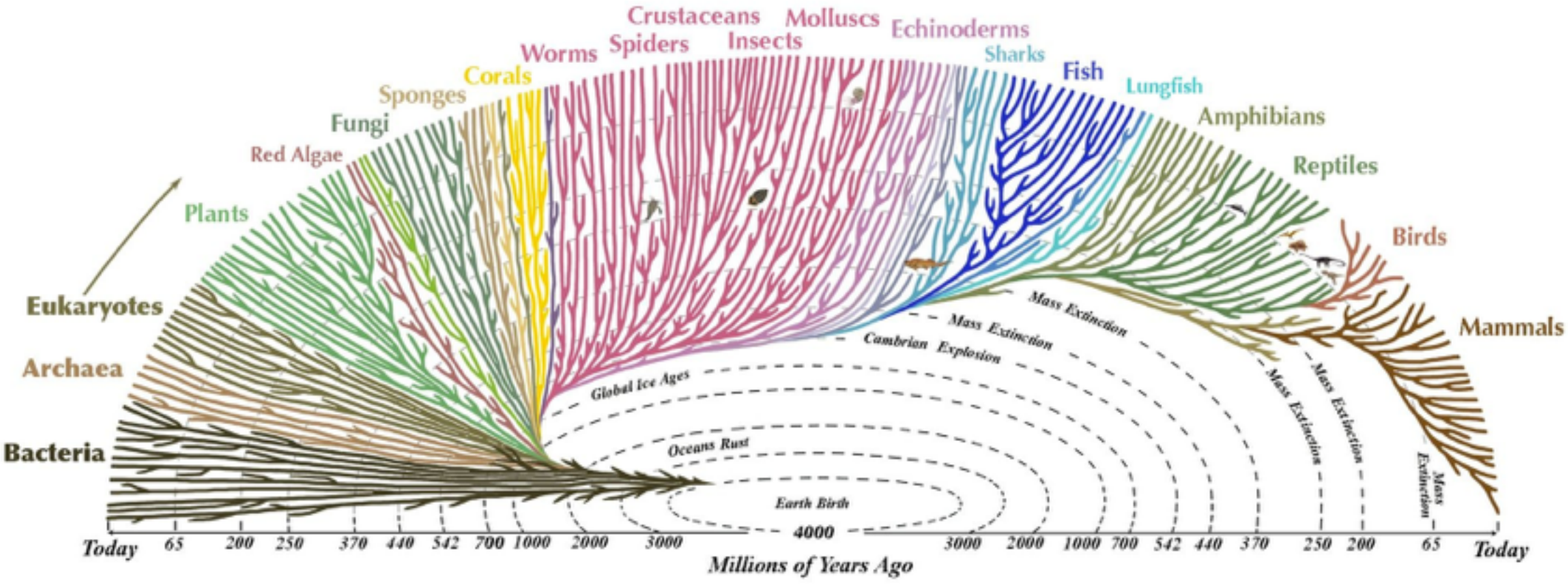
Reconstruction of species trees from gene trees using ASTRAL

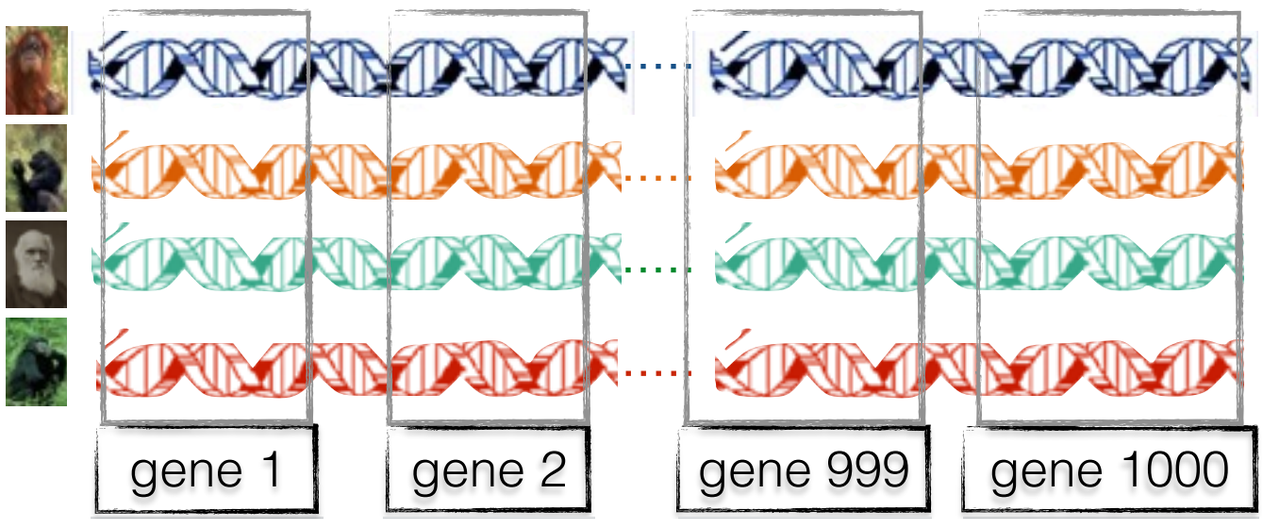
Siavash Mirarab
University of California, San Diego (ECE)

