Species Tree Estimation

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

kubatko.2@osu.edu
twitter: Laura_Kubatko

July 23, 2017

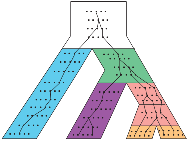
Relationship between population genetics and phylogenetics

- 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]



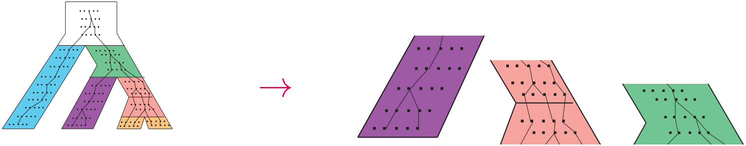
Recall several facts from Peter's lecture

- 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".

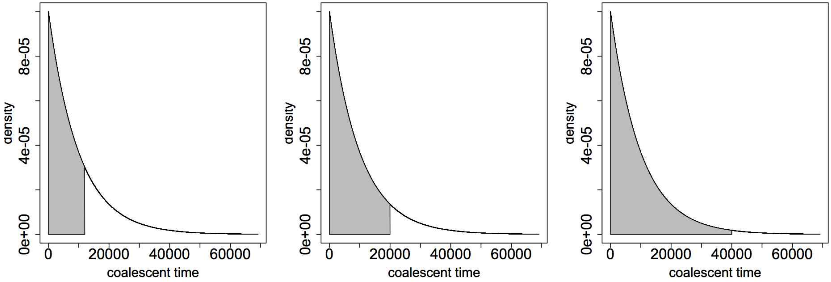
Fitting population trees into a phylogeny



- 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

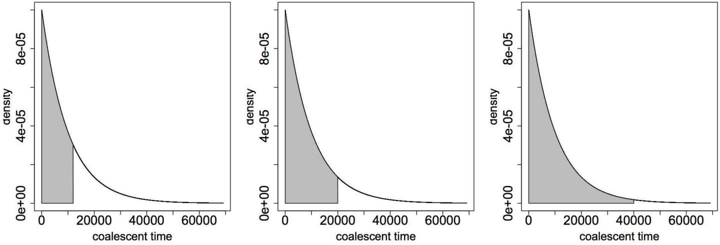


Fitting population trees into a phylogeny

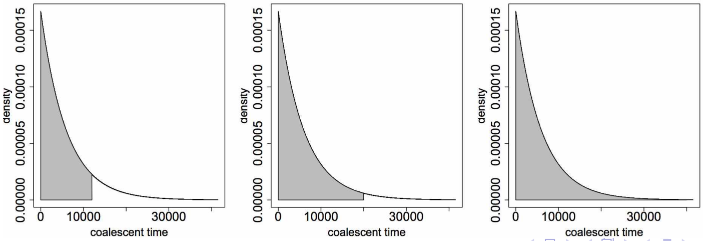
- What happens if we change the population size, N?
- Recall that we have an exponential distribution with rate $\frac{1}{2N}$ and mean $2N$.
- Now suppose $N = 3,000$ and look at the same speciation interval lengths.

Fitting population trees into a phylogeny

- $N = 5,000$

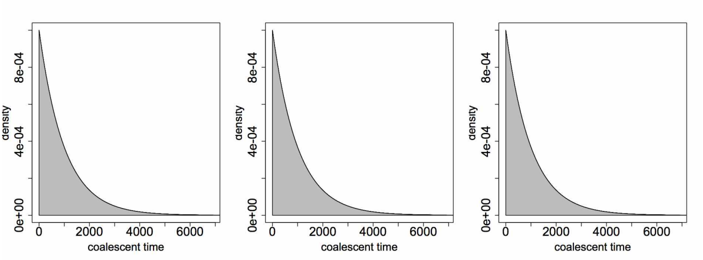


- $N = 3,000$



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 $2N$)

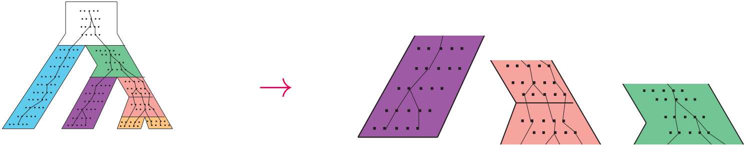


Fitting population trees into a phylogeny

- 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)

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

Fitting population trees into a phylogeny

- 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]

