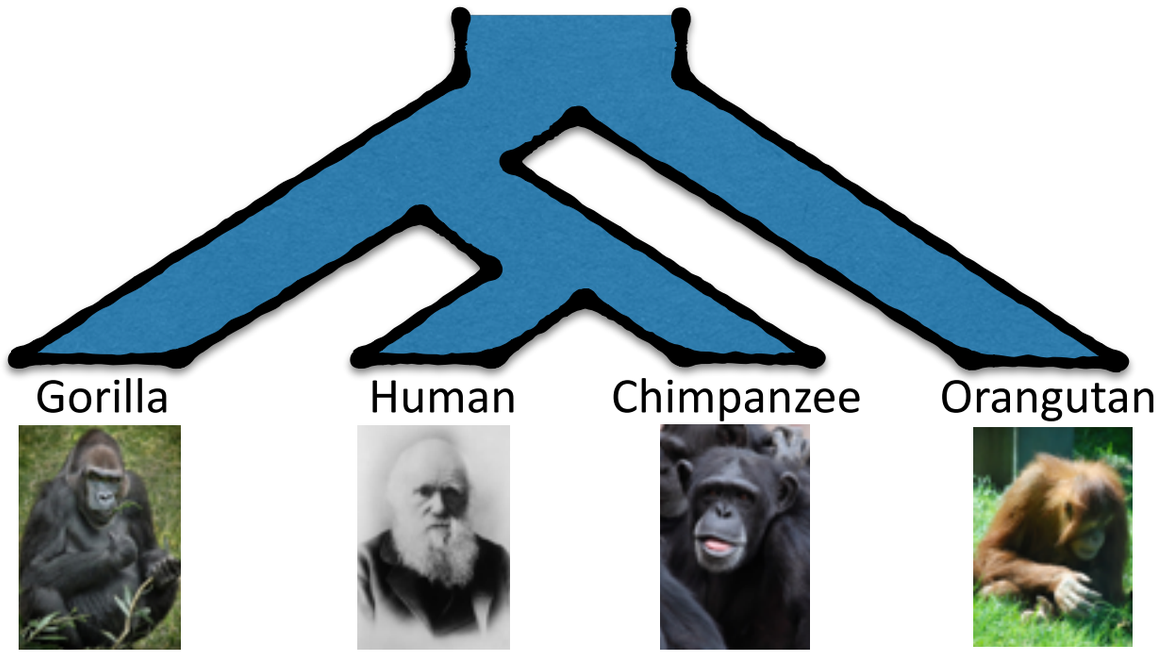
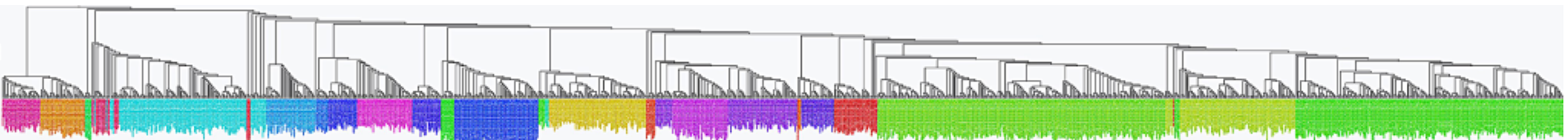
Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes

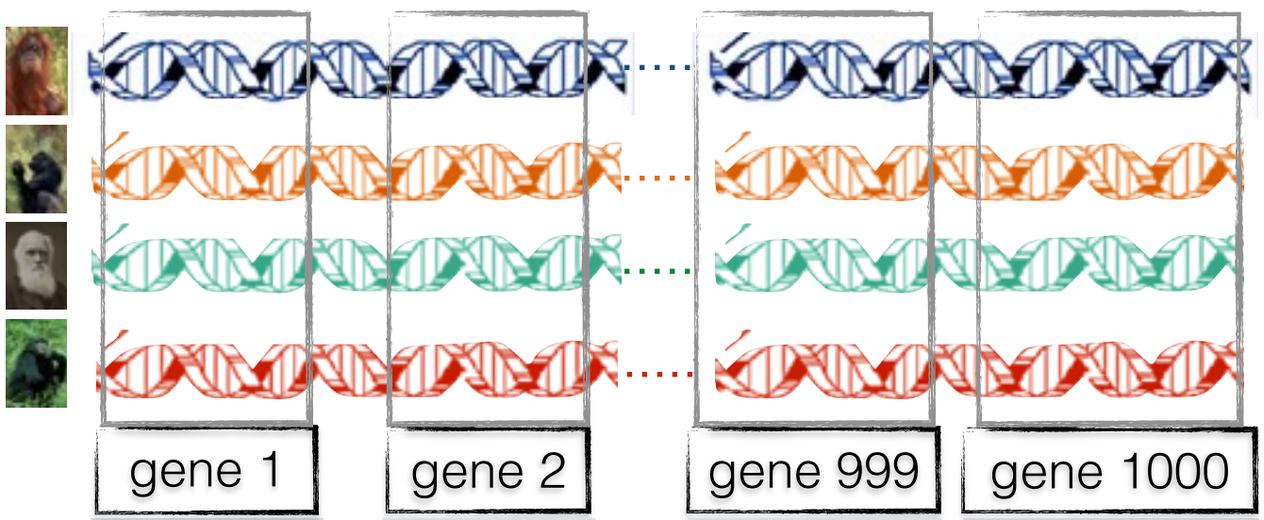
Siavash Mirarab[12], Tandy Warnow[3]

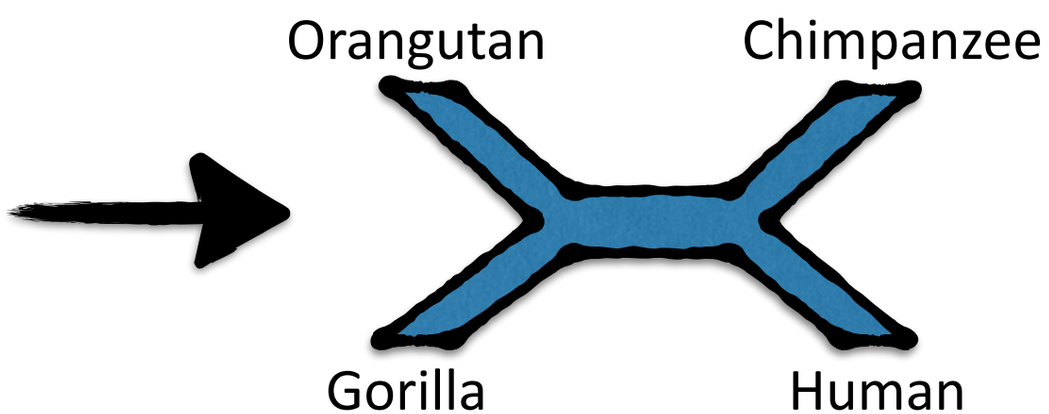
[1]University of Texas at Austin, [2]University of California San Diego,
[3]University of Illinois at Urbana-Champaign





phylogenomics

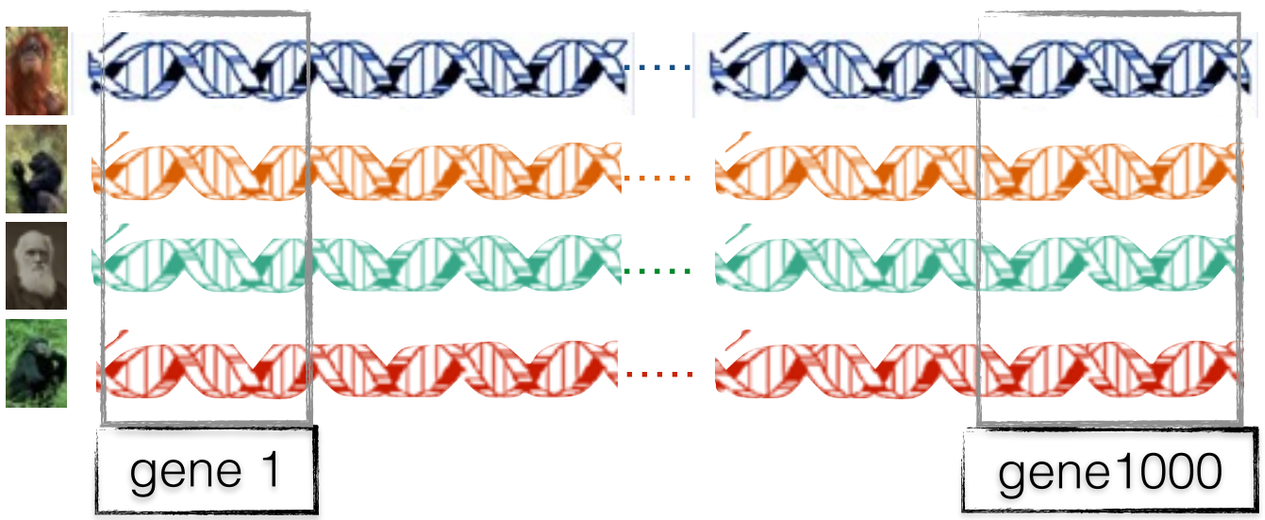




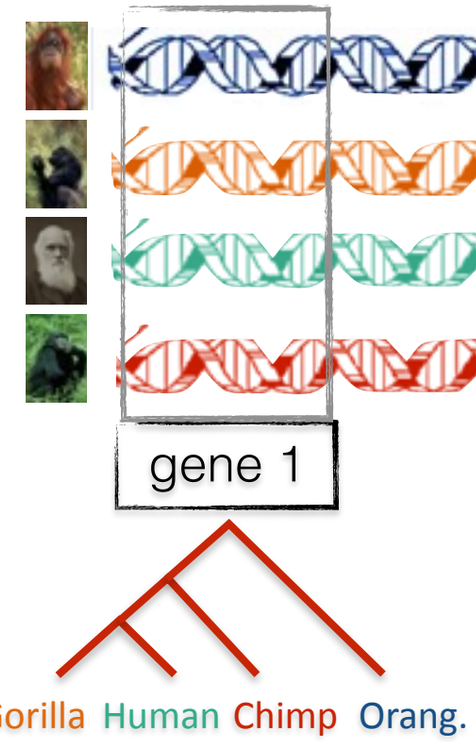

```
ACTGCACACCG   CTGAGCATCG     AGCAGCATCGTG CAGGCACGCACGAA
ACTGC-CCCCG   CTGAGC-TCG ... AGCAGC-TCGTG AGC-CACGC-CATA
AATGC-CCCCG   ATGAGC-TC-     AGCAGC-TC-TG ATGGCACGC-C-TA
-CTGCACACGG   CTGA-CAC-G     C-TA-CACGGTG AGCTAC-CACGGAT
```

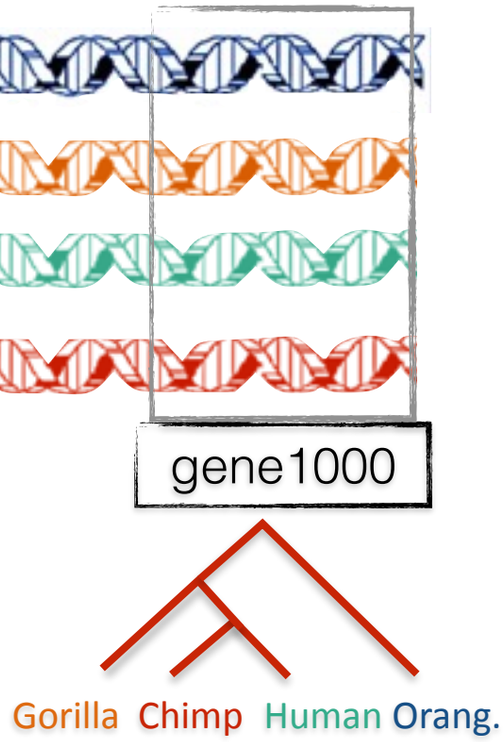
I'll use the term "gene" to refer to "c-genes":
recombination-free orthologous stretches of the genome

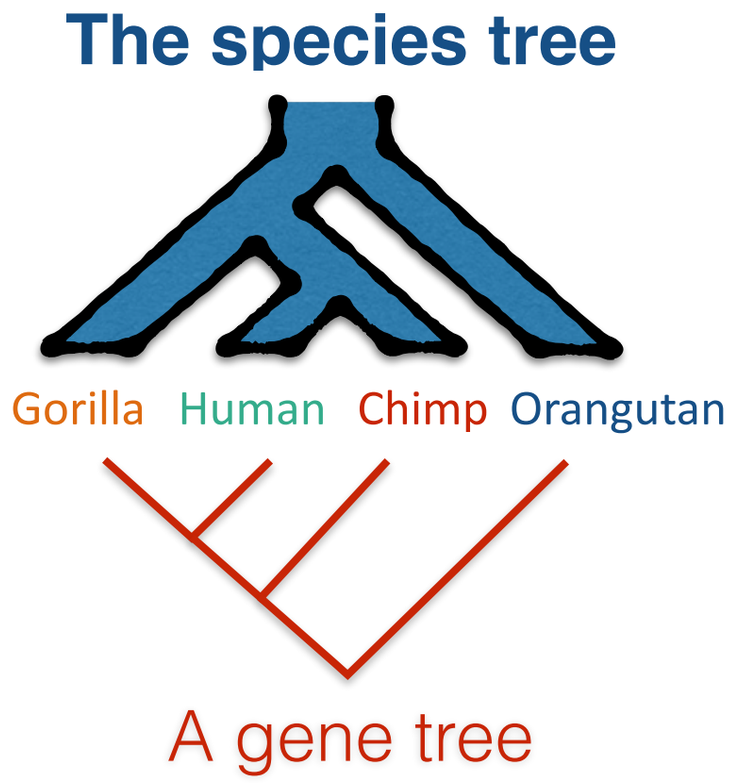
Gene tree discordance



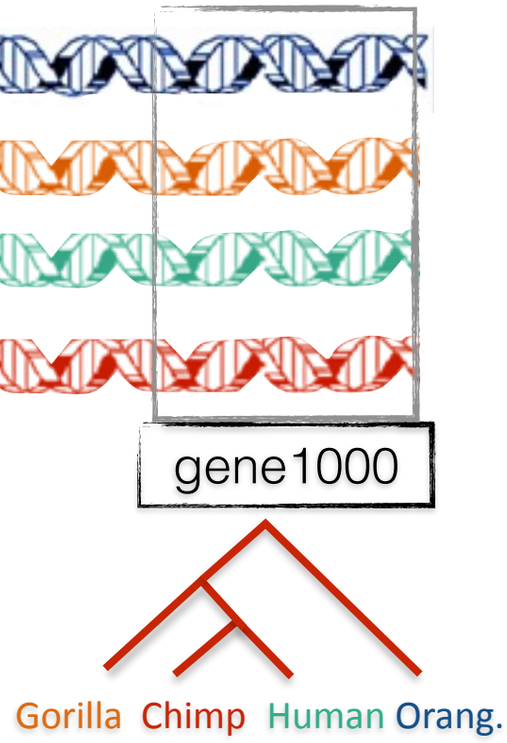
Gene tree discordance

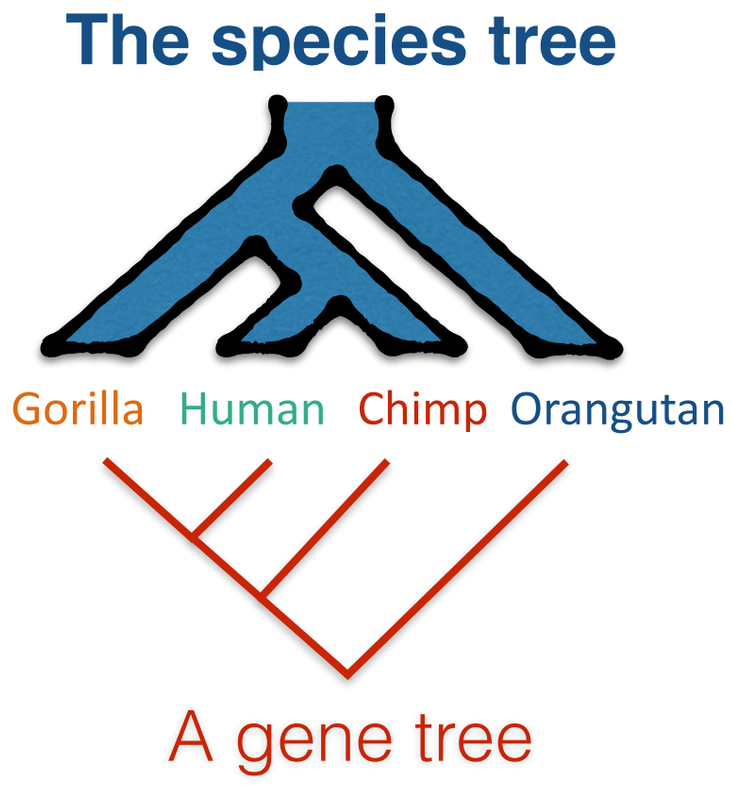






Gene tree discordance



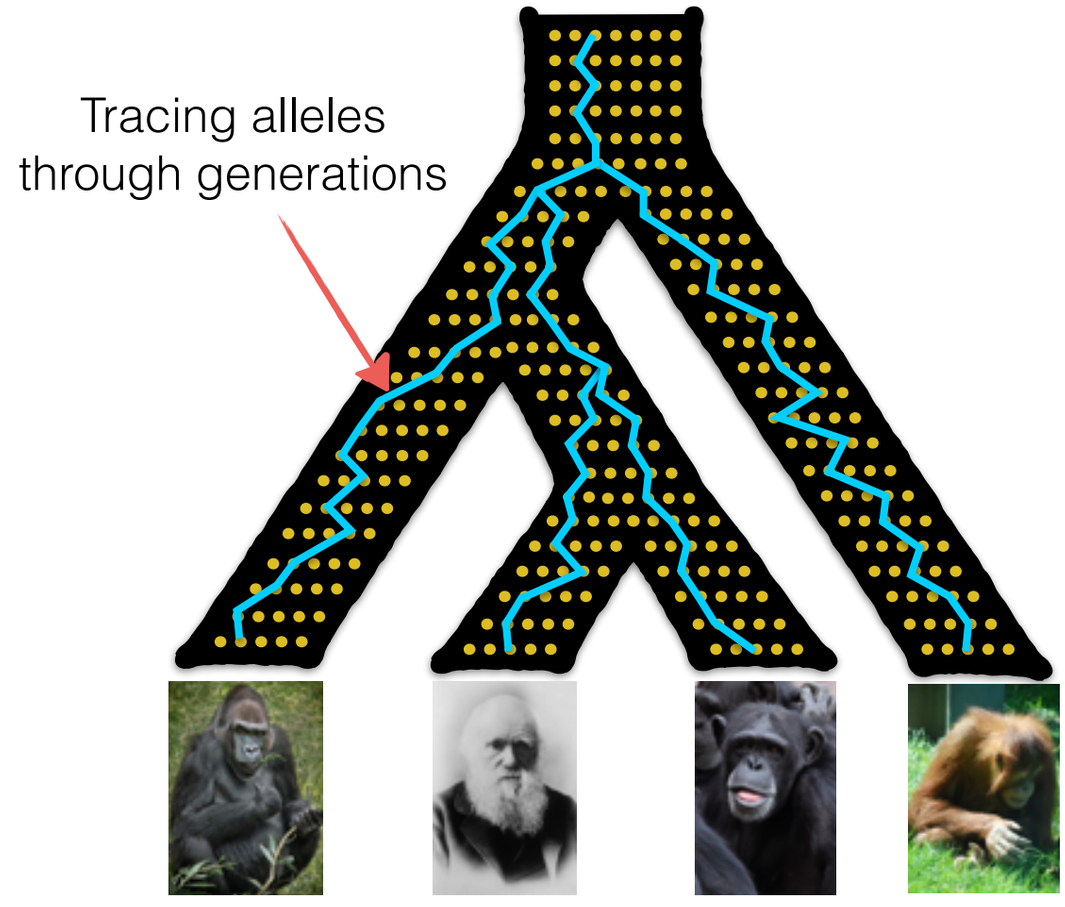


Causes of gene tree discordance include:

- Incomplete Lineage Sorting (ILS)
- Duplication and loss
- Horizontal Gene Transfer (HGT)

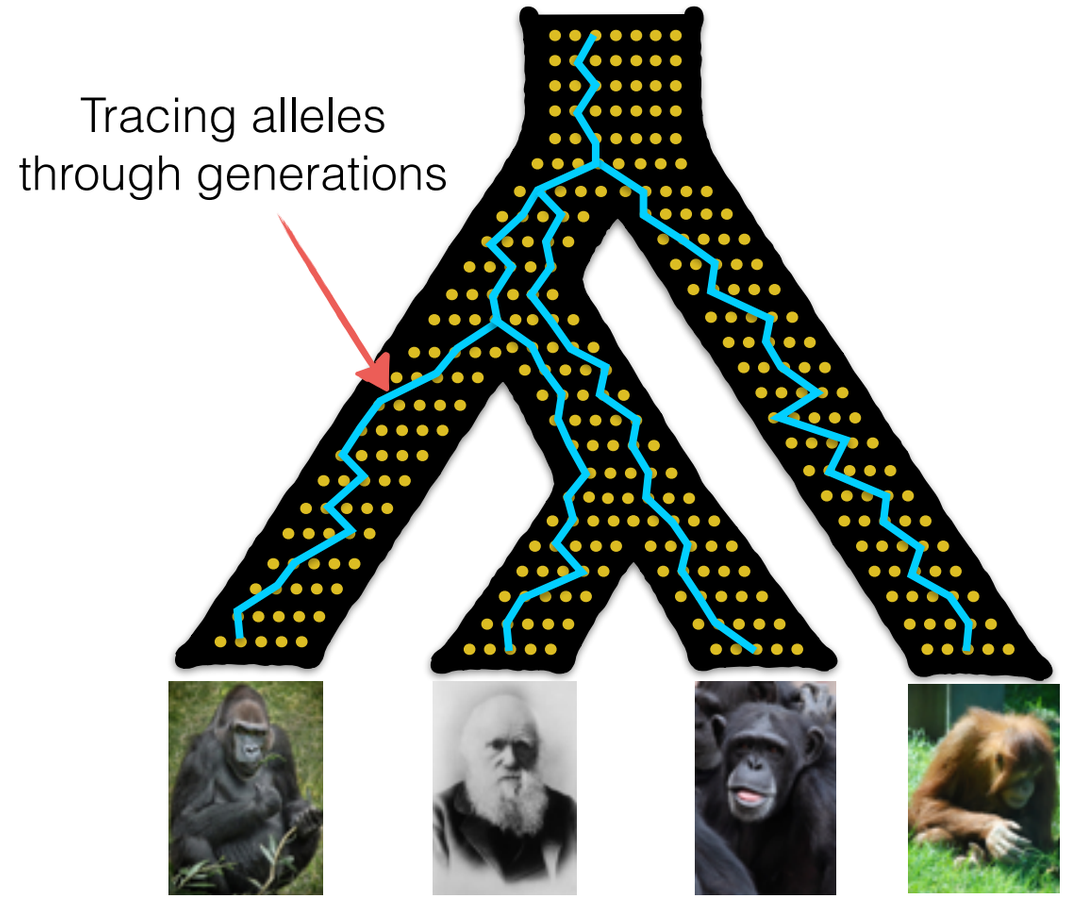
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations



Incomplete Lineage Sorting (ILS)

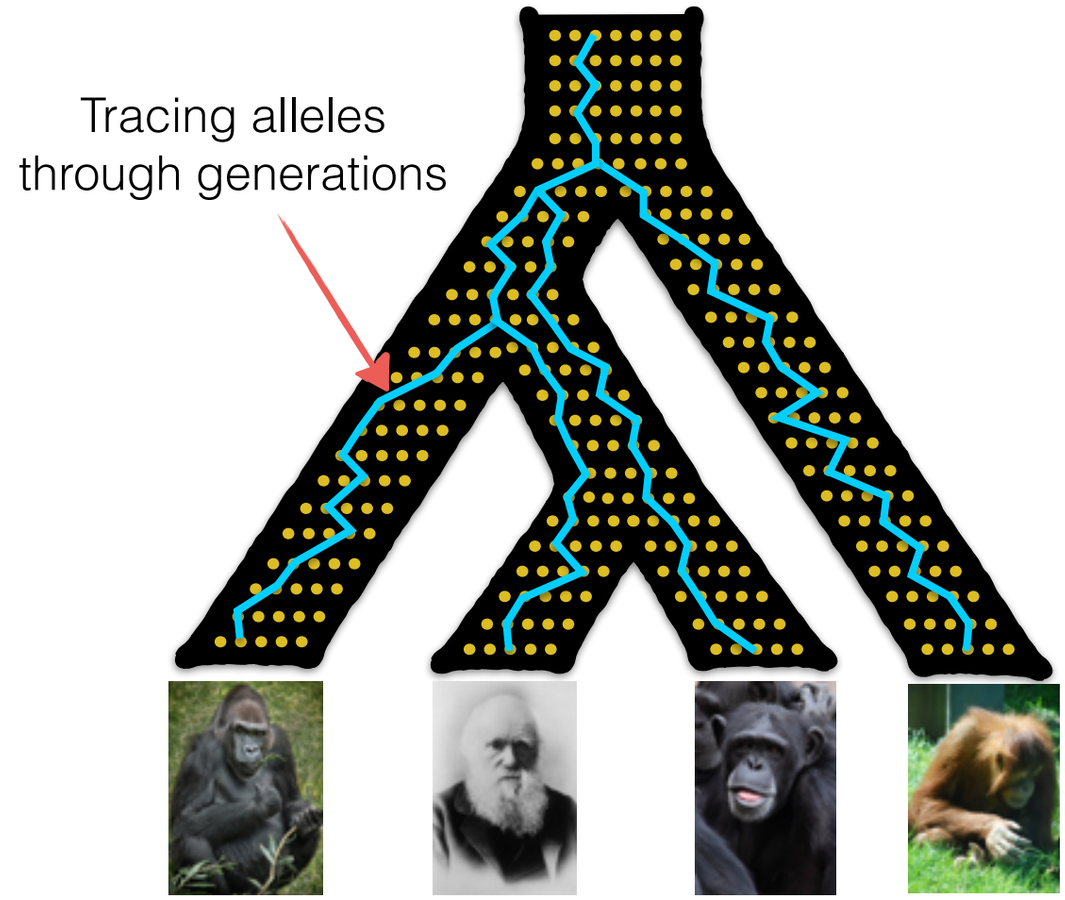
- A random process related to the coalescence of alleles across various populations



Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations

- Omnipresent; most likely for short branches or large population sizes



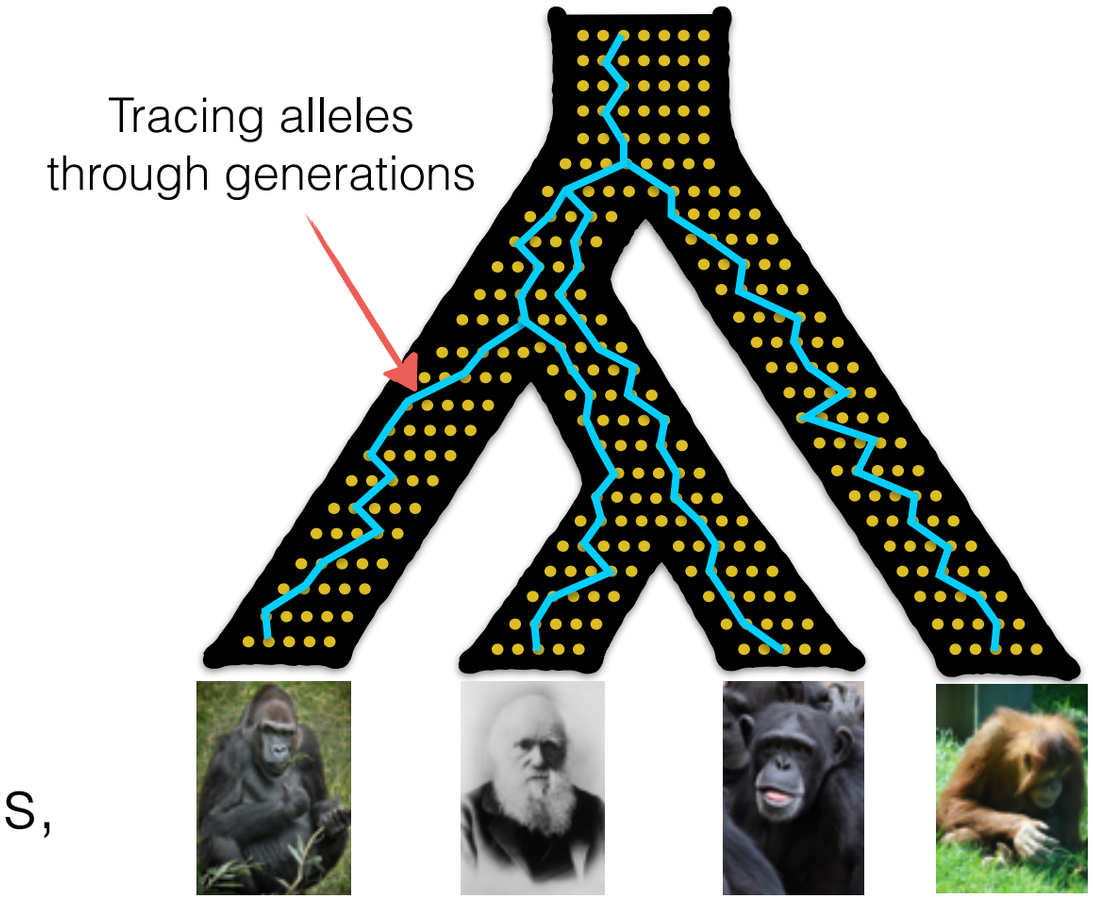
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations

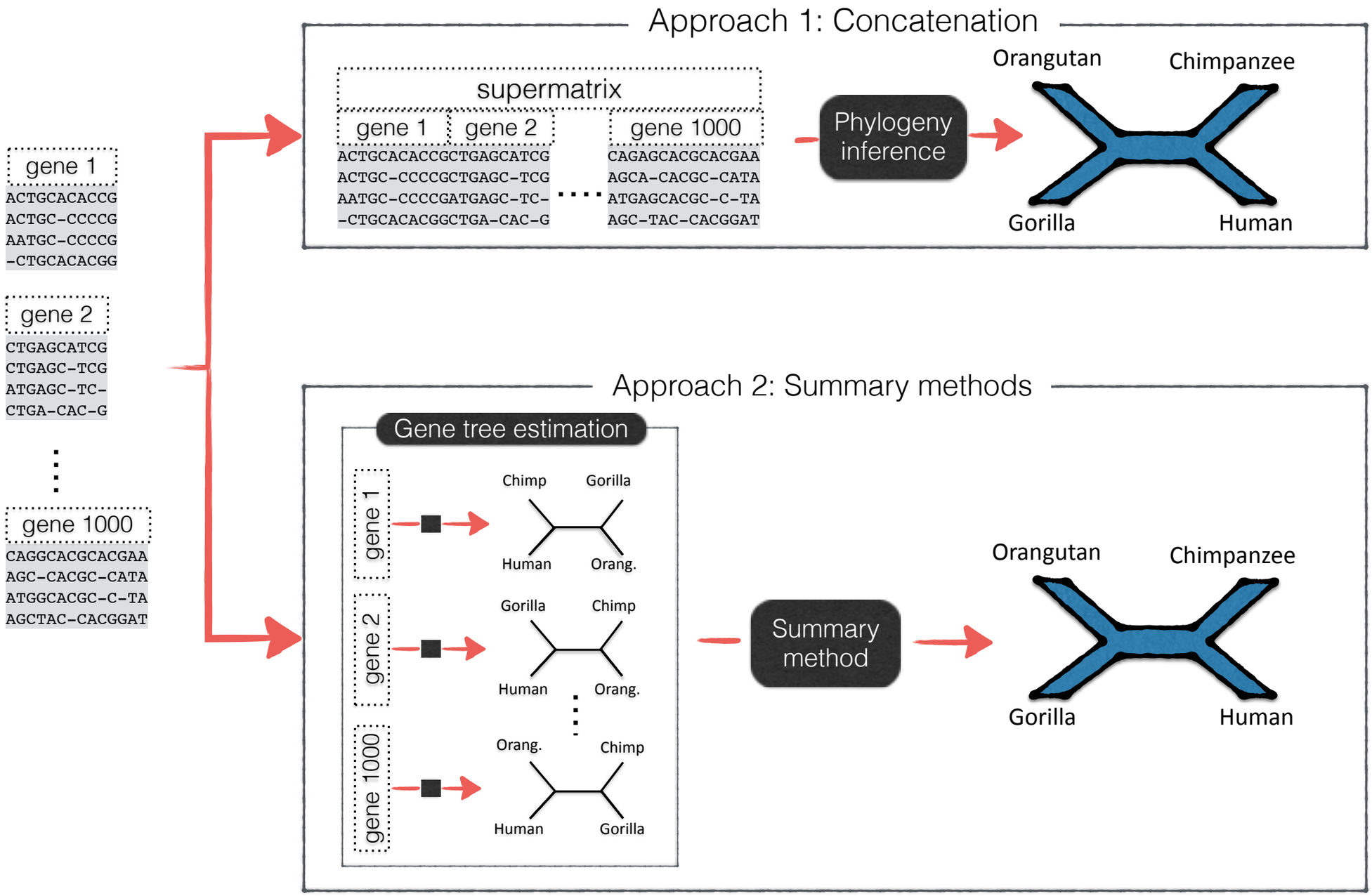
- Omnipresent; most likely for short branches or large population sizes

- We have statistical models of ILS (multi-species coalescent)

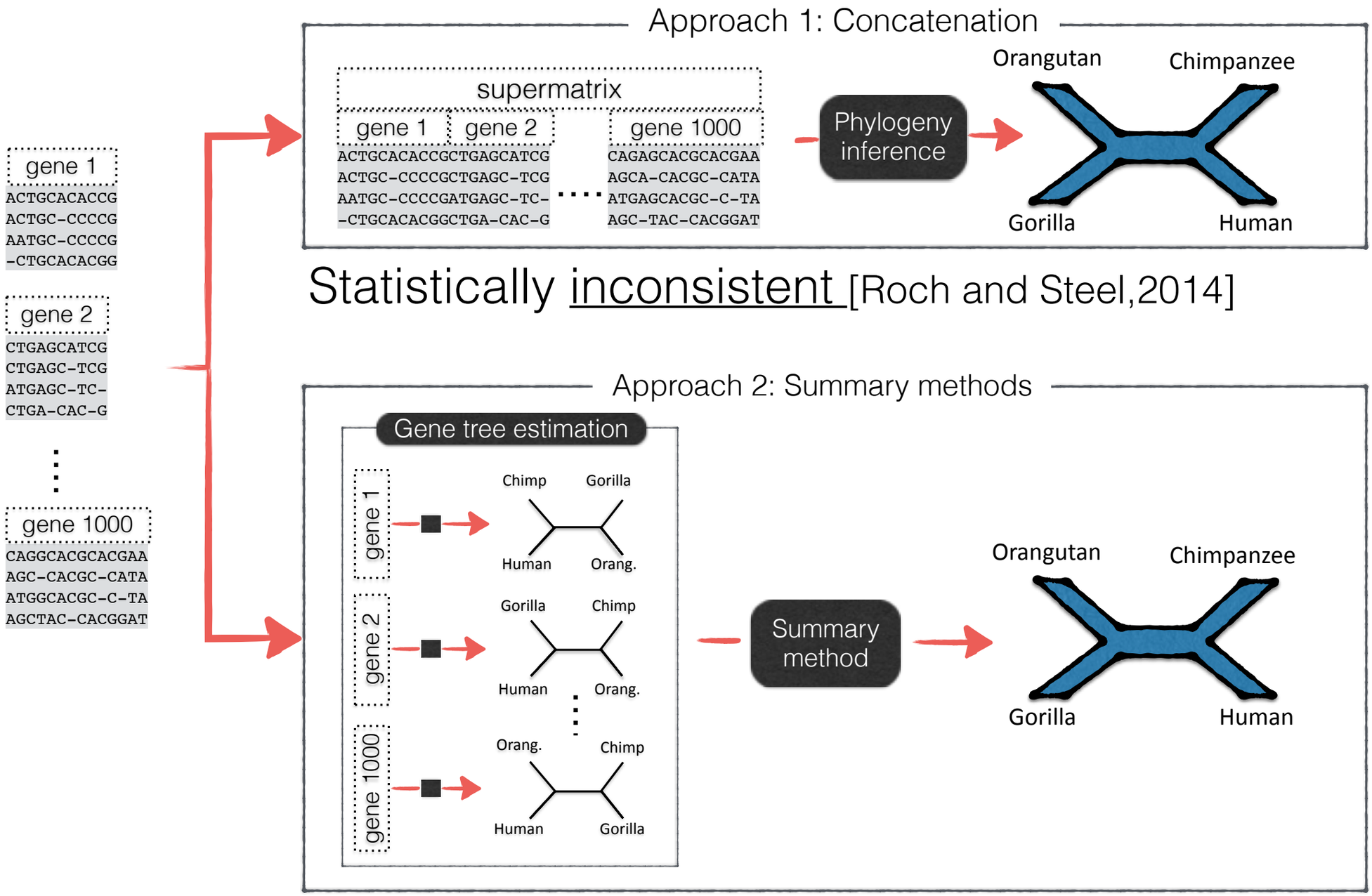
    - The species tree defines the probability distribution on gene trees, and is identifiable from the distribution on gene trees
      [Degnan and Salter, Int. J. Org. Evolution, 2005]



Multi-gene species tree estimation



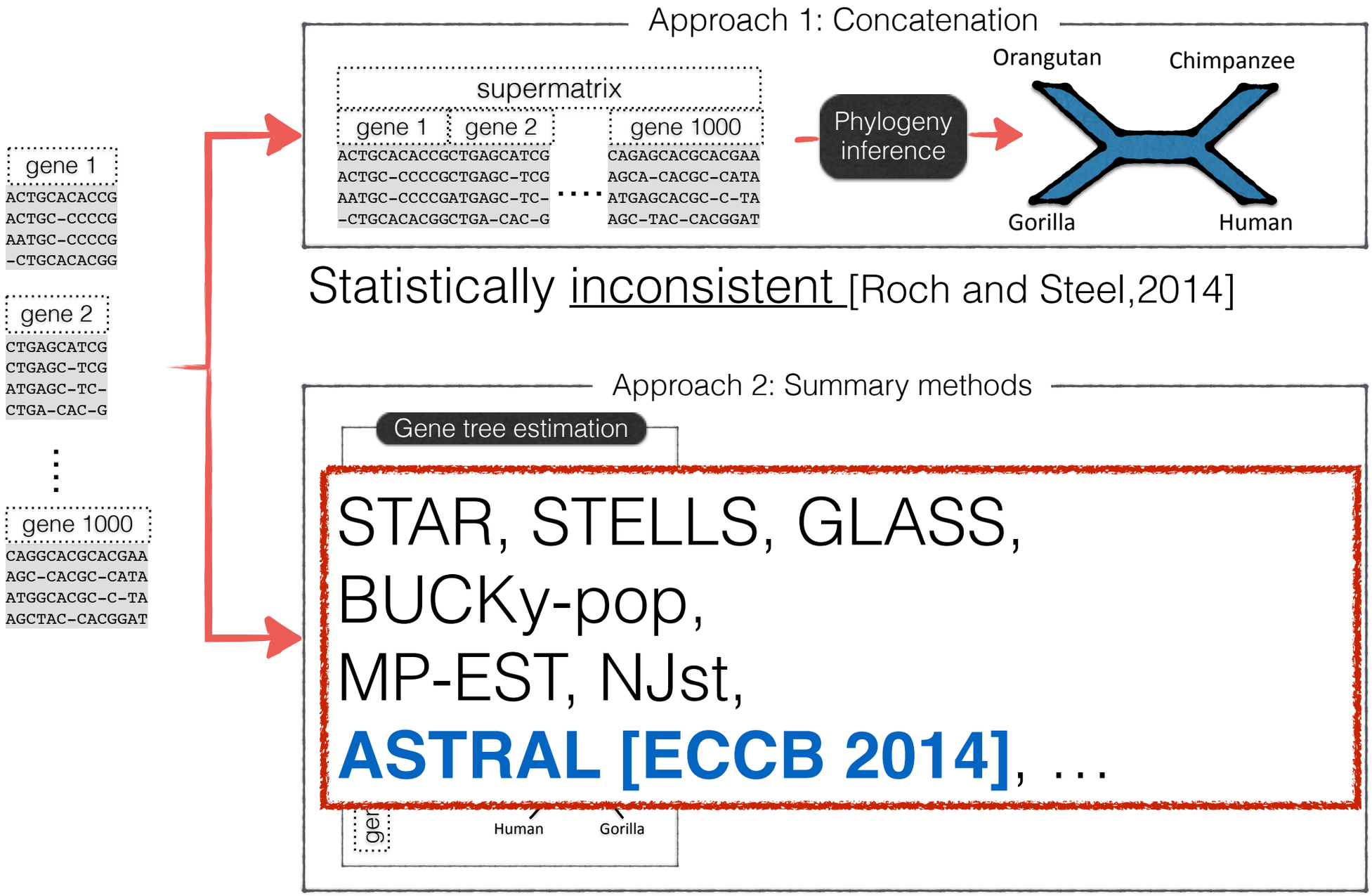
Multi-gene species tree estimation



Statistically <u>inconsistent</u> [Roch and Steel,2014]

Can be statistically consistent given true gene trees

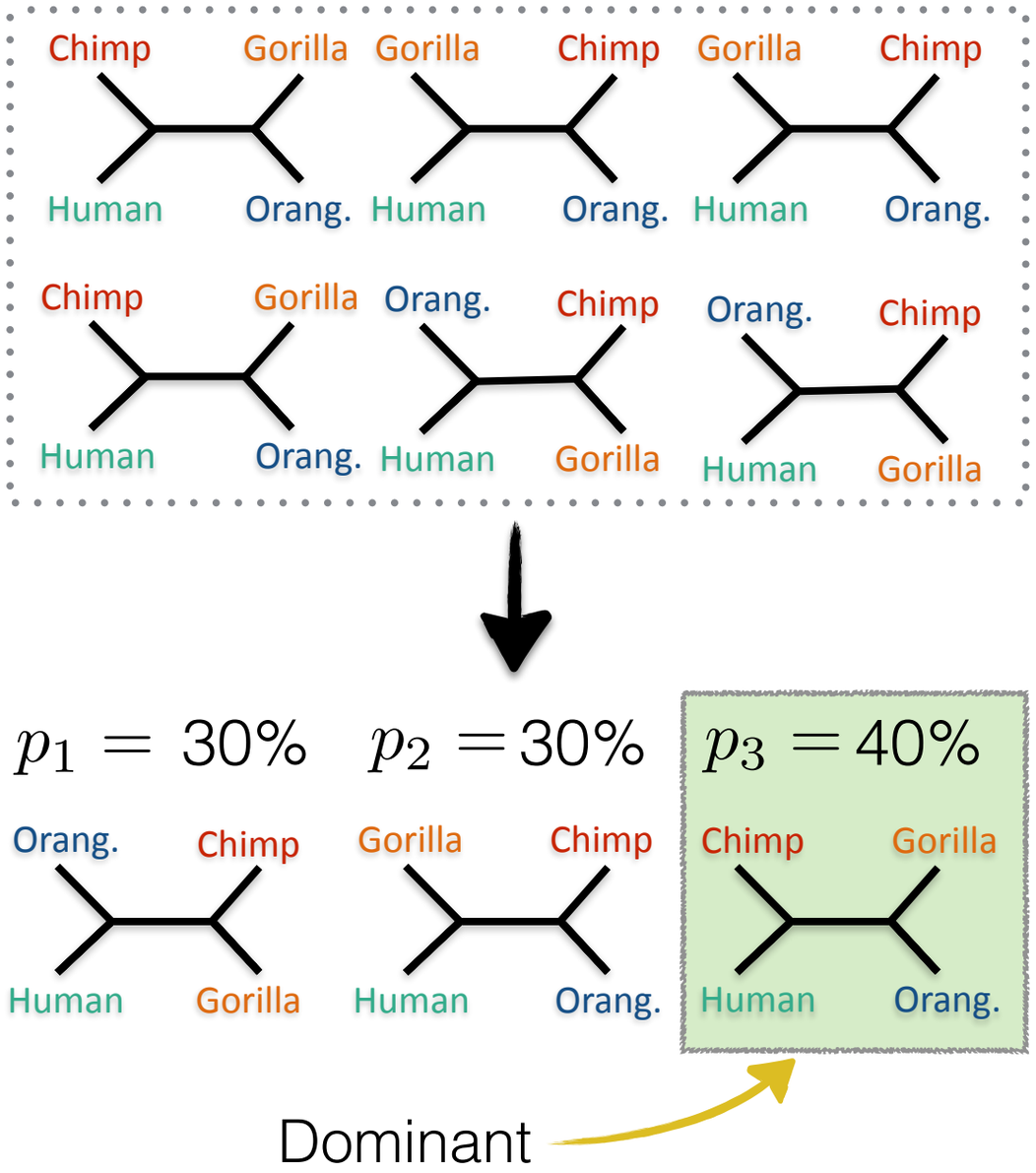
Multi-gene species tree estimation



Statistically <u>inconsistent</u> [Roch and Steel,2014]

