Phylogenomics

- Genome-scale evolution (ILS, GDL, etc.)
- The multi-species coalescent, and gene tree incongruence due to incomplete lineage sorting (ILS)
- Estimating species trees in the presence of ILS
- Open problems
Phylogenomics

(Phylogenetic estimation from whole genomes)
Species Tree

From the Tree of the Life Website,
University of Arizona
The Tree of Life: *Multiple Challenges*

Large datasets:  
100,000+ sequences  
10,000+ genes  
“BigData” complexity

Large-scale statistical phylogeny estimation  
Ultra-large multiple-sequence alignment  
Estimating species trees from incongruent gene trees  
Supertree estimation  
Genome rearrangement phylogeny  
Reticulate evolution  
Visualization of large trees and alignments  
Data mining techniques to explore multiple optima
The Tree of Life: **Multiple Challenges**

Large datasets:
- 100,000+ sequences
- 10,000+ genes

“BigData” complexity

Large-scale statistical phylogeny estimation
Ultra-large multiple-sequence alignment

**Estimating species trees from incongruent gene trees**

Supertree estimation
Genome rearrangement phylogeny

Reticulate evolution
Visualization of large trees and alignments

Data mining techniques to explore multiple optima

This talk
Topics

• Gene tree estimation and statistical consistency
• Gene tree conflict due to incomplete lineage sorting
• The multi-species coalescent model
  – Identifiability and statistical consistency
  – The challenge of gene tree estimation error
  – The challenge of dataset size
• New methods for coalescent-based estimation
  – Statistical binning (Mirarab et al., 2014, Bayzid et al. 2014) – used in Avian tree
  – ASTRAL (Mirarab et al., 2014, Mirarab and Warnow 2015) – used in Plant tree
DNA Sequence Evolution (Idealized)
Markov Model of Site Evolution

Simplest (Jukes-Cantor, 1969):

- The model tree $T$ is binary and has substitution probabilities $p(e)$ on each edge $e$.
- The state at the root is randomly drawn from $\{A,C,T,G\}$ (nucleotides)
- If a site (position) changes on an edge, *it changes with equal probability to each of the remaining states.*
- The evolutionary process is Markovian.

The different sites are assumed to evolve *i.i.d.* (independently and identically) down the tree (with rates that are drawn from a gamma distribution).

More complex models (such as the General Markov model) are also considered, often with little change to the theory.
Maximum Likelihood Phylogeny Estimation

Input: Sequence set $S$
Output: Jukes-Cantor model tree $T$ (with substitution probabilities on edges) such that $\Pr(S|T)$ is maximized

ML tree estimation is usually performed under other more realistic models (e.g., the Generalized Time Reversible model)
Quantifying Error

TRUE TREE

FN: false negative (missing edge)
FP: false positive (incorrect edge)

50% error rate

DNA SEQUENCES

INFERRED TREE

S₁: ACAATTAGAAC
S₂: ACCCTTAGAAC
S₃: ACCATTCCAAC
S₄: ACCAGACCAAC
S₅: ACCAGACCGGA
Maximum Likelihood is Statistically Consistent
Maximum Likelihood is Statistically Consistent

Data

Data are sites in an alignment
Species Tree

Orangutan | Gorilla | Chimpanzee | Human

From the Tree of the Life Website, University of Arizona
Sampling multiple genes from multiple species

From the Tree of the Life Website, University of Arizona
Using multiple genes

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S7</th>
<th>S8</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>TCTAATGGAA</td>
<td>GCTAAGGGAA</td>
<td>TCTAAGGGAA</td>
<td>TCTAACGGAA</td>
<td>TCTAATGGAC</td>
</tr>
<tr>
<td>gene 2</td>
<td>S4</td>
<td>S5</td>
<td>S6</td>
<td>S7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGTAACCCTC</td>
<td>GCTAAACCTC</td>
<td>GGTGACCATC</td>
<td>GCTAAACCTC</td>
<td></td>
</tr>
<tr>
<td>gene 3</td>
<td>S1</td>
<td>S3</td>
<td>S4</td>
<td>S7</td>
<td>S8</td>
</tr>
<tr>
<td></td>
<td>TATTGATACCA</td>
<td>TCTTGATACCCTC</td>
<td>TAGTGATGCAA</td>
<td>TAGTGATGCAA</td>
<td>CATTCATACCA</td>
</tr>
</tbody>
</table>
## Concatenation

<table>
<thead>
<tr>
<th></th>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>TCTAATGGAA</td>
<td>??????????</td>
<td>TATTGATACA</td>
</tr>
<tr>
<td>$S_2$</td>
<td>GCTAAGGGAA</td>
<td>??????????</td>
<td>??????????</td>
</tr>
<tr>
<td>$S_3$</td>
<td>TCTAAGGGAA</td>
<td>??????????</td>
<td>TCTTGATACC</td>
</tr>
<tr>
<td>$S_4$</td>
<td>TCTAACGGA</td>
<td>GGTAACCCCTC</td>
<td>TAGTGATGCA</td>
</tr>
<tr>
<td>$S_5$</td>
<td>??????????</td>
<td>GCTAAACCTC</td>
<td>??????????</td>
</tr>
<tr>
<td>$S_6$</td>
<td>??????????</td>
<td>GGTGACCATC</td>
<td>??????????</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>GCTAAACCTC</td>
<td>TAGTGATGCA</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGA</td>
<td>??????????</td>
<td>CATTTCATACC</td>
</tr>
</tbody>
</table>

