Species Tree Estimation

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

lkubatko@stat.osu.edu
twitter: Laura_Kubatko

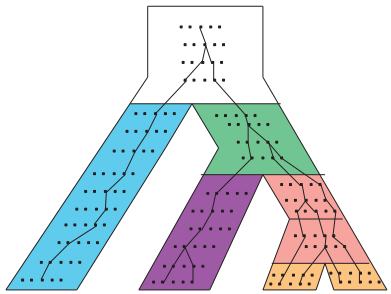
February 3, 2015

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)

Relationship between population genetics and phylogenetics

- Given current technology, we can do much more:
  - Sample many individuals within each taxon (species, population, etc.)
  - Sequence many genes for all individuals
- Need models at two levels:
  - Model what happens within each population [population genetics – coalescent model]
  - Link each within-population model on a phylogeny [phylogenetics]



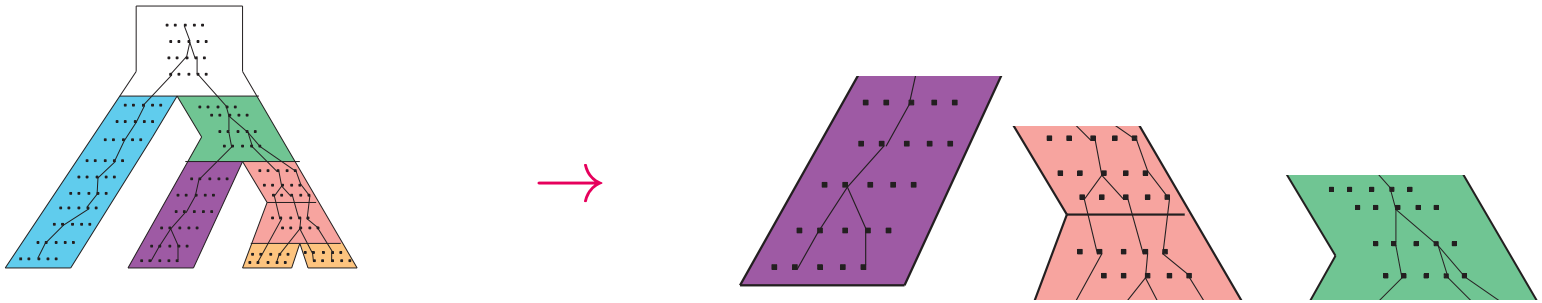
Recall several facts from Peter's lecture

- Under the Wright-Fisher model, the number of generations back into the past until two lineages coalesce $\sim$ Geometric($\frac{1}{2N}$)

- 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 $\frac{4N}{k(k-1)}$

- Peter showed us how to use this model to compute the probability density of a "population tree".

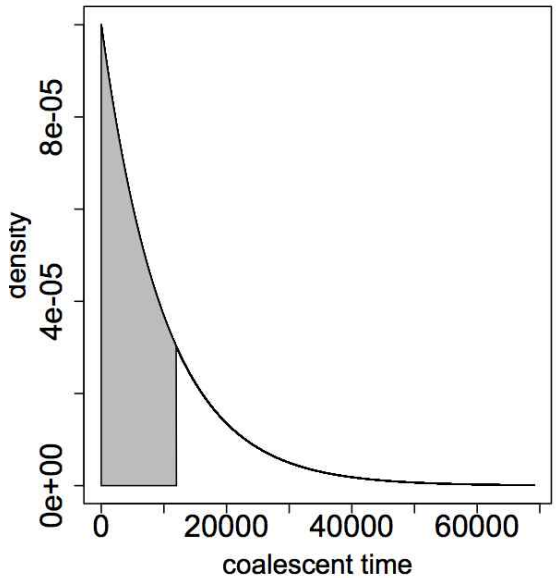
Fitting population trees into a phylogeny

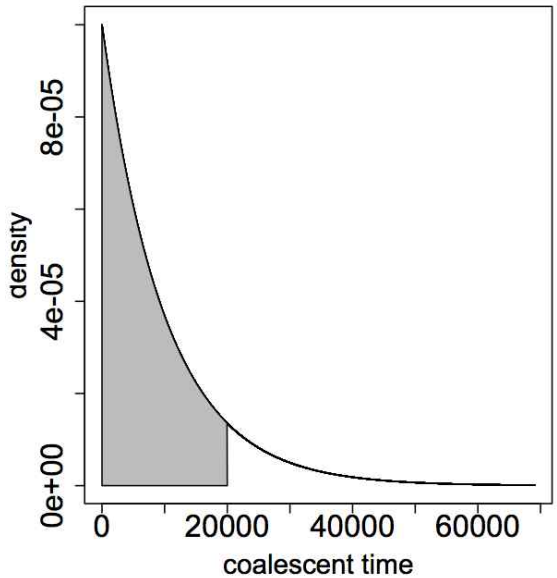


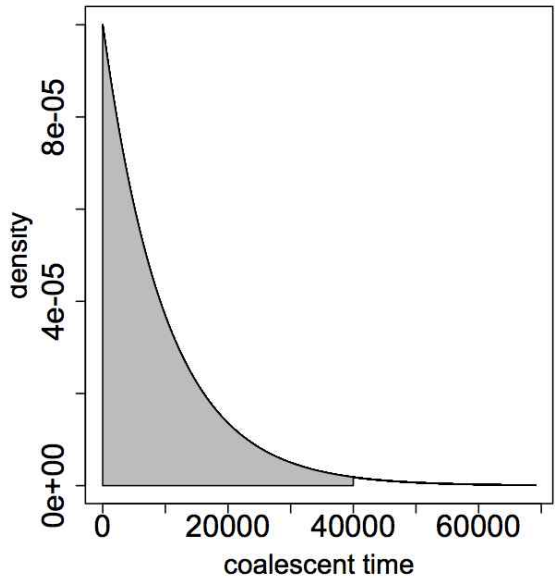
- Focus on just one speciation interval and a sample of $k = 2$ lineages.
- Then, $\binom{k}{2} = 1$ and we have an exponential distribution with rate $\frac{1}{2N}$ and mean $2N$.
- Suppose $N = 5,000$. Let's find the probability that the two lineages coalesce in an interval of a particular length.

Fitting population trees into a phylogeny

- $N = 5,000$ and consider the times: 12,000, 20,000 and 40,000 generations





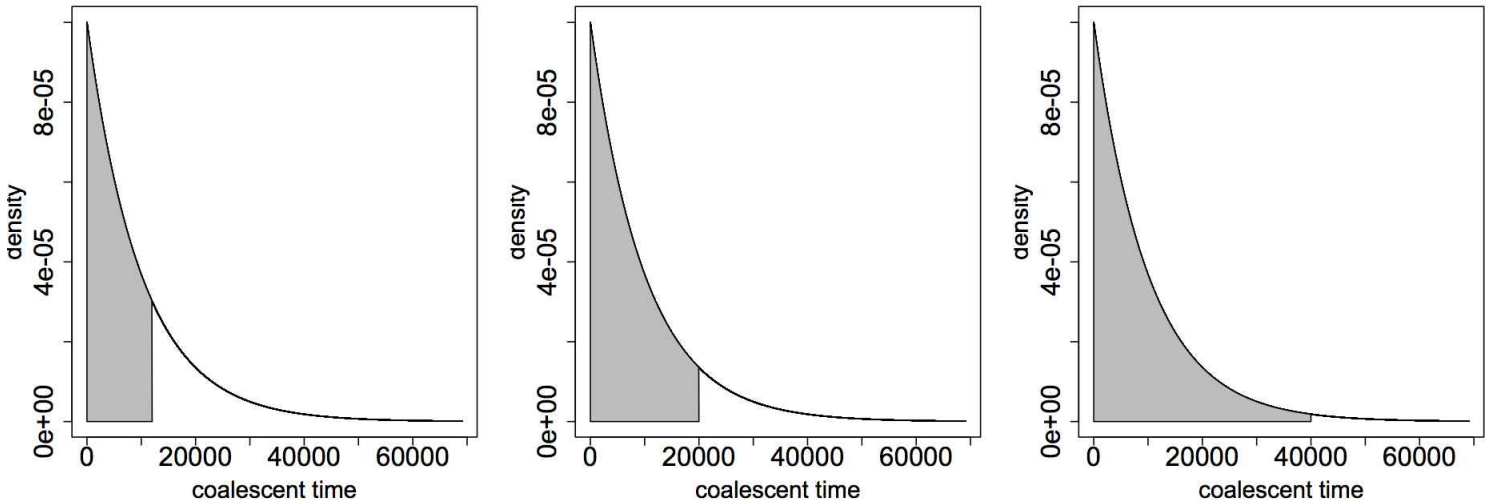


Fitting population trees into a phylogeny

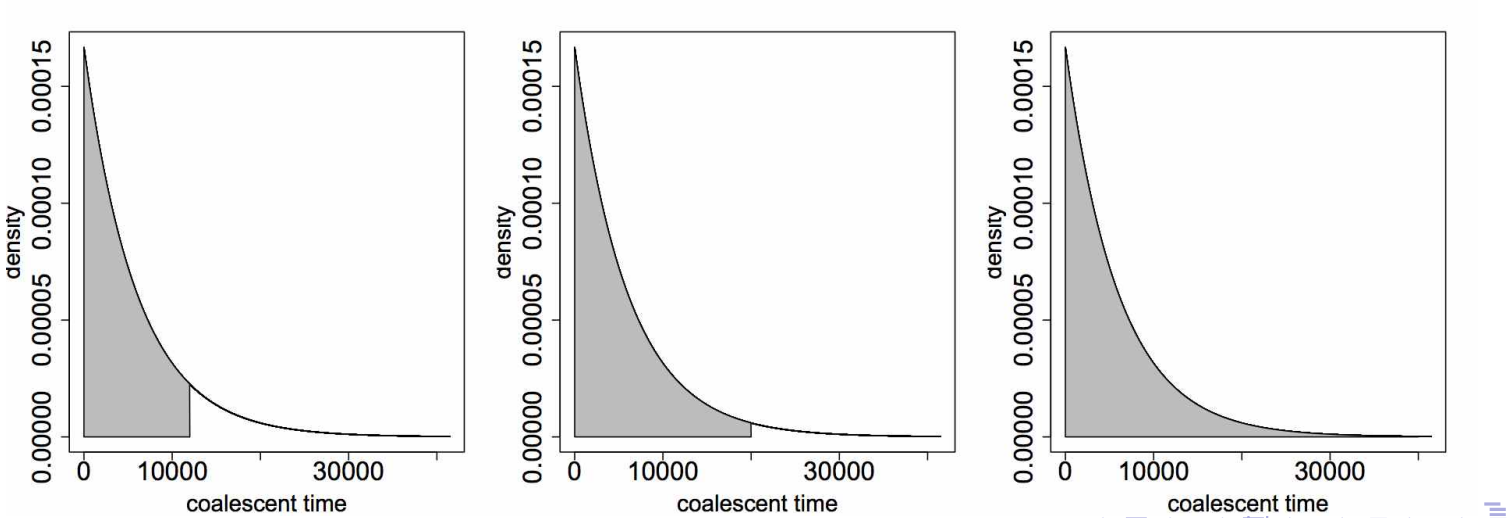
- What happens if we change the population size, N?
- Recall that we have an exponential distribution with rate $\frac{1}{2N}$ and mean $2N$.
- Now suppose $N = 3,000$ and look at the same speciation interval lengths.

Fitting population trees into a phylogeny

- $N = 5,000$

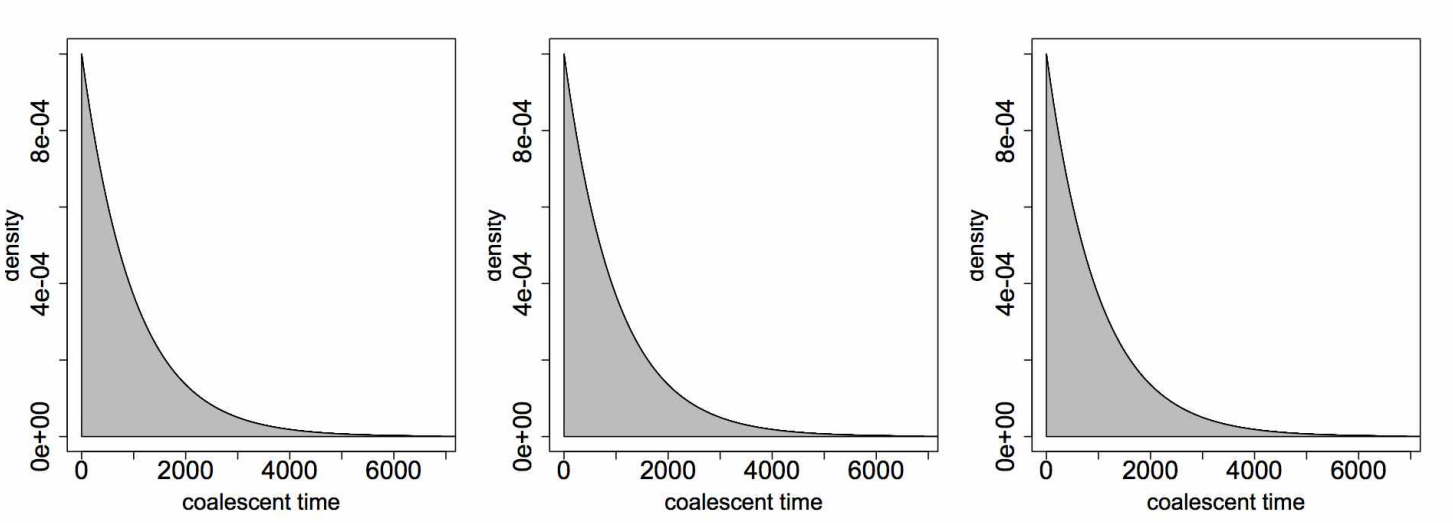


- $N = 3,000$



Fitting population trees into a phylogeny

- What about the effect of sample size, k?
- Consider $N = 5,000$ again, but now use $k = 5$.
  - Rate is $\binom{5}{2}\frac{1}{2N} = \frac{10}{2N}$ (was $\frac{1}{2N}$)
  - Mean is $\frac{4N}{k(k-1)} = \frac{2N}{10}$ (was $2N$)

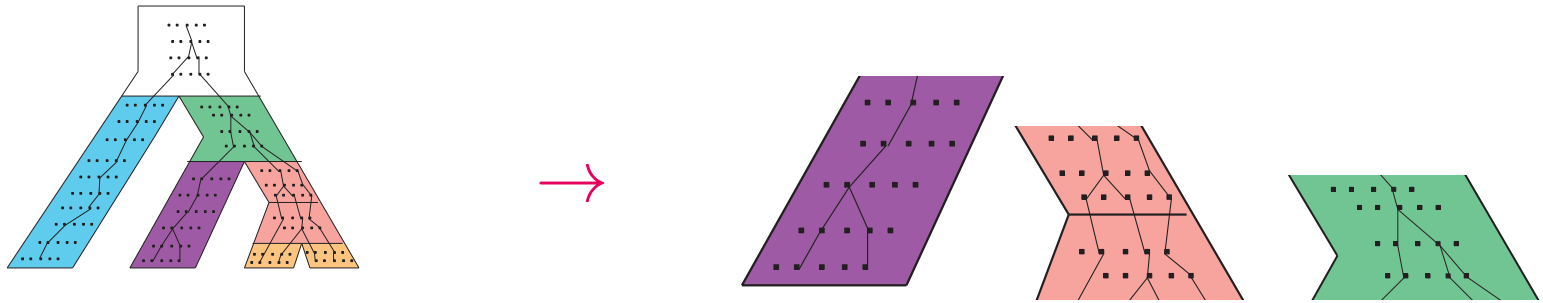


Fitting population trees into a phylogeny

- 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.
- What does this mean for species trees estimation ???

