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 25, 2016
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 $\sim \text{Geometric}(\frac{1}{2N})$

- 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 $g(u) = \binom{k}{2} \frac{1}{2N} e^{-\left(\binom{k}{2}\frac{u}{2N}\right)}$, 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”.
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)

\[
P_{uv}(t) = \sum_{j=1}^{v} \binom{u}{j} \exp\left(-\frac{t}{2j}\right) j! \frac{(v+j)!}{(v+j-y)!} \frac{1}{(u+y)!} \left(\frac{(v+y)}{u+y}\right)^{v+y}
\]

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 $2N$)
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???
Recall our goal to integrate the population process with the phylogeny:

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)

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

Can use our previous results to get the following:
When \( u \) and \( v \) are small, these are easy to compute. For example,

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

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

Similarly,

\[
P_{22}(t) = \text{prob. of no coalescence in time } t \text{ for 2 lineages}
\]
\[
= P(T > t)
\]
\[
= \int_t^\infty e^{-x} \, dx = e^{-t}
\]
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 = $\frac{1}{3} e^{-t}$
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.
What are these probabilities like as a function of $t$, the length of time between speciation events?

- $\text{prob} = 1 - \exp(-t)$
- $\text{prob} = \frac{1}{3}\exp(-t)$
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).

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

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 $S$ is given by

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

- 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


- 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

![Tree Diagram]

- Chimp
  - Human
  - Gorilla
  - Orangutan

- Gorilla
  - Human
  - Chimp
  - Orangutan

- Gorilla
  - Chimp
  - Human
  - Orangutan
Applications of the topology distribution - example 1

Observed proportions of each
gene tree among ML phylogenies

76.6%  11.4%  11.5%
Applications of the topology distribution - example 1

- Observed proportions of each gene tree among ML phylogenies
- Predicted proportions using parameters from Rannala & Yang, 2003.

Tree 1:
- Chimp
- Human
- Gorilla
- Orangutan

Values: 76.6%, 79.1%

Tree 2:
- Gorilla
- Human
- Chimp
- Orangutan

Values: 11.4%, 9.9%

Tree 3:
- Gorilla
- Human
- Orangutan
- Chimp

Values: 11.5%, 9.9%
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
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

y=1, x=1

y=0.01, x=0.01

y=1, x=0.001

A B C D
Applications of the topology distribution - example 2
Applications of the topology distribution - example 2

y=1, x=1

y=0.01, x=0.01

y=1, x=0.001
Applications of the topology distribution - example 2

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


Rosenberg and Tao, *Systematic Biology*, 2008
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?

With A LOT of work, the probability distribution can be computed!

Different history distributions with identical topology distributions

Different history distributions with identical topology distributions

\[ \theta_A = 1 \]
\[ \theta_B = \theta_C = 0.5 \]
\[ \theta_{AB} = 0.2 \]

\[ \theta_A = \theta_B = \theta_C = \theta_{AB} = 1 \]
What about gene flow?

What do we learn from this?

- Many different choices of gene flow and coalescent parameters lead to an identical distribution on the *gene tree topologies*. The rate of gene flow is not identifiable from the topology distribution!

- The distribution of *gene tree histories* appears to be distinct for different parameter choices. 
  Conjecture: The gene tree history distribution identifies the rate of gene flow.

- For some choices for the rates of gene flow, all three gene tree topologies are equally likely (and thus the species tree cannot be identified).
  Contrast this with the situation in the absence of gene flow

Implications for triplet-based species tree inference methods

Reference: Tian and Kubatko, in revision.
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.
What about mutation?

Given this model, how should inference be carried out?

- As more data (genes) are added, the process of estimating species trees from concatenated data can be statistically inconsistent.

- May fail to converge to any single tree topology if there are many equally likely trees.

- May converge to the wrong tree when a gene tree that is topologically incongruent with the species tree has the highest probability.

- The bootstrap may be positively misleading – show strong support for an incorrect clade.