The diagram shows a tree structure connecting the sequences from gene 1, gene 2, and gene 3.
Red gene tree ≠ species tree
(green gene tree okay)
Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

G Zhang, BGI

T. Warnow
UT-Austin

S. Mirarab
UT-Austin

Md. S. Bayzid, UT-Austin

Plus many many other people...

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Science 2014, Jarvis, Mirarab, et al.
1KP: Thousand Transcriptome Project

- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- Multi-institutional project (10+ universities)
- iPLANT (NSF-funded cooperative)
- Gene sequence alignments and trees computed using SATe (Liu et al., Science 2009 and Systematic Biology 2012)

Proceedings of the National Academy of Sciences, Wickett, Mirarab et al., 2014
Gene Tree Incongruence

- Gene trees can differ from the species tree due to:
  - Duplication and loss
  - Horizontal gene transfer
  - Incomplete lineage sorting (ILS)
Incomplete Lineage Sorting (ILS)

• Confounds phylogenetic analysis for many groups:
  – Hominids
  – Birds
  – Yeast
  – Animals
  – Toads
  – Fish
  – Fungi

• There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS.
Species tree estimation: difficult, even for small datasets!

From the Tree of the Life Website, University of Arizona
The Coalescent
Gene tree in a species tree

Courtesy James Degnan
Lineage Sorting

• Population-level process, also called the “Multi-species coalescent” (Kingman, 1982)

• Gene trees can differ from species trees due to short times between speciation events or large population size; this is called “Incomplete Lineage Sorting” or “Deep Coalescence”.
# Using multiple genes

<table>
<thead>
<tr>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>TCTAATGGAA</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>GCTAAGGGAA</td>
<td></td>
</tr>
<tr>
<td>$S_3$</td>
<td>TCTAAGGGAA</td>
<td></td>
</tr>
<tr>
<td>$S_4$</td>
<td>TCTAACGGAA</td>
<td></td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td></td>
</tr>
<tr>
<td>$S_4$</td>
<td>GGTAACCTTC</td>
<td></td>
</tr>
<tr>
<td>$S_5$</td>
<td>GCTAAACCTTC</td>
<td></td>
</tr>
<tr>
<td>$S_6$</td>
<td>GGTGACCATC</td>
<td></td>
</tr>
<tr>
<td>$S_7$</td>
<td>GCTAAACCTTC</td>
<td></td>
</tr>
<tr>
<td>$S_1$</td>
<td>TATTGATACA</td>
<td></td>
</tr>
<tr>
<td>$S_3$</td>
<td>TCTTTGATACC</td>
<td></td>
</tr>
<tr>
<td>$S_4$</td>
<td>TAGTGATGCA</td>
<td></td>
</tr>
<tr>
<td>$S_7$</td>
<td>TAGTGATGCA</td>
<td></td>
</tr>
<tr>
<td>$S_8$</td>
<td>CATTGATACC</td>
<td></td>
</tr>
</tbody>
</table>
How to compute a species tree?
Inconsistent methods

• MDC (Parsimony-style method, Minimize Deep Coalescence)
• Greedy Consensus
• MRP (supertree method)
• Concatenation under maximum likelihood

In other words, all the usual approaches are not consistent – and some can be positively misleading!
Species tree estimation

1- **Concatenation**: statistically inconsistent (Roch & Steel 2014)

2- **Summary methods**: can be statistically consistent

3- **Co-estimation methods**: too slow for large datasets
Two competing approaches

Summary Method

Concatenation

Analyse separately

Species
gene 1 gene 2 ... gene k

Summary Method
Key observation:
Under the multi-species coalescent model, the species tree defines a probability distribution on the gene trees, and is identifiable from the distribution on gene trees.

Courtesy James Degnan
How to compute a species tree?
How to compute a species tree?

Techniques:
Most frequent gene tree?
Under the multi-species coalescent model, the species tree defines a probability distribution on the gene trees.