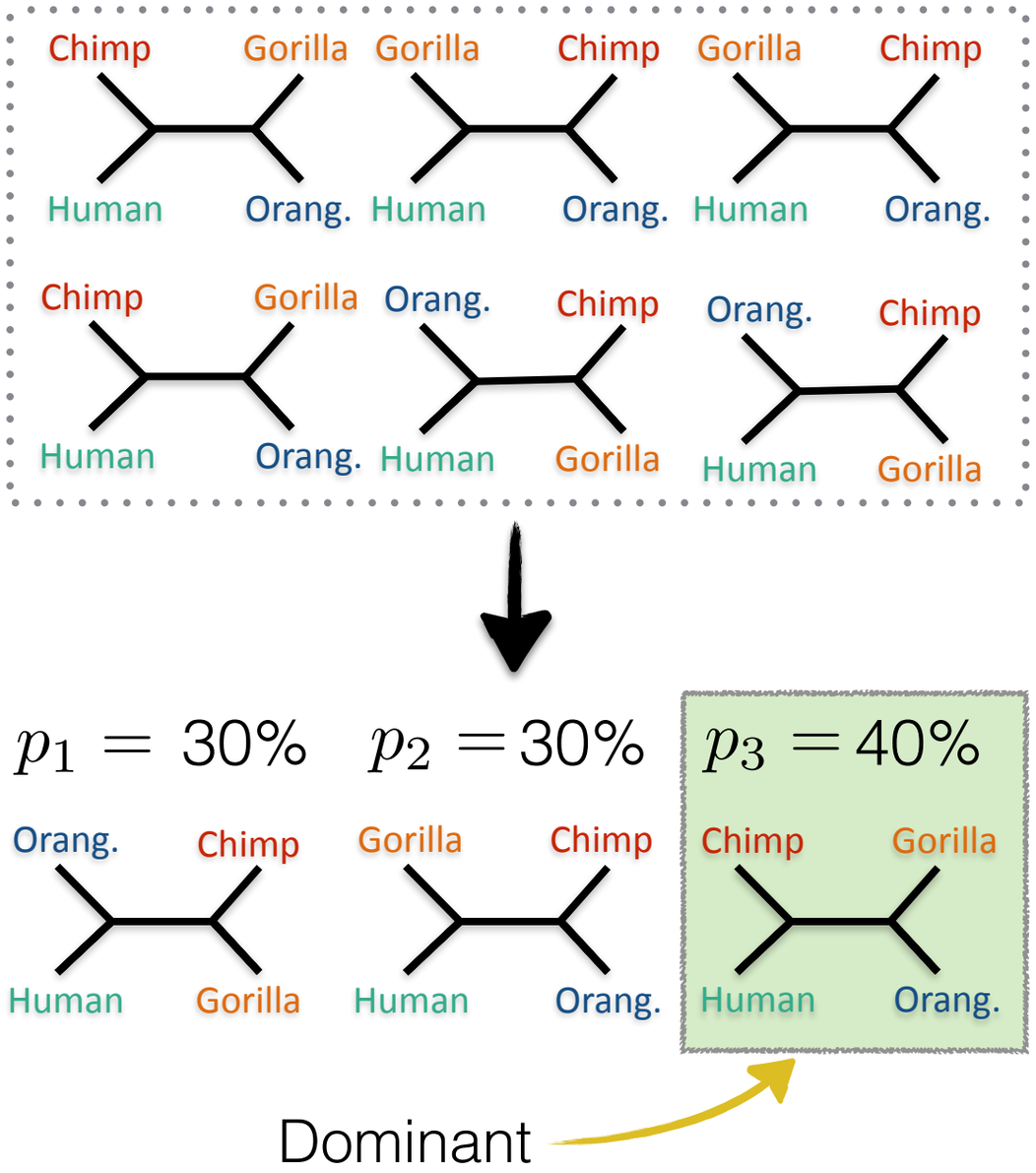
Can be statistically consistent given true gene trees

Properties of quartet trees in presence of ILS



- For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]

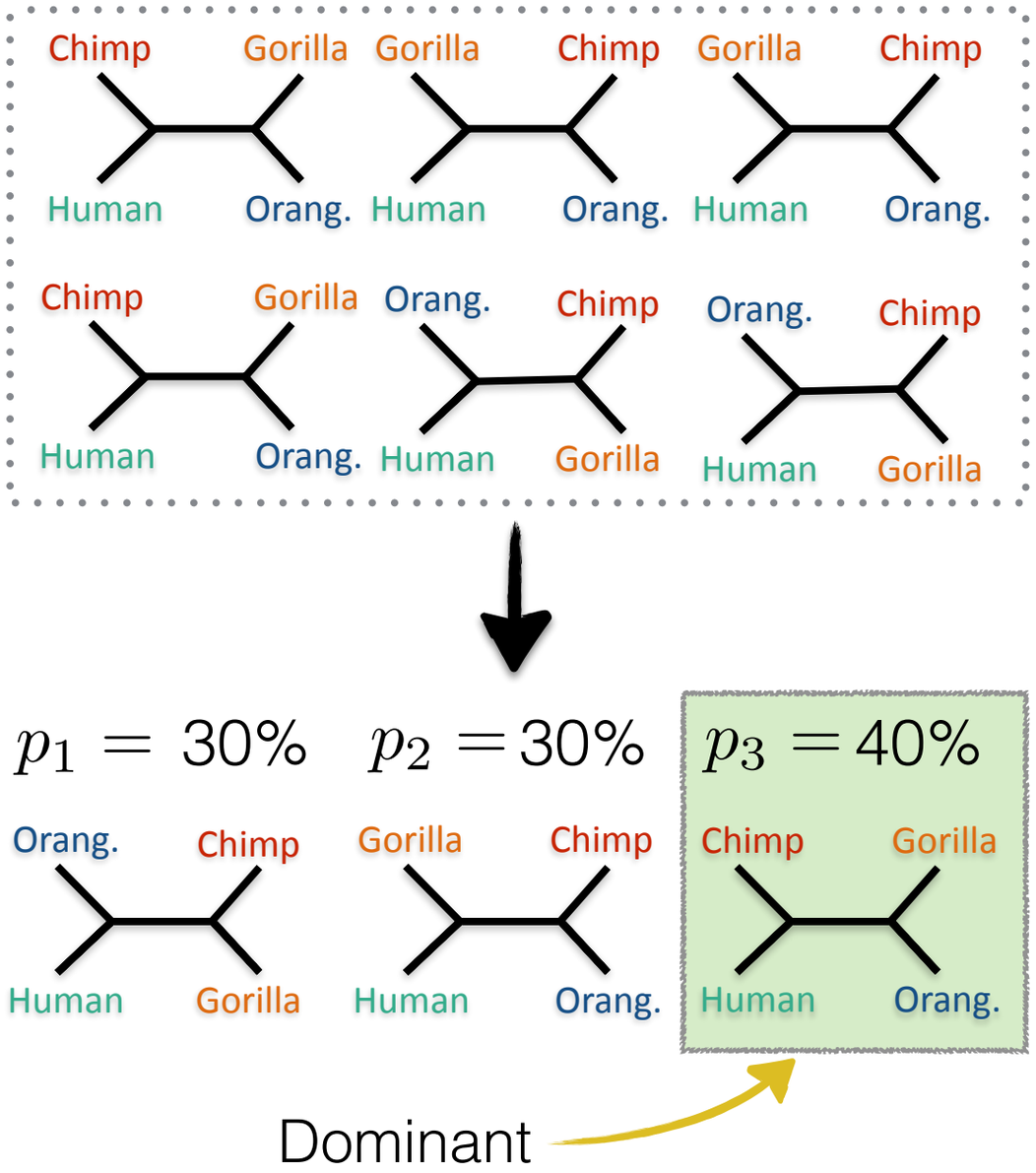
Properties of quartet trees in presence of ILS



- For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]
- For >4 species, the dominant topology may be different from the species tree [Degnan and Rosenberg, 2006]

1. Breakup input each gene tree into $\binom{n}{4}$ trees on 4 taxa (quartet trees)
2. Find all $\binom{n}{4}$ dominant quartet topologies
3. Combine dominant quartet trees

Properties of quartet trees in presence of ILS



- For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]
- For >4 species, the dominant topology may be different from the species tree [Degnan and Rosenberg, 2006]
  1. Breakup input each gene tree into $\binom{n}{4}$ trees on 4 taxa (quartet trees)
  2. Find all $\binom{n}{4}$ dominant quartet topologies
  3. Combine dominant quartet trees
- Alternative: weight 3 $\binom{n}{4}$ quartet topology by their frequency and find the optimal tree

Maximum Quartet Support Species Tree

[Mirarab, et al., ECCB, 2014]

- Optimization Problem (suspected NP-Hard):

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

$$Score(T) = \sum_{t \in \mathcal{T}} |Q(T) \cap Q(t)|$$

Set of quartet trees induced by T

a gene tree

all input gene trees

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

ASTRAL-I

[Mirarab, et al., ECCB, 2014]

- ASTRAL solves the problem exactly using dynamic programming:
  - Exponential running time (feasible for <18 species)

ASTRAL-I

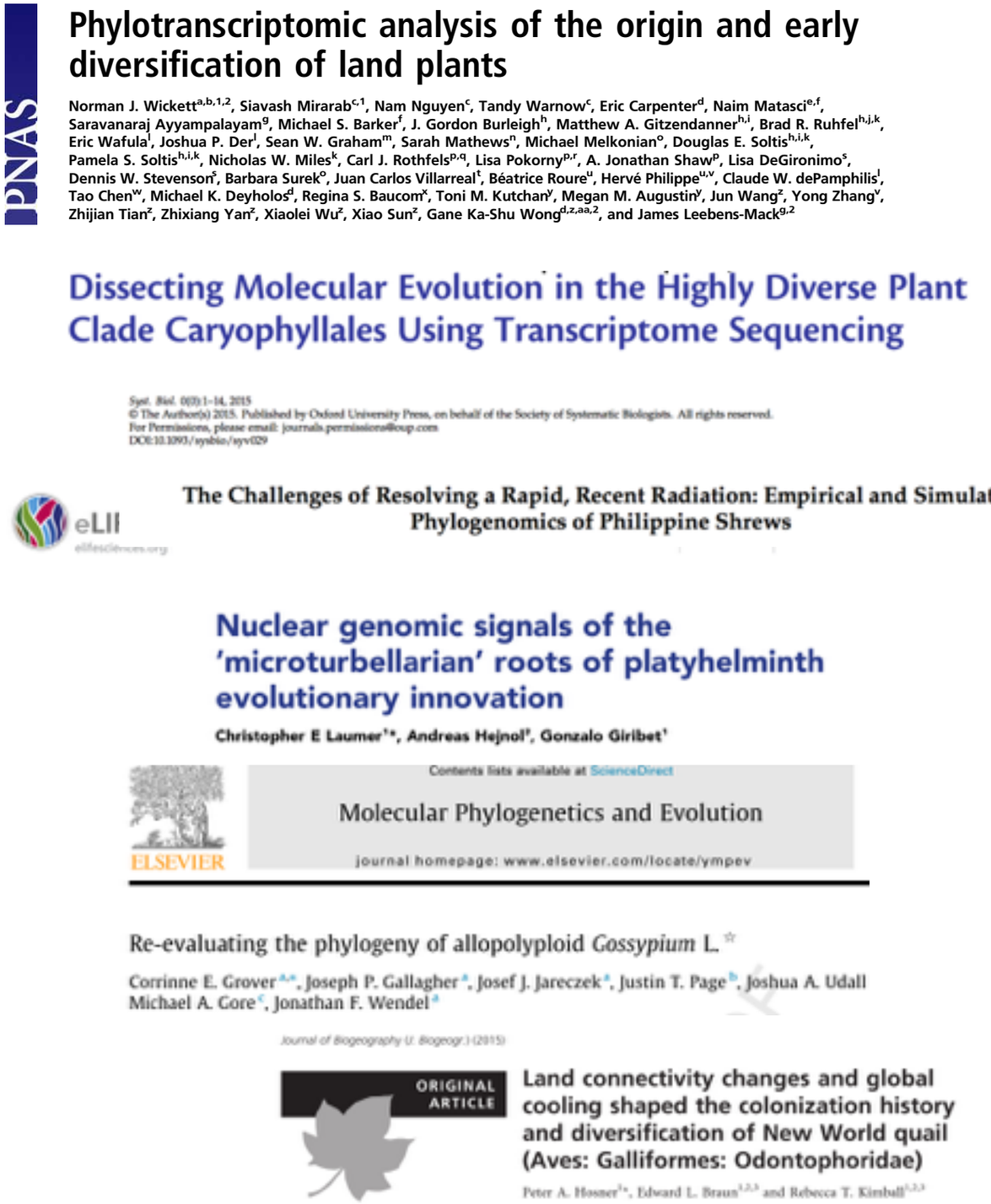
[Mirarab, et al., ECCB, 2014]

- ASTRAL solves the problem exactly using dynamic programming:
  - Exponential running time (feasible for <18 species)

- Introduced a constrained version of the problem
  - Draws the set of branches in the species tree from a given set $\mathcal{X}$ = {all bipartitions in all gene trees}
  - Given many genes, each species tree branch likely appears in at least one of the gene trees
  - Theorem: the constrained version remains statistically consistent
  - Running time: $O(n^2 k|\mathcal{X}|^2)$ for *n* species and *k* species

ASTRAL-I on biological datasets

- 1KP: **103** plant species, 400-800 genes
- Yang, et al. **96** Caryophyllales species, 1122 genes
- Dentinger, et al. **39** mushroom species, 208 genes
- Giarla and Esselstyn. **19** Philippine shrew species, 1112 genes
- Laumer, et al. **40** flatworm species, 516 genes
- Grover, et al. **8** cotton species, 52 genes
- Hosner, Braun, and Kimball. **28** quail species, 11 genes
- Simmons and Gatesy. **47** angiosperm species, 310 genes



Future datasets

- **1200** plants with ~ 400 genes (1KP consortium)
- **250** avian species with 2000 genes (with LSU, UF, and Smithsonian)
- **200** avian species with whole genomes (with Genome 10K, international)
- **250** suboscine species (birds) with ~2000 genes (with LSU and Tulane)
- **140** Insects with 1400 genes (with U. Illinois at Urbana-Champaign)

Shortcomings of ASTRAL-I

- Even the constrained version was too slow for more than about 200 species and hundreds of genes

- The constraint set $\mathcal{X}$ did not include true species tree branches for some challenging datasets, resulting in low accuracy in some cases

- Input gene trees could not have polytomies

ASTRAL-II

[Mirarab and Warnow, ISMB/ECCB, 2015]

1. Faster calculation of the score function inside DP

   - $O(nk|\mathcal{X}|^2)$ instead of $O(n^2k|\mathcal{X}|^2)$ for *n* species and *k* genes

   - Post-order traversal of input trees instead of set operations

ASTRAL-II

[Mirarab and Warnow, ISMB/ECCB, 2015]

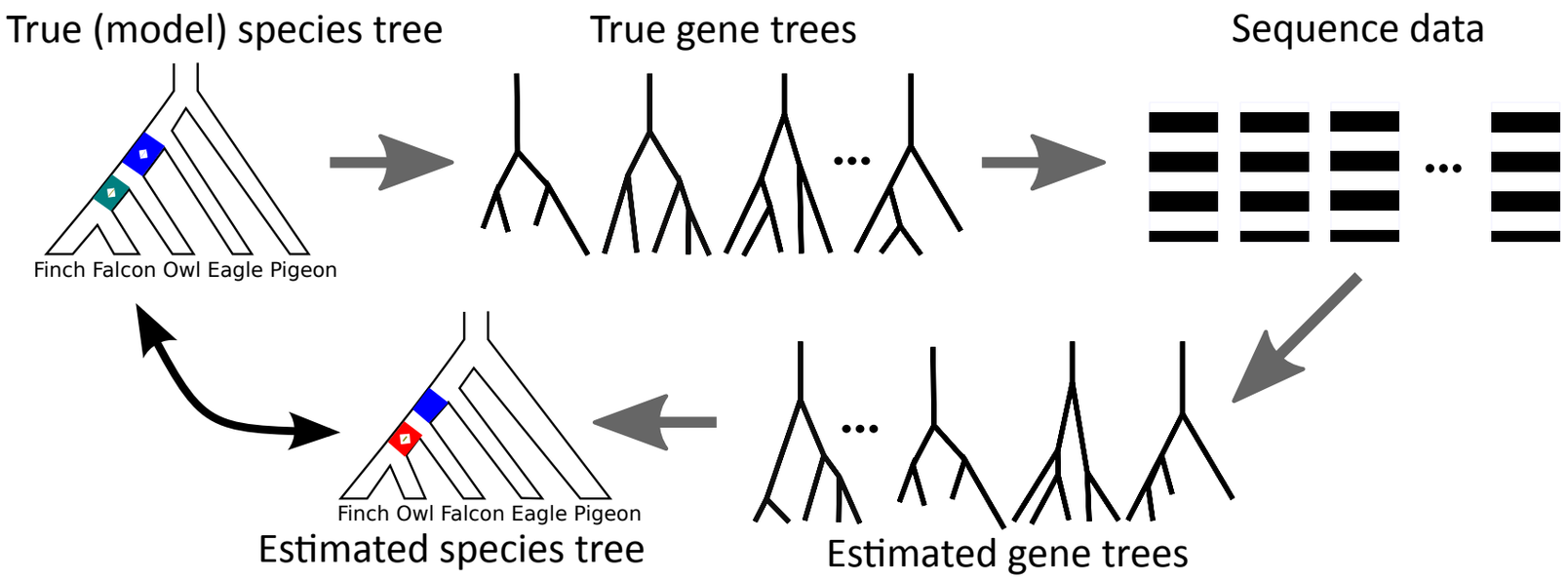
1. Faster calculation of the score function inside DP
   - $O(nk|\mathcal{X}|^2)$ instead of $O(n^2k|\mathcal{X}|^2)$ for *n* species and *k* genes
   - Post-order traversal of input trees instead of set operations
2. Add extra bipartitions to the set $\mathcal{X}$ using heuristic approaches
   - Resolving consensus trees by subsampling taxa
   - Using quartet-based distances to find likely branches

ASTRAL-II

[Mirarab and Warnow, ISMB/ECCB, 2015]

1. Faster calculation of the score function inside DP
   - $O(nk|\mathcal{X}|^2)$ instead of $O(n^2k|\mathcal{X}|^2)$ for *n* species and *k* genes
   - Post-order traversal of input trees instead of set operations
2. Add extra bipartitions to the set $\mathcal{X}$ using heuristic approaches
   - Resolving consensus trees by subsampling taxa
   - Using quartet-based distances to find likely branches
3. Ability to take as input gene trees with polytomies

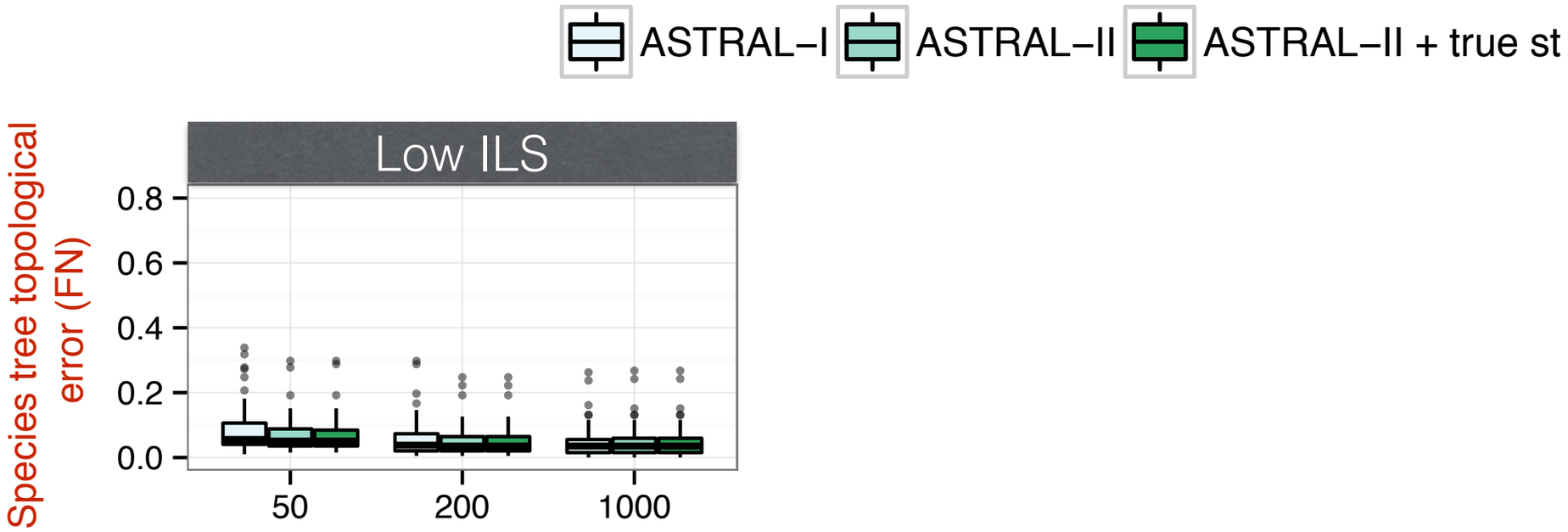
Simulation study



- Using SimPhy. Vary many parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation

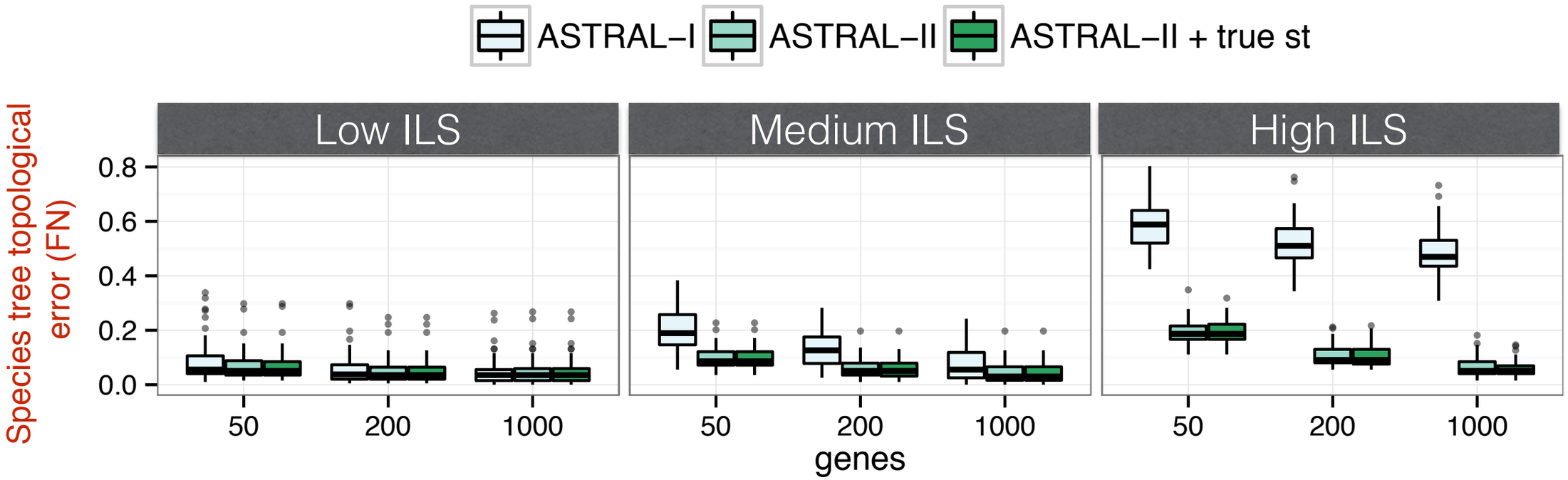
- 11 model conditions (50 replicas each) with heterogenous gene tree error
- Compare to NJst, MP-EST, concatenation (CA-ML) using FastTree-II
- Evaluate accuracy using FN rate: the percentage of branches in the true tree that are missing from the estimated tree

ASTRAL-I versus ASTRAL-II



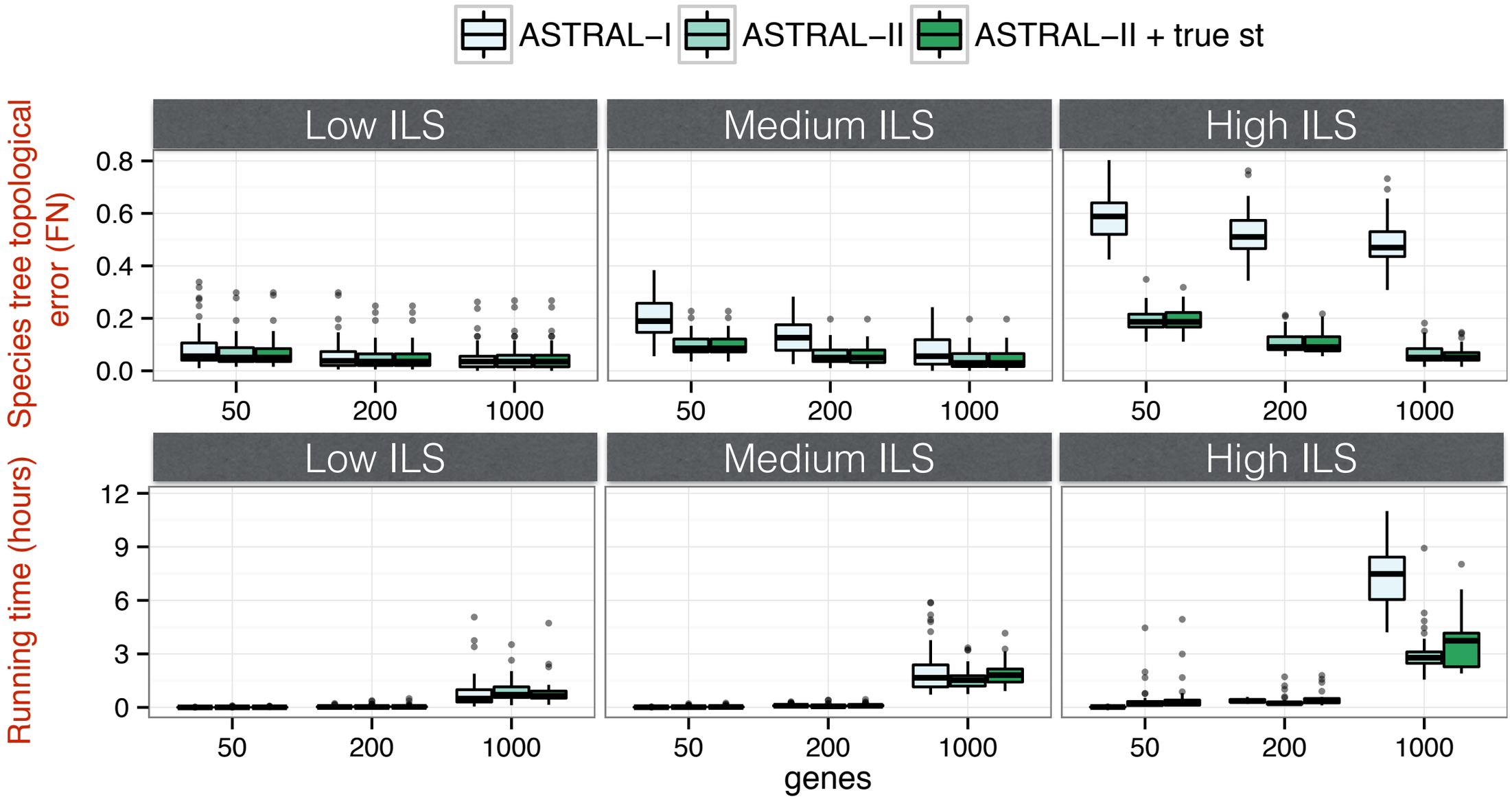
200 species, deep ILS

ASTRAL-I versus ASTRAL-II



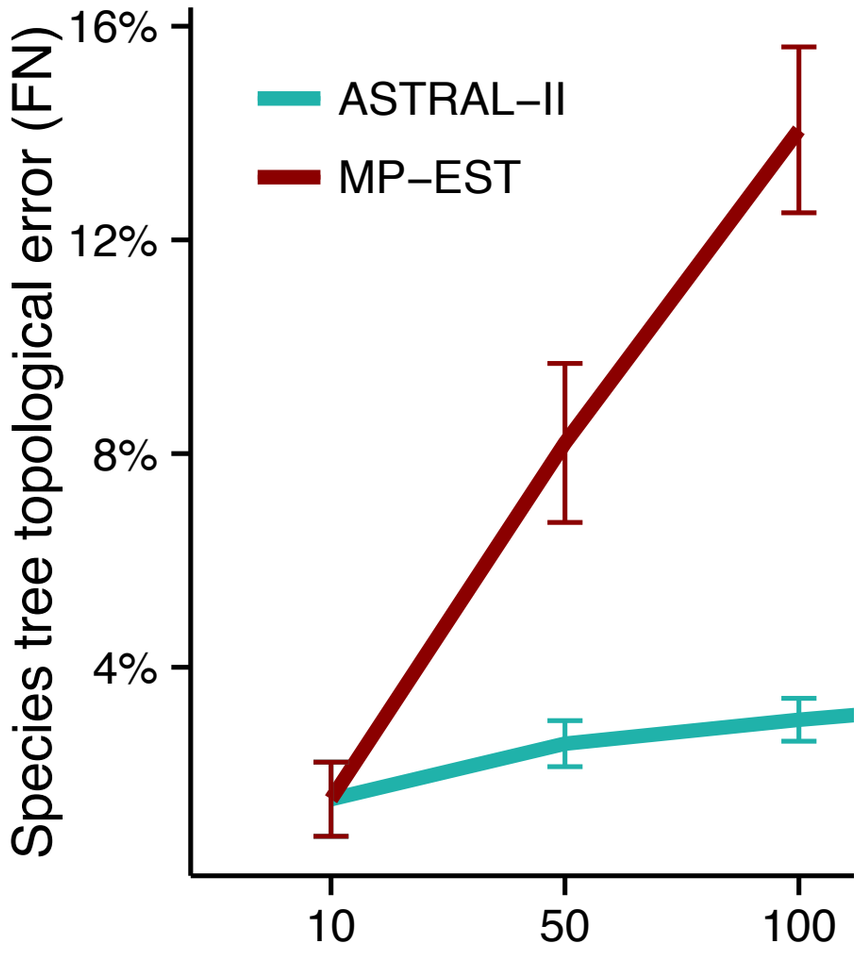
200 species, deep ILS

ASTRAL-I versus ASTRAL-II



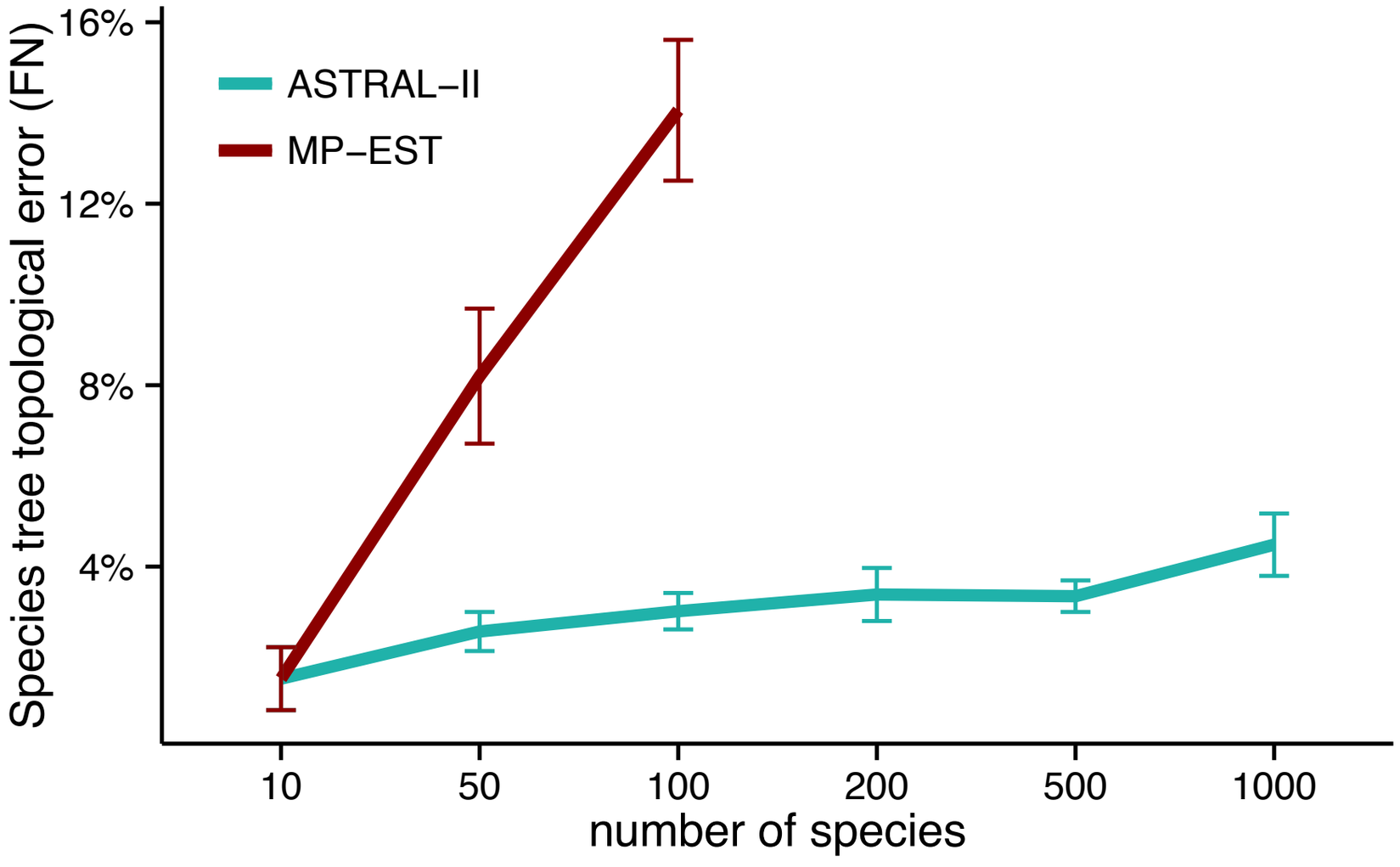
200 species, deep ILS

Tree error, varying # of species



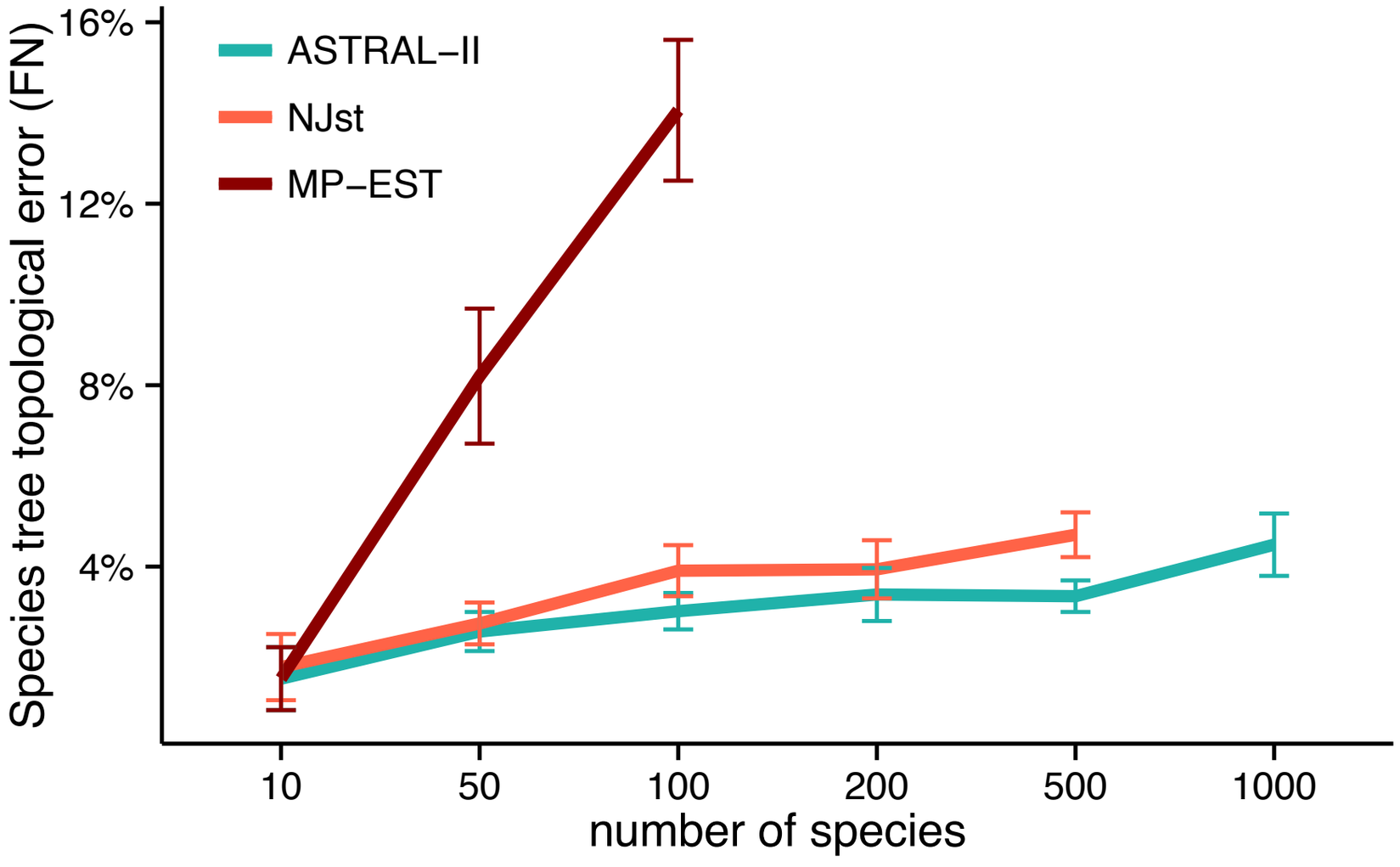
1000 genes, “medium” levels of recent ILS

Tree error, varying # of species



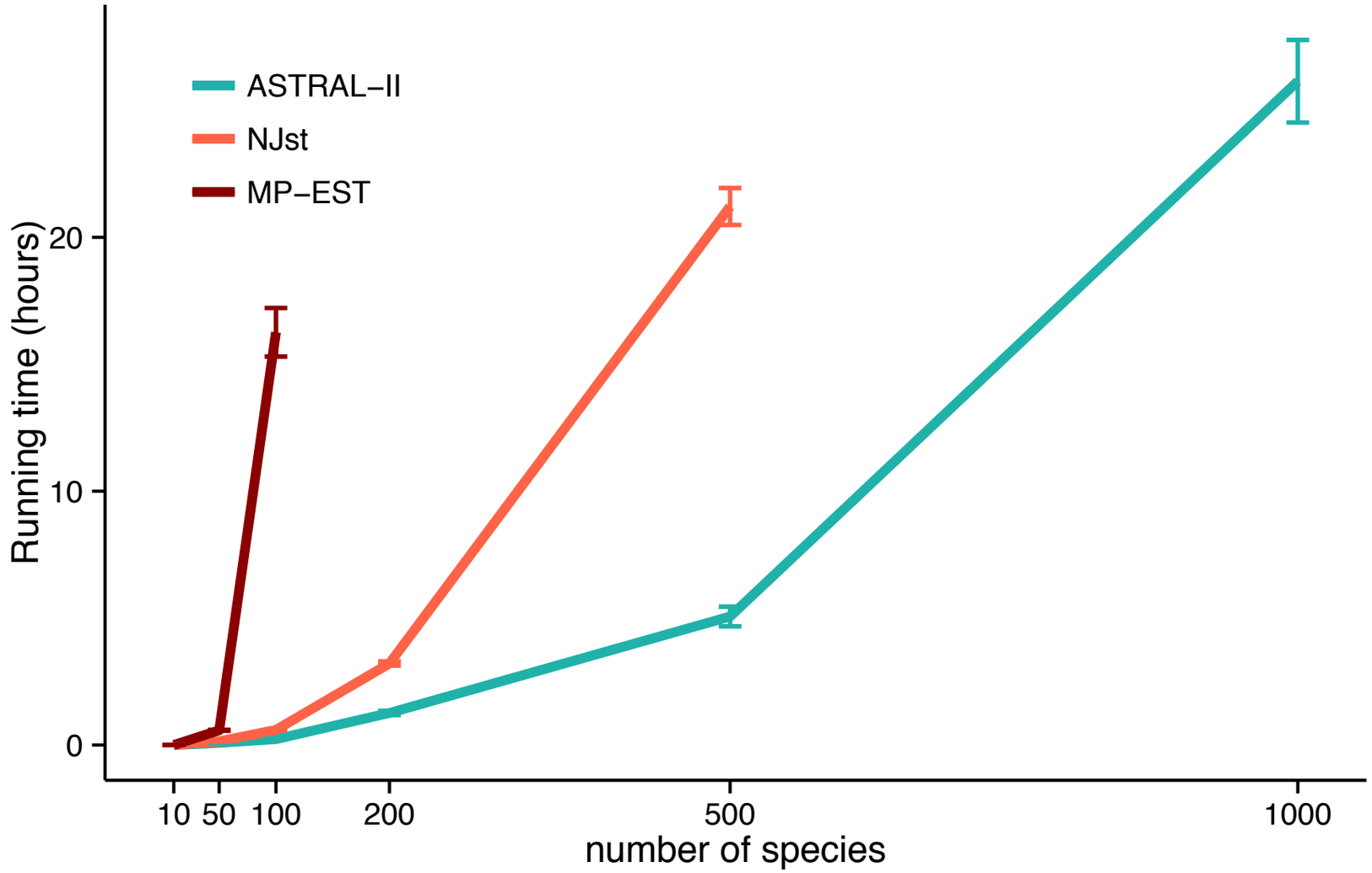
1000 genes, “medium” levels of recent ILS