in collaboration with
Warnow lab (UIUC)

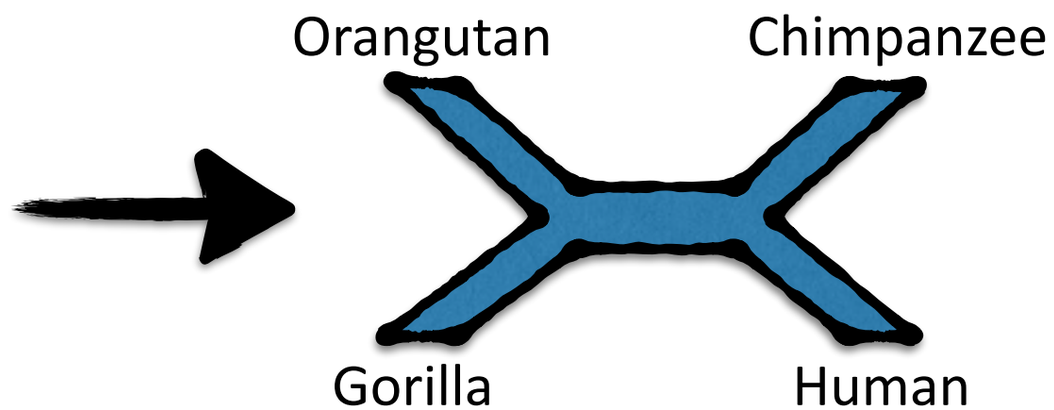


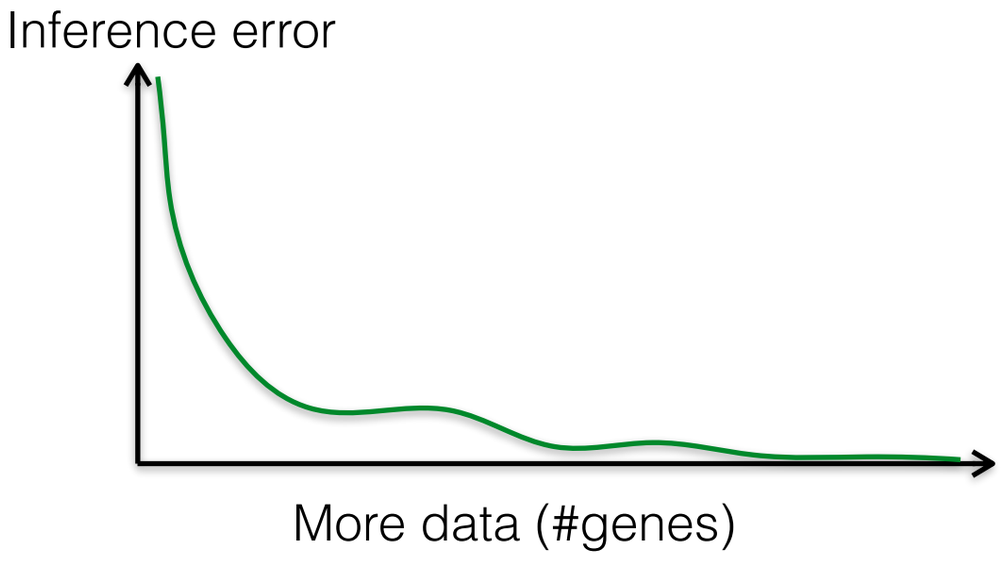
source: http://www.evogeneao.com/

Phylogenomics

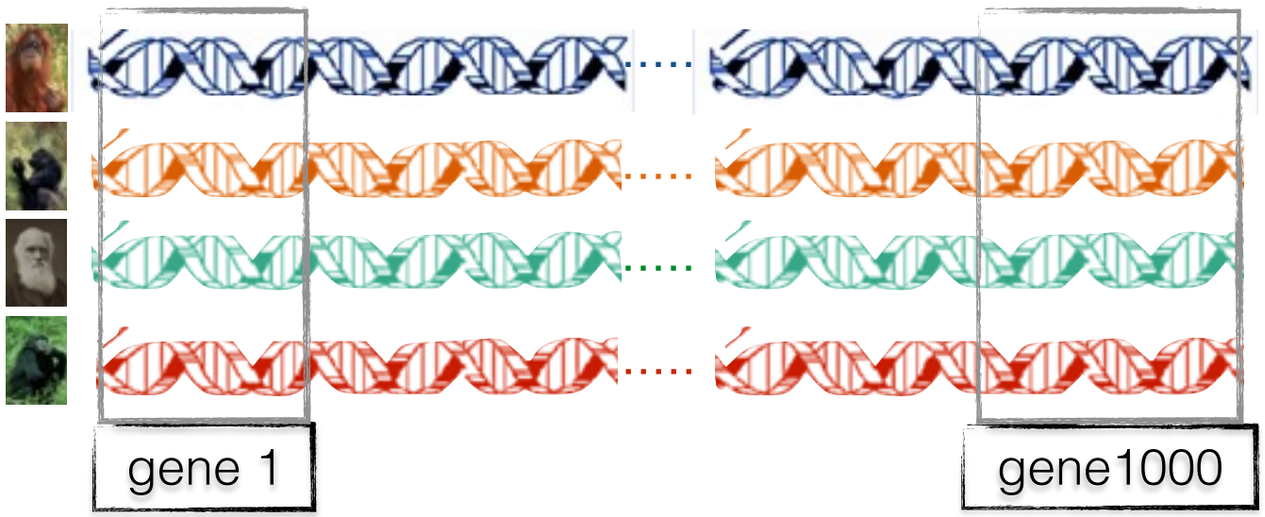


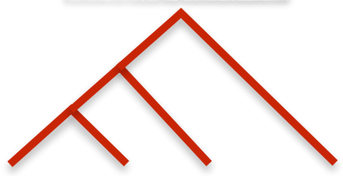

```
ACTGCACACCG   CTGAGCATCG     AGCAGCATCGTG  CAGGCACGCACGAA
ACTGC-CCCCG   CTGAGC-TCG ... AGCAGC-TCGTG  AGC-CACGC-CATA
AATGC-CCCCG   ATGAGC-TC-     AGCAGC-TC-TG  ATGGCACGC-C-TA
-CTGCACACGG   CTGA-CAC-G     C-TA-CACGGTG  AGCTAC-CACGGAT
```