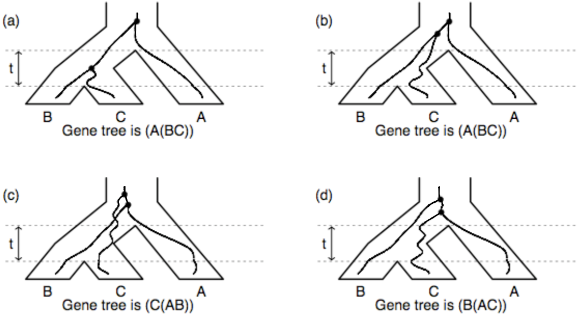
- Similarly,

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

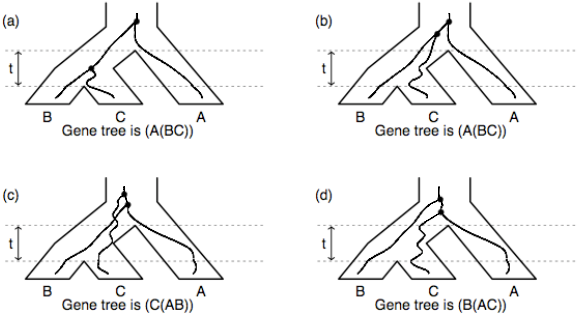
Putting it together . . . the coalescent model along a species tree

- 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.

Phylogenetic coalescent model

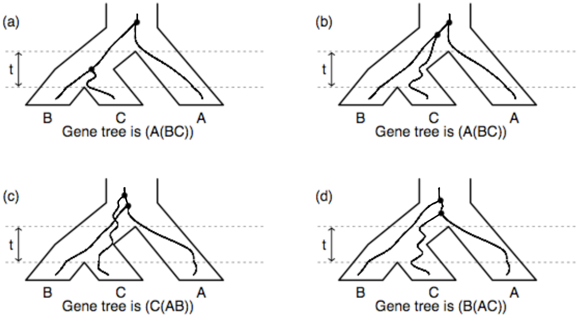


Phylogenetic coalescent model



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

Phylogenetic coalescent model



Gene tree is (A(BC))

Gene tree is (A(BC))

Gene tree is (C(AB))

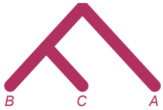
Gene tree is (B(AC))

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

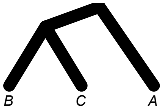
Example: 1.2 coalescent units for an organisms with population size $N = 10,000$ and a generation time of 3 years $= 1.2 \times 20,000 \times 3 = 72,000$ years

Phylogenetic coalescent model

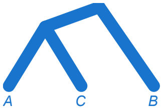
Probabilities of each gene tree history are shown below them
t = length of interval between speciation events



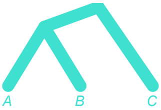
$1 - e^{-t}$



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



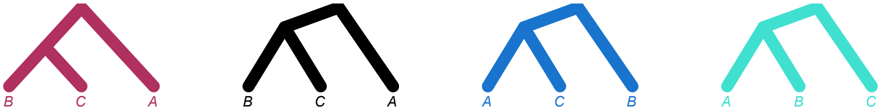
$\frac{1}{3}e^{-t}$



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

Phylogenetic coalescent model

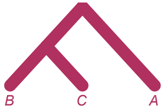
t = length of interval between coalescent events $= 1.0$

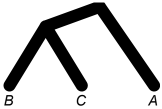


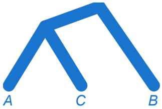
$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12

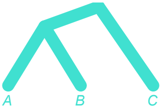
Phylogenetic coalescent model

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





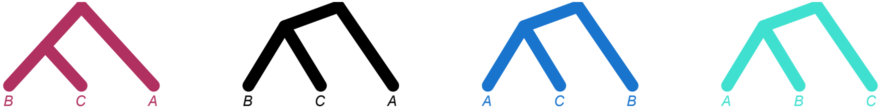




$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20

Phylogenetic coalescent model

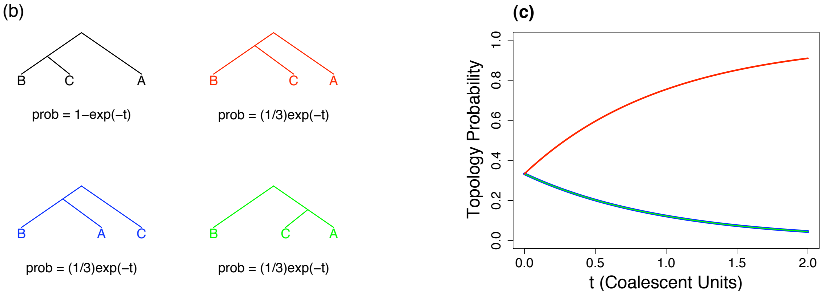
$t = \text{length of interval between coalescent events} = 1.0 = 0.5 = 2.0$