Fitting population trees into a phylogeny

- Recall our goal to integrate the population process with the phylogeny:



- Can use our previous results to get the following:
  - The probability that u lineages coalesce into v lineages in time t is given by (Tavare, 1984; Watterson, 1984; Takahata and Nei, 1985; Rosenberg, 2002)

$$P_{uv}(t) = \sum_{j=v}^{u} e^{-j(j-1)t/2} \frac{(2j-1)(-1)^{j-v}}{v!(j-v)!(v+j-1)} \prod_{y=0}^{j-1} \frac{(v+y)(u-y)}{u+y}$$

- When u and v are small, these are easy to compute. For example,

$$\begin{aligned} P_{21}(t) &= \text{probability that 2 lineages coalesce to 1 lineage in time } t \\ &= \text{probability of 1 coalescent event in time } t \text{ when } k = 2 \\ &= P(T \leq t), \text{ where } T \sim Exp(\mu = 1) \\ &= \int_0^t e^{-x} dx = 1 - e^{-t} \end{aligned}$$

[Note: this is the formula for the gray area in the graphs]

- Similarly,

$$\begin{aligned} P_{22}(t) &= \text{prob. of no coalescence in time } t \text{ for 2 lineages} \\ &= P(T > t) \\ &= \int_t^\infty e^{-x} dx = e^{-t} \end{aligned}$$

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

- Assumptions:
  - 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.

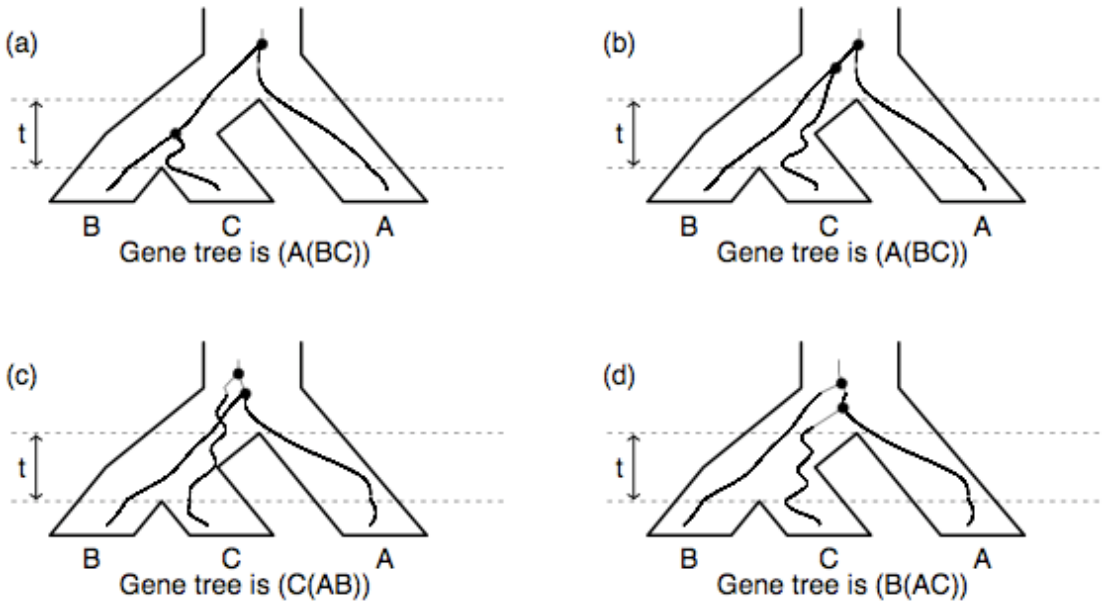
Putting it together . . . the coalescent model along a species tree

- When talking about gene tree distributions, there are two cases of interest:
  - The gene tree topology distribution
  - The joint distribution of topologies and branch lengths
- Start with the simple case of 3 species with 1 lineage sampled in each and look at the gene tree topology distribution

Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

Example of gene tree probability computation:

(a) Prob $= 1 - e^{-t}$; (b), (c), (d) Prob $= \frac{1}{3}e^{-t}$

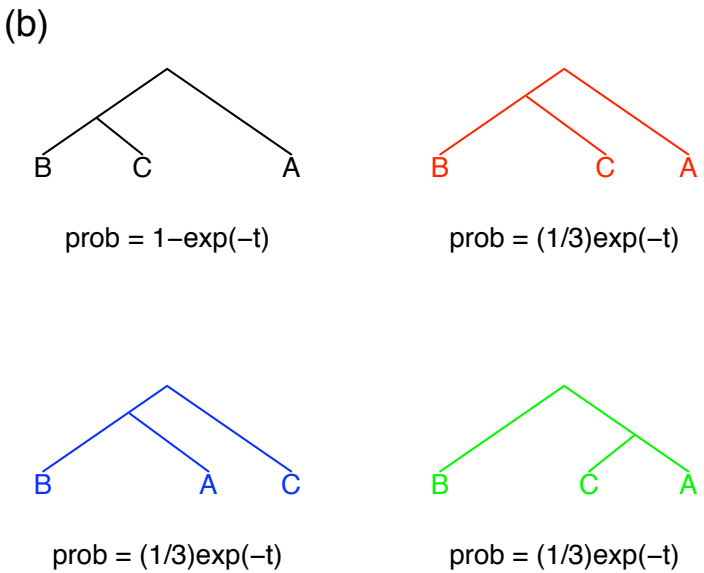


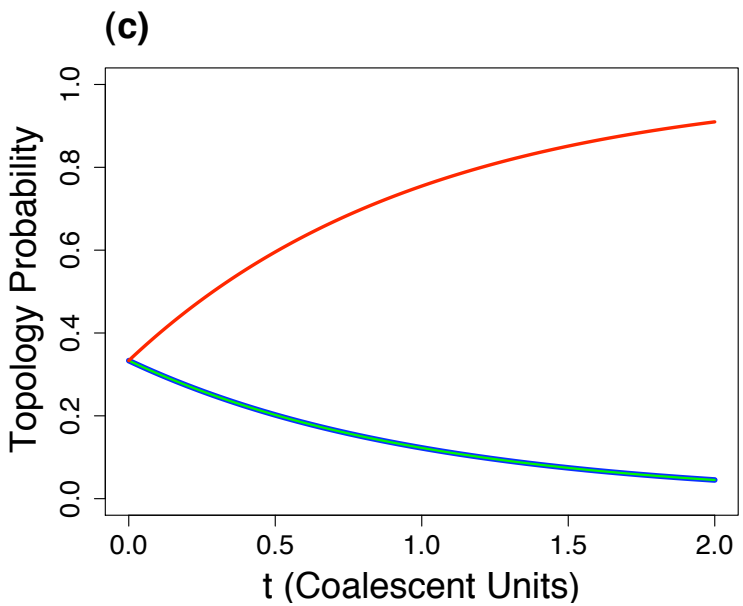
Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

- Thus, we have the following probabilities:
  - Gene tree (A,(B,C)): prob $= 1 - e^{-t} + \frac{1}{3}e^{-t} = 1 - \frac{2}{3}e^{-t}$
  - Gene tree (B,(A,C)): prob $= \frac{1}{3}e^{-t}$
  - Gene tree (C,(A,B)): prob $= \frac{1}{3}e^{-t}$
- Note: There are two ways to get the first gene tree. We call these histories.
- The probability associated with a gene tree topology will be the sum over all histories that have that topology.

Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

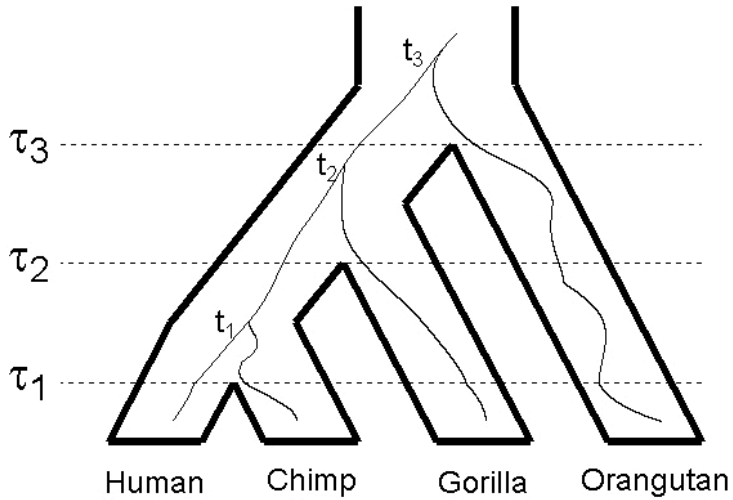
- What are these probabilities like as a function of t, the length of time between speciation events?





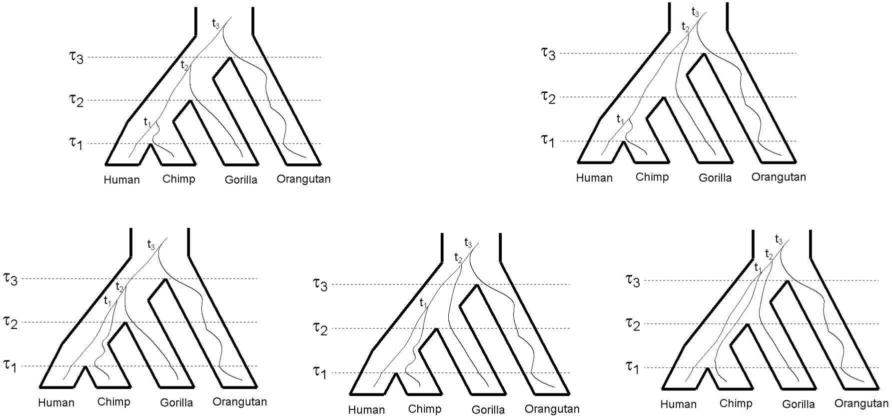
Example: a slightly larger case

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



Coalescent histories for the 4-taxon example

- There are 5 possibilities 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

Computing the Topology Distribution by Enumerating Histories

- In the general case, we have the following:

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

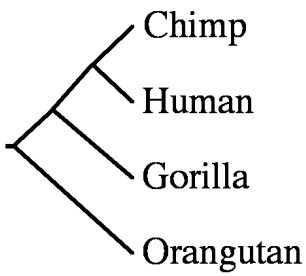
$$P\{G = g|\mathcal{S}\} = \sum_{histories} P\{G = g, history|\mathcal{S}\}$$

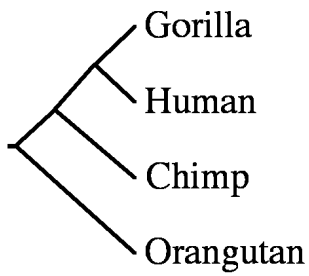
- Implemented in the software COAL (Degnan and Salter, *Evolution*, 2005)
- A more efficient method has been proposed (Wu, *Evolution*, 2012)

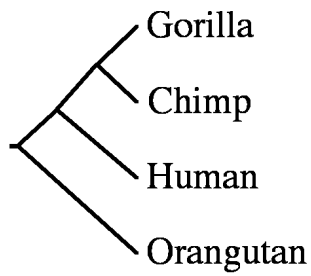
Applications of the topology distribution - example 1

- 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.

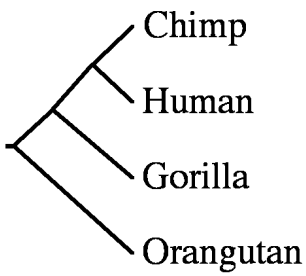
Applications of the topology distribution - example 1



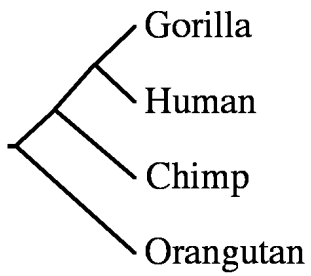




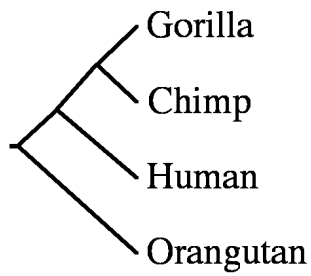
Applications of the topology distribution - example 1



76.6%



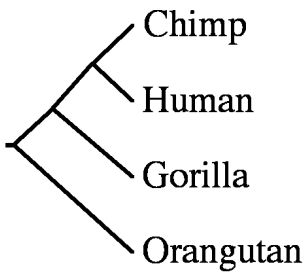
11.4%

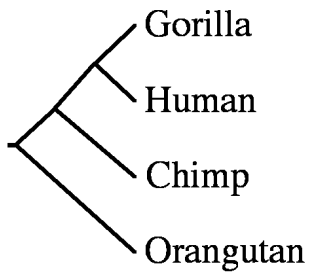


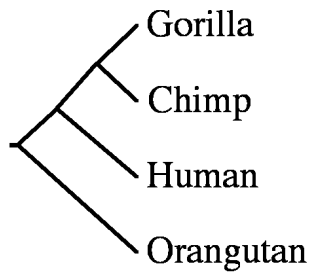
11.5%

Observed proportions of each gene tree among ML phylogenies

Applications of the topology distribution - example 1







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

Observed proportions of each gene tree among ML phylogenies

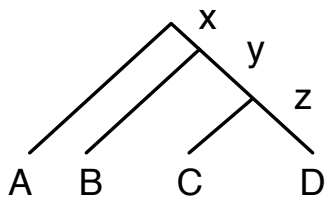
Predicted proportions using parameters from Rannala & Yang, 2003.