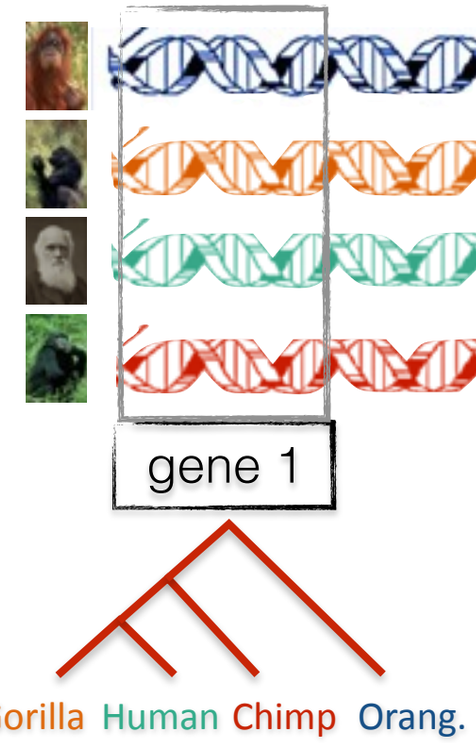


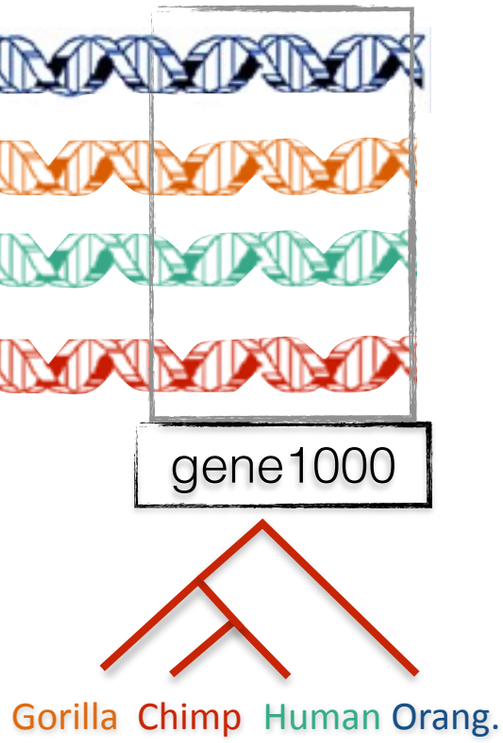
Gene tree discordance

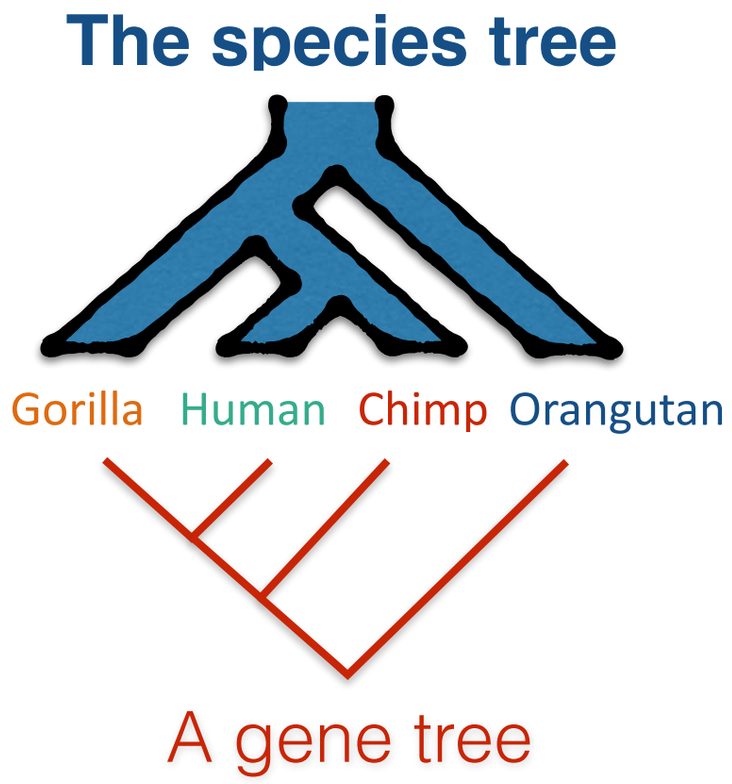