Theorem (Degnan et al., 2006, 2009): Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on \( \{A, B, C\} \) is identical to the rooted species tree induced on \( \{A, B, C\} \).
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology \(((A,B),C)\), and probability \(p^*\) of coalescence on the edge \(e^*\) above LCA\((A,B)\), with \(0<p^*<1\). We wish to show that

\[
P(gt= ((A,B),C)) > P(gt= ((A,C),B)) = P(gt= ((B,C),A)).
\]
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology ((A,B),C), and probability $p^*$ of coalescence on the edge $e^*$ above LCA(A,B), with $0 < p^* < 1$. We wish to show that

$Pr(gt=((A,B),C)) > Pr(gt=((A,C),B)) = Pr(gt=((B,C),A)).$

(1) If the lineages from A and B coalesce on $e^*$, then the gene tree is topologically identical to the species tree.
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology ((A,B),C), and probability $p^*$ of coalescence on the edge $e^*$ above LCA(A,B), with $0<p^*<1$. We wish to show that

$$\Pr(gt=((A,B),C)) > \Pr(gt=((A,C),B)) = \Pr(gt=((B,C),A)).$$

(1) If the lineages from A and B coalesce on $e^*$, then the gene tree is topologically identical to the species tree.

(2) If they do not coalesce, then all three lineages enter the edge above the root – and any pair can coalesce with equal probability.
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology $((A,B),C)$, and probability $p^*$ of coalescence on the edge $e^*$ above $\text{LCA}(A,B)$, with $0<p^*<1$. We wish to show that

$$\Pr(gt=((A,B),C)) > \Pr(gt=((A,C),B)) = \Pr(gt=((B,C),A)).$$

(1) If the lineages from A and B coalesce on $e^*$, then the gene tree is topologically identical to the species tree.
(2) If they do not coalesce, then all three lineages enter the edge above the root – and any pair can coalesce with equal probability.

Hence, $\Pr(gt=((B,C),A)) = \Pr(gt=((A,C),B)) = (1-p^*)/3 < 1/3$. 
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology \(((A,B),C)\), and probability \(p^*\) of coalescence on the edge \(e^*\) above LCA\((A,B)\), with \(0<p^*<1\). We wish to show that

\[
\Pr(gt=((A,B),C)) > \Pr(gt=((A,C),B)) = \Pr(gt=((B,C),A))
\]

(1) If the lineages from A and B coalesce on \(e^*\), then the gene tree is topologically identical to the species tree.

(2) If they do not coalesce, then all three lineages enter the edge above the root – and any pair can coalesce with equal probability.

Hence,

\[
\Pr(gt=((B,C),A)) = \Pr(gt=((A,C),B)) = (1-p^*)/3 < 1/3
\]

And also,

\[
\Pr(gt = ((A,B),C)) = p^* + (1-p^*)/3 > 1/3
\]
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Courtesy James Degnan
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.
Let the rooted species tree have topology \(((A,B),(C,D))\), and probabilities \(p1, p2\) of coalescence on the edges \(e1\) and \(e2\) above LCA(A,B) and LCA(C,D), respectively, with \(0<p1,p2<1\). We wish to show that as *unrooted* gene trees,

\[
\Pr(gt= ((A,B),(C,D))) > \Pr(gt= ((A,C),(B,D))) = \Pr(gt= ((B,C),(A,D))).
\]
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.
Let the rooted species tree have topology ((A,B),(C,D)), and probabilities \(p_1, p_2\) of coalescence on the edges \(e_1\) and \(e_2\) above LCA(A,B) and LCA(C,D), respectively, with \(0<p_1,p_2<1\). We wish to show that as *unrooted* gene trees,
\[
\Pr(gt=((A,B),(C,D))) > \Pr(gt=((A,C),(B,D))) = \Pr(gt=((B,C),(A,D))).
\]

(1) If the lineages from A and B coalesce on \(e_1\), then the gene tree is topologically identical to the species tree as an unrooted tree.
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology \(((A,B),(C,D))\), and probabilities \(p_1, p_2\) of coalescence on the edges \(e_1\) and \(e_2\) above LCA(A,B) and LCA(C,D), respectively, with \(0<p_1,p_2<1\). We wish to show that as *unrooted* gene trees,

\[
\Pr(gt=((A,B),(C,D))) > \Pr(gt=((A,C),(B,D))) = \Pr(gt=((B,C),(A,D))).
\]

(1) If the lineages from A and B coalesce on \(e_1\), then the gene tree is topologically identical to the species tree as an unrooted tree.

(2) If the lineages from C and D coalesce on \(e_2\), then the gene tree is topologically identical to the species tree as an unrooted tree.
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology ((A,B),(C,D)), and probabilities p1, p2 of coalescence on the edges e1 and e2 above LCA(A,B) and LCA(C,D), respectively, with 0<p1,p2<1. We wish to show that as *unrooted* gene trees,

\[ \Pr(gt=((A,B),(C,D))) > \Pr(gt=((A,C),(B,D))) = \Pr(gt=((B,C),(A,D))). \]

(1) If the lineages from A and B coalesce on e1, then the gene tree is topologically identical to the species tree as an unrooted tree.

(2) If the lineages from C and D coalesce on e2, then the gene tree is topologically identical to the species tree as an unrooted tree.