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

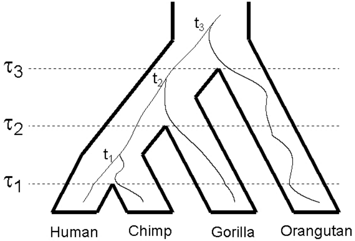
Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

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



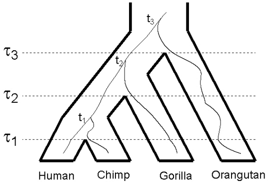
Example: a slightly larger case

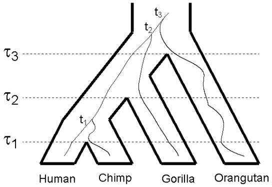
- Consider 4 taxa – the human-chimp-gorilla problem

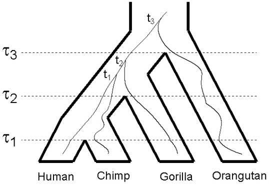


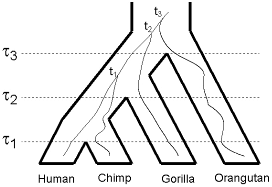
Coalescent histories for the 4-taxon example

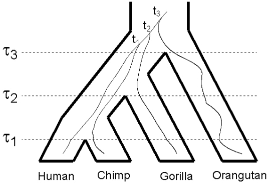
- There are 5 possible histories for this example:











Enumerating Histories

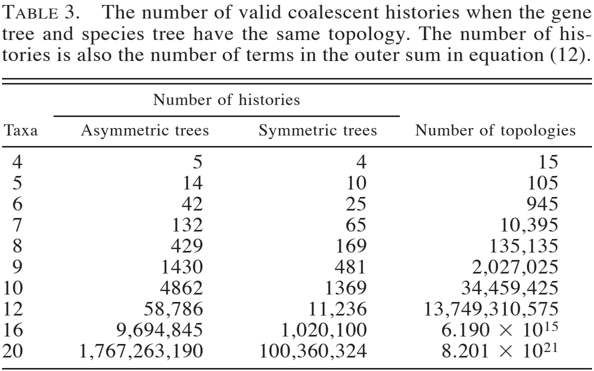


TABLE 3. The number of valid coalescent histories when the gene tree and species tree have the same topology. The number of histories is also the number of terms in the outer sum in equation (12).

	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

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

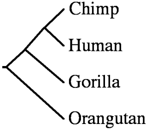
  $$P\{G = g|\mathcal{S}\} = \sum_{histories} P\{G = g, history|\mathcal{S}\}$$

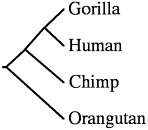
- Implemented in the software COAL (Degnan and Salter, *Evolution*, 2005)
- A more efficient method has been proposed (Wu, *Evolution*, 2012)

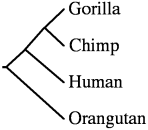
Applications of the topology distribution - example 1

- 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.

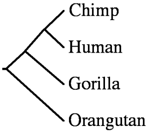
Applications of the topology distribution - example 1



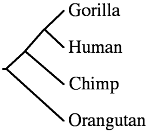




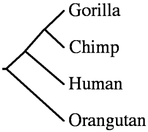
Applications of the topology distribution - example 1



76.6%



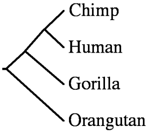
11.4%



11.5%

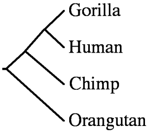
Observed proportions of each gene tree among ML phylogenies

Applications of the topology distribution - example 1



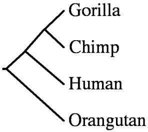
76.6%

79.1%



11.4%

9.9%



11.5%

9.9%

Observed proportions of each gene tree among ML phylogenies

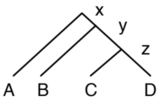
Predicted proportions using parameters from Rannala & Yang, 2003.

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

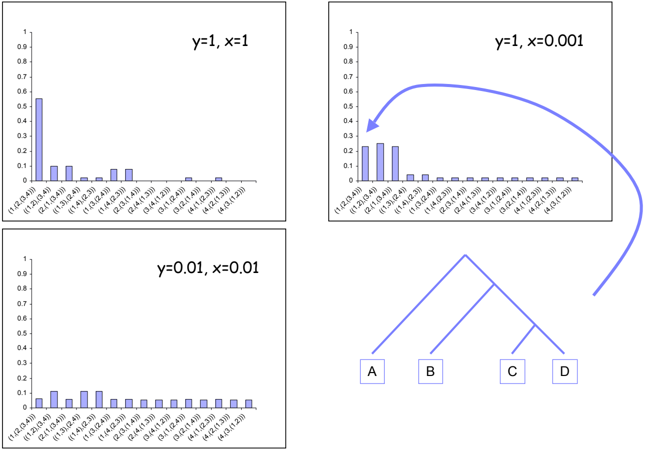
Applications of the topology distribution - example 2