Tree error, varying # of species



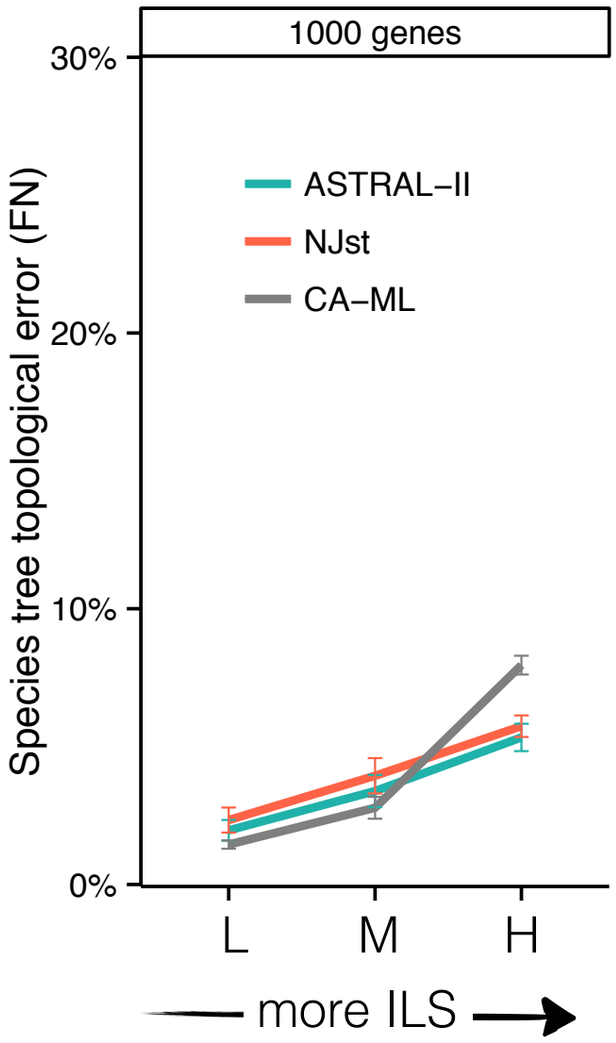
1000 genes, “medium” levels of recent ILS

Running time, varying # of species



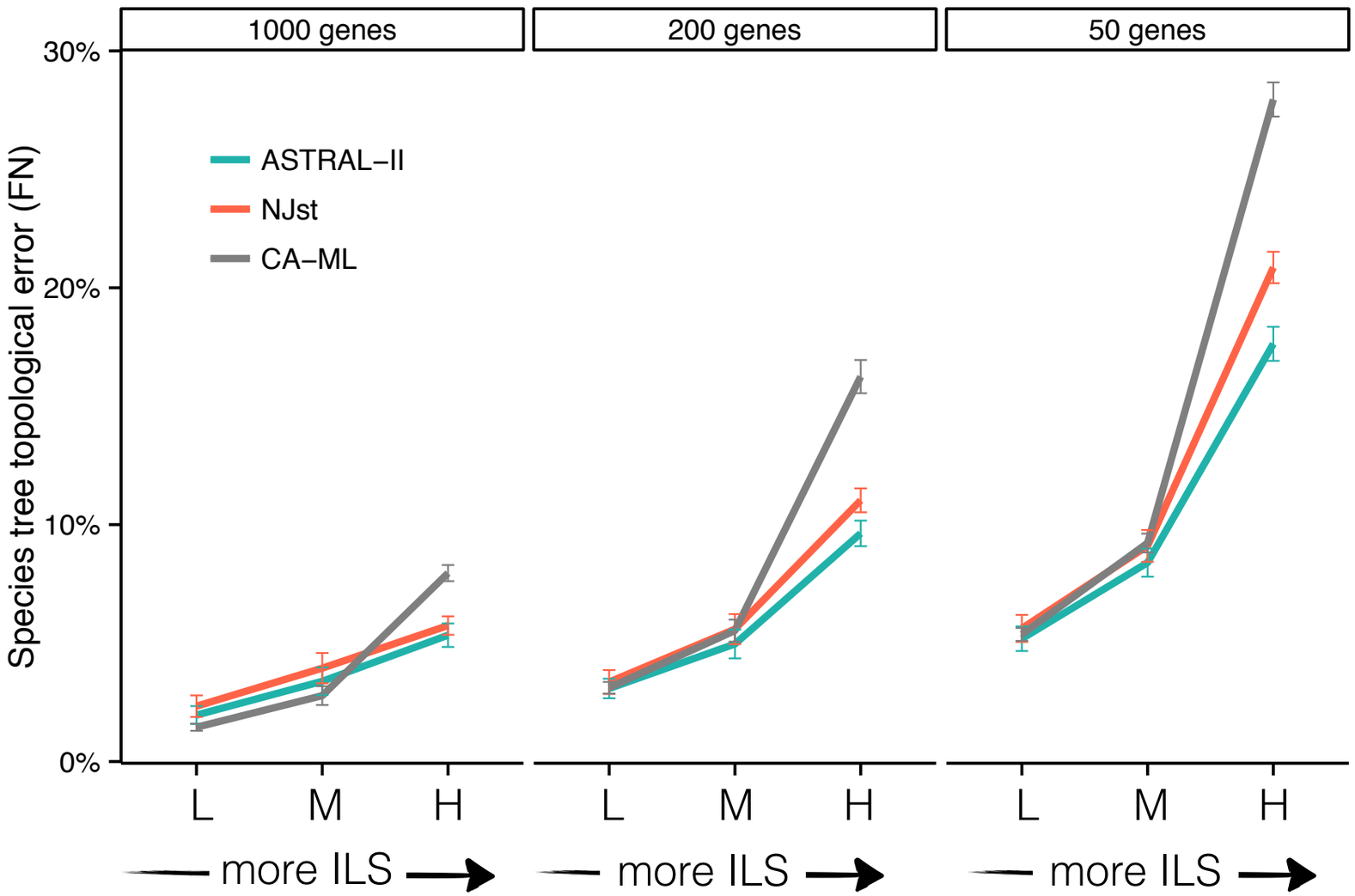
1000 genes, “medium” levels of recent ILS

Tree error, varying level of ILS



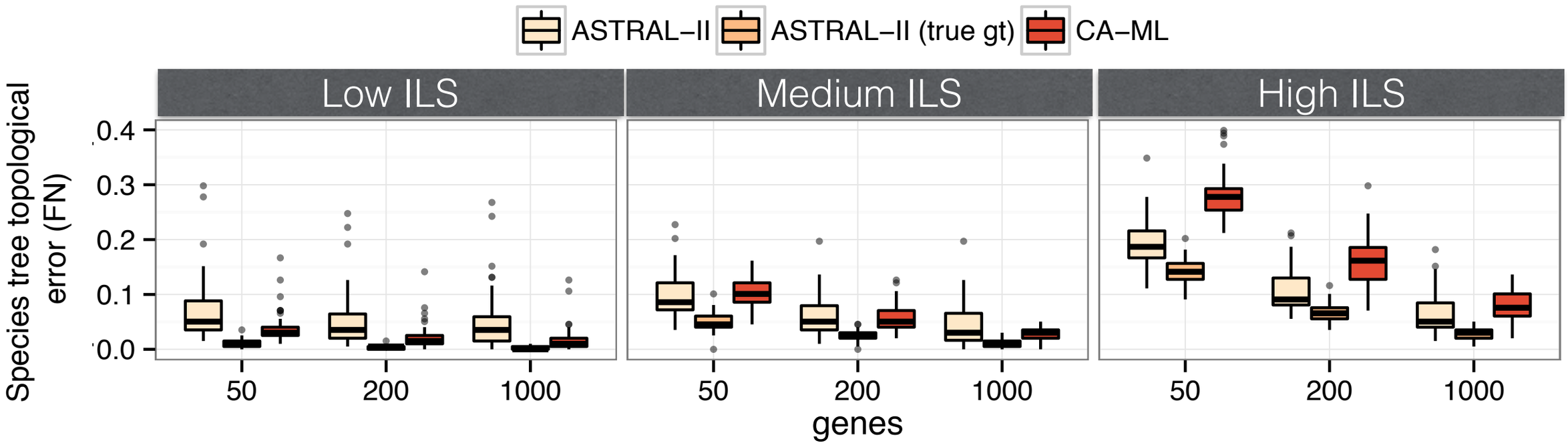
200 species, recent ILS

Tree error, varying level of ILS

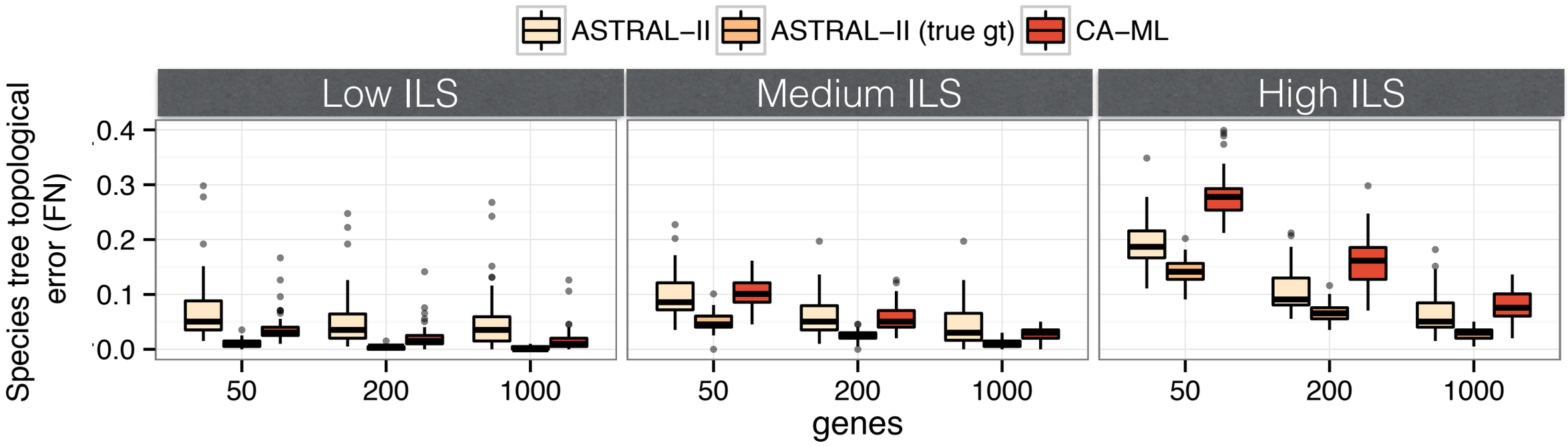


200 species, recent ILS

Impact of gene tree error (using true gene trees)



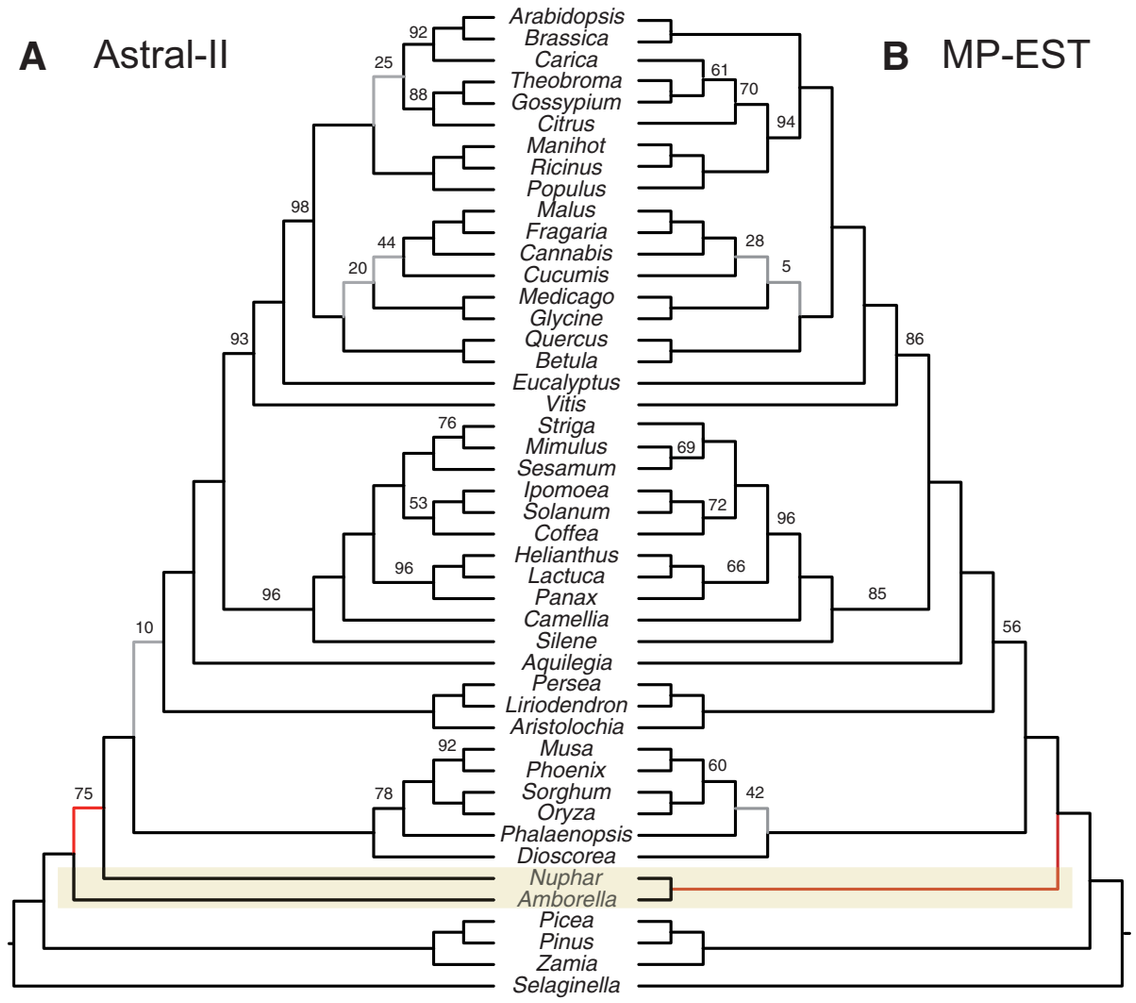
Impact of gene tree error (using true gene trees)



- When we divide our 50 replicates into low, medium, or high gene tree estimation error, ASTRAL tends to be better with low error

Insights on biological data

- Main question: The placement of Amborella at the base of angiosperms
- Xi et al. (2014) used a collection of 310 genes sampled from 46 species.
- Conflicting results:
  - Concatenation puts Amborella at the base (H1)
  - MP-EST puts Amobrella+water lilies at the base (H2)
  - Xi et al. conclude ILS is the cause
- ASTRAL like many other recent studies (e.g., 1KP) recovers H1
  - ILS is not necessarily the cause



Summary

- Genome-scale data provides a wealth of information for resolving long-standing phylogenetic questions.
- ASTRAL-II improves on ASTRAL-I in terms of both accuracy and running time.
- ASTRAL-II can handle datasets with 1000 genes from 1000 taxa in a day of single cpu running time.
- ASTRAL dominates other summary methods. However, Concatenation is better when gene trees have high error.
- In the future, we need to further explore the impact of model violations, recombination, missing data, and multiple sources of gene tree discordance (e.g., HGT).

Acknowledgments

Tandy Warnow

Keshav Pingali

Jim Leebens-mack
(UGA)

Norman Wickett
(U Chicago)

Gane Wong
(U of Alberta)

. . .

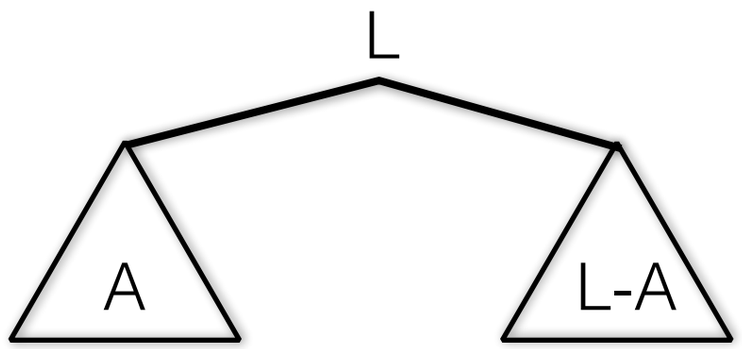
S.M. Bayzid

Théo
Zimmermann

Travel funding to ISMB/ECCB 2015
was generously provided by akamia.

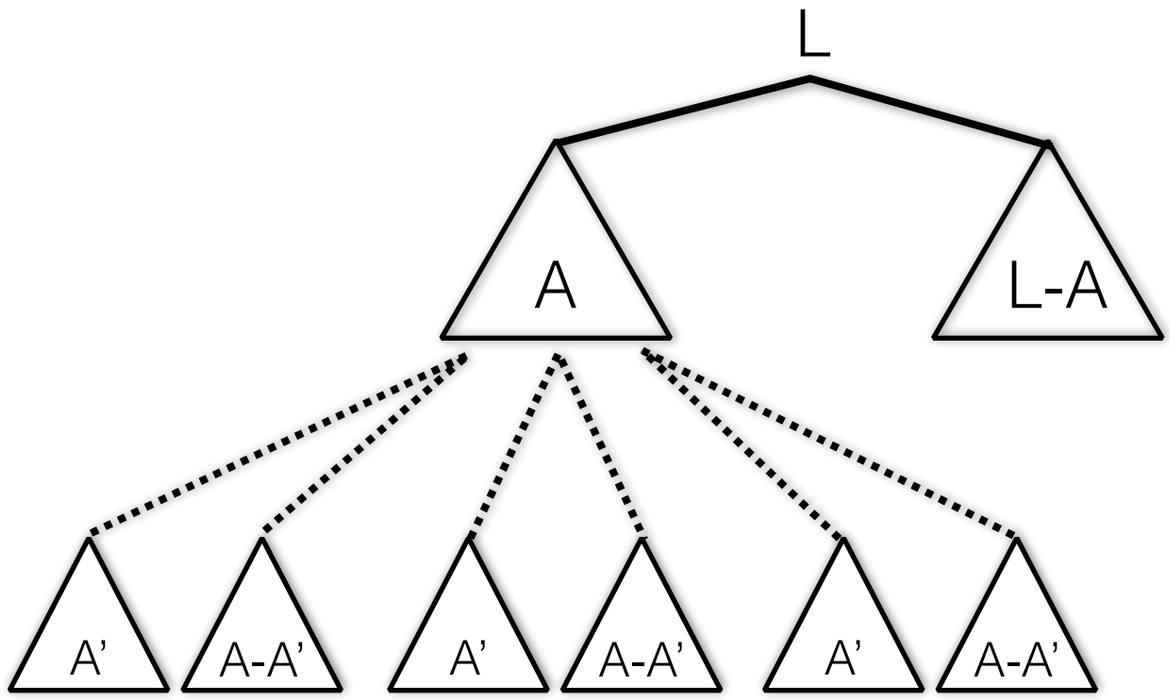
Dynamic programming

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$



Dynamic programming

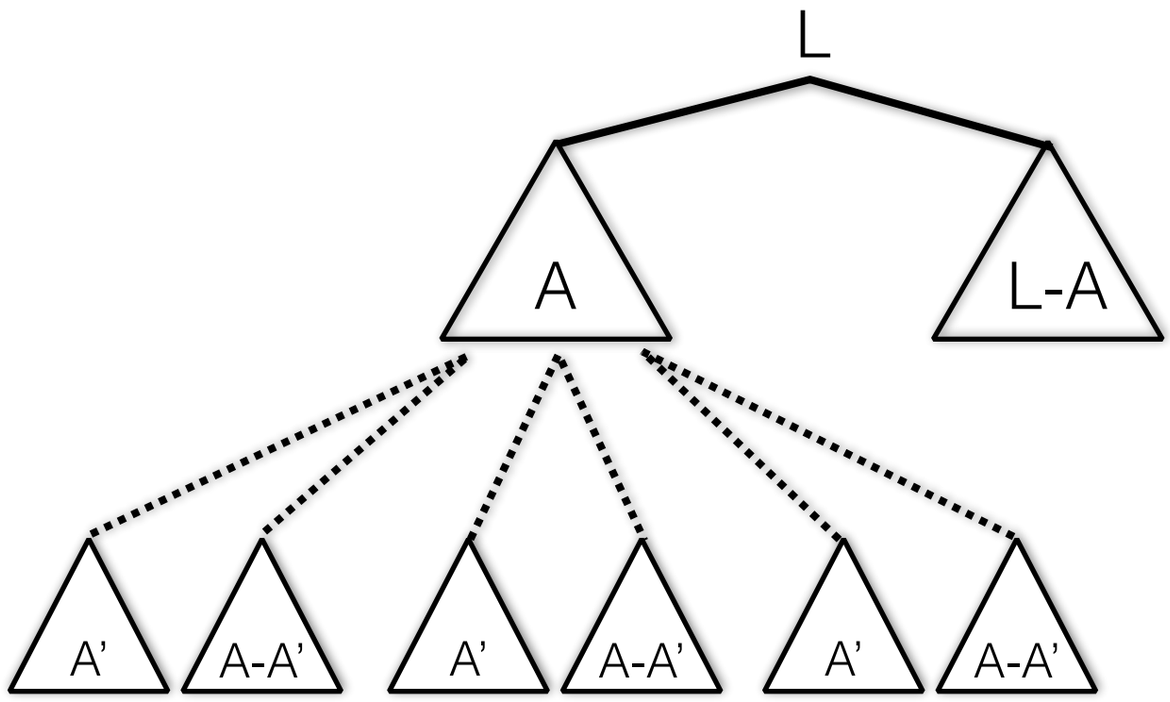
$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$

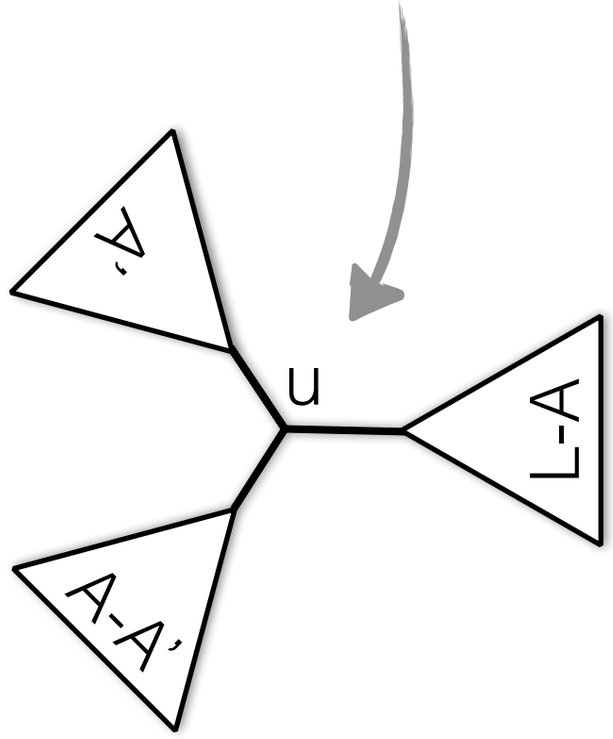


- Recursively break subsets of species into smaller subsets

Dynamic programming

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}' | \mathcal{A} - \mathcal{A}' | \mathcal{L} - \mathcal{A})\}$$