Applications of the topology distribution - example 2

- In the previous example, one topology is clear preferred
- Must the distribution always look this way?
- Examine the entire distribution when the number of taxa is small

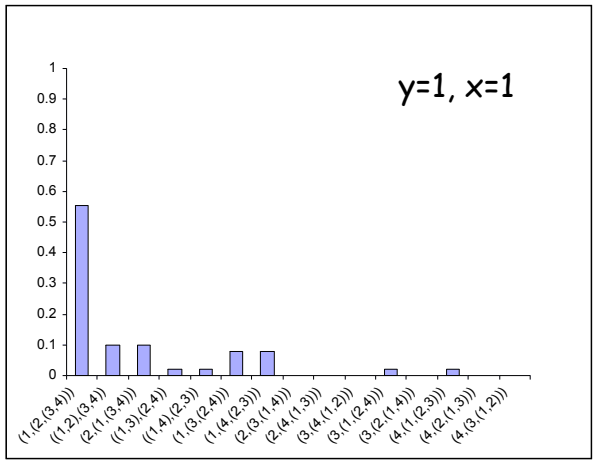
Applications of the topology distribution - example 2

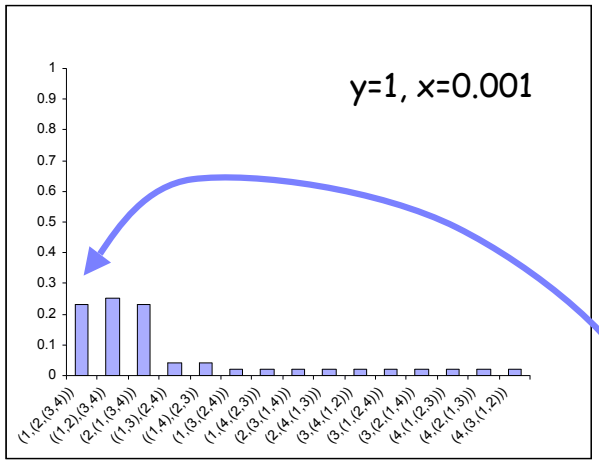
- Consider 4 taxa: A, B, C, and D
- Species tree:

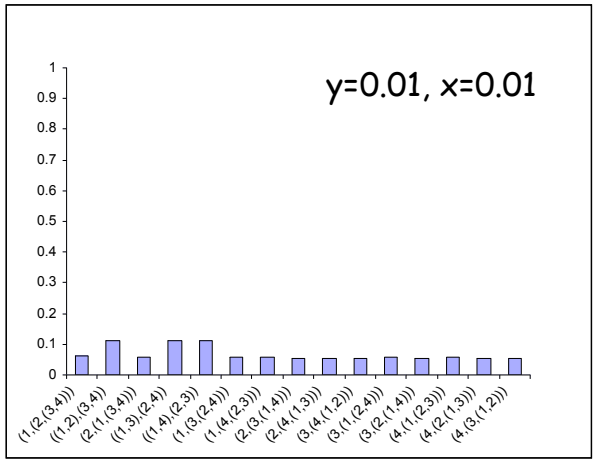


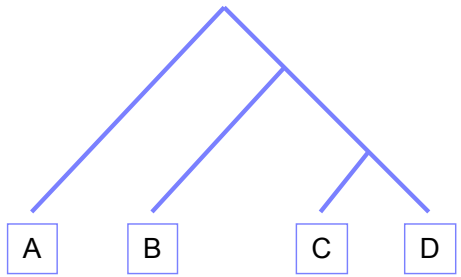
- Look at probabilities of all 15 tree topologies for values of x, y, and z

Applications of the topology distribution - example 2

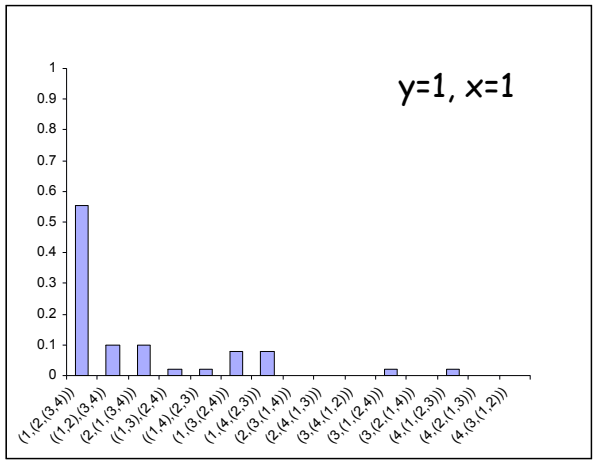


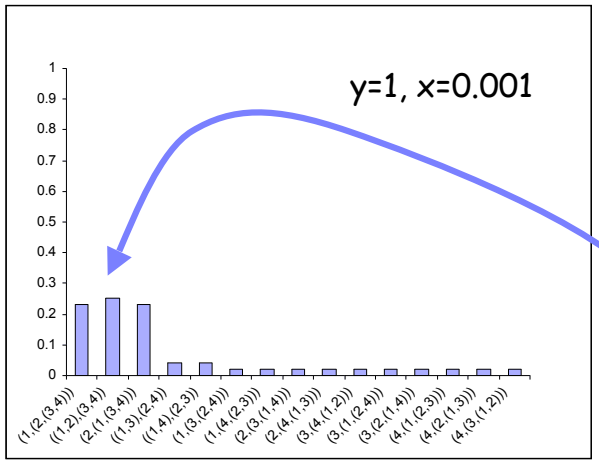


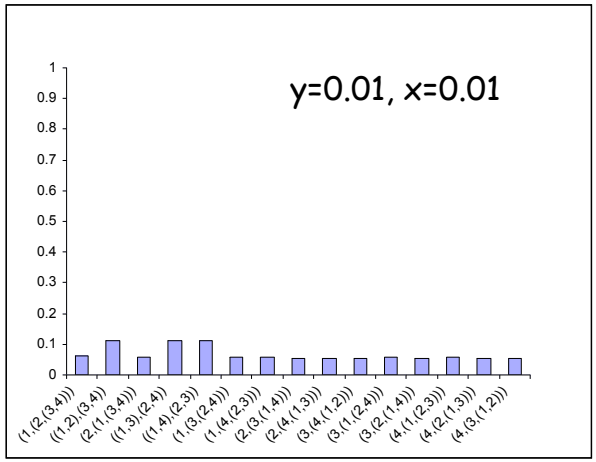


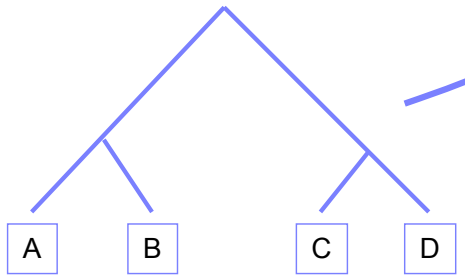


Applications of the topology distribution - example 2

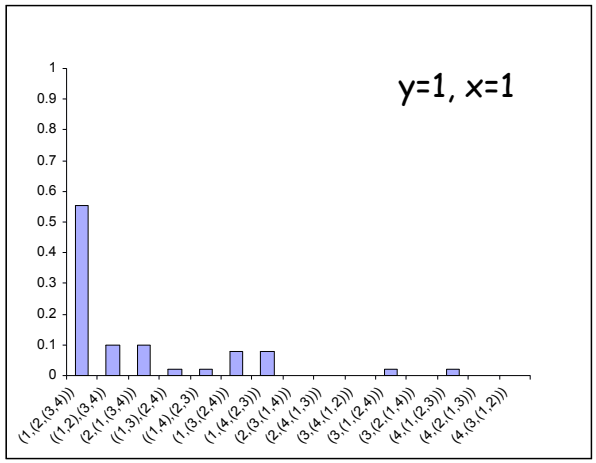


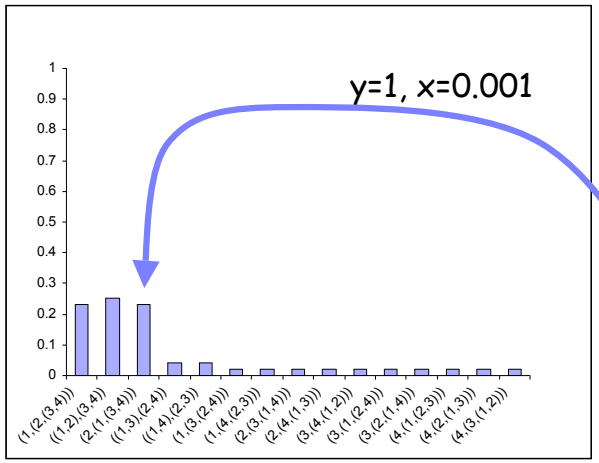


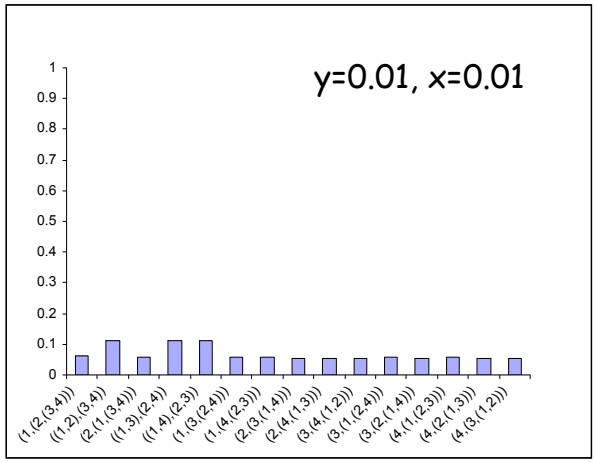


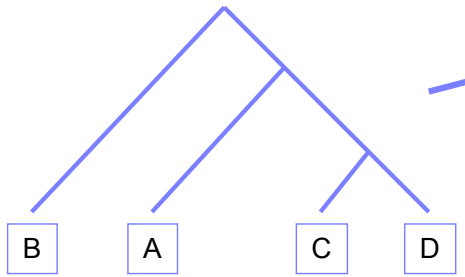


Applications of the topology distribution - example 2

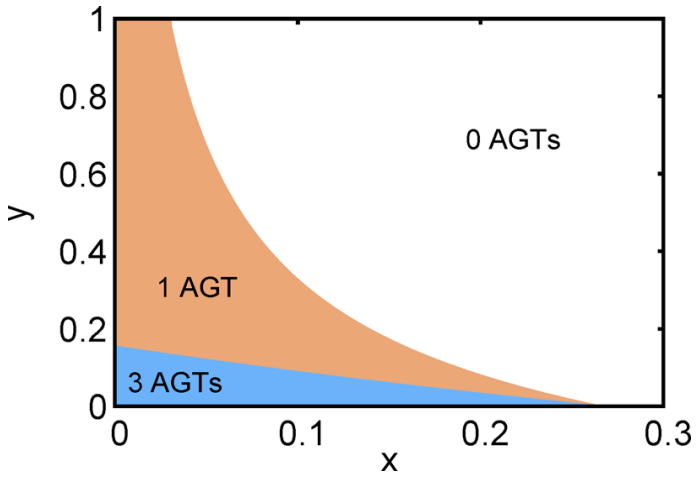








Applications of the topology distribution - example 2



- 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

Applications of the topology distribution - example 3

- 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.

Applications of the topology distribution - example 3

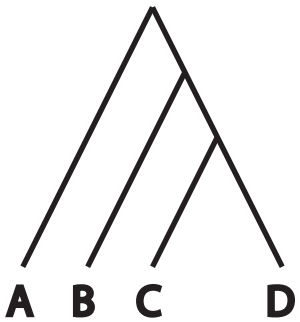
- Given this model, how should inference be carried out?

Applications of the topology distribution - example 3

- Given this model, how should inference be carried out?
- Hypothesis: 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.