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

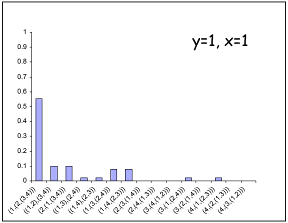


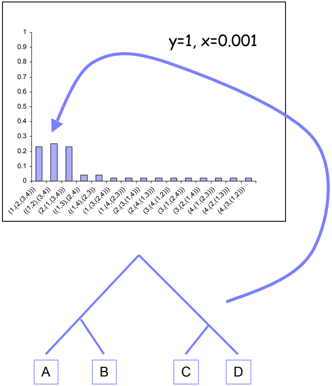
- Look at probabilities of all 15 tree topologies for values of x, y, and z

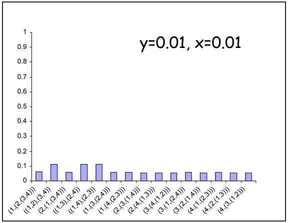
Applications of the topology distribution - example 2



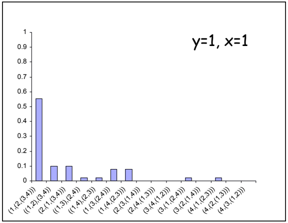
Applications of the topology distribution - example 2

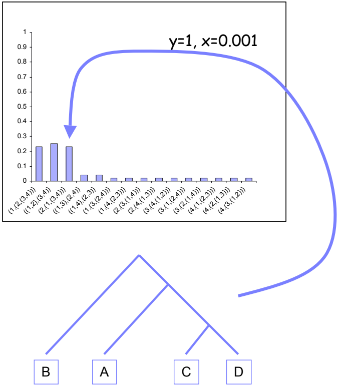


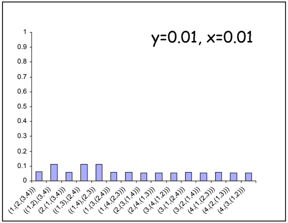




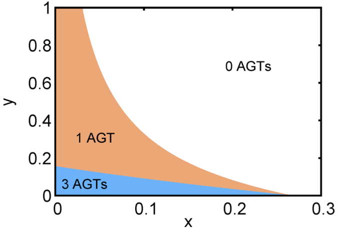
Applications of the topology distribution - example 2







Applications of the topology distribution - example 2



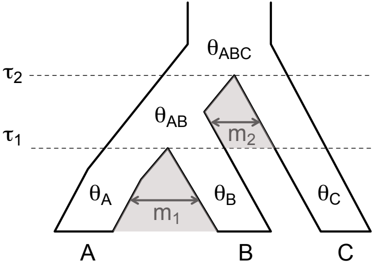
Degnan and Rosenberg, *PLoS Genetics*, 2006

Rosenberg and Tao, *Systematic Biology*, 2008

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

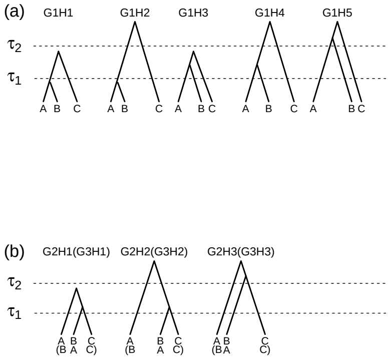
What about gene flow?

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



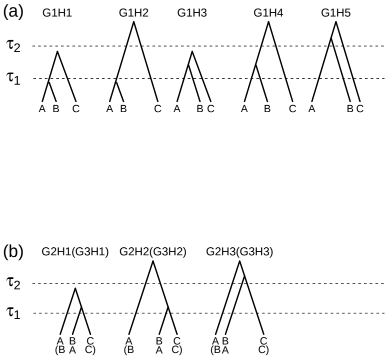
What about gene flow?

Complication: More histories are possible, because coalescent events can happen “before” speciation



What about gene flow?

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

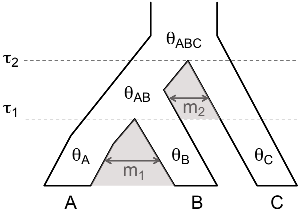


Anomalous gene trees are possible, even for three taxa!