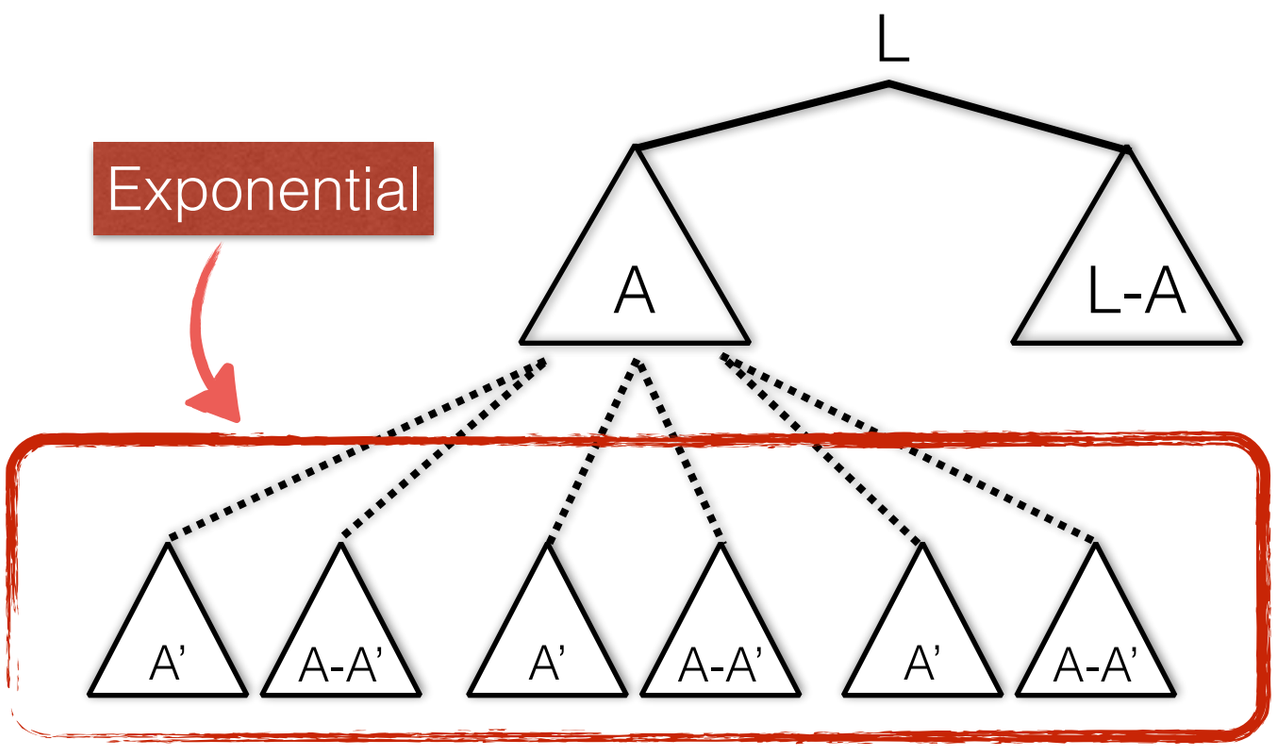


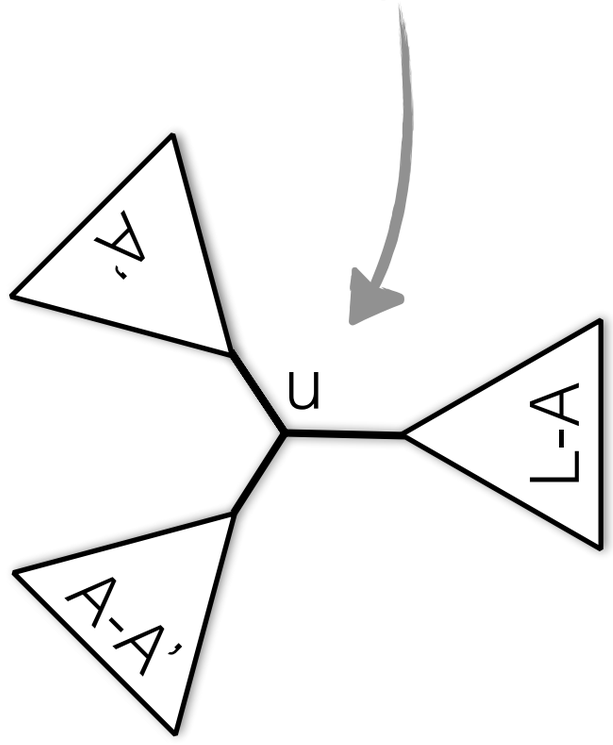
- Recursively break subsets of species into smaller subsets

- $w(u)$: Compare u against input gene trees and compute quartets from gene trees satisfied by u

Dynamic programming

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$



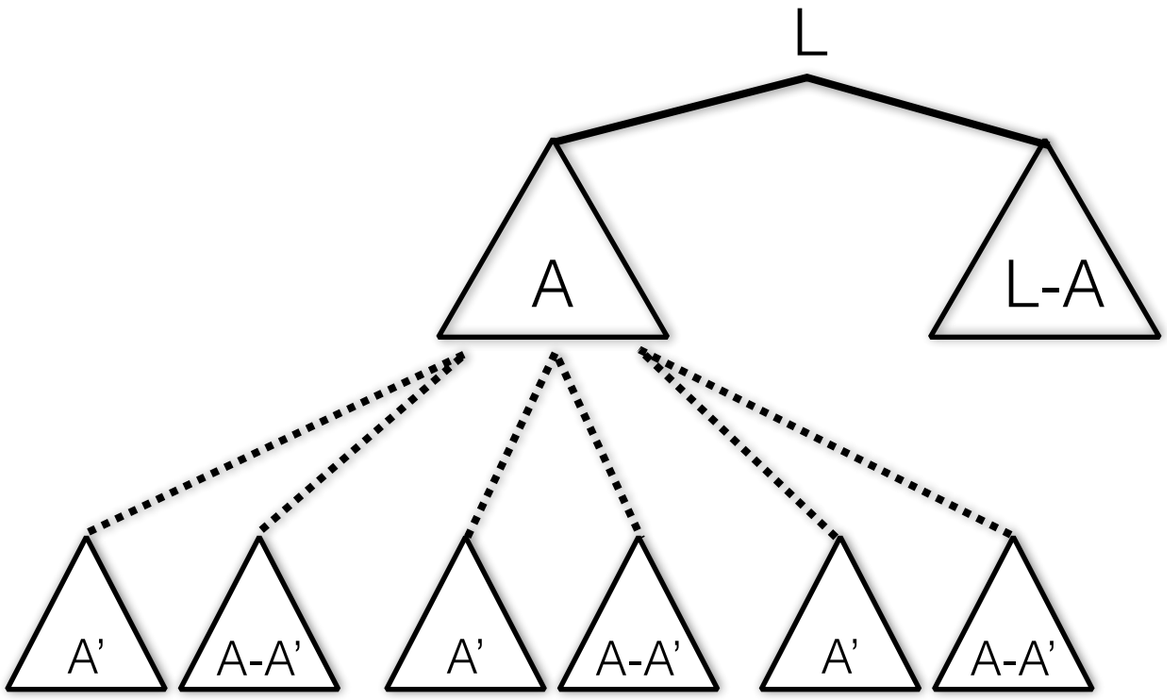


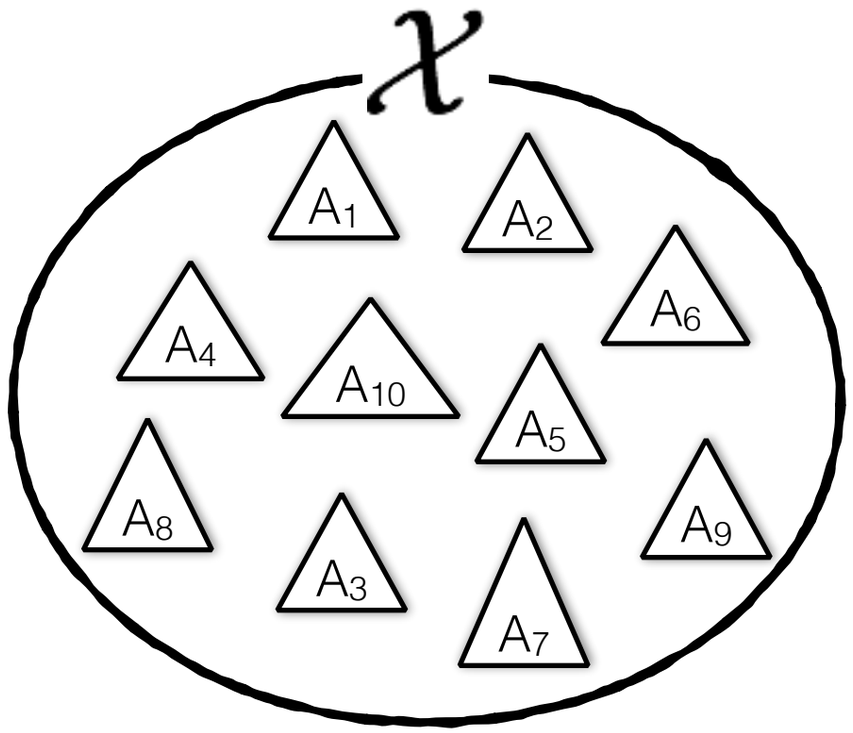
- Recursively break subsets of species into smaller subsets

- $w(u)$: Compare u against input gene trees and compute quartets from gene trees satisfied by u

Constrained version

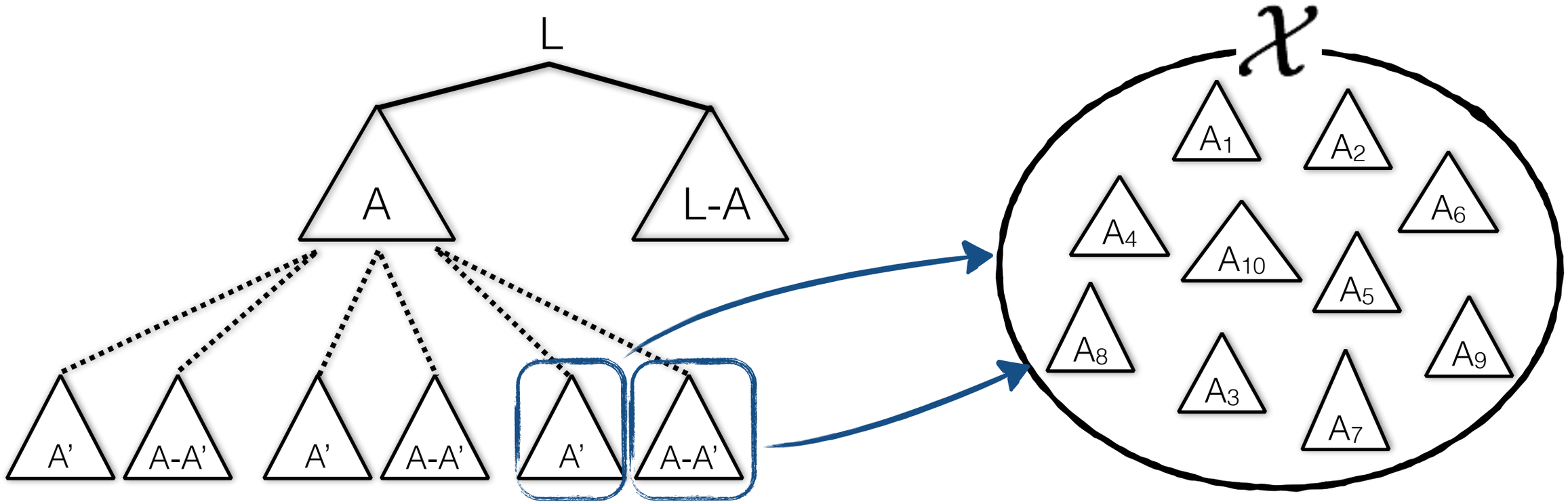
$$S(\mathcal{A}) = \max_{\{\mathcal{A}', \mathcal{A}-\mathcal{A}'\} \subset \mathcal{X}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$





Constrained version

$$S(\mathcal{A}) = \max_{\{\mathcal{A}', \mathcal{A}-\mathcal{A}'\} \subset \mathcal{X}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$



- Restrict “branches” in the species tree to a given constraint set $\mathcal{X}$.

Proof (sketch)

$$Score(T) = \sum_{t \in \mathcal{T}} (Q(T) \cap Q(t)) = \sum_{q} F_{\mathcal{T}}(T|q)$$

number of gene trees that induce a quartet topology

tree *T* restricted to *q*

a quartet of taxa

- Recall [Allman, et al. 2010]: for any 4 species q, the most probable unrooted quartet tree is the one induced by the species tree T^*

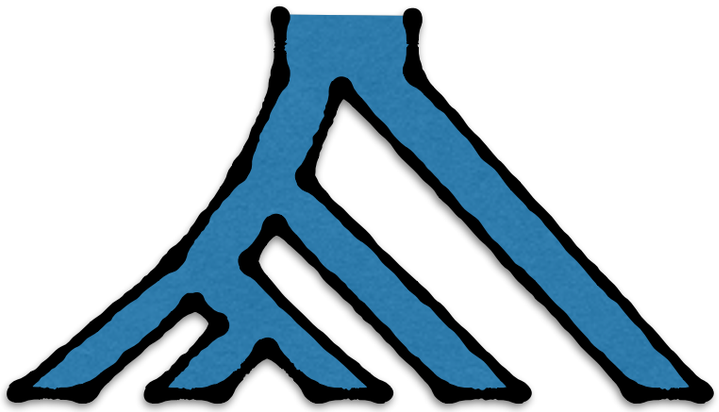
The species tree

- For any q, as the number of genes increases, the dominant topology converges to the species tree topology restricted to q

- $\forall\ T \neq T^*$: $F_{\mathcal{T}}(T^*|q) \geq F_{\mathcal{T}}(T|q)$, and thus: $Score(T^*) > Score(T)$

- Hence, as the number of genes increases, with probability converging to 1, ASTRAL returns the true species tree T^*

More than 4 species



1. Break gene trees into $\binom{n}{4}$ quartets of species

2. Find the dominant tree for all quartets of taxa

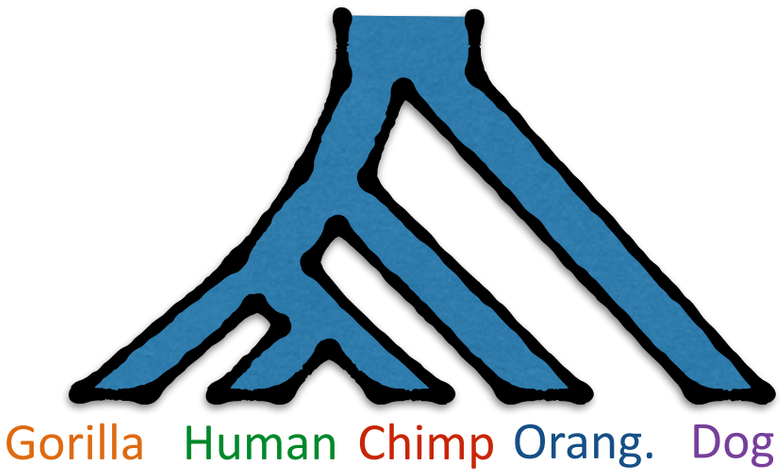
3. Combine quartet trees

Example: BUCKy-pop. [Larget, et al., Bioinformatics, 2010]

(probabilities are made-up just as an example)

Gorilla Human / Orangutan Chimp	Orang. Chimp / Human Gorilla	Gorilla Chimp / Human Orang.
50%	25%	25%
Gorilla Human / Dog Chimp	Dog Chimp / Human Gorilla	Gorilla Chimp / Human Dog
55%	21%	24%
Gorilla Human / Orangutan Dog	dog Gorilla / Human Orang.	Orang. dog / Human Gorilla	Gorilla dog / Human Orang.
	7%	87%	6%
Gorilla Dog / Orangutan Chimp	Chimp Gorilla / Dog Orang.	Orang. Chimp / Dog Gorilla	Gorilla Chimp / Dog Orang.
	6%	88%	6%
Dog Human / Orangutan Chimp	Chimp Dog / Human Orang.	Orang. Chimp / Human Dog	Dog Chimp / Human Orang.
	95%	2%	3%

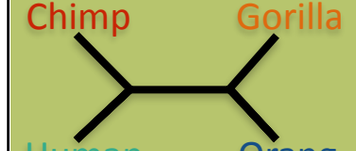
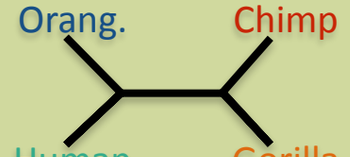
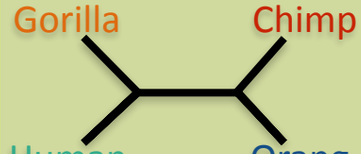
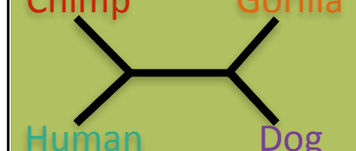
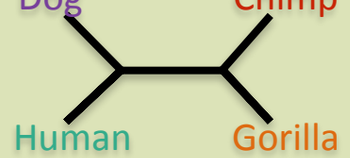
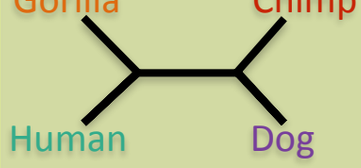
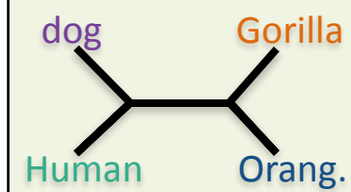
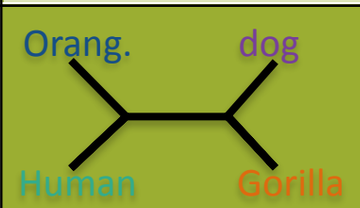
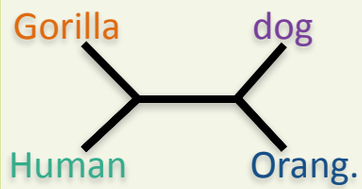
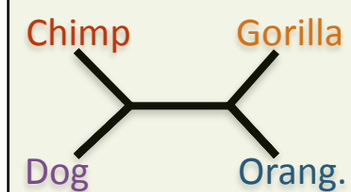
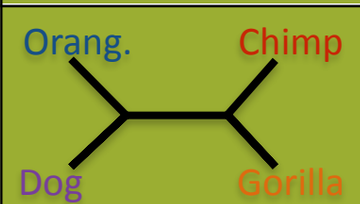
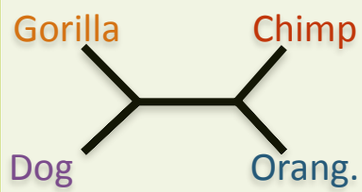
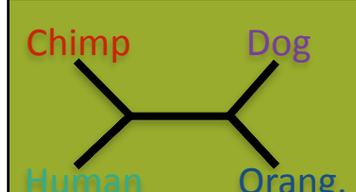
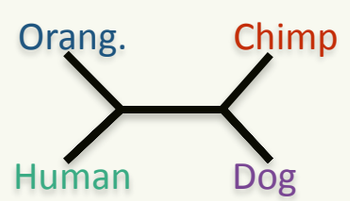
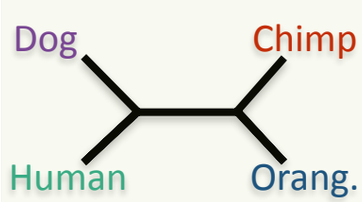
ASTRAL: weighting by frequency



1. Break gene trees into $\binom{n}{4}$ quartets of species

2. Find the tree that "satisfies" the maximum number of weighted quartets from gene trees

(probabilities are made-up just as an example)

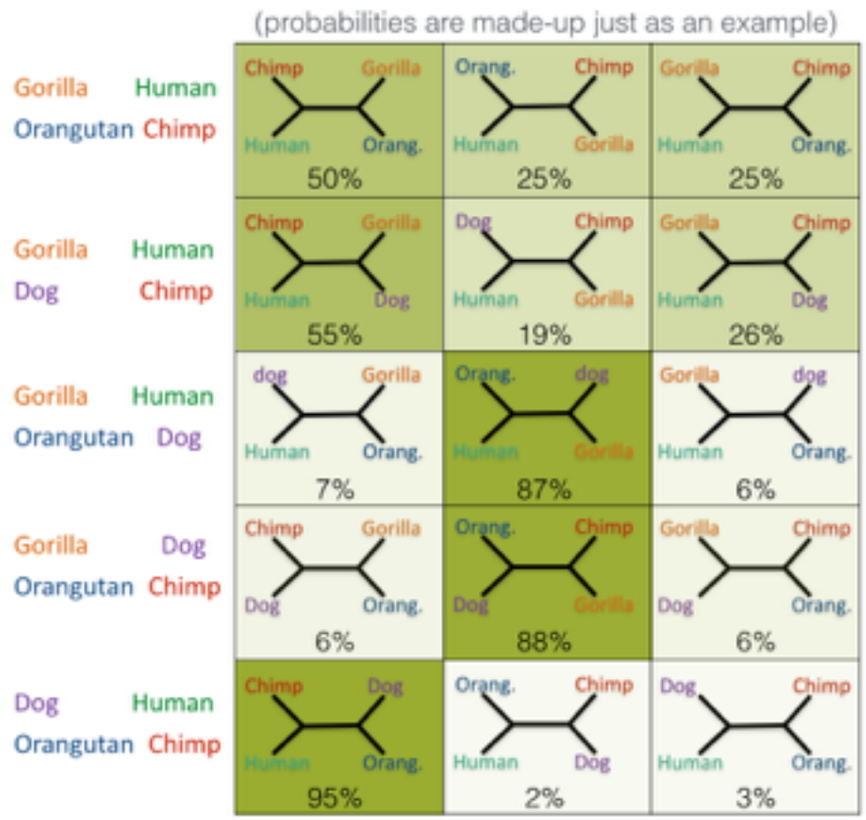
Gorilla Human Orangutan Chimp	 50%	 25%	 25%
Gorilla Human Dog Chimp	 55%	 19%	 26%
Gorilla Human Orangutan Dog	 7%	 87%	 6%
Gorilla Dog Orangutan Chimp	 6%	 88%	 6%
Dog Human Orangutan Chimp	 95%	 2%	 3%