Applications of the topology distribution - example 3

Species Tree

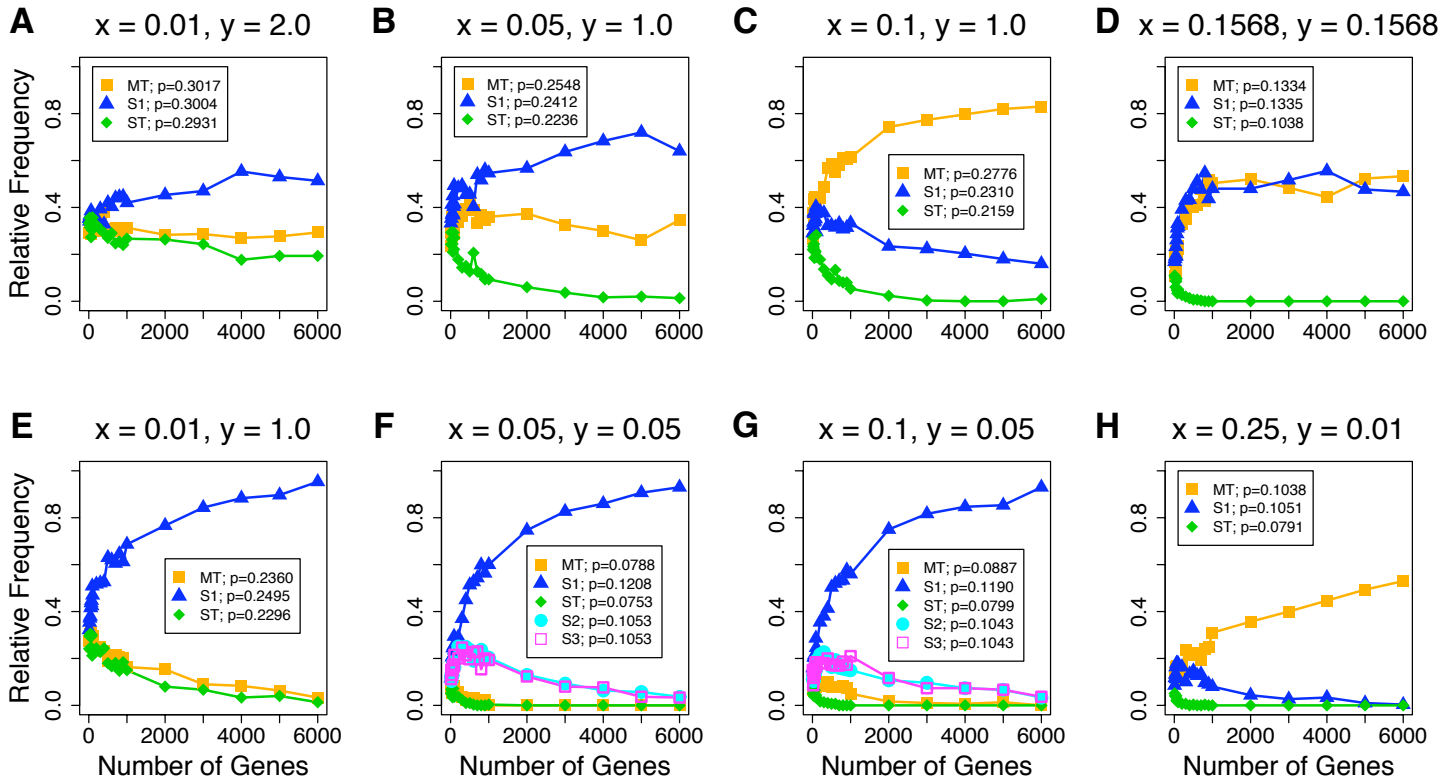


→ Generate gene trees in COAL → Generate sequence data in Seq-Gen → Estimate tree using concatenation

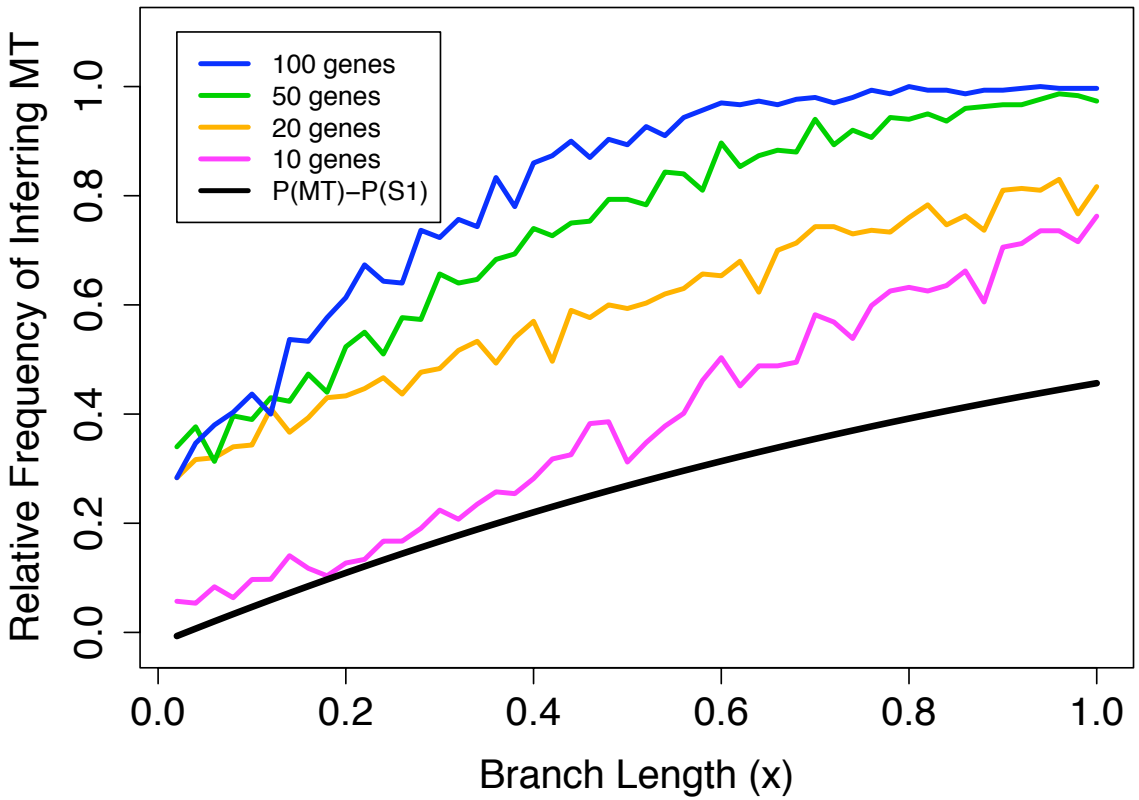
Repeat 100 times

Applications of the topology distribution - example 3

Simulation Study 1



Simulation Study 2



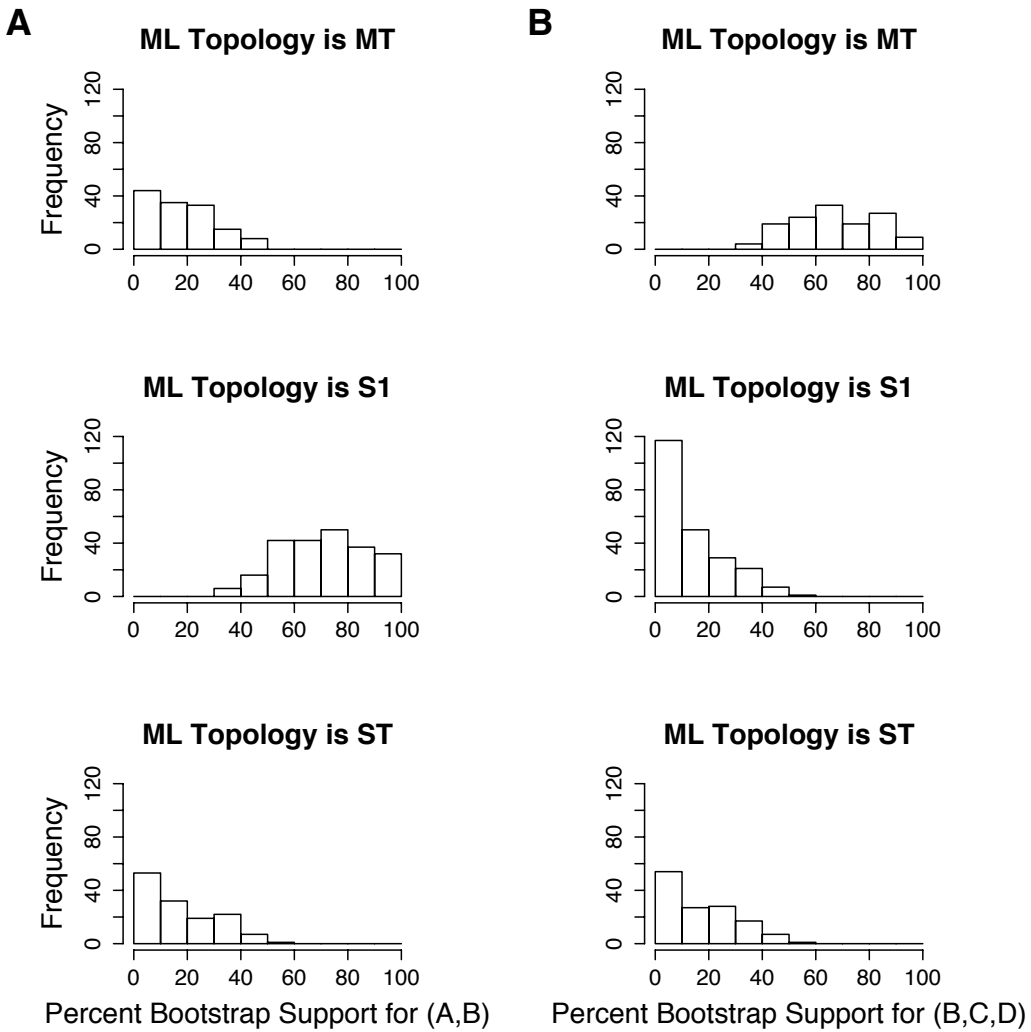
Applications of the topology distribution - example 3

- Performance of the Concatenation Approach:
  - Can be statistically inconsistent when branch lengths in the species phylogeny are sufficiently small
  - May perform poorly even when branch lengths are only moderately short
  - Bootstrap procedure can be positively misled in this situation
- Question: How does the bootstrap perform in these cases?

The concatenation approach – performance of the bootstrap

- Hypothesis: The bootstrap may provide strong support for the incorrect tree when gene trees that are incongruent with the species trees are fairly probably
- Simulation study to examine the performance of the bootstrap:
  - n=100 loci
  - x=0.01, y=1.0
  - θ=0.001
  - B=200 bootstrap samples per repetition
  - Repeated 500 times

The concatenation approach – performance of the bootstrap

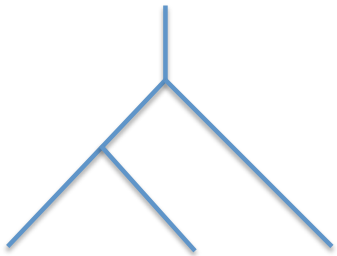


- The bootstrap can be positively misleading – show strong support for an incorrect clade
- Important note: This is NOT a failing of the bootstrap methodology; the observed "poor" performance is due to the use of an incorrect model (concatenation)
- Question: Is there a better way to estimate species phylogenies?

Explicitly model the coalescent process!

Model Underlying Coalescent-based Species Tree Inference

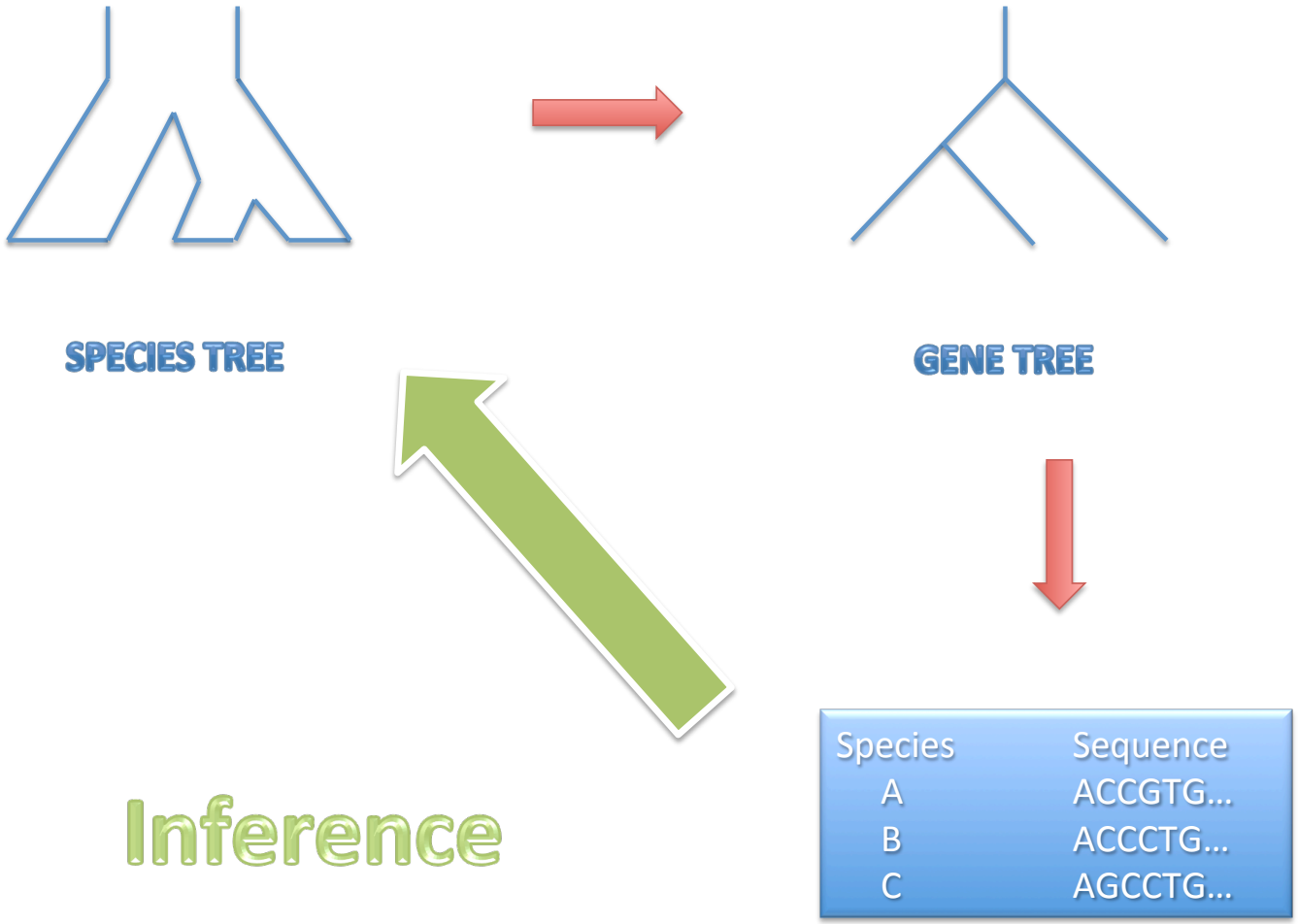
SPECIES TREE



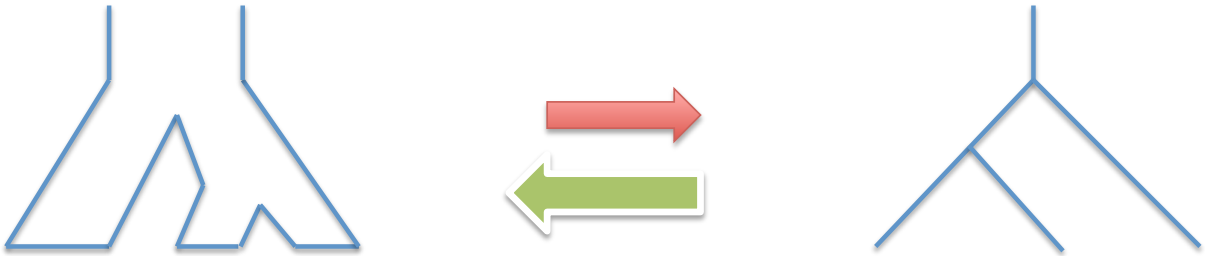
GENE TREE

Species	Sequence
A	ACCGTG…
B	ACCCTG…
C	AGCCTG…

Model Underlying Coalescent-based Species Tree Inference



Model Underlying Coalescent-based Species Tree Inference



SPECIES TREE

GENE TREE

Summary statistics methods

Species	Sequence
A	ACCGTG...
B	ACCCTG...
C	AGCCTG...

