

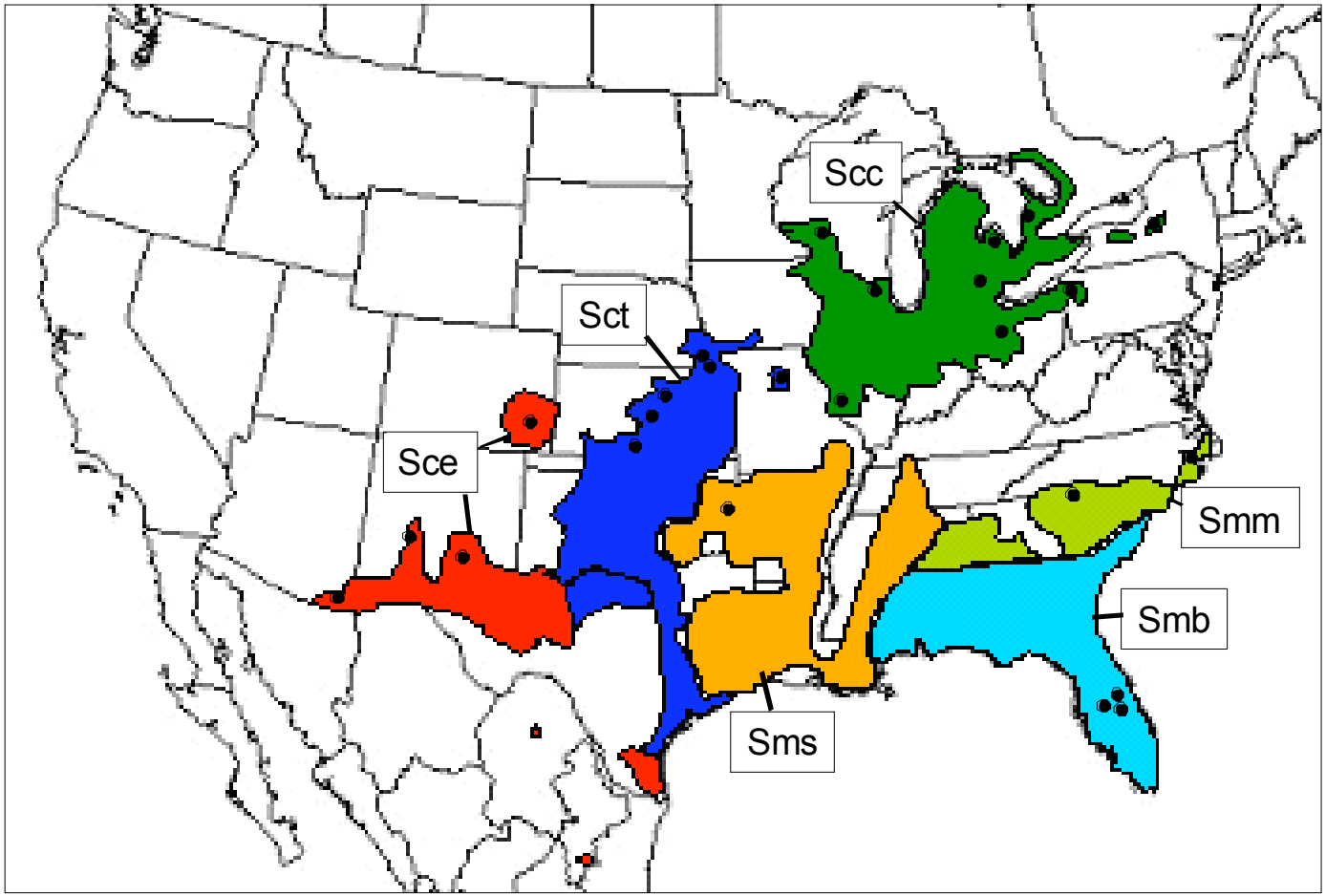
- Please note!
  - This is NOT a comprehensive list of methods
  - The methods discussed here largely deal ONLY with the phenomenon of incomplete lineage sorting that is modeled by the multispecies coalescent
  - Other processes – e.g., horizontal gene transfer, gene duplication and loss – are often important, too, and can be modeled
  - The methods discussed here apply to sexually-reproducing organisms for which variation in gene history along a chromosome arises
- Now on to empirical examples!