Anomalous three-taxon gene trees in the presence of gene flow

Recent results (Colby Long):

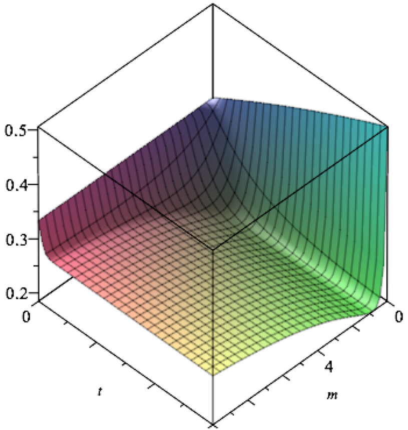
- When $\theta_{AB} = \theta_C$, there are no anomalous gene trees
- When $\theta_{AB} \neq \theta_C$ and $m_2 > 0$, anomalous gene trees are possible – the probability of the gene tree matching the species tree could be as low as $\frac{1}{9}$ (leaving probability $\frac{4}{9}$ for each of the other two gene trees)



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

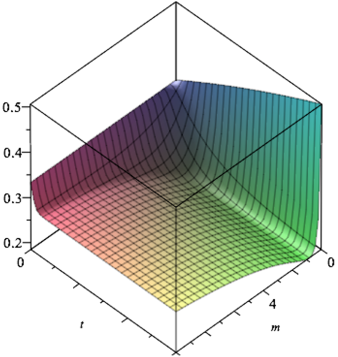
Anomalous three-taxon gene trees in the presence of gene flow

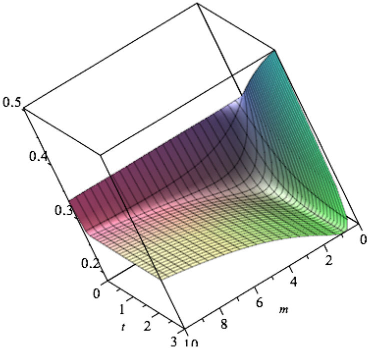
The anomaly zone is much more complicated in this case ....



Anomalous three-taxon gene trees in the presence of gene flow

The anomaly zone is much more complicated in this case ....



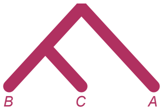


What about mutation?

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

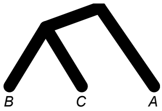
Phylogenetic coalescent model with mutation

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



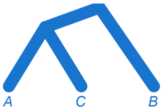
$1 - e^{-t}$

0.63



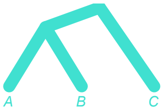
$\frac{1}{3}e^{-t}$

0.12



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

0.12



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

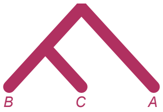
0.12

Phylogenetic coalescent model with mutation

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

Phylogenetic coalescent model with mutation

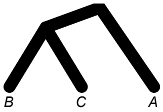
Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}



$1 - e^{-t}$

0.63

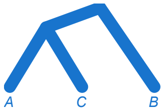
$p^{1a}_{TGT} = 0.1$



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

0.12

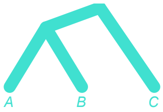
$p^{1b}_{TGT} = 0.025$



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

0.12

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



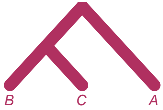
$\frac{1}{3}e^{-t}$

0.12

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

Phylogenetic coalescent model with mutation

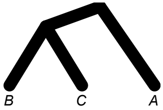
Example: Want to compute the probability that taxon A has nucleotide T, taxon B has nucleotide G and taxon C has nucleotide T – call this p_{TGT}



$1 - e^{-t}$

0.63

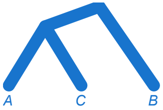
$p^{1a}_{TGT} = 0.05$



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

0.12

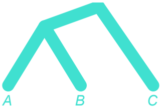
$p^{1b}_{TGT} = 0.025$



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

0.12

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



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

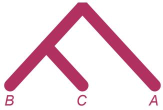
0.12

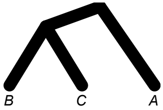
$p^{3}_{TGT} = 0.025$

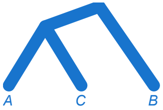
$$p_{TGT} = 0.63 \times 0.05 + 0.12 \times 0.025 + 0.12 \times 0.2 + 0.12 \times 0.025 = 0.0615$$

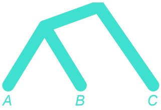
Phylogenetic coalescent model with mutation

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









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

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

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

Given this model, how should inference be carried out?

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

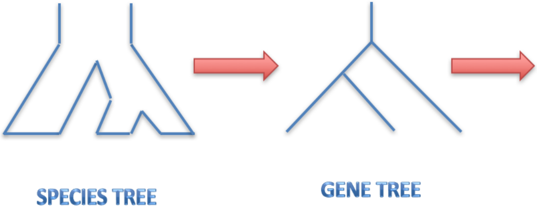
  Important note: This is NOT a failing of the bootstrap methodology; the observed “poor” performance is due to the use of an incorrect model (concatenation)

Kubatko and Degnan, 2007; Roch and Steel, 2015

Is there a better way to estimate species phylogenies?