- 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)

- Methods that utilize the full data: Input is aligned sequences
  - BEST (Liu and Pearl 2007)
  - *BEAST (Heled and Drummond 2010)
  - SNAPP (Bryant et al. 2012)
  - SVDquartets (Chifman and Kubatko 2014)

- Comparison of approaches:
  - Summary statistics methods
    - ★ Advantage: Quick
    - ★ Disadvantage: Ignore information in the data
    - ★ Most current implementations do not easily allow assessment of uncertainty
  - Full data methods
    - ★ Advantage: Fully model-based framework
    - ★ Disadvantage: Computationally intensive, sometimes prohibitively so
    - ★ BEST, *BEAST, and SNAPP utilize a Bayesian framework and involve MCMC

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

Likelihood function

- 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 he inner sum (which is really an integral) [recall Peter's "galaxy slide"]

- To deal with this, either assume gene trees are known (summary statistics methods), use Bayesian techniques (full data approaches), or think about small problems ☹.

STEM: The gene tree-species tree likelihood function

- A simpler problem is to suppose that our data consist of a set of gene trees
- Let $g_1, t_2, \ldots, g_N$ be a set of N gene trees with branch lengths
- Consider a species tree, S (topology and branch lengths)
- The likelihood function is

$$L(S|D_1, D_2, \ldots, D_N) = \prod_{j=1}^{N} f(g_j|S)$$

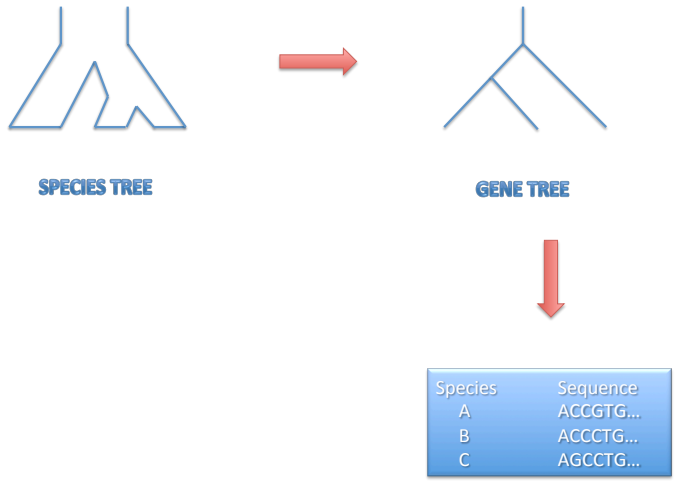
where $f(g|S)$ is given by Rannala and Yang (2003).

Maximum likelihood estimate of the species tree

- Liu et al. (2009) showed that the ML estimate of the species tree can be computed by sequentially clustering minimum observed divergence times between pairs of species across genes.
- They have shown that when gene trees are known without error, the ML species tree is a consistent estimator.
- A similar result was obtained by Roch & Mossel (2010) – they call their estimator the GLASS tree (an acronym for Global LAteSt Split, based on the algorithm they developed to compute it).
- STEM computes the ML estimate of the species tree this way.
- Note the important and undesirable assumption of STEM (and other summary statistics methods) that the gene trees are known without error!

Full data methods: BEST, *BEAST, SNAPP

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



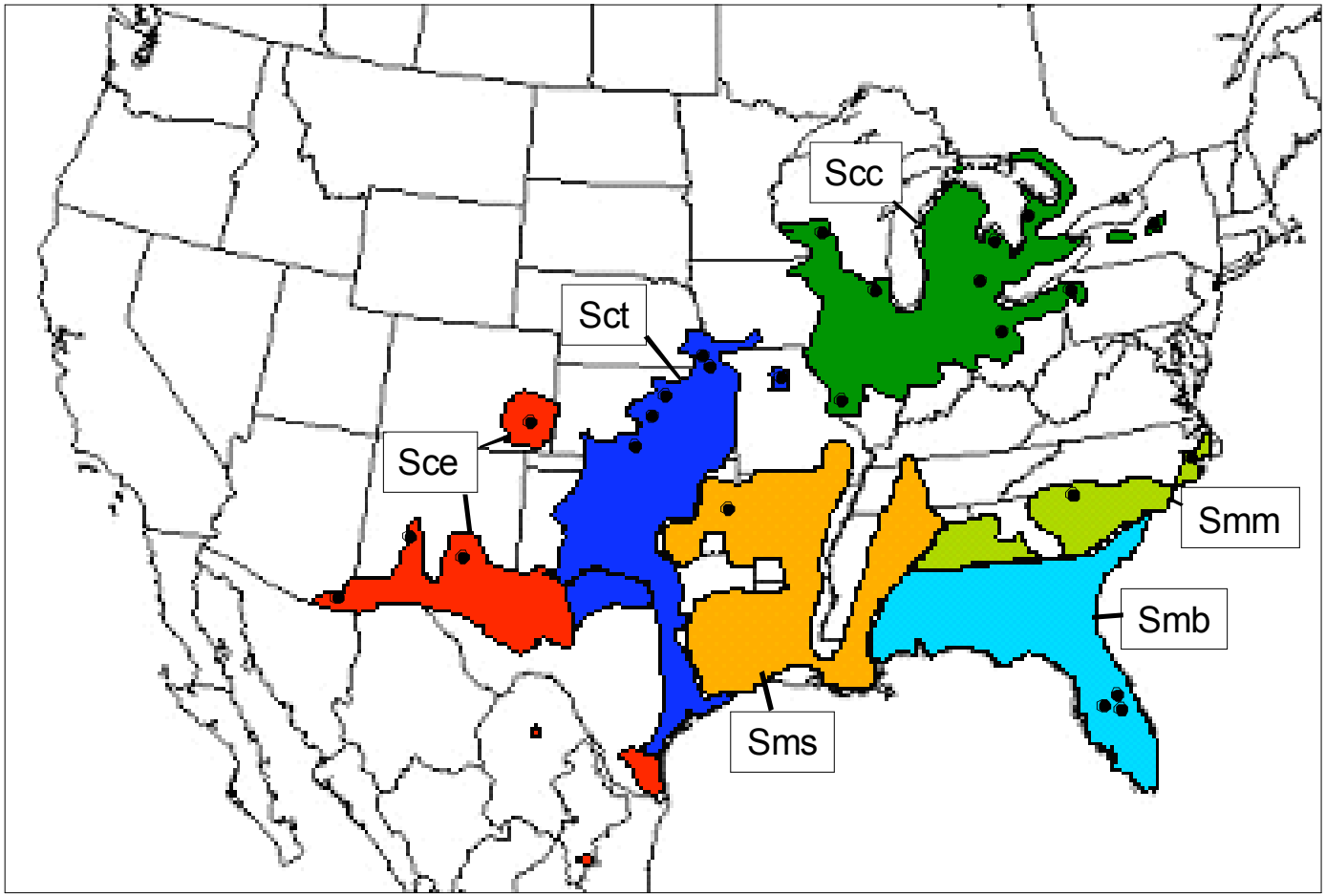
- BEST and *BEAST use MCMC by considering both gene trees and the species tree, but their implementations are different
- SNAPP uses a clever two-step peeling algorithm to carry out the integration over gene trees, allowing it to consider a reduced space – but currently limited to biallelic data.

An Empirical Example: 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

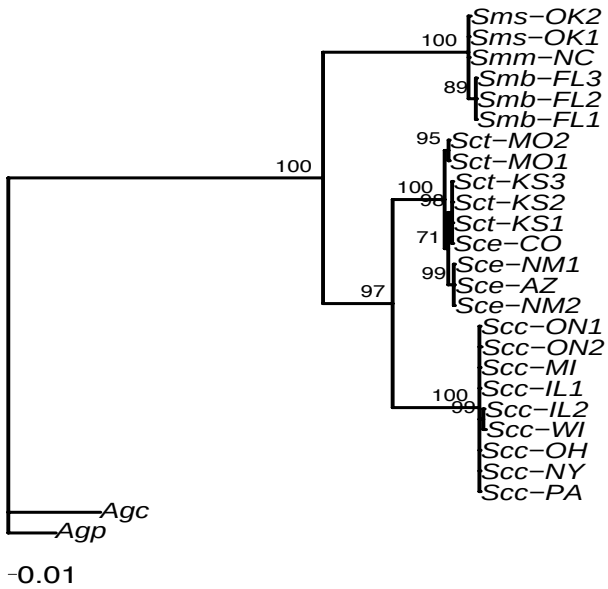


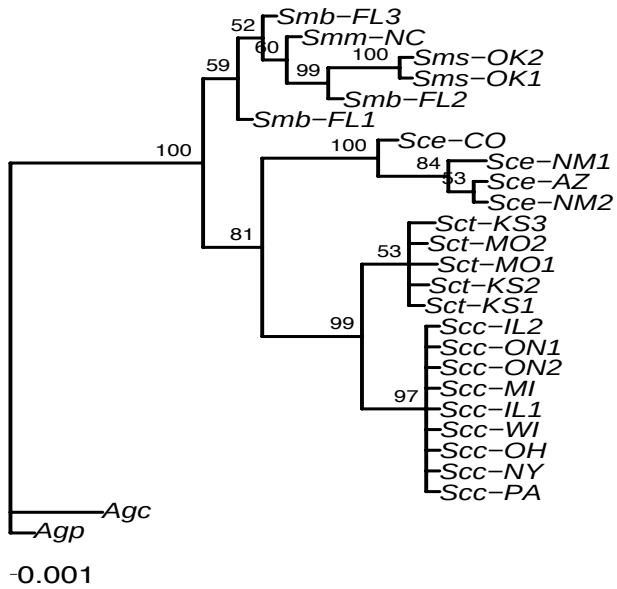
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

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

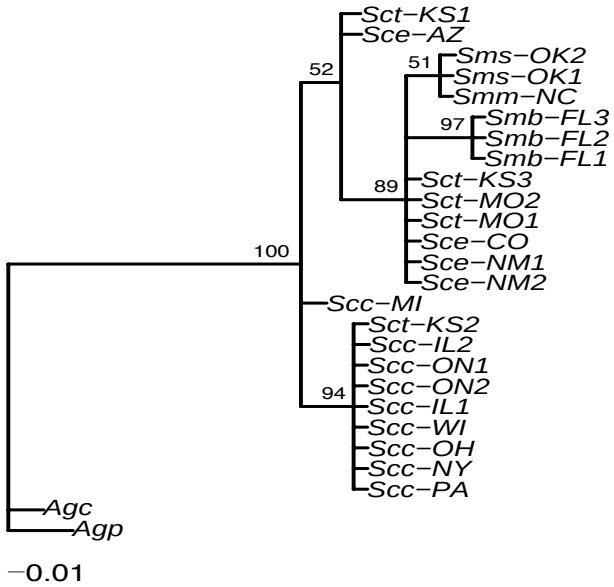
Some are very informative:

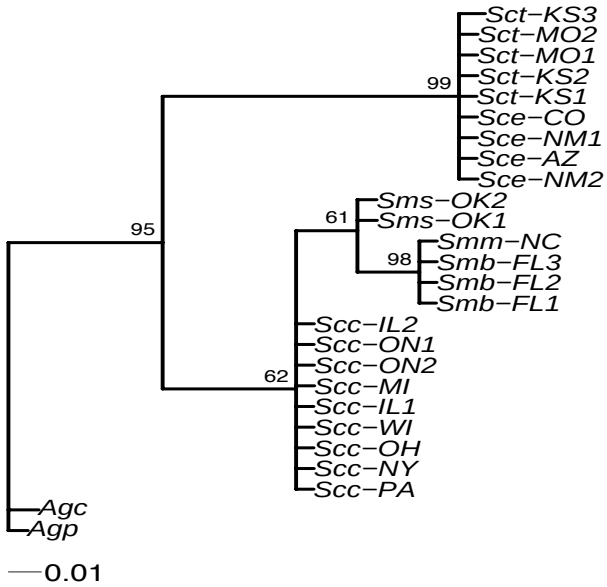




Individual Gene Tree Estimates

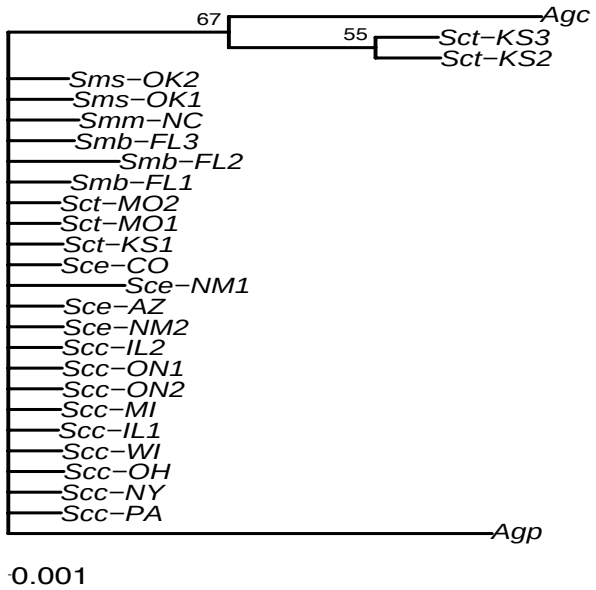
Some are a little informative:

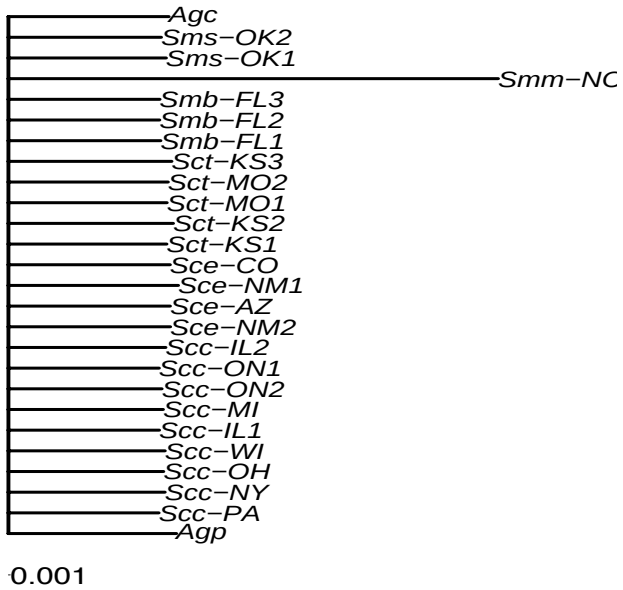




Individual Gene Tree Estimates

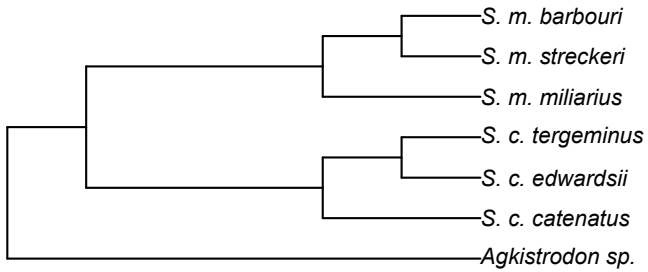
And then there are others .....