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

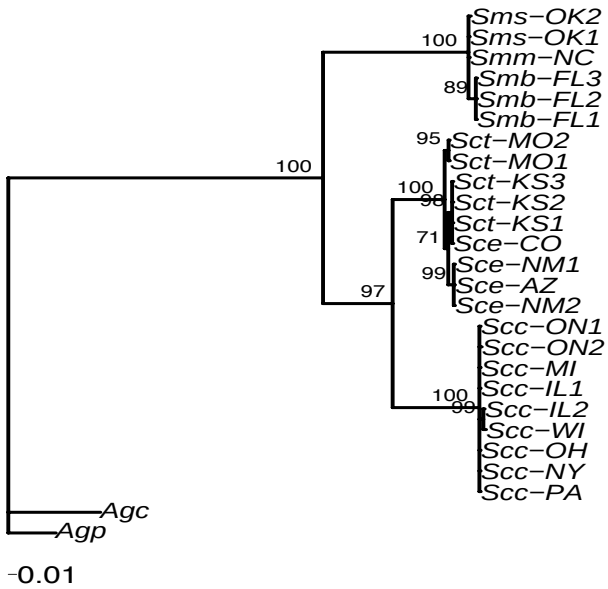


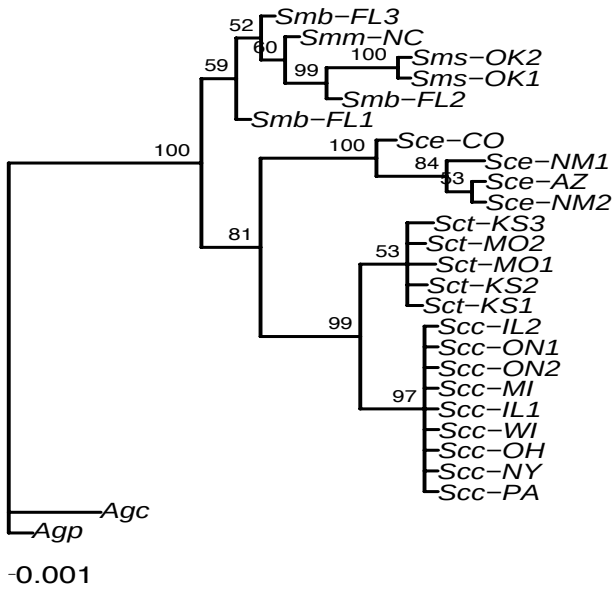
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

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

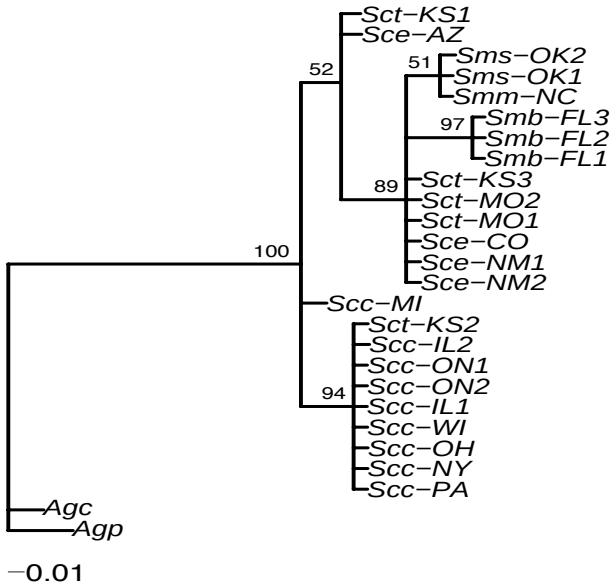
Some are very informative:

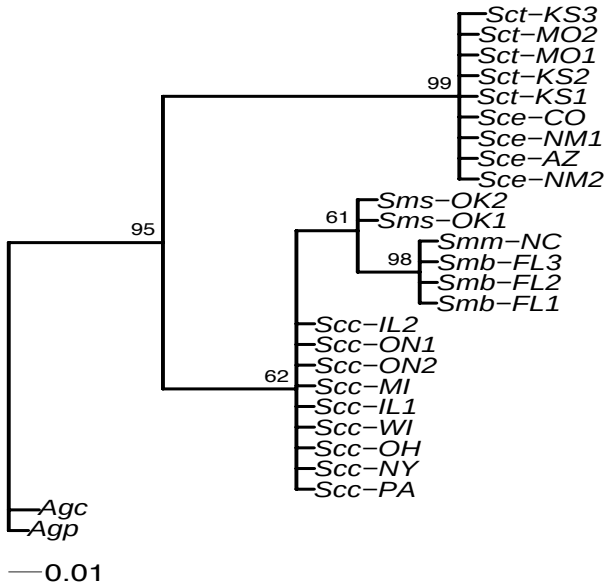




Individual Gene Tree Estimates

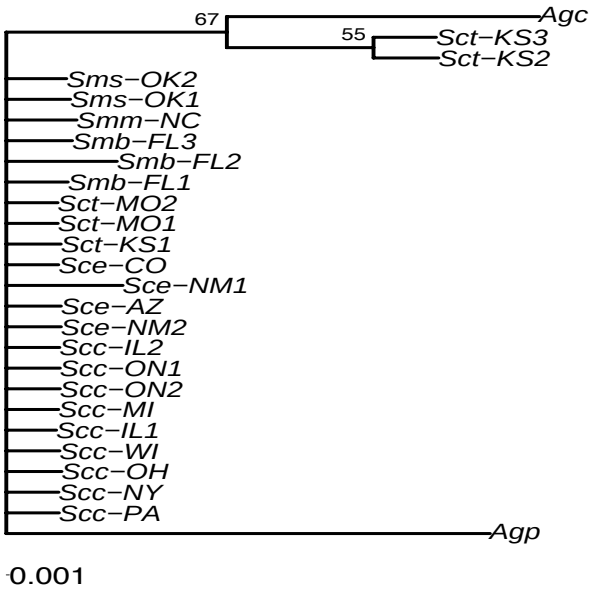
Some are a little informative:

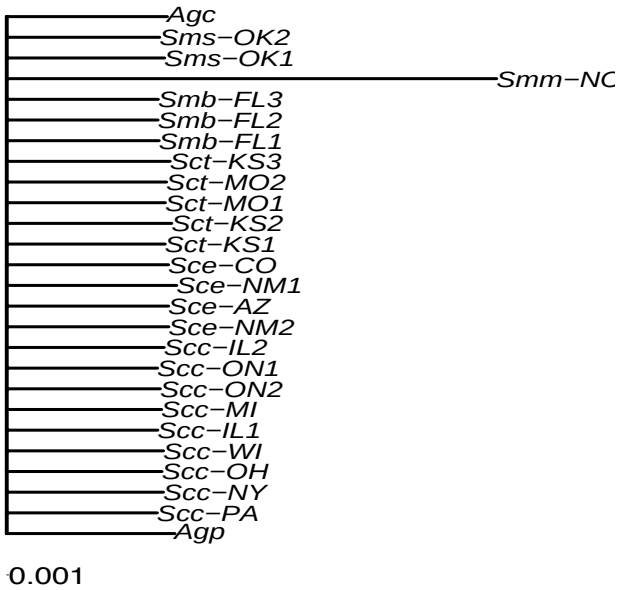




Individual Gene Tree Estimates

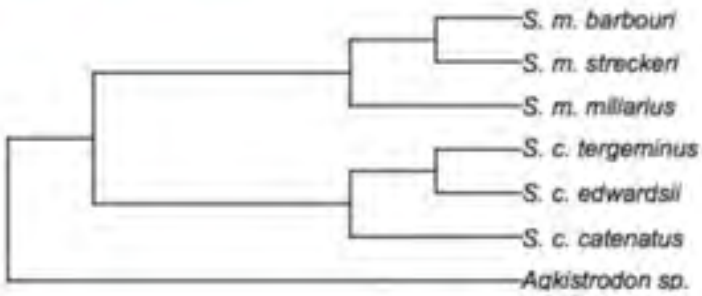
And then there are others .....





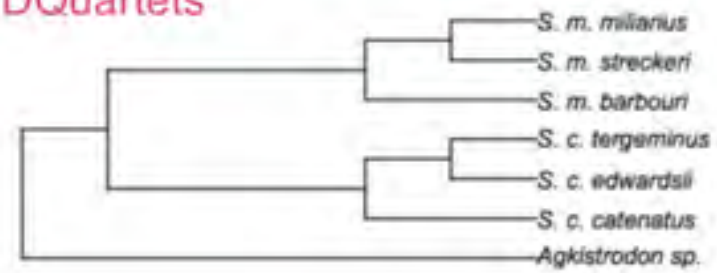
Example 1: Sistrurus rattlesnakes

STEM, STEAC

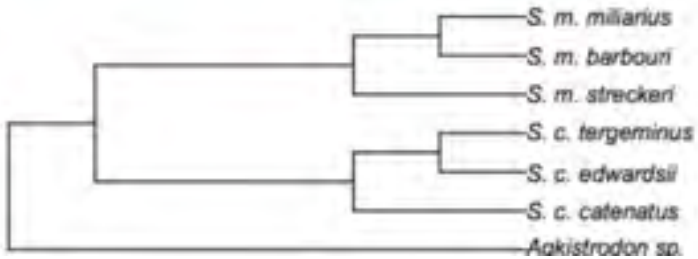


BEAST (concatenated data), *BEAST

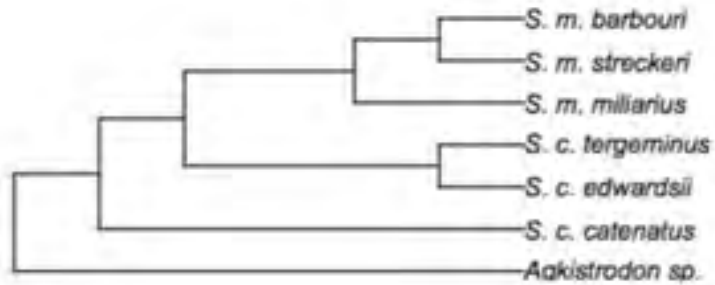
SVDQuartets



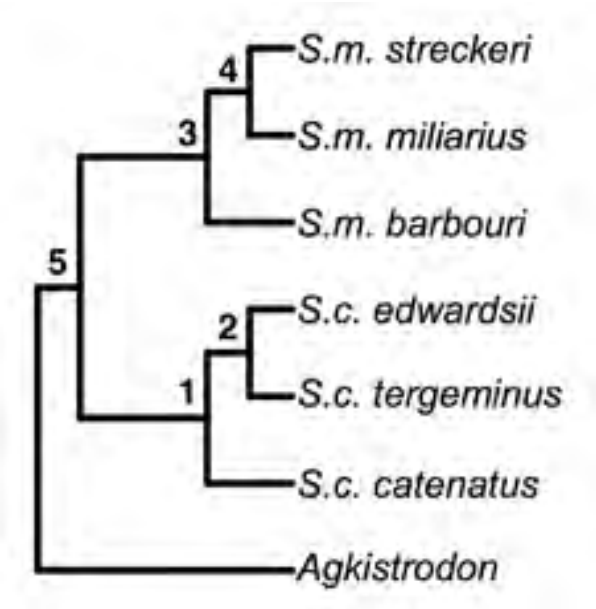
BEST, Parsimony & MrBayes (concatenated data), Astral