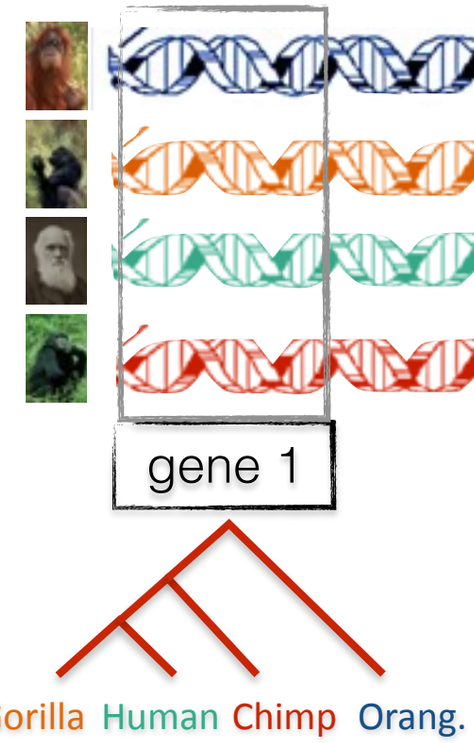
Gene tree discordance

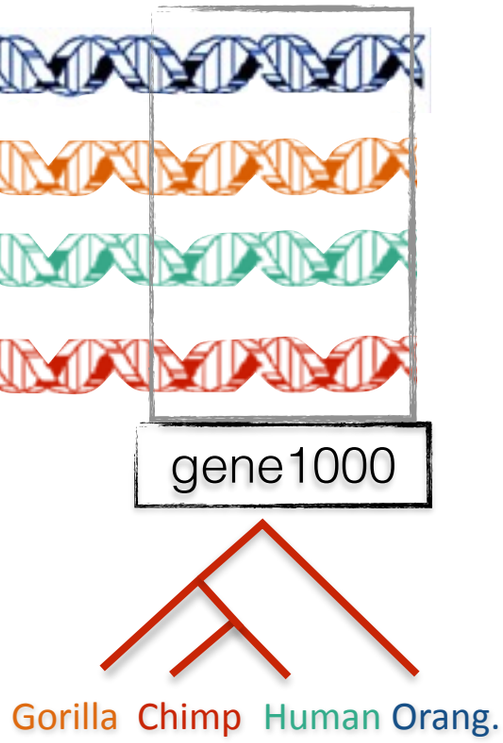