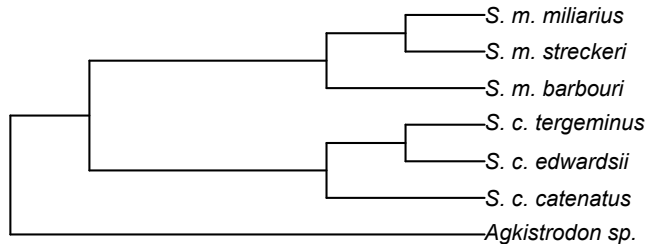


Example: Sistrurus rattlesnakes

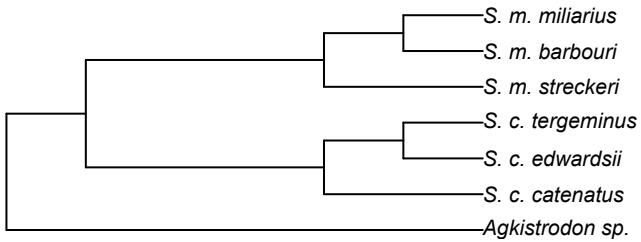
STEM, STEAC



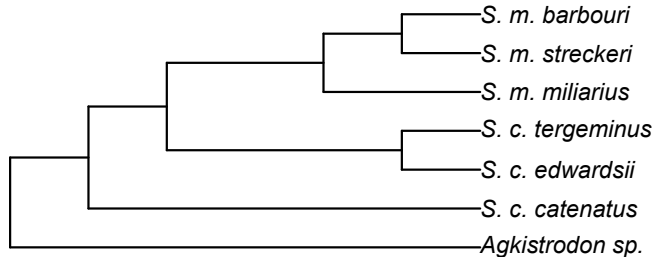
BEAST (concatenated data), *BEAST



BEST, Parsimony & MrBayes (concatenated data)



PhyloNet, STAR



Example: Sistrurus rattlesnakes

- Some observations:
  - Estimate from PhyloNet places S. c. catenatus as sister to the entire clade – it turns out this is due to only two gene trees. If those genes are removed, the estimate agrees with STEM.
  - The portion of the tree that differs between STEM, *BEAST, and BEST is the arrangement of the S. miliarius subspecies – all three arrangements are observed.
  - Both BEST and *BEAST have trouble converging: BEST did not converge in the branch length parameters, while *BEAST did not converge in the effective population size parameters, especially for the tip species (same problem?).
  - *BEAST was much faster than BEST (days vs. months for $\sim$ 350 million iterations) – but with an older version of BEST.

Goal of this work:

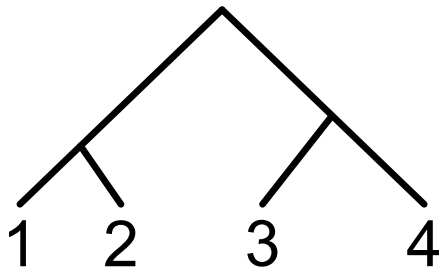
Develop a full data approach that is computationally feasible for large-scale data

Goal of this work:

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

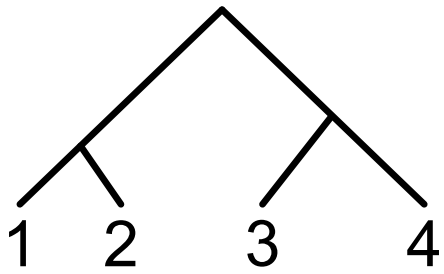
How?

- 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



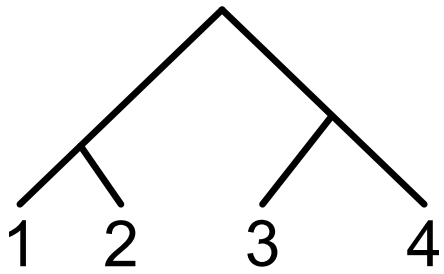
Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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



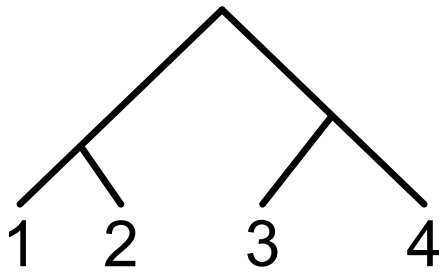
Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

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



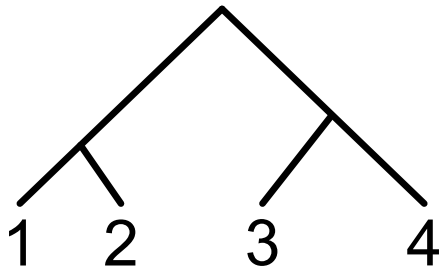
Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

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



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

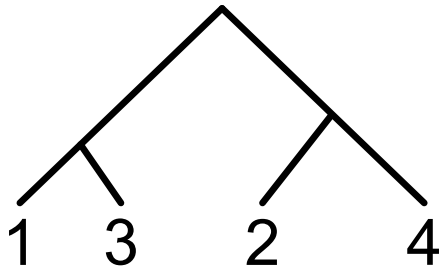
$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\ [AG] & p_{AGAA} &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

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

What about the incorrect tree?



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

$$\mathbf{Flat}_{13|24}(\mathbf{P}) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\ [AG] & p_{AGAA} & p_{\mathbf{AGAC}} & p_{AGAG} & p_{AGAT} & p_{\mathbf{AGCA}} & \cdots \\ [AT] & p_{ATAA} & p_{\mathbf{ATAC}} & p_{ATAG} & p_{ATAT} & p_{\mathbf{ATCA}} & \cdots \\ [CA] & p_{CAAA} & p_{\mathbf{CAAC}} & p_{CAAG} & \mathbf{2} & p_{\mathbf{CACA}} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

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

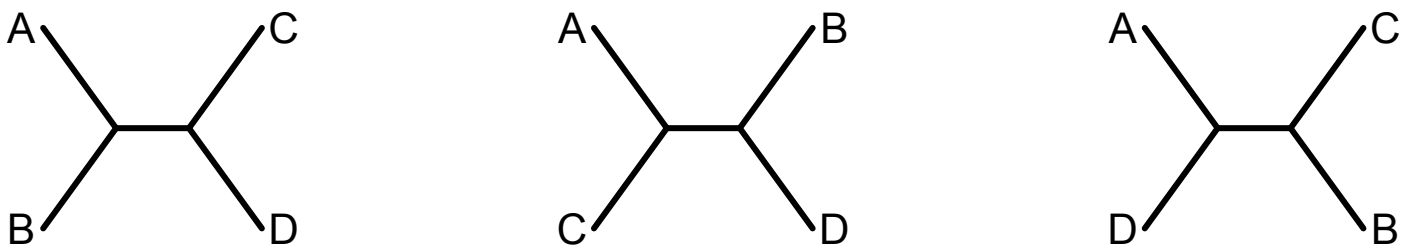
- **Basic idea:**
  - Data: aligned DNA sequences for multiple loci or for a collection of SNPs
  - Construct the flattening matrix
  - Compute some measure of how close the observed flattening matrix is to a matrix with rank 10

    We use singular value decomposition (SVD) of the flattening matrix – define the SVD score for a split $A|B$ to be

$$SVDscore(Flat_{A|B}(\hat{P})) = \sqrt{\sum_{i=11}^{16} \sigma_i^2}$$

    where σ_i^2 is the i^{th} singular value of the matrix $Flat_{A|B}(\hat{P})$.
  - Pick tree relationships that give the best value of the measure in the previous step

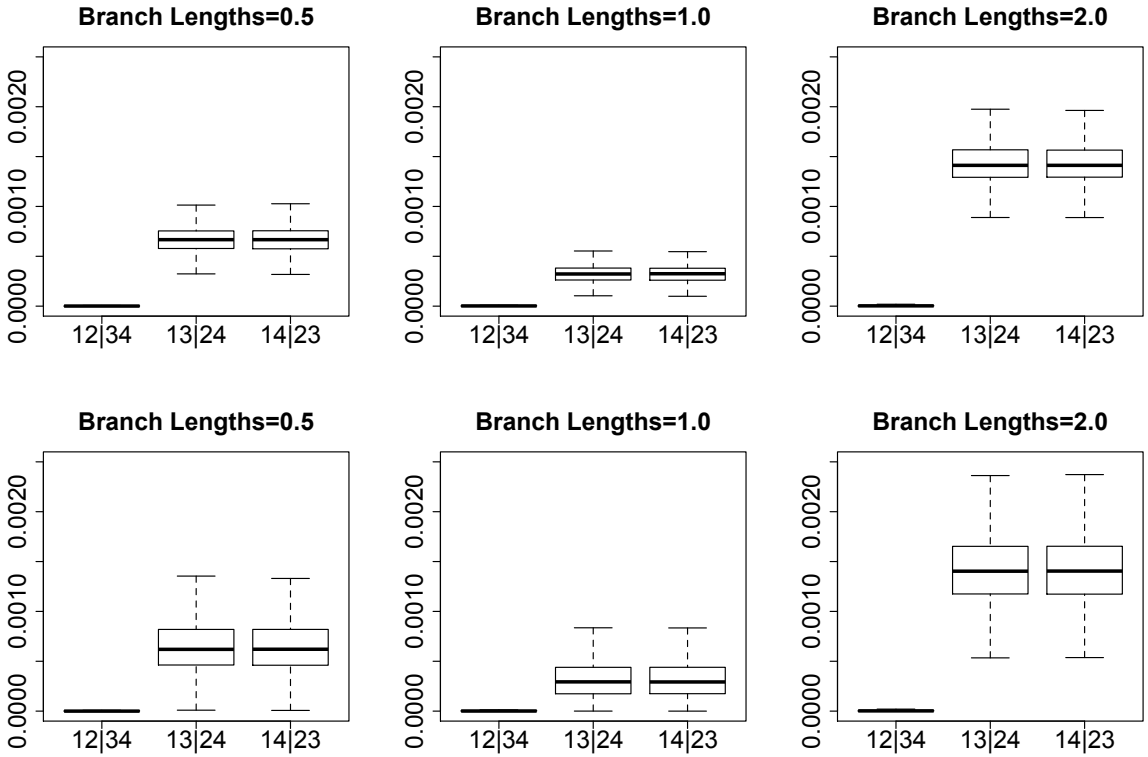
Main idea: use the observed site pattern distribution to provide information about which of the three possible splits for a set of four taxa is the true split.



The program SVDquartets computes a score for each split in a given quartet of taxa and chooses the split with the best (lowest) score.

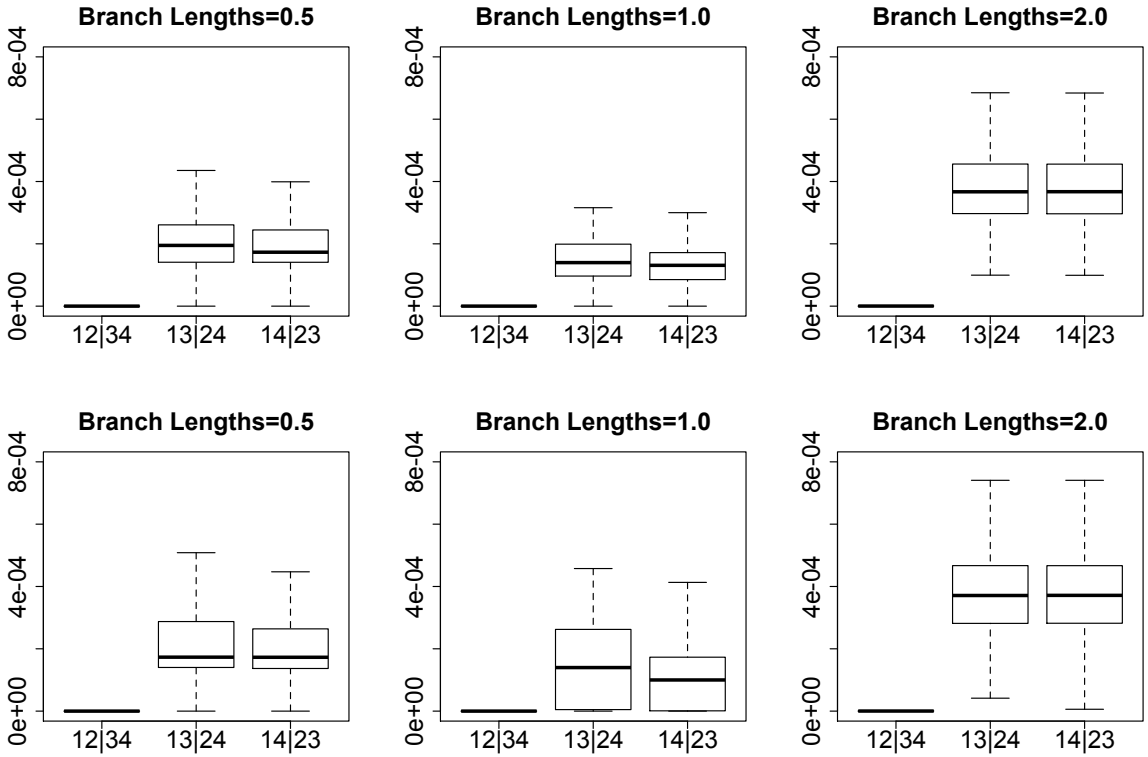
Simulation study 1 – can we detect the correct split?