Explicitly model the coalescent process!

Phylogenetic coalescent model with mutation



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

The likelihood function

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

Likelihood function

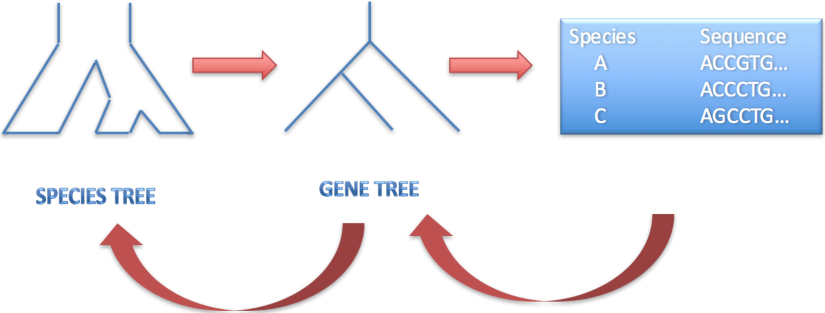
- Recall the Felsenstein equation from Peter's lecture, except that now we replace θ with S, the species tree. Use this to form the species tree likelihood for a multi-locus data set:

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

  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"]

Inference option 1: Summary statistics methods



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

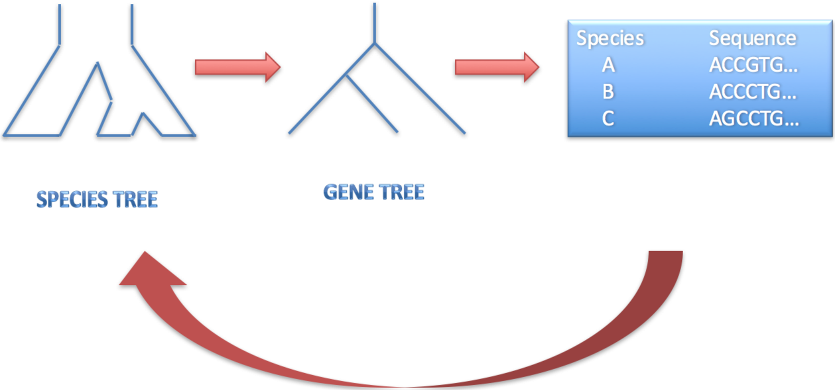
Coalescent-based methods for species tree inference

- 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)

Coalescent-based methods for species tree inference

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

Inference option 2: Full data methods



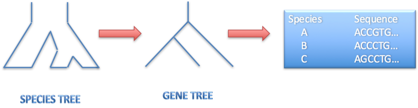
Species	Sequence
A	ACCGTG...
B	ACCCTG...
C	AGCCTG...

SPECIES TREE

GENE TREE

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

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

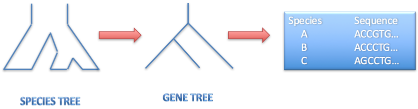


Species	Sequence
A	ACCGTG...
B	ACCCTG...
C	AGCCTG...

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

Full data methods II: SVDQuartets

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



SPECIES TREE

GENE TREE

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

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

Coalescent-based method for species tree inference

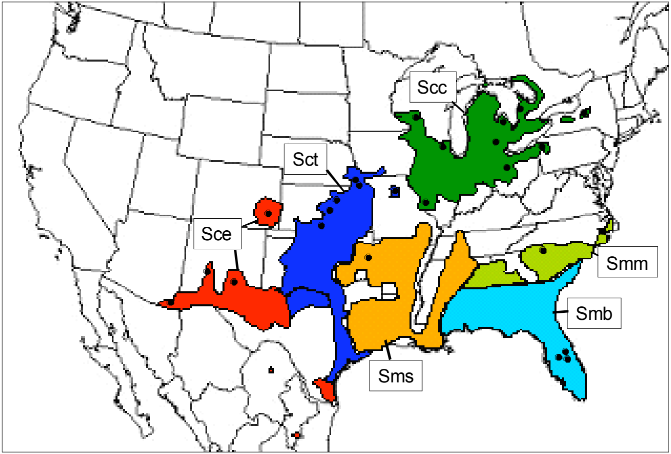
- Comparison of approaches:
  - Summary statistics methods
    - ⋆ Advantage: Quick
    - ⋆ Disadvantage: Ignore information in the data
    - ⋆ Most current implementations do not easily allow assessment of uncertainty (but bootstrap can be used, at the expense of computational efficiency)
  - Full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