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

Kubatko and Degnan, 2007
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

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ACCGTG...</td>
</tr>
<tr>
<td>B</td>
<td>ACCCTG...</td>
</tr>
<tr>
<td>C</td>
<td>AGCCTG...</td>
</tr>
</tbody>
</table>
Model Underlying Coalescent-based Species Tree Inference

Species 'Sequence'
'''''A' '' ACCGTG…'
'''''B' '' ACCCTG…'
'''''C' '' AGCCTG…'

Inference

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ACCGTG…</td>
</tr>
<tr>
<td>B</td>
<td>ACCCTG…</td>
</tr>
<tr>
<td>C</td>
<td>AGCCTG…</td>
</tr>
</tbody>
</table>
Why is this so hard?

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 $\theta$ with $S$, the species tree. Use this to form the species tree likelihood for a multi-locus data set:

$$L(S|D_1, D_2, \ldots D_N) = \prod_{i=1}^{N} P(D_i|S) \text{ [loci conditionally independent]}$$

$$= \prod_{i=1}^{N} \sum_{g_j=1}^{G} P(D_i|g_j)f(g_j|S)$$

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

- This likelihood is difficult to evaluate directly, because of the dimension of the inner sum (which is really an integral) [recall Peter’s “galaxy slide”]

- To deal with this, either assume gene trees are known (summary statistics methods), use Bayesian techniques (most full data approaches), or think about small problems ☹. But new methods based on different ways of summarizing data are being developed!
Model Underlying Coalescent-based Species Tree Inference

Summary statistics methods

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ACCGTG...</td>
</tr>
<tr>
<td>B</td>
<td>ACCCTG...</td>
</tr>
<tr>
<td>C</td>
<td>AGCCTG...</td>
</tr>
</tbody>
</table>
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)
MP-EST

- Idea: Use a set of input gene trees to form a set of rooted triples

- We have seen that, in the absence of gene flow, the true triplet relationship has the highest probability under the coalescent model

- Use the rooted triples with the highest frequency of the three choices to construct the species tree

- Branch lengths are can also be estimated in this way (recall our figure about gene tree topology probabilities as a function of $t$)

- Potential problem: assumes that the gene trees are correctly estimated and are error-free [but can use bootstrap]

- Scott will discuss this method tomorrow
Idea: Use a set of input gene trees to define a set of observed quartets, $X$

For a given species tree, define a score based on the number of quartets that agree with those in set of the input gene trees.

Find the species tree that maximizes this score (i.e., that agrees with the largest number of quartets from the input gene trees); heuristic search methods are used for large datasets.

Potential problem: assumes that the gene trees are correctly estimated and are error-free [but can use bootstrap]
Summary of Model Underlying Coalescent-based Species Tree Inference

Full Data Methods

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ACCGTG...</td>
</tr>
<tr>
<td>B</td>
<td>ACCCTG...</td>
</tr>
<tr>
<td>C</td>
<td>AGCCTG...</td>
</tr>
</tbody>
</table>
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

- 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.
Goal: Estimate the joint posterior distribution of species tree and gene trees (and model parameters), given the data

MCMC algorithm:
- Step 1: From a current state of Markov chain \((g_i, S_i)\), propose a new set of gene trees, \(g_{i+1}\)
- Step 2: Propose a new species tree, \(S_{i+1}\), that follows the constraints induced by the proposed gene trees from Step 1
  * Specifically, compute the Maximum Tree and modify a Poisson-distributed number of its nodes
- Step 3: Determine whether to accept the newly proposed state, \((g_{i+1}, S_{i+1})\)
  * An optional simulated annealing step can be used here to improve mixing

Implemented from within MrBayes
Goal: Estimate the joint posterior distribution of species tree and gene trees (and model parameters), given the data

Jointly propose gene trees and a species tree at each step of the algorithm and determine whether to accept the newly proposed state

Also allows variation in population size across branches of the species tree

Implemented in BEAST
Goal: Estimate the joint posterior distribution of species tree and gene trees (and model parameters), given the data

Jointly propose gene trees and a species tree at each step of the algorithm and determine whether to accept the newly proposed state

Currently, only the JC69 model is implemented

Can also be used for species delimitation

Similar to the paml program
SNAPP (SNP and AFLP Phylogenies)