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



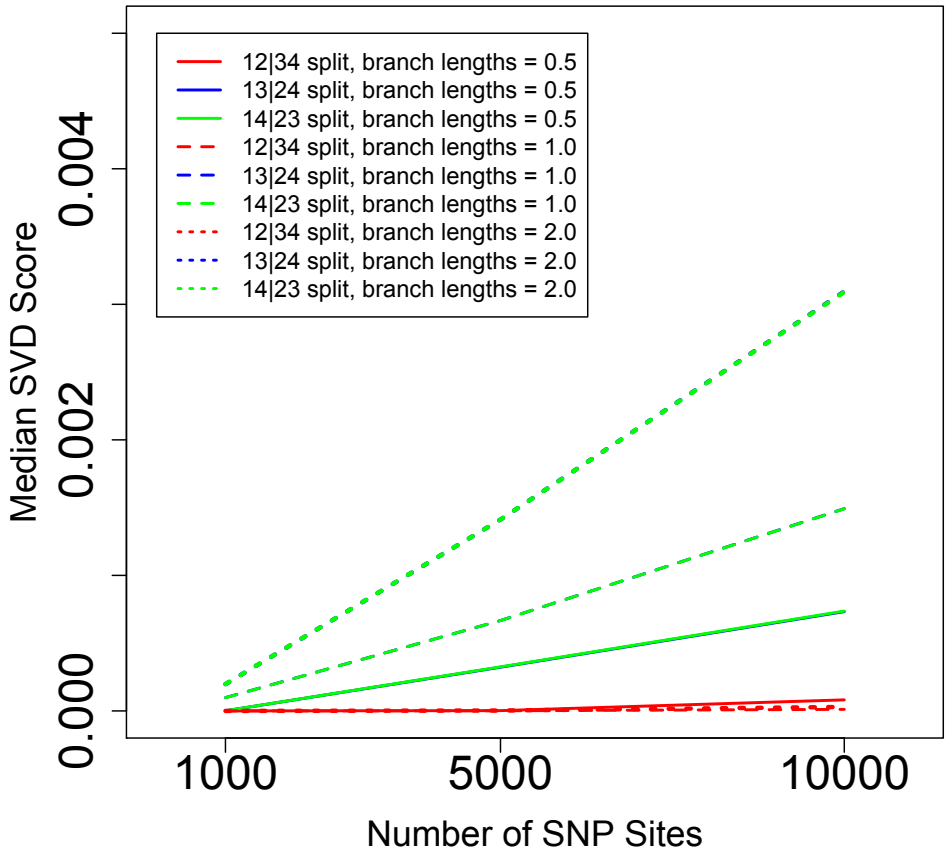
Simulation study 1 – can we detect the correct split?

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

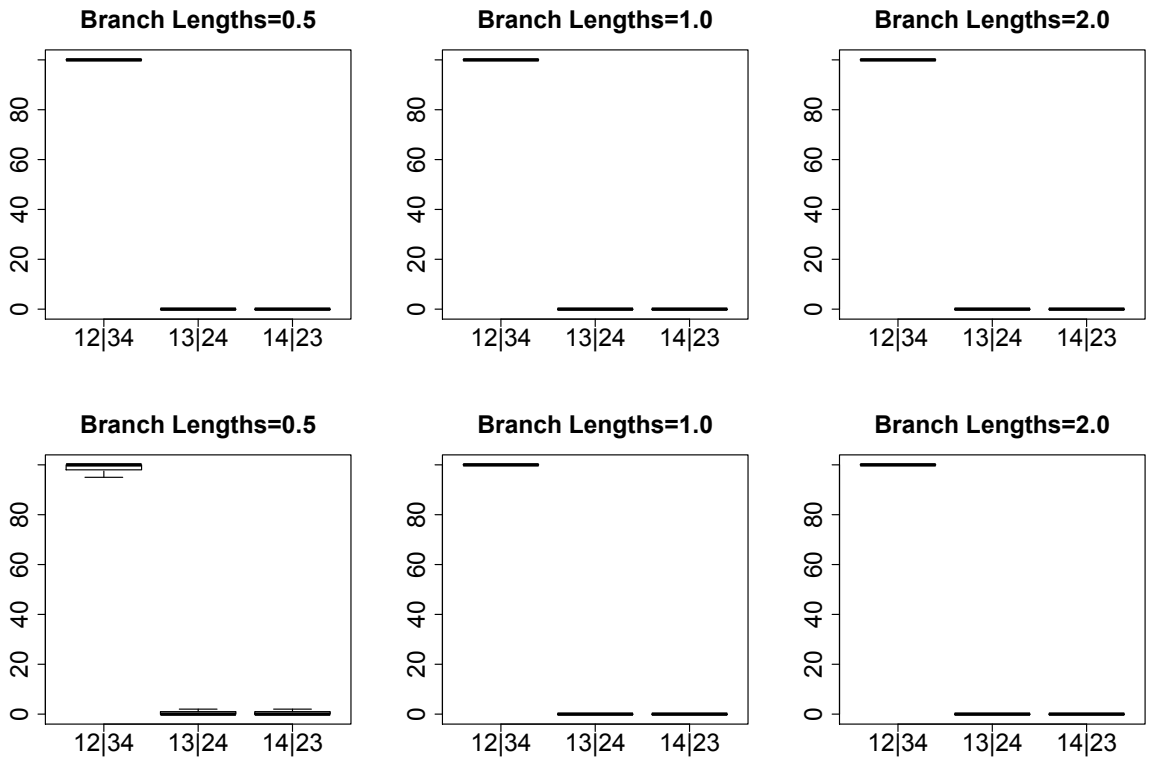


How do we assess variability?

- 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

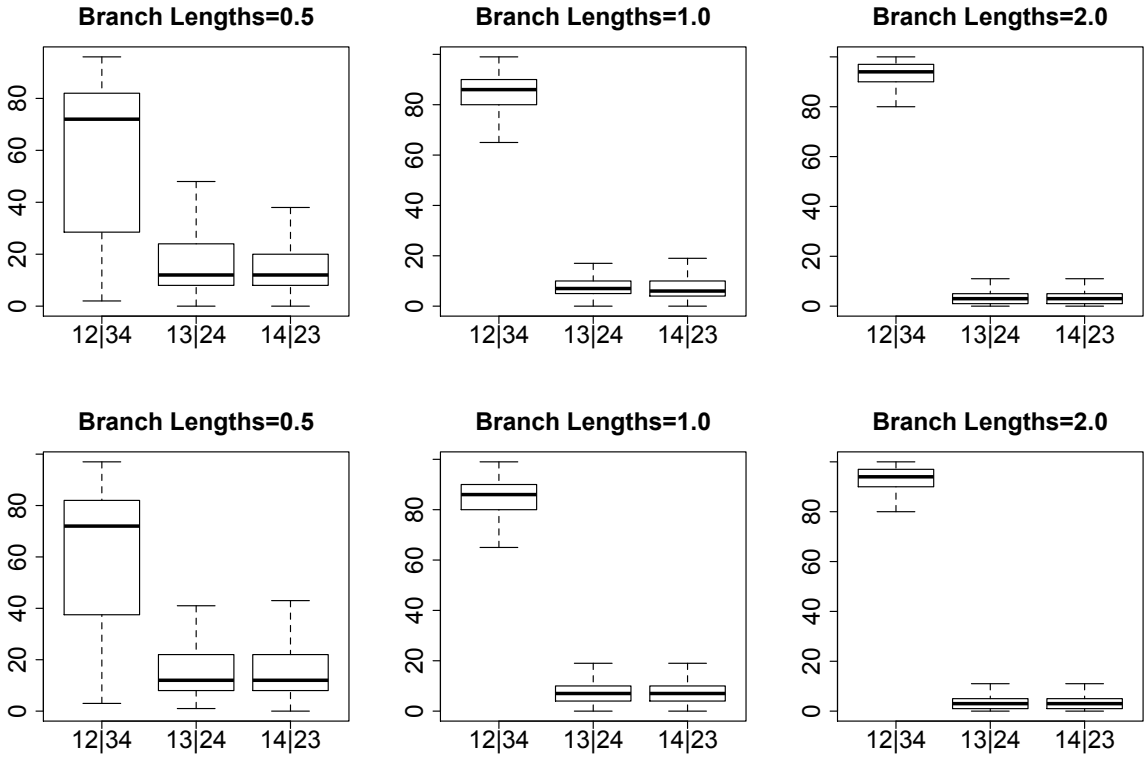
Assessing support using the bootstrap

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



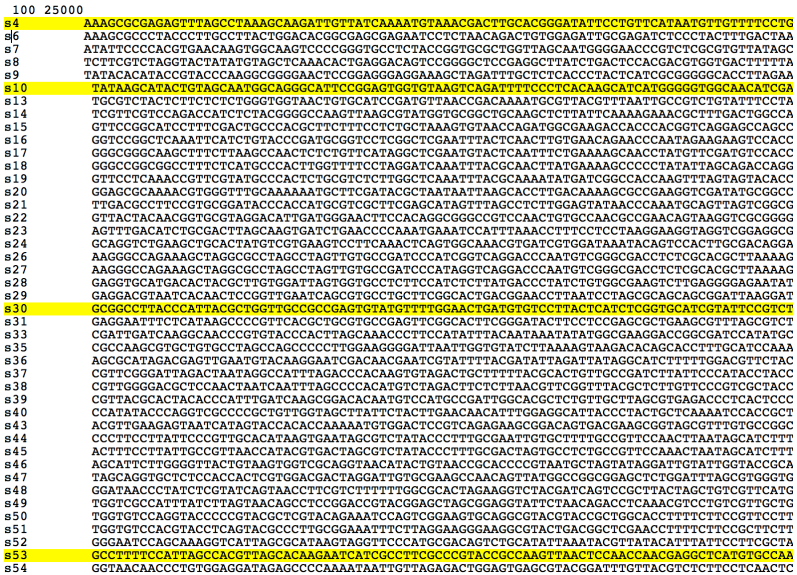
Assessing support using the bootstrap

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



Algorithm

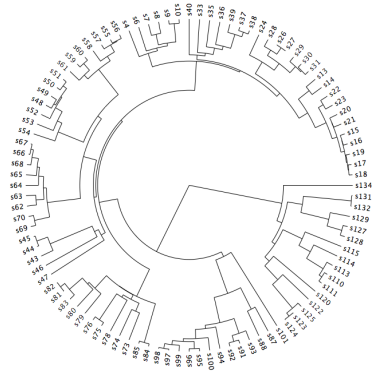
1. Generate all quartets (small problems) or sample quartets (large problems)
2. Estimate the correct quartet relationship for each sampled quartet
3. Use a quartet assembly method to build the tree



$\rightarrow$

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

$\rightarrow$

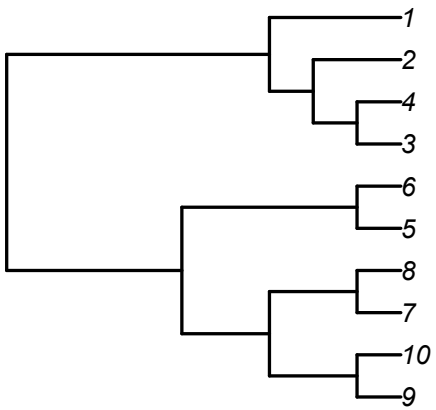


Extension to larger trees

- 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)

Simulation study 2 – larger trees

average RF distance (range 0 - 14)

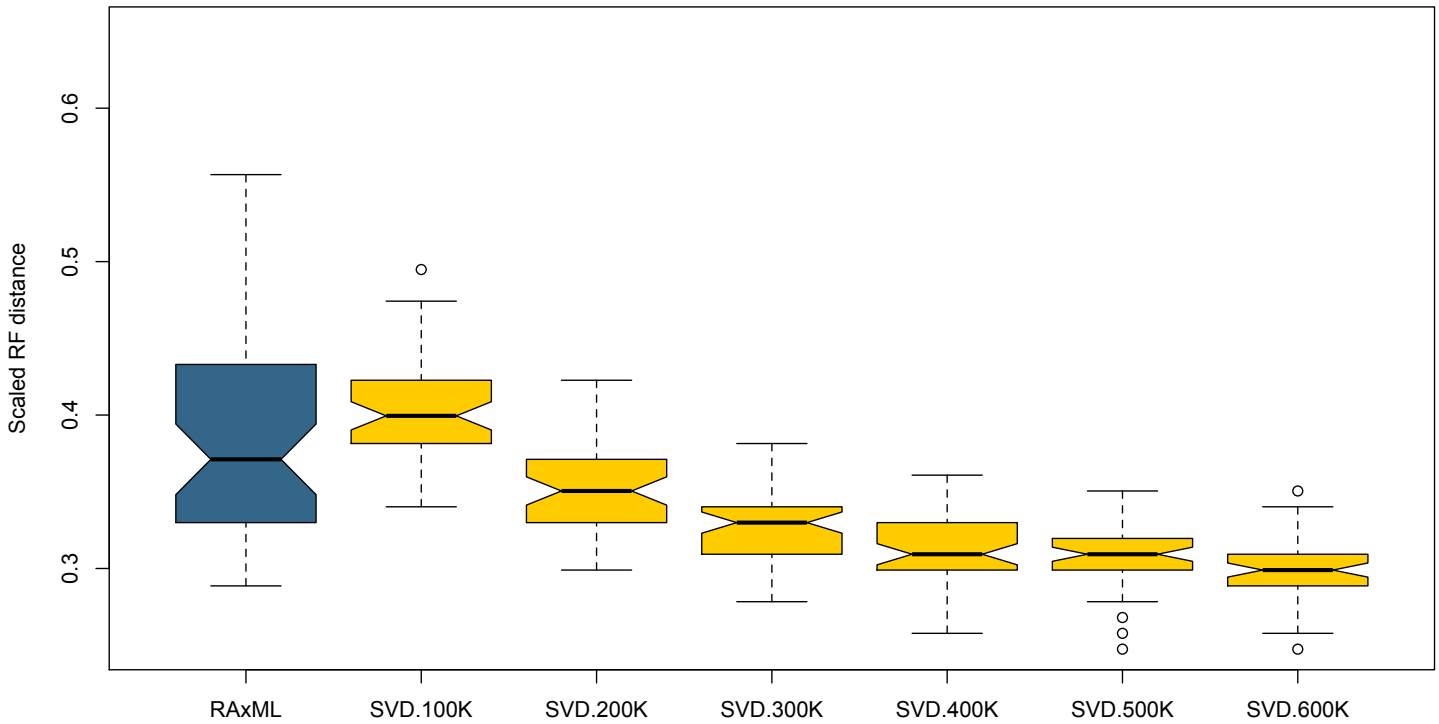


black = 500 bp / gene
red = 2,000 bp / gene
blue = No. genes × 500 SNPs

	10 genes	20 genes	50 genes	100 genes
Short (0.5)	4.51 3.31 3.48	3.55 1.94 1.74	1.04 0 0.32	0.2 0 0.16
Medium (1.0)	1.59 0.80 0.76	0.56 0.16 0.14	0 0 0.16	0 0 0.32
Long (2.0)	0.34 0.04 0.18	0.04 0 0.04	0 0 0	0 0 0