Multilocus data 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

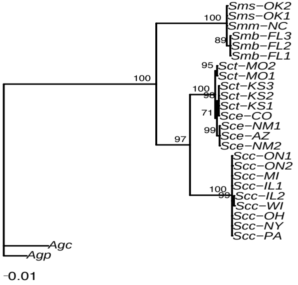


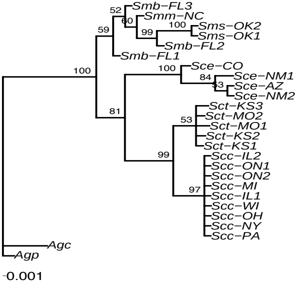
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

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

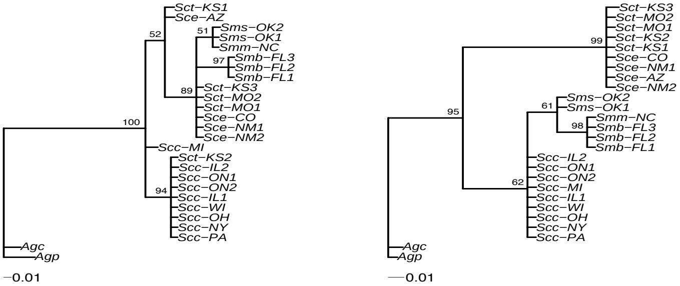
Some are very informative:





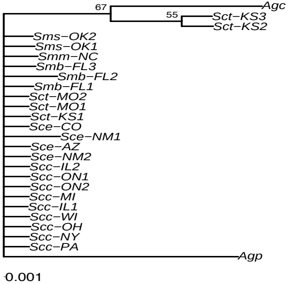
Individual Gene Tree Estimates

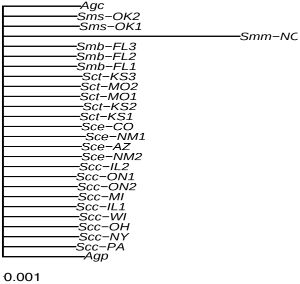
Some are a little informative:



Individual Gene Tree Estimates

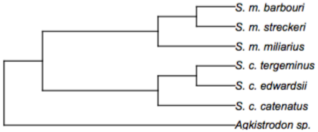
And then there are others .....



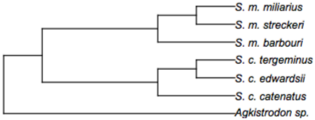


Example: Sistrurus rattlesnakes

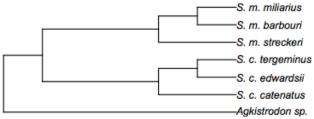
STEM, STEAC



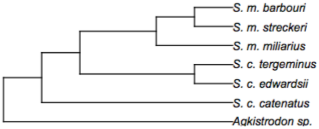
BEAST (concatenated data), *BEAST



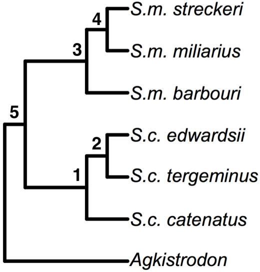
BEST, Parsimony & MrBayes (concatenated data), Astral



PhyloNet, STAR



Example 1: Sistrurus rattlesnakes

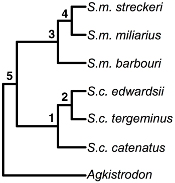


Node	1	2	3	4	5
*BEAST	100	100	100	46*	100
BPP	100	99	100	33*	100
SVDQ	93	100	100	46	100

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

Example 1: Sistrurus rattlesnakes

Very rough ideas of computational time ...