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

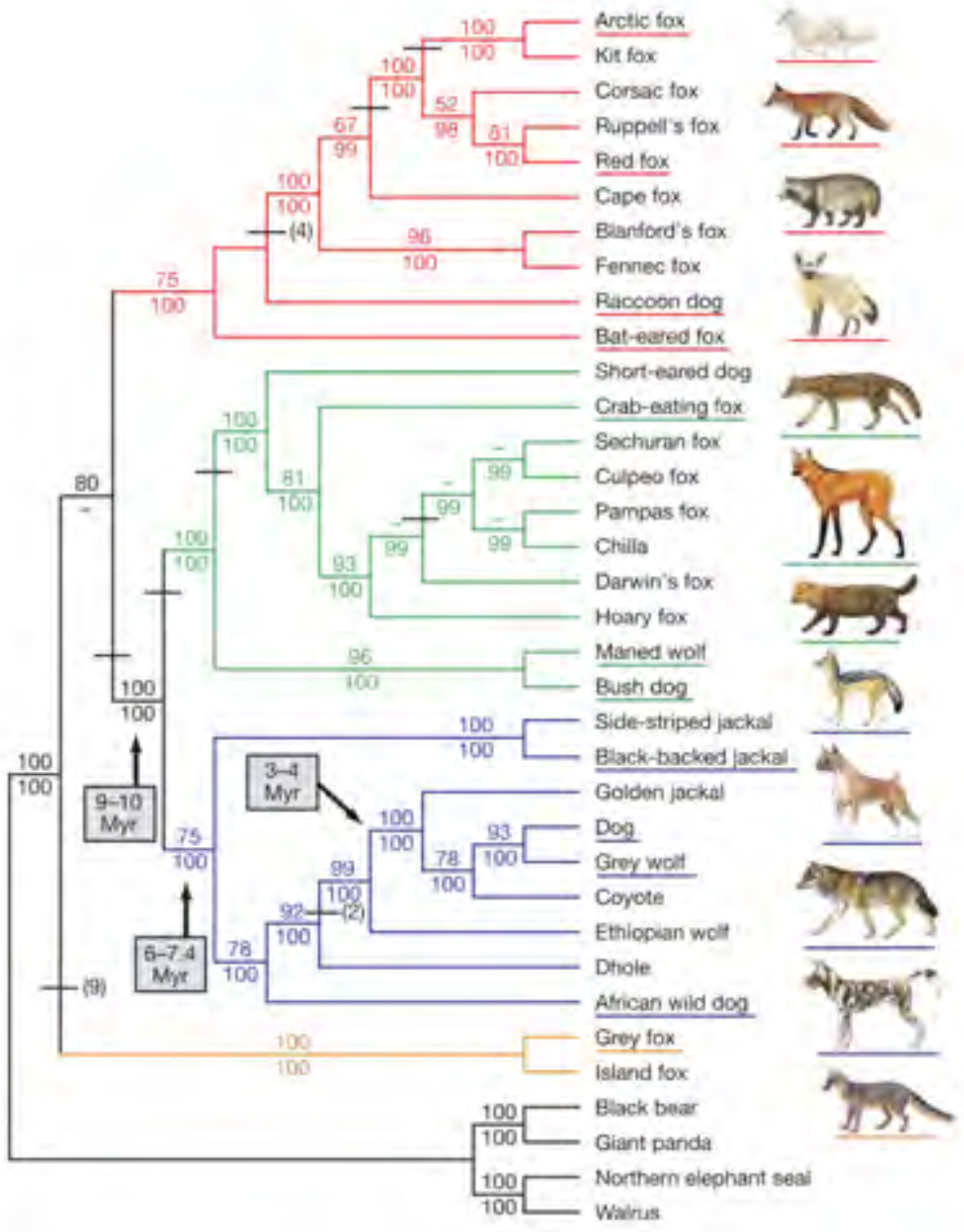
* = This clade was not in the maximum clade credibility (*S. m. miliarius* and *S. m. barbouri* received 48.78% posterior probability with *BEAST and 59% posterior probability with BPP)