(3) If neither pair coalesces on these edges, then all four lineages enter the edge above the root – and any pair can coalesce with equal probability.
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology ((A,B),(C,D)), and probabilities \(p_1, p_2\) of coalescence on the edges \(e_1\) and \(e_2\) above LCA(A,B) and LCA(C,D), respectively, with \(0<p_1,p_2<1\). We wish to show that as *unrooted* gene trees,

\[
\Pr(gt=((A,B),(C,D))) > \Pr(gt=((A,C),(B,D))) = \Pr(gt=((B,C),(A,D))).
\]

(1) If the lineages from A and B coalesce on \(e_1\), then the gene tree is topologically identical to the species tree as an unrooted tree.

(2) If the lineages from C and D coalesce on \(e_2\), then the gene tree is topologically identical to the species tree as an unrooted tree.

(3) If neither pair coalesces on these edges, then all four lineages enter the edge above the root – and any pair can coalesce with equal probability.

Hence, \(\Pr(gt=((B,C),(A,D))) = \Pr(gt=((A,C),(B,D))) = (1-p_1)(1-p_2)/3 < 1/3\).

And so \(\Pr(gt=((A,B),(C,D))) > 1/3\)
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.
Theorem: The most probable **rooted** gene tree on **four** species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))), and probability $p_1$, $p_2$ of coalescence on the edges $e_1$ and $e_2$ above $LCA(C,D)$ and $LCA(B,C)$, respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1$, $p_2$ so that as *rooted* gene trees, $Pr(gt=((A,B),(C,D))) > Pr(gt=T^*=(A,(B,(C,D))))$. 
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree T* have topology ((A,(B,(C,D))), and probability p1, p2 of coalescence on the edges e1 and e2 above LCA(C,D) and LCA(B,C), respectively, with 0<p1,p2<1. We wish to show that we can pick p1, p2 so that as *rooted* gene trees, \[ \text{Pr}(gt=((A,B),(C,D))) > \text{Pr}(gt=T*=(A,(B,(C,D)))) \].

(1) Let \( \epsilon > 0 \) be given. Pick p1 and p2 small enough so that with probability at least 1-\( \epsilon \), all four lineages enter the edge above the root (i.e., \( (1-p1)(1-p2) > 1-\epsilon \)).
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A, (B, (C, D))),$ and probability $p_1, p_2$ of coalescence on the edges $e_1$ and $e_2$ above $\text{LCA}(C, D)$ and $\text{LCA}(B, C),$ respectively, with $0 < p_1, p_2 < 1.$ We wish to show that we can pick $p_1, p_2$ so that as *rooted* gene trees, $\Pr(\text{gt} = ((A, B), (C, D))) > \Pr(\text{gt} = T^* = (A, (B, (C, D))))$.

(1) Let $\varepsilon > 0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1 - \varepsilon,$ all four lineages enter the edge above the root (i.e., $(1 - p_1)(1 - p_2) > 1 - \varepsilon$).

(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))),$ and probability $p_1, p_2$ of coalescence on the edges $e_1$ and $e_2$ above $\text{LCA}(C,D)$ and $\text{LCA}(B,C),$ respectively, with $0<p_1,p_2<1.$ We wish to show that we can pick $p_1, p_2$ so that as *rooted* gene trees, $\Pr(gt=((A,B),(C,D))) > \Pr(gt=T^*=(A,(B,(C,D))))$.

(1) Let $\epsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\epsilon,$ all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\epsilon$).

(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –

(1) $(A,(B,(C,D)))$ is produced by the sequence of coalescences $C&D,$ then $B&CD,$ then $A&BCD.$
Theorem: The most probable **rooted** gene tree on **four** species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D)))$, and probability $p_1$, $p_2$ of coalescence on the edges $e_1$ and $e_2$ above LCA(C,D) and LCA(B,C), respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1$, $p_2$ so that as *rooted* gene trees, $\Pr(gt=((A,B),(C,D))) > \Pr(gt=T^*=(A,(B,(C,D))))$.

1. Let $\epsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\epsilon$, all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\epsilon$).
2. Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –
   1. $((A,B),(C,D)))$ is produced by the sequence of coalescences C&D, then B&CD, then A&BCD.
   2. $((A,B),(C,D)))$ is produced by two sequences of coalescence events – A&B, then C&D, then AB&CD, and also by C&D, then A&B, then AB&CD.
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))))$, and probability $p_1, p_2$ of coalescence on the edges $e_1$ and $e_2$ above $\text{LCA}(C,D)$ and $\text{LCA}(B,C)$, respectively, with $0 < p_1, p_2 < 1$. We wish to show that we can pick $p_1, p_2$ so that as *rooted* gene trees, $\Pr(\text{gt} = ((A,B),(C,D))) > \Pr(\text{gt} = T^* = (A,(B,(C,D))))$.