Asymptotic running time

- $O(nk|\mathcal{X}|^2)$ for *k* genes of *n* species

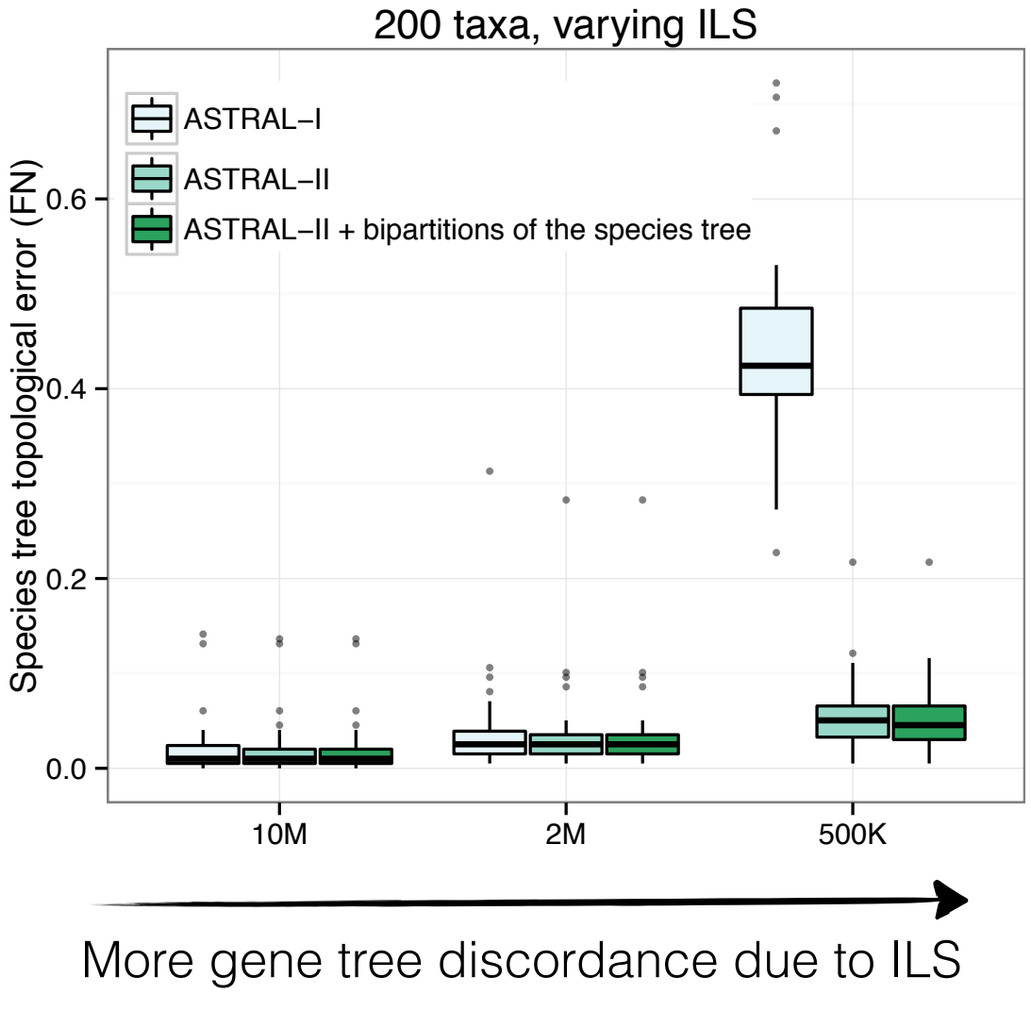
Asymptotic running time

- $O(nk|\mathcal{X}|^2)$ for *k* genes of *n* species

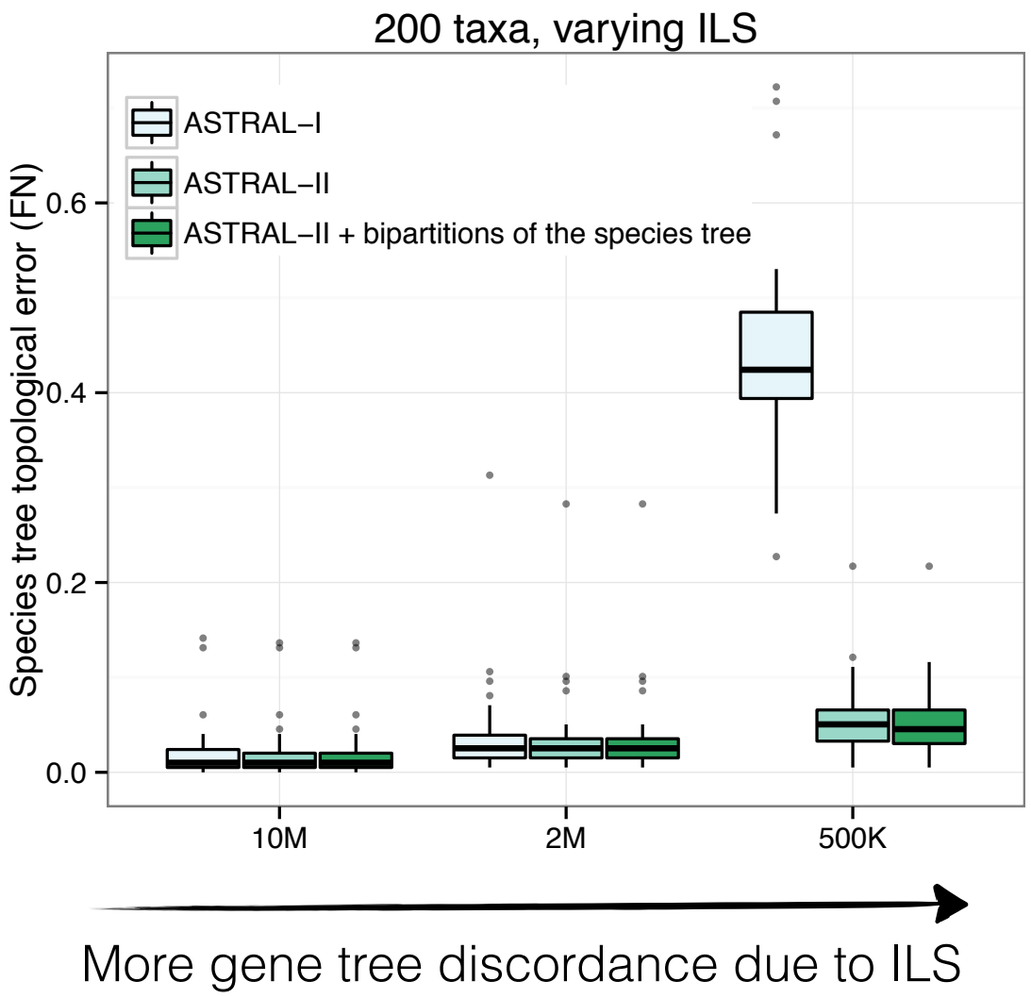


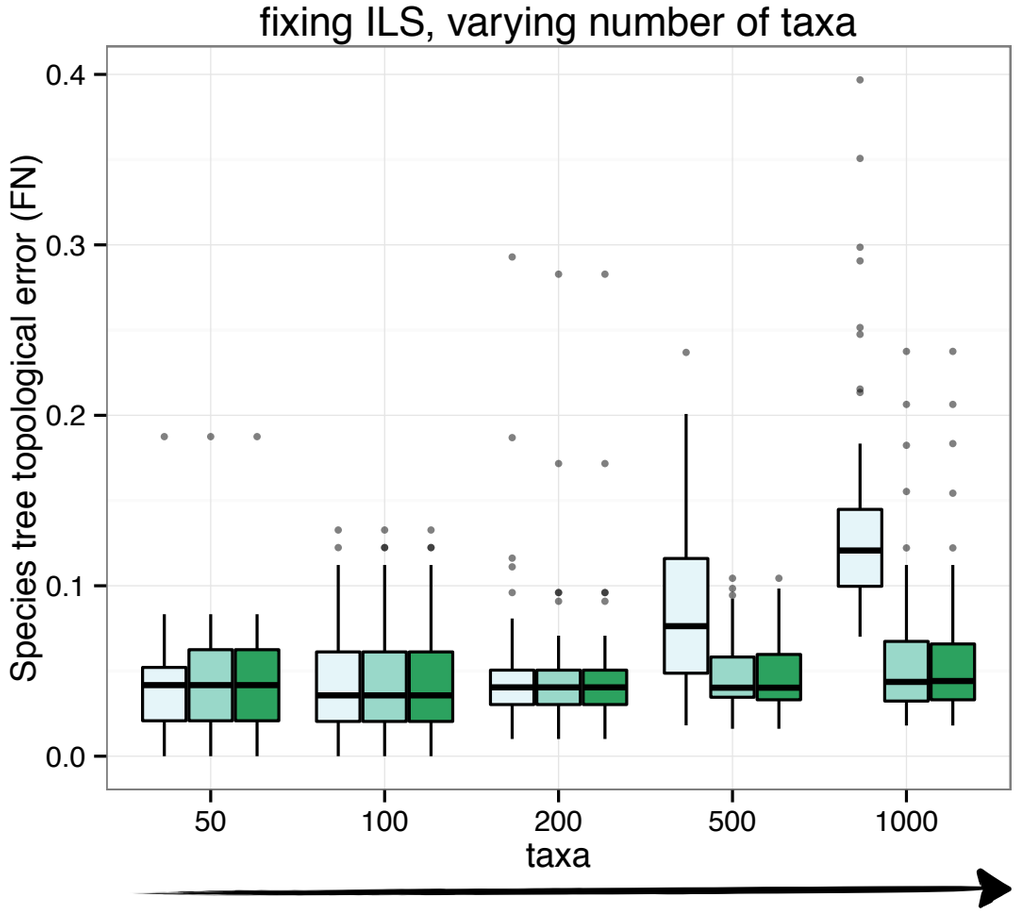
- Surprise: running time is better than $\Theta(n^4)$
- Don't we have to at least list all $\binom{n}{4}$ quartets?
- No! we calculate scores without listing $\binom{n}{4}$ quartets

ASTRAL-I versus ASTRAL-II

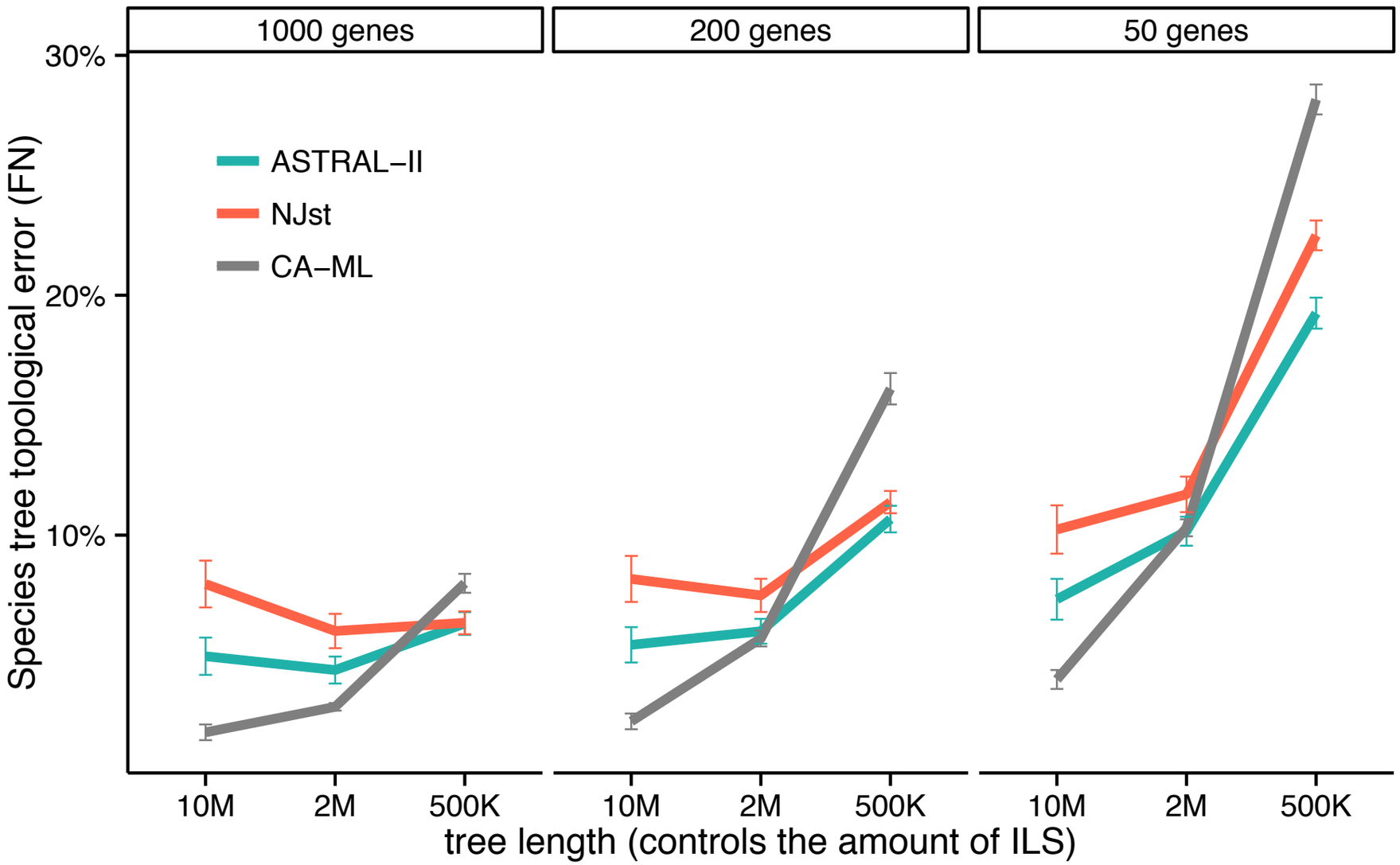


ASTRAL-I versus ASTRAL-II



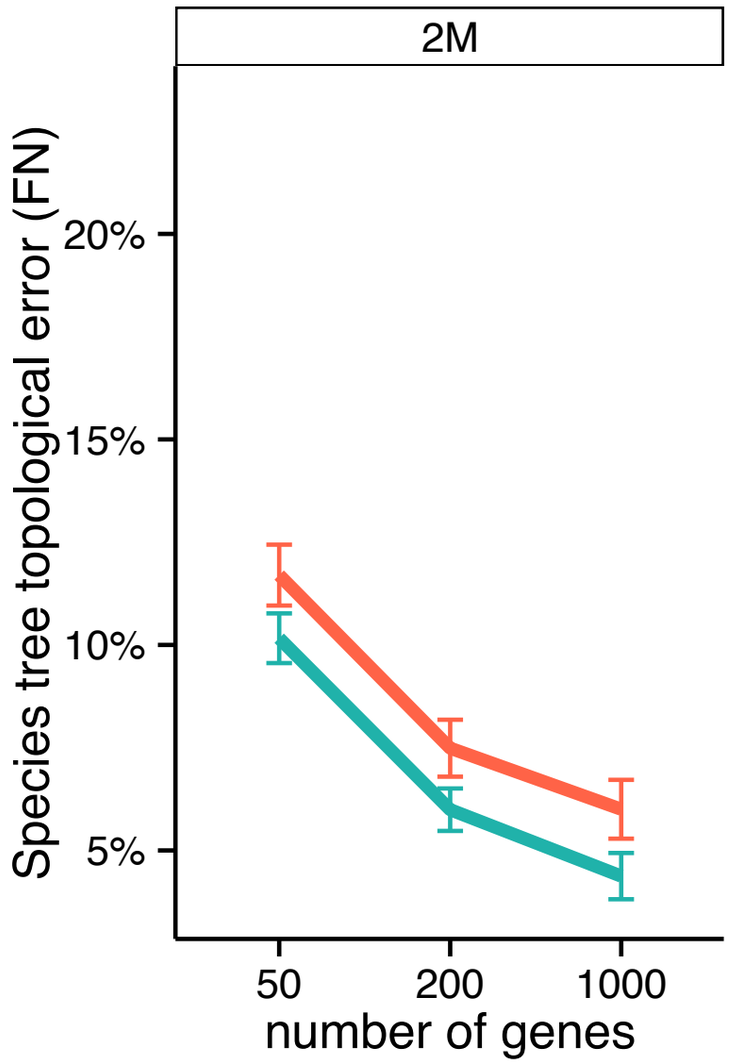


Deep ILS

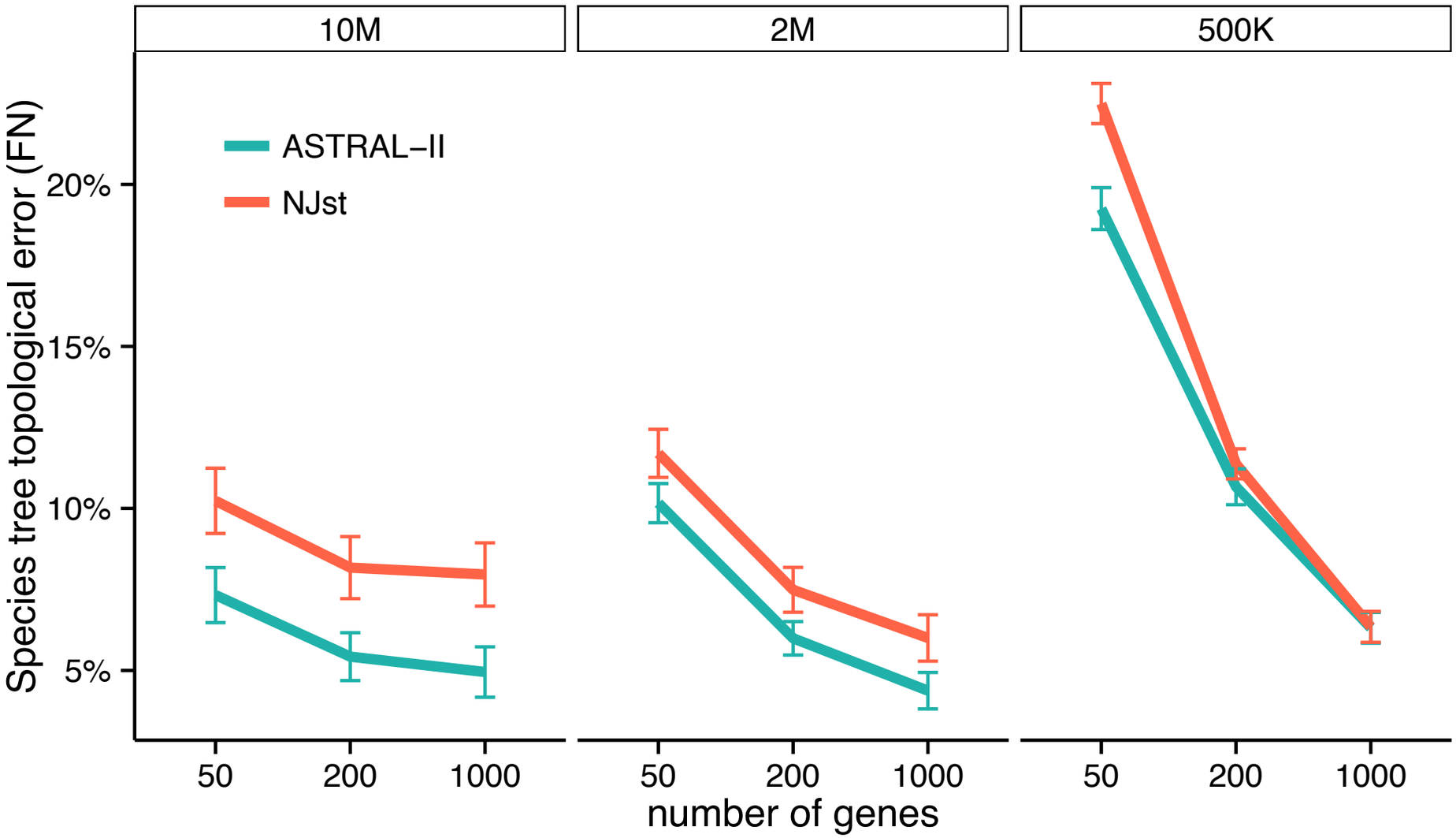


200 species, simulated species trees, deep ILS
[Mirarab and Warnow, ISMB, 2015]