- **Goal:** Estimate the posterior distribution of species trees, given the data

- **Data:** Biallelic SNPs or AFLPs

- Rather than also considering gene trees, integrate over gene trees to get likelihood of a species tree

- How? Limit consideration to two-state data and use a peeling algorithm

- Implemented in BEAST
Full data methods II: SVDQuartets

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

SVDQuartets is implemented in PAUP*

SVDQuartets will be discussed in detail in this afternoon’s lab
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: 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

<table>
<thead>
<tr>
<th>Species</th>
<th>Location</th>
<th>No. of individuals per gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. catenatus catenatus</td>
<td>Eastern U.S. and Canada</td>
<td>9</td>
</tr>
<tr>
<td>S. c. edwardsii</td>
<td>Western U.S.</td>
<td>4</td>
</tr>
<tr>
<td>S. c. tergeminus</td>
<td>Western and Central U.S.</td>
<td>5</td>
</tr>
<tr>
<td>S. miliarius miliarius</td>
<td>Southeastern U.S.</td>
<td>1</td>
</tr>
<tr>
<td>S. m. barbouri</td>
<td>Southeastern U.S.</td>
<td>3</td>
</tr>
<tr>
<td>S. m. streckerii</td>
<td>Southeastern U.S.</td>
<td>2</td>
</tr>
<tr>
<td>Agkistrodon sp. (outgroup)</td>
<td>U.S.</td>
<td>2</td>
</tr>
</tbody>
</table>
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: Sistrurus rattlesnakes

*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: *Sistrurus* rattlesnakes

Very rough ideas of computational time ...

<table>
<thead>
<tr>
<th>Program</th>
<th>Time</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEST</td>
<td>∼3 days</td>
<td>11,770,000 iterations (not converged)</td>
</tr>
<tr>
<td><em>BEAST</em></td>
<td>16.8 hours</td>
<td>100,000,000 iterations all ESS &gt; 200 except 1 (＞100)</td>
</tr>
<tr>
<td>BPP</td>
<td>4 days</td>
<td>500,000 iterations</td>
</tr>
<tr>
<td>SVDQ</td>
<td>11 minutes</td>
<td>all quartets sampled 100 bootstrap reps</td>
</tr>
<tr>
<td>ASTRAL</td>
<td>2.215 sec</td>
<td>given gene trees! also need bootstrap</td>
</tr>
</tbody>
</table>
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.”

- 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
<table>
<thead>
<tr>
<th>Software</th>
<th>Data Type</th>
<th>Measure of Uncertainty</th>
<th>Computation Time</th>
<th>Models Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEST</td>
<td>multilocus</td>
<td>posterior probability</td>
<td>long; can be run in parallel</td>
<td>coalescent; all reversible substitution models</td>
</tr>
<tr>
<td>*BEAST</td>
<td>multilocus</td>
<td>posterior probability</td>
<td>intermediate; can be run in parallel</td>
<td>coalescent; all reversible substitution models; relaxed clock; variable population sizes</td>
</tr>
<tr>
<td>BPP</td>
<td>multilocus</td>
<td>posterior probability</td>
<td>long</td>
<td>coalescent; JC69 model only; species delimitation</td>
</tr>
<tr>
<td>SVDQ</td>
<td>multilocus; SNP</td>
<td>bootstrap</td>
<td>short</td>
<td>coalescent; all reversible substitution models; parameter estimation ?</td>
</tr>
<tr>
<td>SNAPP</td>
<td>biallelic SNP; AFLP</td>
<td>posterior probability</td>
<td>long; can be run in parallel</td>
<td>coalescent; two-state substitution model; Bayes factor delimitation</td>
</tr>
<tr>
<td>ASTRAL</td>
<td>unrooted gene trees</td>
<td>bootstrap</td>
<td>short given gene trees</td>
<td>no specific model assumed</td>
</tr>
<tr>
<td>MP-EST</td>
<td>rooted gene trees</td>
<td>bootstrap</td>
<td>short given gene trees</td>
<td>coalescent model</td>
</tr>
</tbody>
</table>
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!