(1) Let $\epsilon > 0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1 - \epsilon$, all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\epsilon$).
(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –

(1) $(A,(B,(C,D)))$ is produced by the sequence of coalescences $C&D$, then $B&CD$, then $A&BCD$.
(2) $((A,B),(C,D))$ is produced by two sequences of coalescence events – $A&B$, then $C&D$, then $AB&CD$, and also by $C&D$, then $A&B$, then $AB&CD$. Hence, the rooted tree $((A,B),(C,D))$ is twice as likely as the rooted tree $(A,(B,(C,D)))$, which is the species tree $T^*$!
**Theorem:** The most probable **rooted** gene tree on **four** species may not be the rooted species tree.

**Proof:** Let the species tree $T^*$ have topology $((A, (B, (C, D))),$ and probability $p_1,$ $p_2$ of coalescence on the edges $e_1$ and $e_2$ above $LCA(C, D)$ and $LCA(B, C),$ respectively, with $0 < p_1, p_2 < 1.$ We wish to show that we can pick $p_1,$ $p_2$ so that as *rooted* gene trees, $Pr(gt = ((A, B), (C, D))) > Pr(gt = T^* = (A, (B, (C, D))))$.

(1) Let $\varepsilon > 0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1 - \varepsilon,$ all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\varepsilon$).

(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –

(1) $(A, (B, (C, D)))$ is produced by the sequence of coalescences $C&D,$ then $B&CD,$ then $A&BCD.$

(2) $((A, B), (C, D))$ is produced by two sequences of coalescence events – $A&B,$ then $C&D,$ then $AB&CD,$ and also by $C&D,$ then $A&B,$ then $AB&CD.$ Hence, the rooted tree $((A, B), (C, D))$ is twice as likely as the rooted tree $(A, (B, (C, D))),$ which is the species tree $T^*$!

Hence, for some model species tree, the rooted species tree topology is not the most likely rooted gene tree topology.
Let the species tree have topology \((A, (B, (C, D)))\). Show that for any probabilities \(p1\) and \(p2\) of coalescence on the internal edges of the species tree, then considering *unrooted* gene trees,

\[
\text{Pr}(\text{gt}=((A, B), (C, D))) > \text{Pr}(\text{gt}=((A, C), (B, D))) = \text{Pr}(\text{gt}=((B, C), (A, D))).
\]
How to compute a species tree?

Technique:
Most frequent gene tree?
How to compute a species tree?

Technique:
Most frequent gene tree?

**YES** if you have only three species

**YES** if you have four species and are content with the unrooted species tree

Otherwise **NO!**
How to compute a rooted species tree from rooted gene trees?
How to compute a rooted species tree from rooted gene trees?

Theorem (Degnan et al., 2006, 2009):
Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on \{A,B,C\} is identical to the rooted species tree induced on \{A,B,C\}.
How to compute a rooted species tree from rooted gene trees?

Estimate species tree for every 3 species

Theorem (Degnan et al., 2006, 2009):
Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on \{A, B, C\} is identical to the rooted species tree induced on \{A, B, C\}.
How to compute a rooted species tree from rooted gene trees?

Estimate species tree for every 3 species

Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.
How to compute a rooted species tree from rooted gene trees?

Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.
How to compute a rooted species tree from rooted gene trees?

Theorem (Degnan et al., 2009): Under the multi-species coalescent, the rooted species tree can be estimated correctly (with high probability) given a large enough number of true rooted gene trees.
How to compute an unrooted species tree from unrooted gene trees?

(Pretend these are unrooted trees)
How to compute an unrooted species tree from unrooted gene trees?

(Pretend these are unrooted trees)

Theorem (Allman et al., 2011, and others): For every four leaves \{a,b,c,d\}, the most probable unrooted quartet tree on \{a,b,c,d\} is the true species tree.
How to compute an unrooted species tree from unrooted gene trees?

(Pretend these are unrooted trees)  Estimate species tree for every 4 species

Theorem (Allman et al., 2011, and others): For every four leaves \{a,b,c,d\}, the most probable unrooted quartet tree on \{a,b,c,d\} is the true species tree.
How to compute an unrooted species tree from unrooted gene trees?

Theorem (Allman et al., 2011, and others): For every four leaves \{a, b, c, d\}, the most probable unrooted quartet tree on \{a, b, c, d\} is the true species tree.

Use the All Quartets Method to construct the species tree, based on the most frequent gene trees for each set of four species.
How to compute an unrooted species tree from unrooted gene trees?

Theorem (Allman et al., 2011, and others): For every four leaves \{a,b,c,d\}, the most probable unrooted quartet tree on \{a,b,c,d\} is the true species tree.

Use the All Quartets Method to construct the species tree, based on the most frequent gene trees for each set of four species.