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

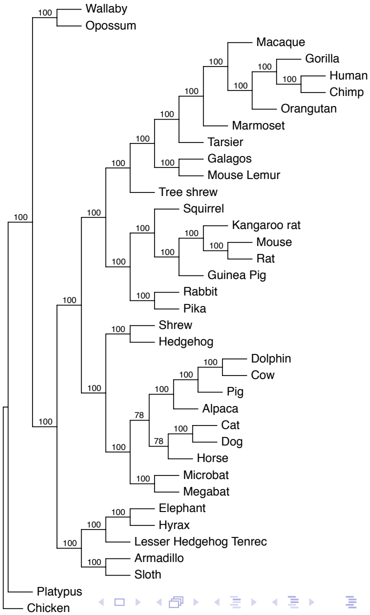
Multilocus data example 2: Mammals

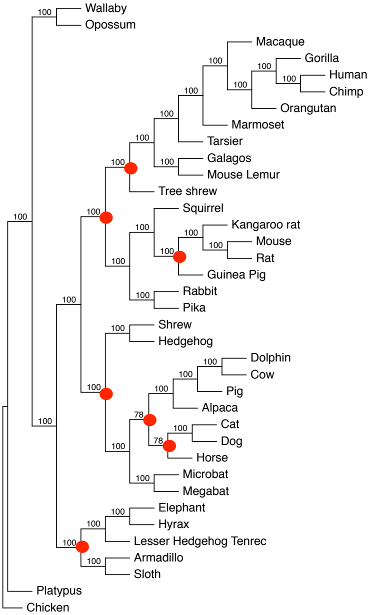
- Series of papers in the literature debating proper phylogenetic relationships among a group of mammals
  - Meredith RW, et al. (Science, 2011) criticized by Song et al. (PNAS, 2012):
    - ★ Amount of data “insufficient” (26 genes, 35,603 bp, 164 mammals)
    - ★ Concatenation not appropriate
  - Response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

Example 2: Mammals

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

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





- "Historically problematic nodes" identified by McCormack et al. (Genome Research, 2012) are identified with a red circle

- Overall, the SVDQ analysis agrees with the analysis of Song et al. (2012), who used the coalescence-based method MP-EST

- The SVDQ analysis differs from analyses based on concatenation for some of the difficult nodes, but agrees with concatenation for the two nodes with lower bootstrap support

SNP data example: *Diospyros*

From Wikipedia:

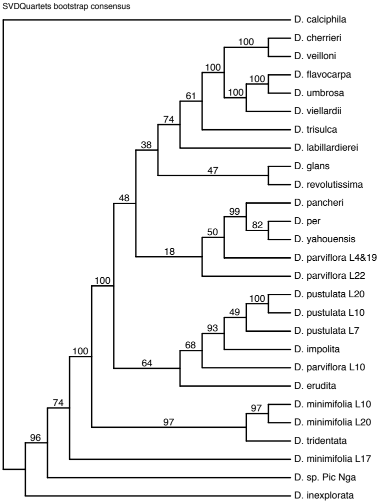
”Diospyros is a genus of over 700 species of deciduous and evergreen trees, shrubs and small bushes. The majority are native to the tropics, with only a few species extending into temperate regions. Depending on their nature, individual species are commonly known as ebony or persimmon trees. Some are valued for their hard, heavy, dark timber, and some for their fruit. Some are useful as ornamentals and many are of local ecological importance.”

© G.Sankowsky

- Data: samples from New Caledonia archipelago – Ovidiu et al., Syst. Biol. 65(2):212-227, 2016
- 84 individuals, sampled from 39 populations, representing 21 species
- 26 tips on species tree
- Data set 1 (PAUP*) : 8,488 SNPs
- Data set 2 (SNAPP) : 1,506 SNPs (one per locus)

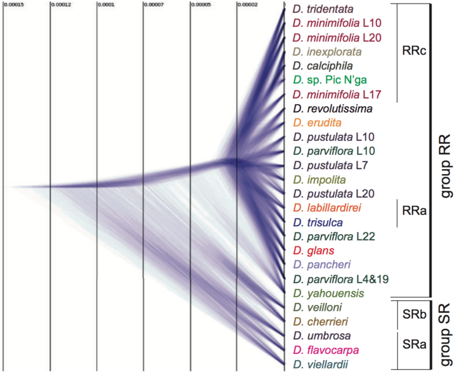
Diospyros data

SVDQ: 15 hours, 27 minutes for 100 bootstrap reps, sampling all quartets



SNAPP: Ovidiu et al. (2016) used 5,000,000 iterations (all ESS >100, most > 200)

My analysis: 1500 iterations took ∼ 2 days



Species Tree Inference Summary – Comparison of Methods

Software	Data Type	Measure of Uncertainty	Computation Time	Models Included
BEST	multilocus	posterior probability	long; can be run in parallel	coalescent; all reversible substitution models
*BEAST	multilocus	posterior probability	intermediate; can be run in parallel	coalesent; all reversible substitution models; relaxed clock; variable population sizes

BPP	multilocus	posterior probability	long	coalescent; JC69 model only;
				molecular clock;
				species delimitation

SVDQ	multilocus; SNP	bootstrap	short	coalescent; all reversible substitution models; non-clock; gene flow; parameter estimation ?
SNAPP	biallelic SNP; AFLP	posterior probability	long; can be run in parallel	coalescent; two-state substitution model; Bayes factor delimitation
ASTRAL	unrooted gene trees	bootstrap	short given gene trees	no specific model assumed
MP-EST	rooted gene trees	bootstrap	short given gene trees	coalescent model

Species Tree Inference Summary

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