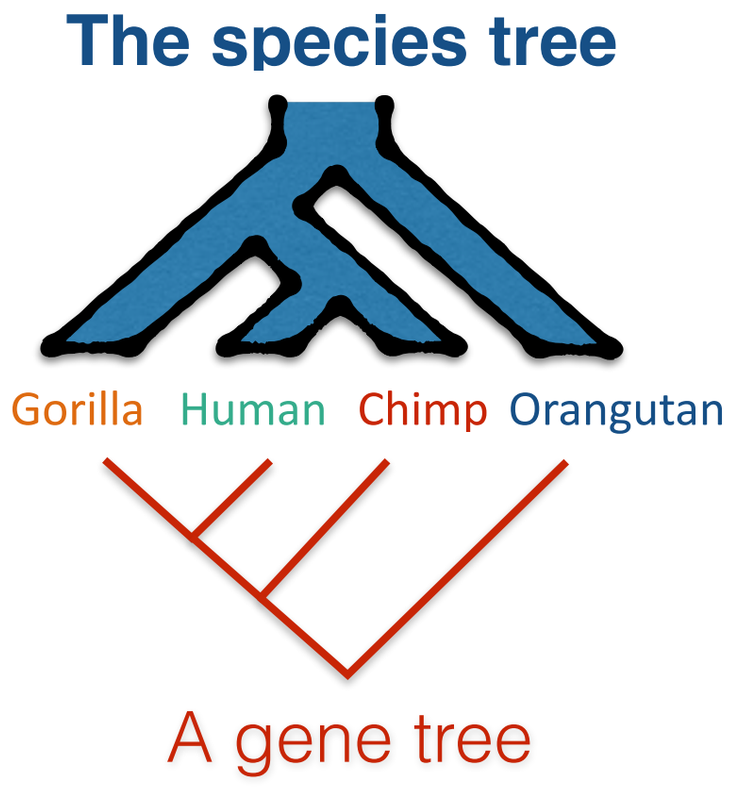




Gene tree discordance





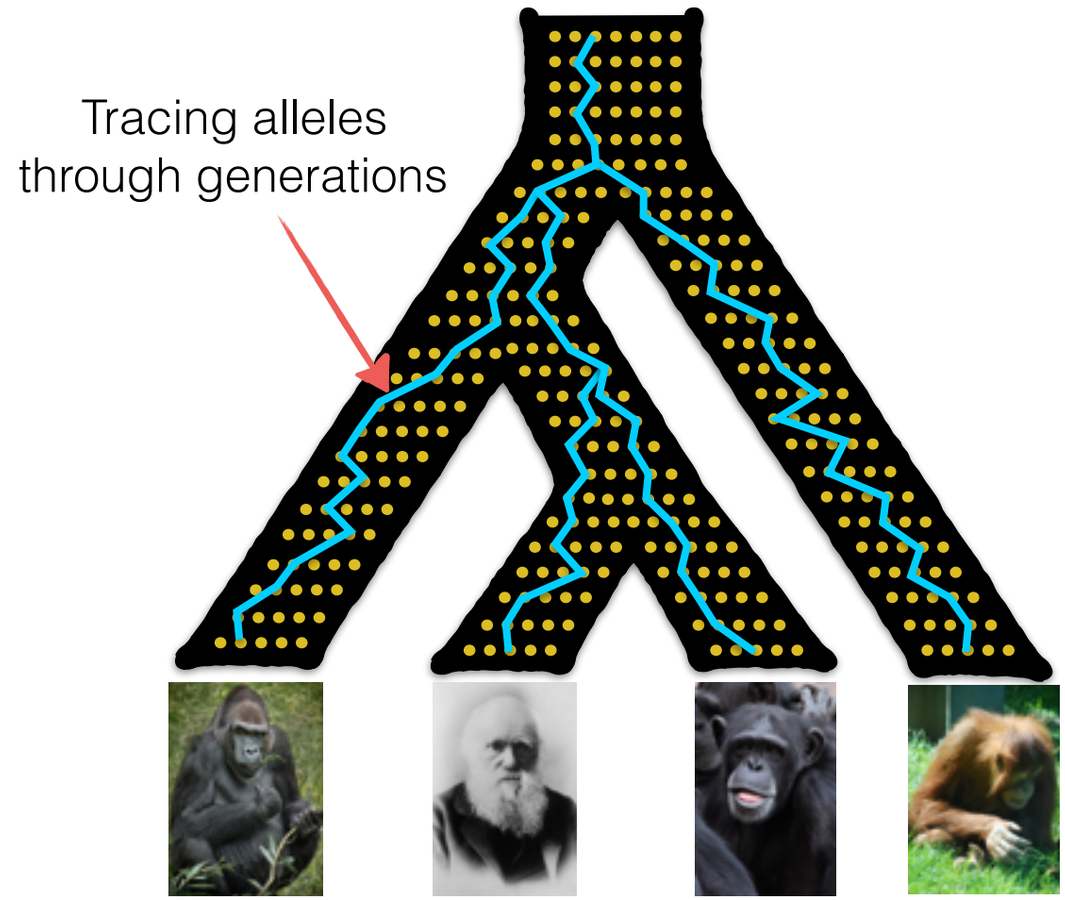


Causes of gene tree discordance include:

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

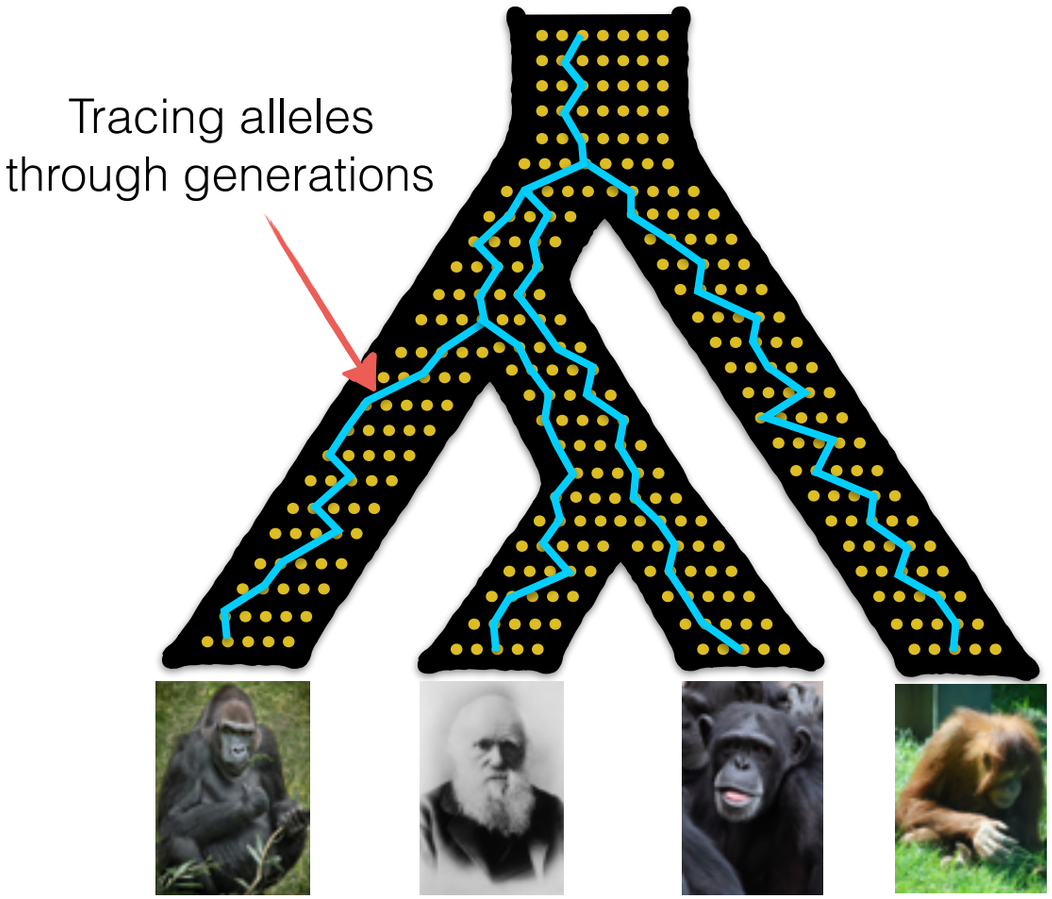
Incomplete Lineage Sorting (ILS)