Combine unrooted 4-taxon trees

(Pretend these are unrooted trees)
Statistically consistent methods

Methods that require rooted gene trees:
- **SRSTE** (Simple Rooted Species Tree Estimation) - see textbook
- **MP-EST** (Liu et al. 2010): maximum pseudo-likelihood estimation

Methods that work from unrooted gene trees:
- **SUSTE** (Simple Unrooted Species Tree Estimation) - see textbook
- **BUCKy-pop** (Ané and Larget 2010): quartet-based Bayesian species tree estimation
- **ASTRAL** (Mirarab et al., 2014) and **ASTRAL-2** (Mirarab and Warnow 2015) – quartet-based estimation
- **ASTRID** (Vachaspati & Warnow 2015), GLASS, **NJst** (Liu and Yu, 2011), and **STEAC** (Liu et al., 2009) - distance-based methods

Methods that work from sequence alignments:
- **BEST** (Liu 2008) and **BEAST** (Heled, and Drummond): Bayesian co-estimation of gene trees and species trees
- **SVDquartets** (Chifman and Kubatko, 2014): quartet-based method
- **SNAPP** (Bryant et al., 2012)
- **METAL** (Dasarathy, Nowak, and Roch 2015)

Note that some of these methods are only statistically consistent under a strict molecular clock, and many are computationally intensive. Some have never been implemented.
Summary methods are statistically consistent species tree estimators.

Here, the “data” are true gene trees.
Results on 11-taxon datasets with weak ILS

*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013
Results on 11-taxon datasets with strong ILS

Average FN rate

5-genes 10-genes 25-genes 50-genes

*BEAST more accurate than summary methods (MP-EST, BUCKy, etc).
CA-ML: (concatenated analysis) also very accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013
*BEAST co-estimation produces more accurate gene trees than Maximum Likelihood

11-taxon weak ILS datasets

17-taxon (very high ILS) datasets

11-taxon datasets from Chung and Ané, Syst Biol 2012
17-taxon datasets from Yu, Warnow, and Nakhleh, JCB 2011

Bayzid & Warnow, Bioinformatics 2013
Impact of Gene Tree Estimation Error on MP-EST

MP-EST has no error on true gene trees, but MP-EST has 9% error on estimated gene trees.

Datasets: 11-taxon strong ILS conditions with 50 genes

Similar results for other summary methods (MDC, Greedy, etc.).
Problem: poor gene trees

• Summary methods combine estimated gene trees, not true gene trees.
Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.
- The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.
- The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
- Species trees obtained by combining poorly estimated gene trees have poor accuracy.
Summary methods combine estimated gene trees, not true gene trees.

The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

Species trees obtained by combining poorly estimated gene trees have poor accuracy.
Addressing gene tree estimation error

- Get better estimates of the gene trees
- Restrict to subset of estimated gene trees
- Model error in the estimated gene trees
- Modify gene trees to reduce error
- Develop methods with greater robustness to gene tree error
Addressing gene tree estimation error

• Get better estimates of the gene trees
• Restrict to subset of estimated gene trees
• Model error in the estimated gene trees
• Modify gene trees to reduce error
• Develop methods with greater robustness to gene tree error
  – ASTRAL. Bioinformatics 2014 (Mirarab et al.)
  – Statistical binning. Science 2014 (Mirarab et al.)
Avian Phylogenomics Project

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Species tree estimated using Statistical Binning with MP-EST (Jarvis, Mirarab, et al., Science 2014)
Avian whole genomes phylogenies
[Jarvis*, Mirarab*, et al., Science, 2014]

- International team of more than 100 researchers
- Whole genomes for 48 bird species (~100 million years of evolution)
- Goal: a phylogeny of major bird lineages
  - Extremely challenging due to rampant gene tree incongruence
  - Implications for traits such as vocal learning
- 14,000 “genes” (typically short and relatively conserved)
Gene trees on the avian dataset

A measure of confidence in estimated gene tree branches

14,000 genes from avian genome-scale data
[Jarvis*, Mirarab*, et al., Science, 2014]
Gene trees on the avian dataset

14,000 genes from avian genome-scale data [Jarvis*, Mirarab*, et al., Science, 2014]

Avian-like simulations (1000 genes) [Mirarab, et al., Science, 2014]
Gene trees on the avian dataset

Gene tree error matters

[Ané, et al, MBE, 2007]
[Patel, et al, MBE, 2013]
[Mirarab, et al., Systematic Biology, 2014]

14,000 genes from avian genome-scale data
[Jarvis*, Mirarab*, et al., Science, 2014]

Avian-like simulations (1000 genes)
• The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
• The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

• Species trees obtained by combining poorly estimated gene trees have poor accuracy.
• The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

• Species trees obtained by combining poorly estimated gene trees have poor accuracy.

• There are no theoretical guarantees for summary methods except for perfectly correct gene trees.
The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

Species trees obtained by combining poorly estimated gene trees have poor accuracy.

There are no theoretical guarantees for summary methods except for perfectly correct gene trees.
Idea: combine best aspects of concatenation and summary methods

- Concatenation (fully partitioned) works fine when the concatenated data evolve under identical (or very similar) trees.

- Some pairs of genes are not discordant (at least in topology).

- Concatenate “combinable” sets of genes into “supergenes” to increase the phylogenetic signal.

- But how do we know which genes are combinable if we cannot estimate them correctly?
Statistical binning

Input: estimated gene trees with bootstrap support, and minimum support threshold $t$

Step 1: partition of the estimated gene trees into sets, so that no two gene trees in the same set are strongly incompatible, and the sets have approximately the same size.