- 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

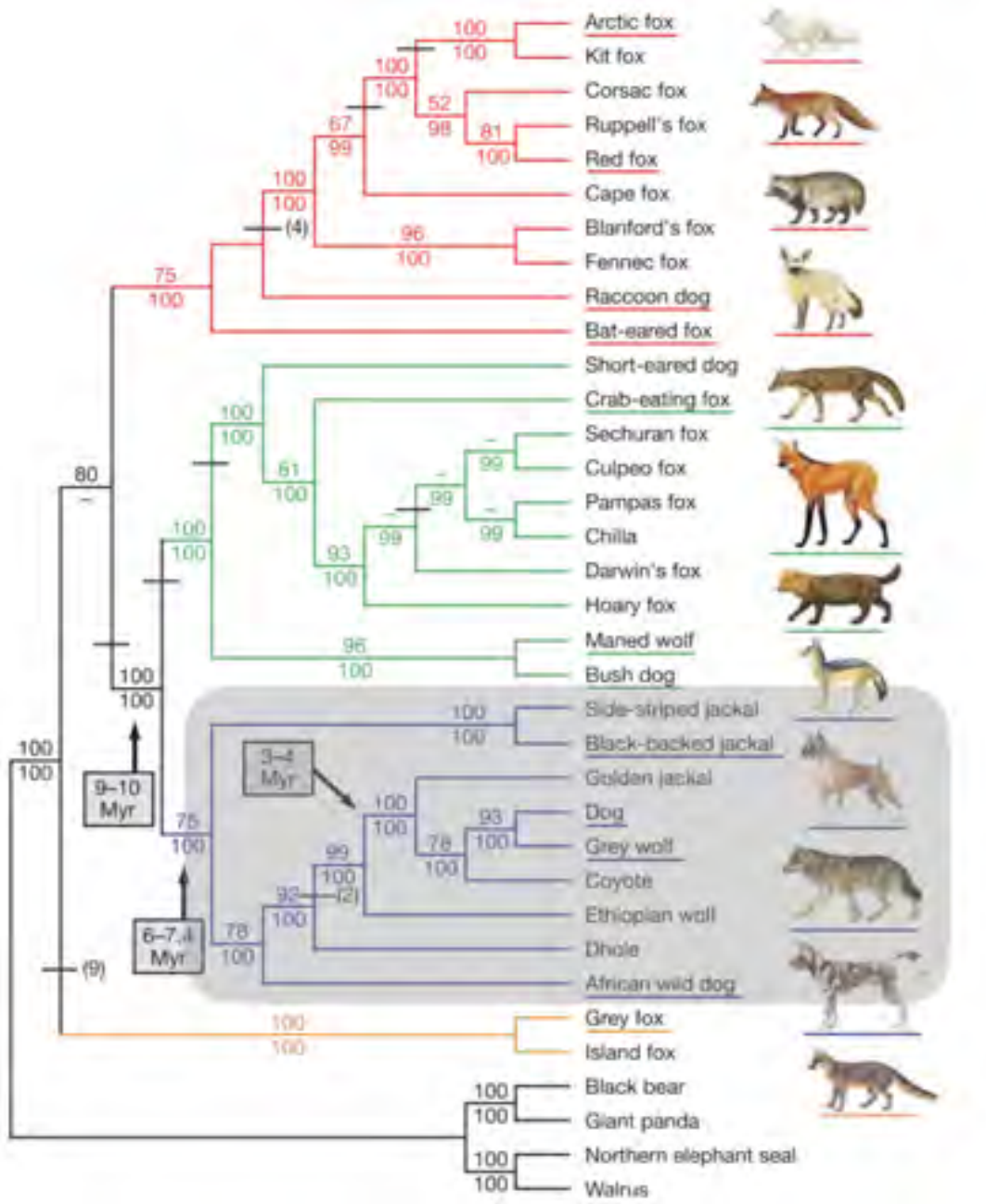
Example 2: Canid phylogeny

- Lindblad-Toh et al. (*Nature* 438: 803-819, 2005) reported a genome sequence for domesticated dog, and used it to construct a phylogeny for dogs and their close relatives
- The phylogeny was based on 16 loci with a total of 15K bp
- Estimated with parsimony in PAUP* (bootstrap frequencies above the nodes) and MrBayes 3 (posterior probabilities are below the nodes)

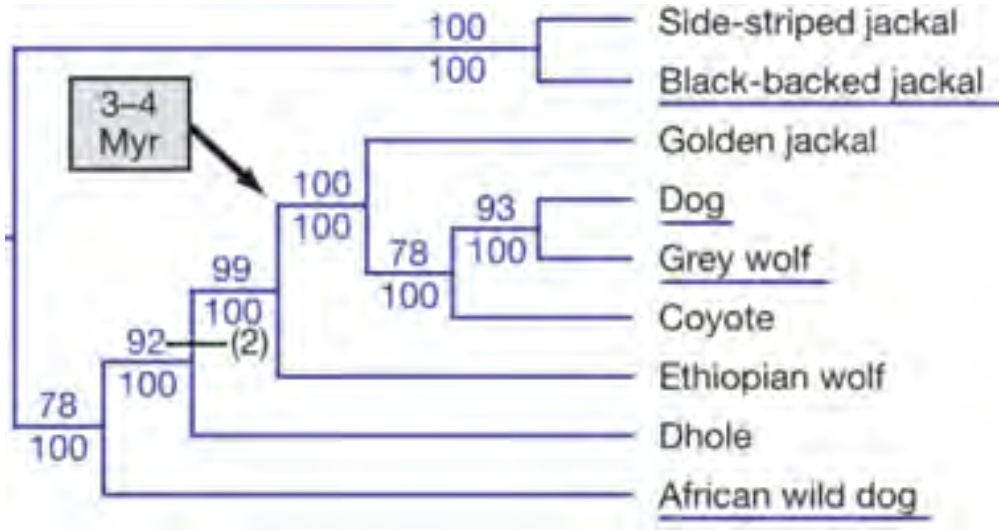


Example 2: Canid phylogeny

- Lindblad-Toh et al. (*Nature* 438: 803-819, 2005) reported a genome sequence for domesticated dog, and used it to construct a phylogeny for dogs and their close relatives
- The phylogeny was based on 16 loci with a total of 15K bp
- Estimated with parsimony in PAUP* (bootstrap frequencies above the nodes) and MrBayes 3 (posterior probabilities are below the nodes)



Example 2: Canid phylogeny



- Species tree estimated with:
  - StarBEAST
  - SVDQuartets (bootstrap consensus tree)
  - BUT, there is much lower support for most nodes than in the concatenated analysis: bootstrap support values are 63 – 81

Summary

- Many evolutionary processes can contribute to variation in the evolutionary histories of individual genes, and modeling such processes can be an important part of species tree inference
- The multispecies coalescent is the most commonly used model of the lineage sorting process
- There are (at least) three reasons to use a species tree inference method when the object of interest is the species tree
- Methods for species tree inference continue to be developed – to improve accuracy, scalability, and realism
- I'll post links to some tutorials in slack.