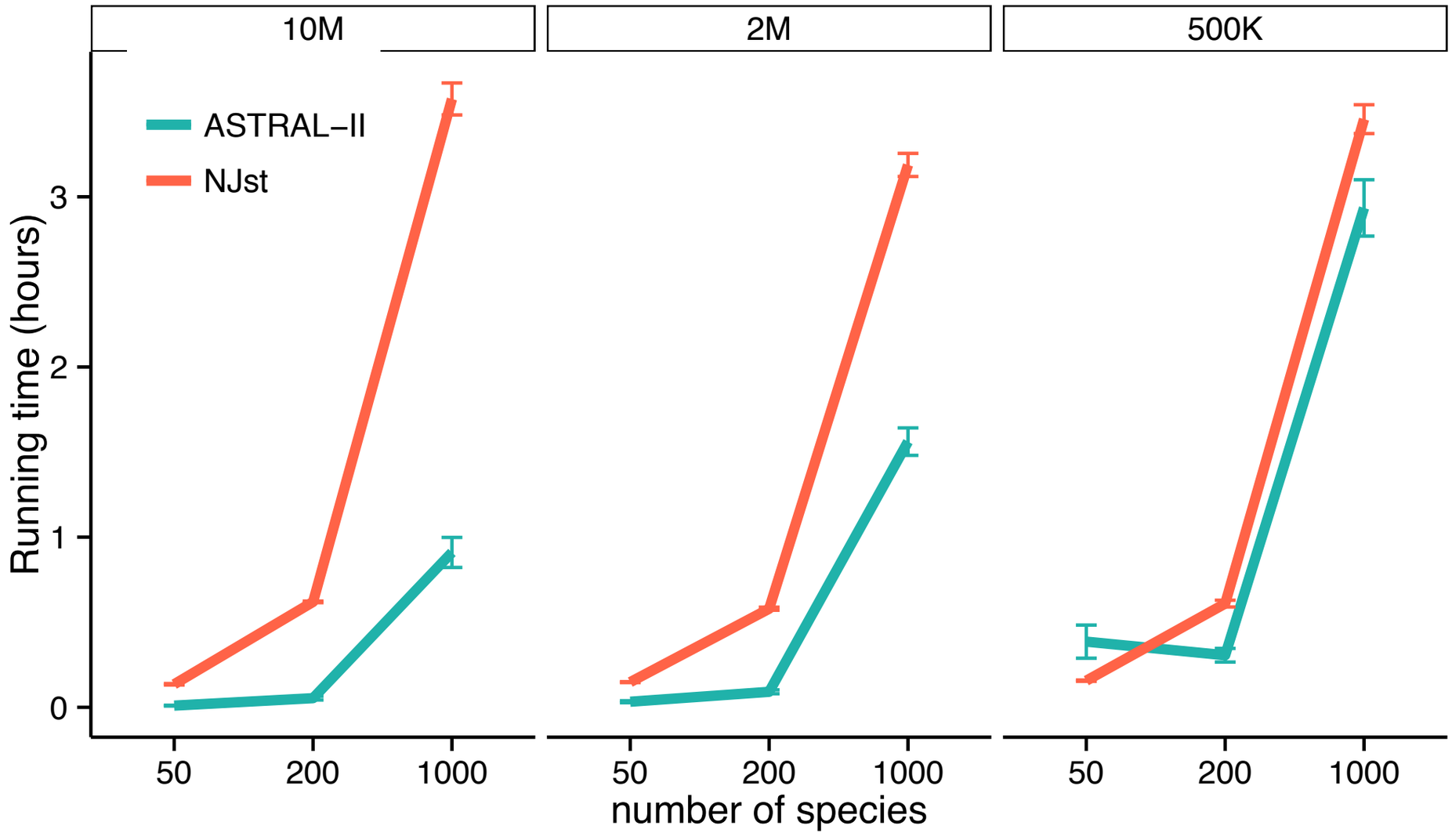
Varying the number of genes



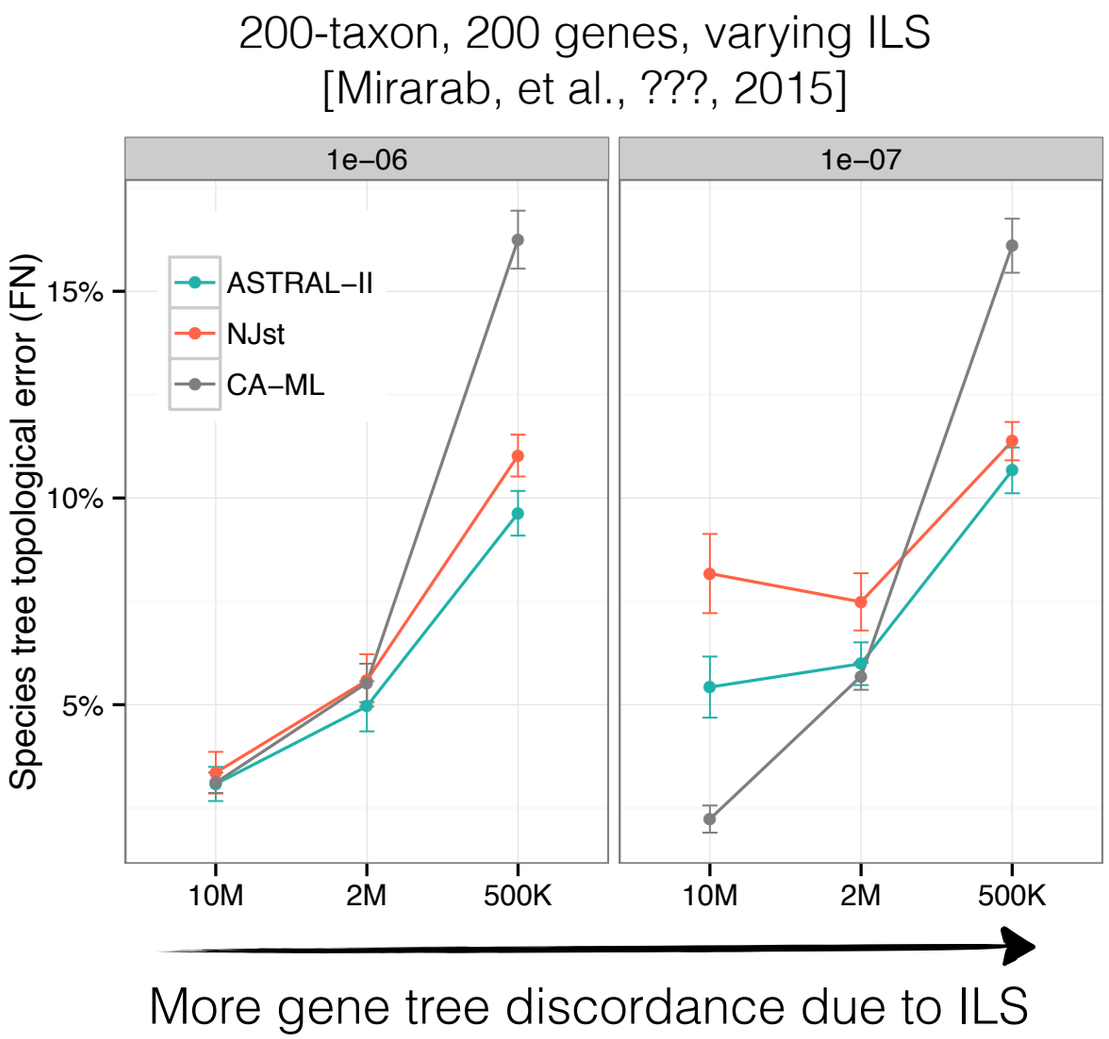
Varying the number of genes



Varying the number of genes

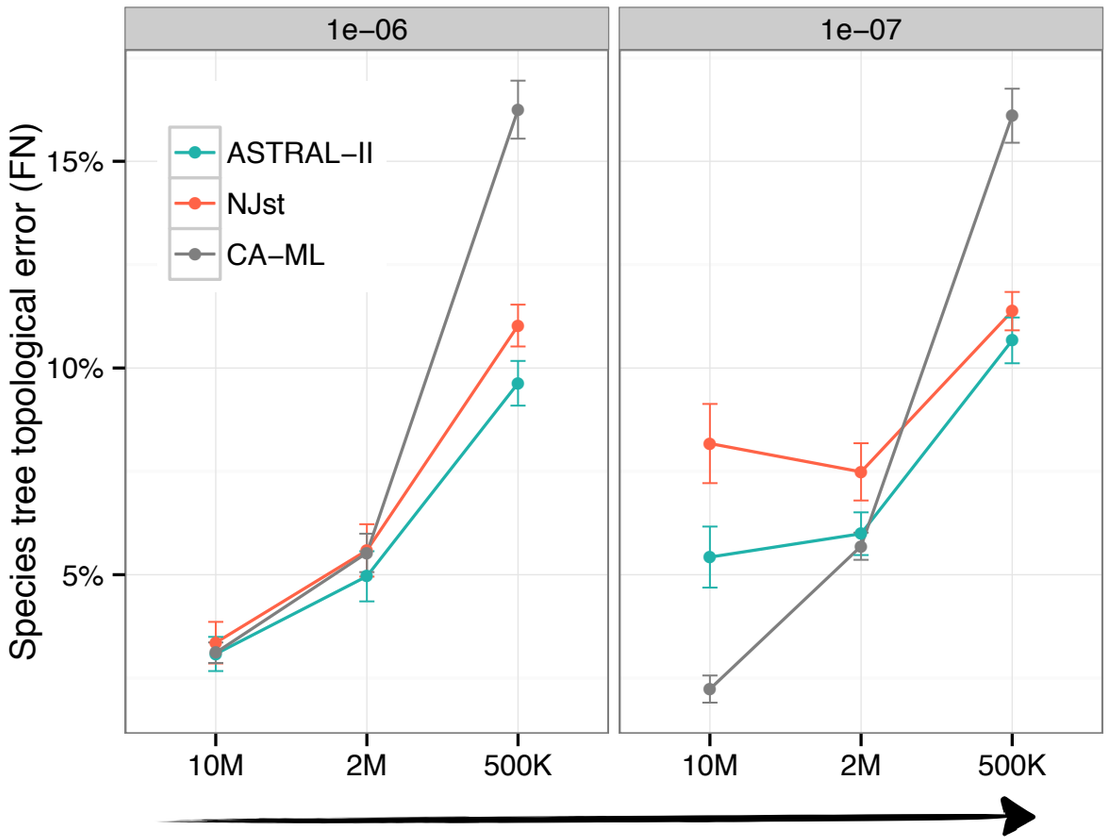


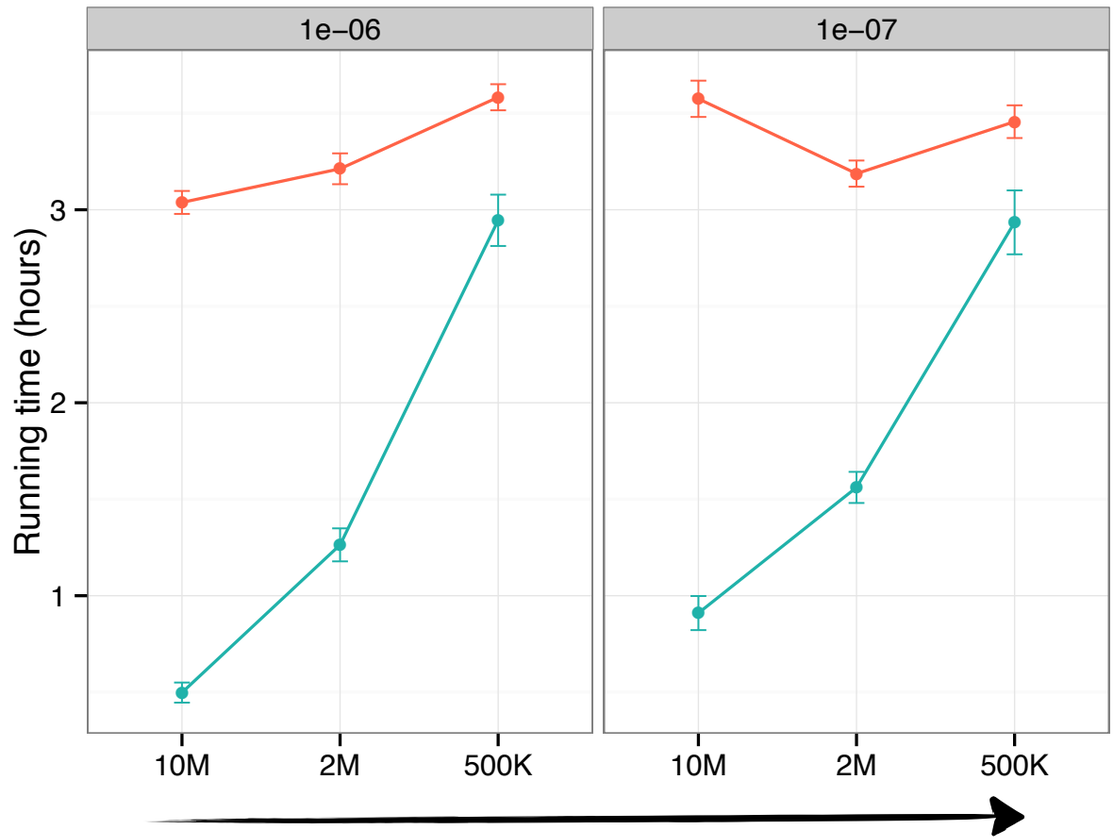
Increasing levels of ILS



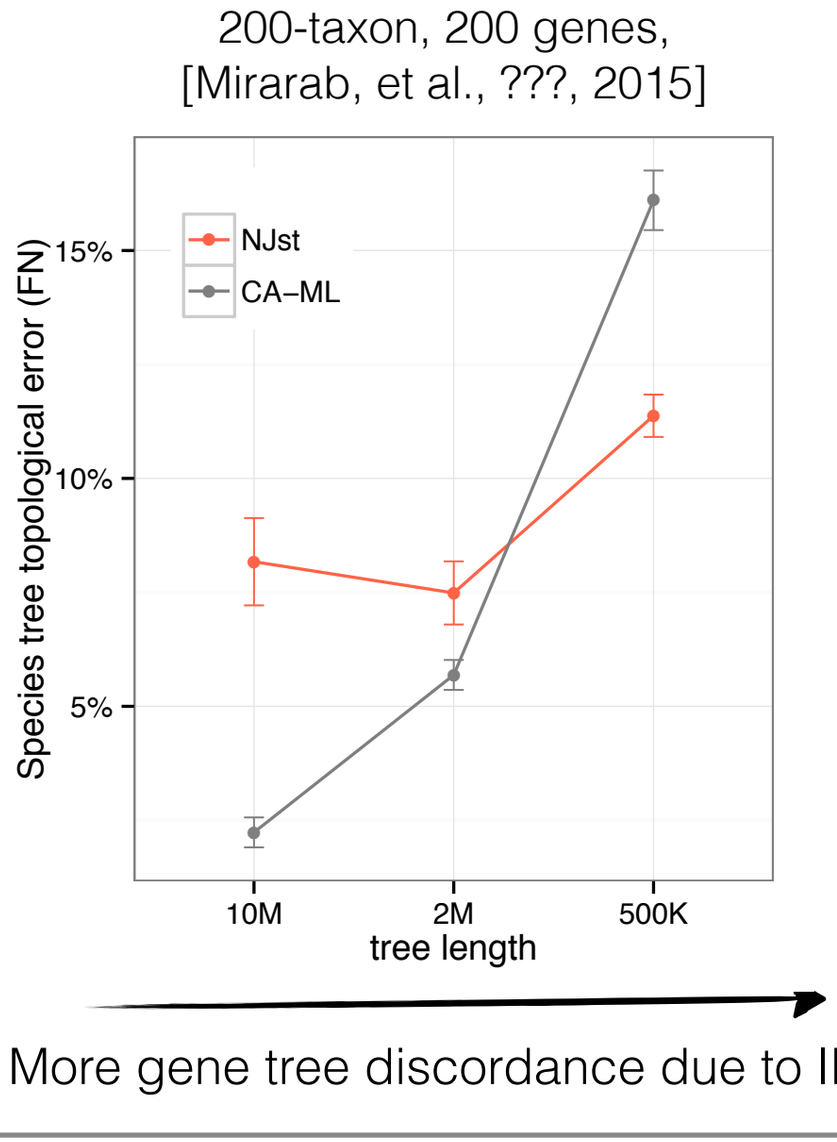
Increasing levels of ILS

200-taxon, 200 genes, varying ILS
[Mirarab, et al., ???, 2015]

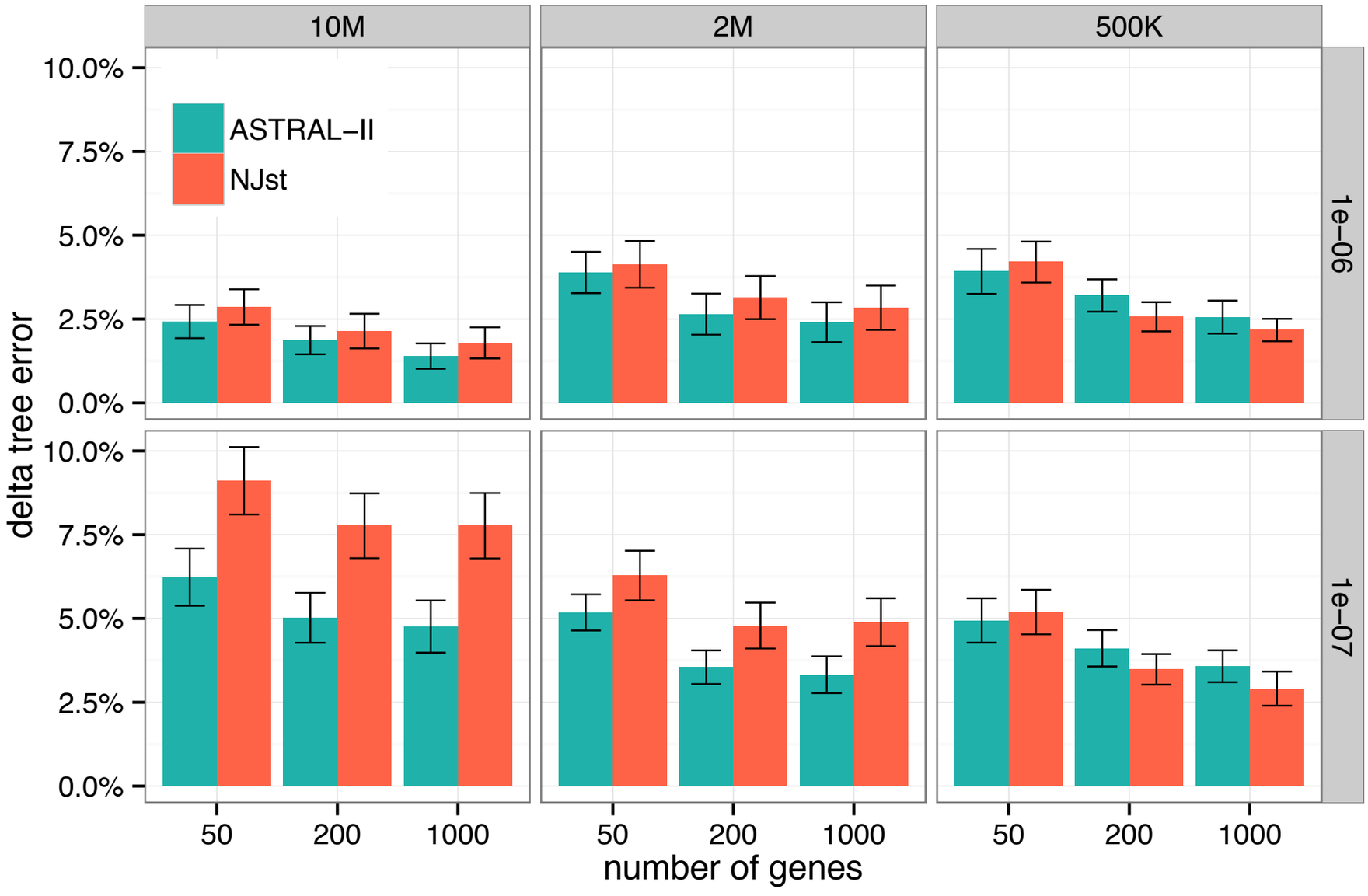




Impact of amount of ILS

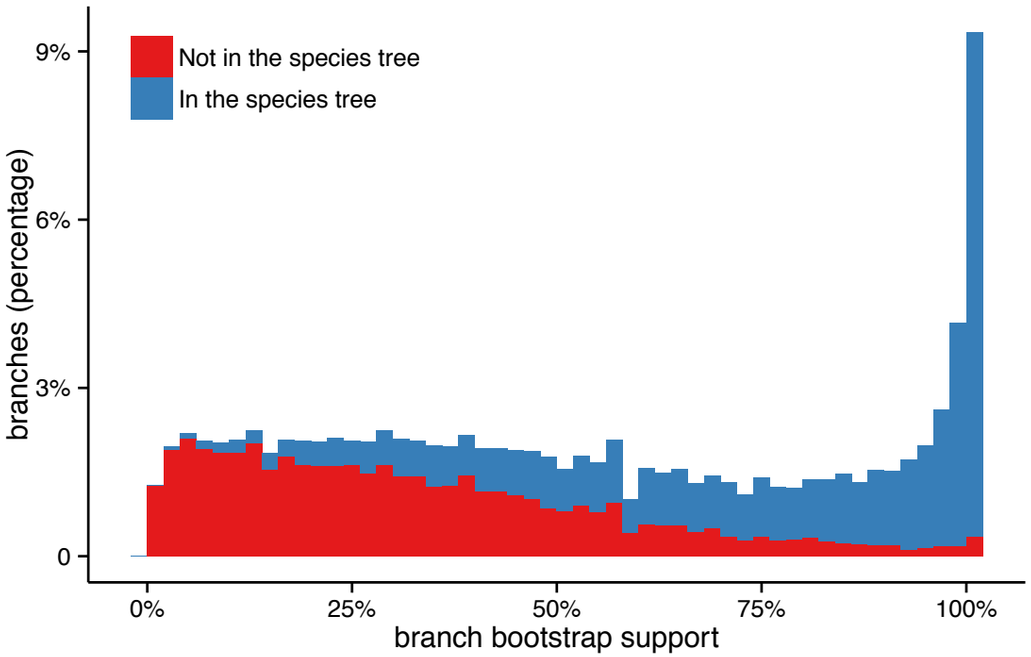


Concatenation does not work well with high levels of ILS


10M
2M
500K
1e−06
1e−07
ASTRAL−II
NJst
10.0%
7.5%
5.0%
2.5%
0.0%
delta tree error
50
200
1000
number of genes

ASTRAL on 1KP

- The ASTRAL tree:
  - High support
  - Similar to concatenation with some interesting differences (e.g., recovered bryophtes)
- ASTRAL took only about 10 minutes on 103 taxa and 600 genes



[Wicket*, Mirarab*, et al., PNAS, 2014]