Step 2: estimate “supergene” trees on each set using concatenation (maximum likelihood)

Step 3: combine supergene trees using coalescent-based method

Note: Step 1 requires solving the NP-hard “balanced vertex coloring problem”, for which we developed a good heuristic (modified 1979 Brelaz algorithm)
Gene alignments

Estimated gene trees

Species tree

Traditional pipeline (unbinned)

Sequence data → Gene alignments → Estimated gene trees → Species tree

Statistical Binning pipeline

Incompatibility Graph → Binned supergene alignments

Weighted

Supergene trees (weighted) → Species tree

Unweighted

Supergene trees (unweighted) → Species tree

Sequence data

Supergene trees (weighted)

Weighted

Supergene trees (unweighted)

Unweighted
Statistical binning vs. unbinned

Miraраб, et al., Science 2014
Binning produces bins with approximate 5 to 7 genes each
Datasets: 11-taxon strongILS datasets with 50 genes, Chung and Ané, Systematic Biology
Avian Simulation: Impact of binning with MP-EST

(a) Varying gene tree estimation error

(b) Varying the level of ILS

- Unbinned
- Binned-uw
- Binned-w
- Concatenation
Comparing Binned and Un-binned MP-EST on the Avian Dataset

Unbinned MP-EST strongly rejects Columbea, a major finding by Jarvis, Mirarab, et al.
Summary so far

- Standard coalescent-based methods (such as MP-EST) have poor accuracy in the presence of gene tree error.
- Statistical binning improves the estimation of gene tree distributions, and so:
  - Improves species tree estimation
Summary so far

• Standard coalescent-based methods (such as MP-EST) have poor accuracy in the presence of gene tree error.

• Statistical binning improves the estimation of gene tree distributions, and so:
  • Improves species tree estimation
  • Improves species tree branch lengths
Summary so far

- Standard coalescent-based methods (such as MP-EST) have poor accuracy in the presence of gene tree error.

- Statistical binning improves the estimation of gene tree distributions, and so:
  - Improves species tree estimation
  - Improves species tree branch lengths
  - Reduces incidence of strongly supported false positive branches
1KP: Plant whole transcriptomes
[Wickett*, Mirarab*, et al., PNAS, 2014]

Phylotranscriptomic analysis of the origin and early diversification of land plants
Dennis W. Stevenson*, Barbara Surek*, Juan Carlos Villarreal*, Béatrice Roure*, Hervé Philippe*,
Gane Ka-Shu Wong*2,3,2, and James Leebens-Mack*2

- Whole transcriptomes for 103 plant species
  - 1,200 in the next phase
- 400-800 single copy “genes”
- Spans ~1 billion years of evolution
- Many unanswered questions about plant evolution
Summary methods on the 1KP data (103 plants)

- Existing summary methods produced species trees with low support and unbelievable relationships
- .. despite having gene trees with relatively high bootstrap support

400 genes from 1KP data
[Wickett*, Mirarab*, et al., PNAS, 2014]
Summary methods on the 1KP data (103 plants)

• Existing summary methods produced species trees with low support and unbelievable relationships

• .. despite having gene trees with relatively high bootstrap support

• Our simulation studies showed that the reason had to do with the number of taxa

400 genes from 1KP data [Wickett*, Mirarab*, et al., PNAS, 2014]

1000 simulated genes, “medium” levels of ILS [Mirarab and Warnow, ISMB, 2015]
Summary methods on the 1KP data (103 plants)

- Existing summary methods produced species trees with low support and unbelievable relationships
- .. despite having gene trees with relatively high bootstrap support
- Our simulation studies showed that the reason had to do with the number of taxa

400 genes from 1KP data [Wickett*, Mirarab*, et al., PNAS, 2014]

1000 simulated genes, “medium” levels of ILS [Mirarab and Warnow, ISMB, 2015]
Summary methods on the 1KP data (103 plants)

- Existing summary methods produced species trees with low support and unbelievable relationships

- .. despite having gene trees with relatively high bootstrap support

- Our simulation studies showed that the reason had to do with the number of taxa

The problem size (# species) matters too!