Simulation study 3 – very large trees

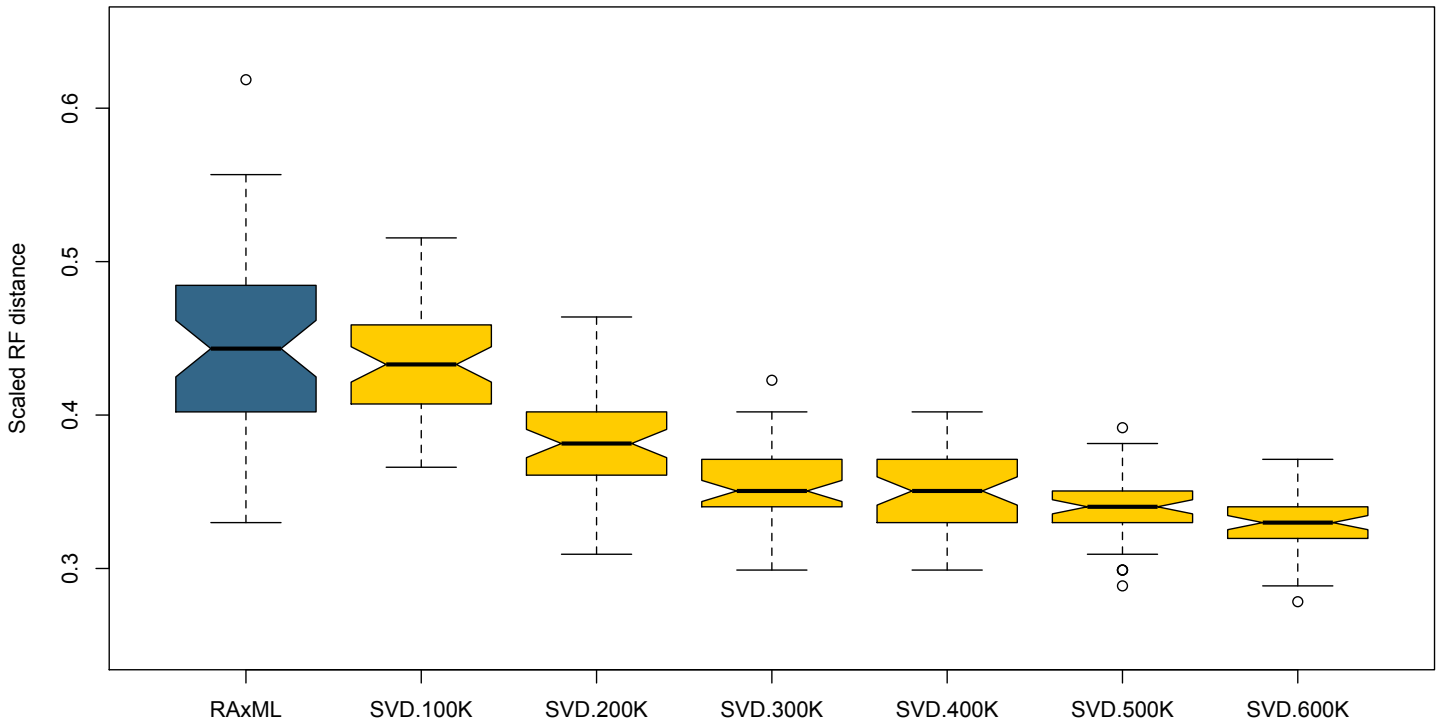
- 100-taxon species tree, 100 loci, 500bp per locus, $\theta = 0.01$
- Look at the effect of number of quartets sampled
- Compare to concatenation



Simulations by Paul Blischak

Simulation study 3 – very large trees

- 100-taxon species tree, 50 loci, 500bp per locus, $\theta = 0.01$
- Look at the effect of number of quartets sampled
- Compare to concatenation



Simulations by Paul Blischak

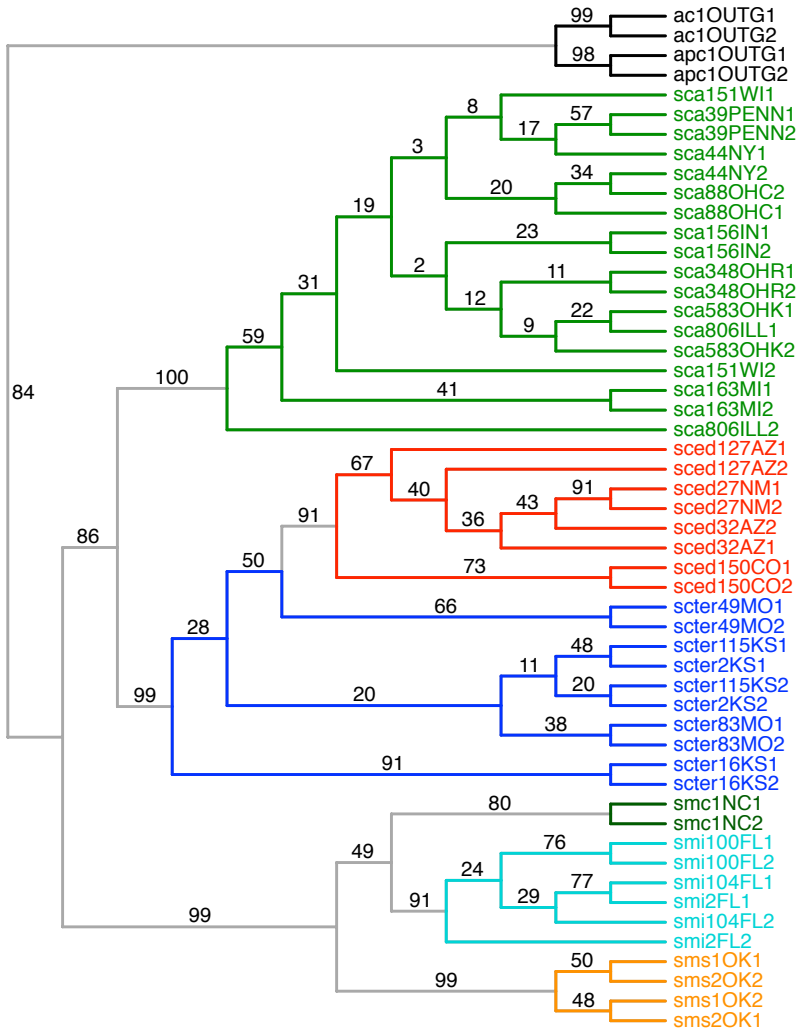
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

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

Using 20,000 quartets and 100 bootstrap replicates

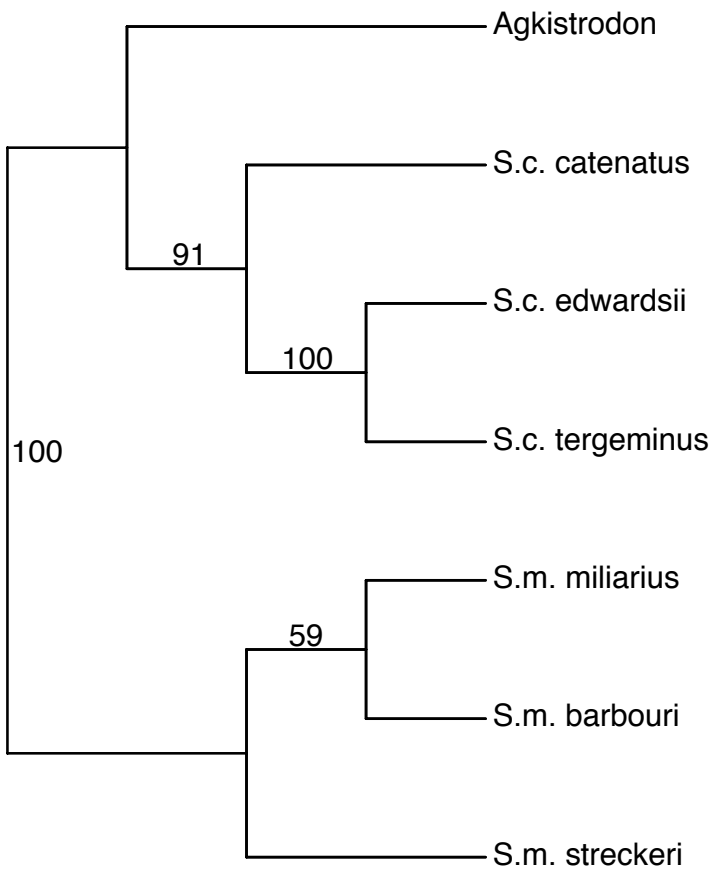
$\sim$ 10 minutes



Empirical example: Sistrurus rattlesnakes

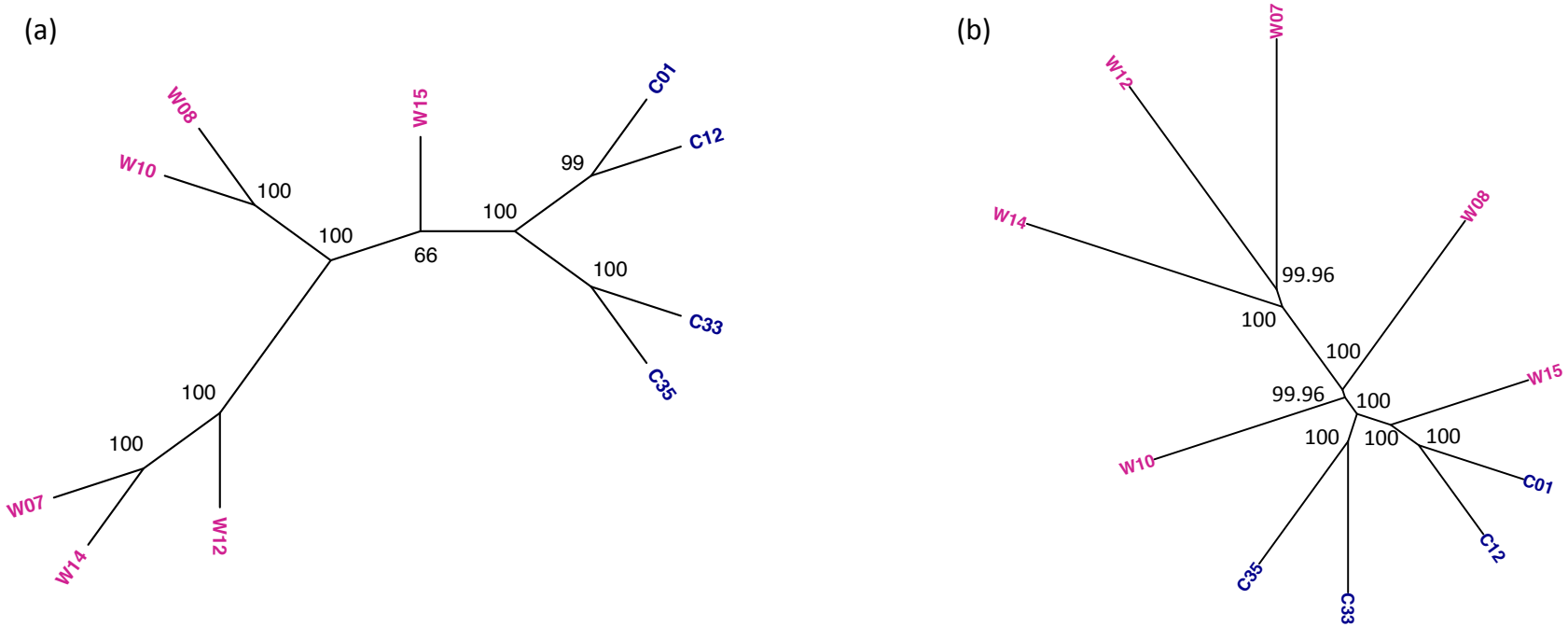
Using 20,000 quartets and 100 bootstrap replicates

$\sim$ 10 minutes



Empirical example 2: soybeans

- 10 soybeans species, 1,027,026 SNPs
- SVDquartets, 20,000 quartets, < 24 hours
- SNAPP, 28 days on 1 processor, 2.23 million iterations



SVDquartets Summary

- Advantages:
  - Quick! And scales well to large taxon sets and next-gen sequencing data
  - Easily parallelized
  - Intuitive method for handling missing data
  - Potential for application to other data types (codons, amino acids, etc.)
- Disadvantages:
  - Gives only an estimate of the unrooted topology

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. – get involved and have fun!