400 genes from 1KP data
[Wickett*, Mirarab*, et al., PNAS, 2014]

1000 simulated genes, “medium” levels of ILS
[Mirarab and Warnow, ISMB, 2015]
1KP: Thousand Transcriptome Project

- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- Gene sequence alignments and trees computed using SATe (Liu et al., Science 2009 and Systematic Biology 2012)

Species tree estimated using ASTRAL (Bioinformatics, 2014)

1KP paper by Wickett, Mirarab et al., PNAS 2014
ASTRAL

• Accurate Species Trees Algorithm
• Mirarab et al., ECCB 2014 and Bioinformatics 2014
• Statistically-consistent estimation of the species tree from unrooted gene trees
ASTRAL’s approach

• Input: set of unrooted gene trees $T_1, T_2, \ldots, T_k$
• Output: Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees

Theorem:

• An exact solution to this problem would be a statistically consistent algorithm in the presence of ILS
ASTRAL’s approach

• Input: set of unrooted gene trees $T_1, T_2, ..., T_k$
• Output: Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees

Theorem:
• An exact solution to this problem is NP-hard
ASTRAL’s approach

• Input: set of unrooted gene trees $T_1, T_2, ..., T_k$ and set $X$ of bipartitions on species set $S$

• Output: Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees, subject to Bipartitions($T^*$) drawn from $X$

Theorem:

• An exact solution to this problem is achievable in polynomial time!
ASTRAL’s approach

• Input: set of unrooted gene trees $T_1, T_2, \ldots, T_k$ and set $X$ of bipartitions on species set $S$

• Output: Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees, subject to $\text{Bipartitions}(T^*)$ drawn from $X$

Theorem:

• Letting $X$ be the set of bipartitions from the input gene trees is statistically consistent and polynomial time.
ASTRAL vs. MP-EST and Concatenation

200 genes, 500bp

Mammalian Simulation Study, Varying ILS level
ASTRAL on plants dataset

- The ASTRAL tree:
  - High support
  - Similar to concatenation with some interesting differences (e.g., recovered bryophytes)
  - ASTRAL took only about 10 minutes (serial running time) on 103 taxa and 400 genes

[Wickett*, Mirarab*, et al., PNAS, 2014]
1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UIUC

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Plus many many other people…

2020 Nature paper Challenges (~1200 species, ~400 loci):

• Species tree estimation under the multi-species coalescent from hundreds of conflicting gene trees on >1000 species;
  we used ASTRAL-3
Summary

- Gene tree estimation error (e.g., due to insufficient phylogenetic signal) is a typical occurrence in phylogenomics projects.
- Standard summary methods for coalescent-based species tree estimation (e.g., MP-EST) are impacted by gene tree error.
- Statistical binning improves the estimation of gene tree distributions – and hence leads to improved species tree estimation when gene trees have insufficient accuracy. Weighted statistical binning provides similar empirical improvements but is theoretically better (with respect to statistical consistency).
- ASTRAL is a statistically consistent summary method that is quartet-based, and can analyze very large datasets (1000 genes and 1000 species).
Papers and Software

- ASTRAL (Mirarab et al., ECCB and Bioinformatics 2014, paper 115)
- ASTRAL-2 (Mirarab and Warnow, ISMB 2015, paper 125)
- Statistical Binning (Mirarab et al., Science 2014, paper 122)
- Weighted statistical binning (Bayzid et al., PLOS One 2015, paper 125)
- SVDQuartets (Chifman and Kubatko, Bioinformatics 2014)
- ASTRID (Vachaspati and Warnow, 2015, paper 131)
- BUCKy (Larget et al., Bioinformatics 26(22):2910–2911, 2010)
- METAL (Dasarathy, Nowak, and Roch, TCBB 12(2):422-432, 2015)
- STEAC (Liu et al., Syst Biol 2009)

Open source software for most methods on github.

Datasets (simulated and biological) available online.
Research Projects

- **Improving quartet-based methods through better quartet tree amalgamation methods.** Many summary methods are quartet-based methods, and so could potentially be improved by using better methods to compute trees from a set of quartet trees (called quartet-amalgamation techniques). Similarly for triplet-based methods.

- **Improving accuracy of distance-based methods.** NJst runs NJ on a distance matrix it computes. ASTRID (paper 131) improves this by using FastME instead of NJ. Perhaps other distance-based methods (e.g., STEAC) could be similarly improved?

- **Improving speed of summary methods.** Many summary methods are not designed for large numbers of species, and so improvements in running time might be easily obtained through better algorithm design (and re-implementation). Perhaps even HPC implementations.

- **Improving the scalability of *BEAST to larger numbers of taxa.** *BEAST co-estimates gene trees and species trees, and can have outstanding accuracy -- but uses an MCMC method that is computationally intensive, and so the method is limited to small numbers of species and genes. BBCA (paper #117) improves the speed of convergence for *BEAST by partitioning the gene set, but does not improve the scalability to large numbers of taxa. Divide-and-conquer approaches (paper #116) could be helpful.

- **Improving co-estimation of gene trees and species trees.** Both *BEAST and Statistical Binning provide improved accuracy over standard summary methods because they are able to get more accurate gene trees. But *BEAST is too computationally intensive, and statistical binning (even with weighting) relies on concatenation to estimate better gene trees. Can we do this in a different way?

- **Testing existing species tree estimation methods.** Little is known about the empirical or theoretical performance of species tree estimation methods except under idealized conditions. In particular, little is known about the impact of missing data, deviation from a molecular clock, limited sequence length per gene, and orthology error (duplication/loss or horizontal gene tree scenarios that also create gene tree discordance) on the accuracy of methods. See papers 127, 128, 129, and 130.