- 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

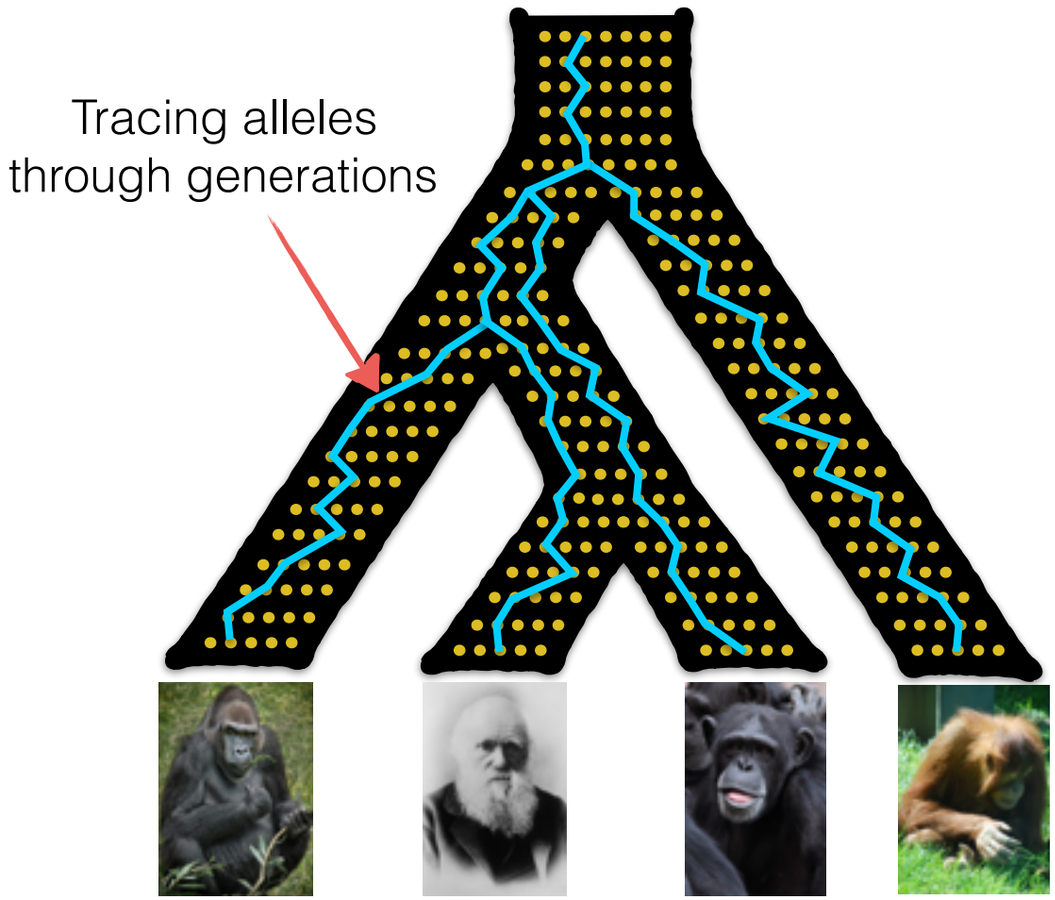


Incomplete Lineage Sorting (ILS)

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

- Omnipresent: possible for every tree

    - Likely for short branches or large population sizes



MSC and Identifiability

- A statistical model called multi-species coalescent (MSC) can generate ILS.

MSC and Identifiability

- A statistical model called multi-species coalescent (MSC) can generate ILS.
- Any species tree defines a unique distribution on the set of all possible gene trees

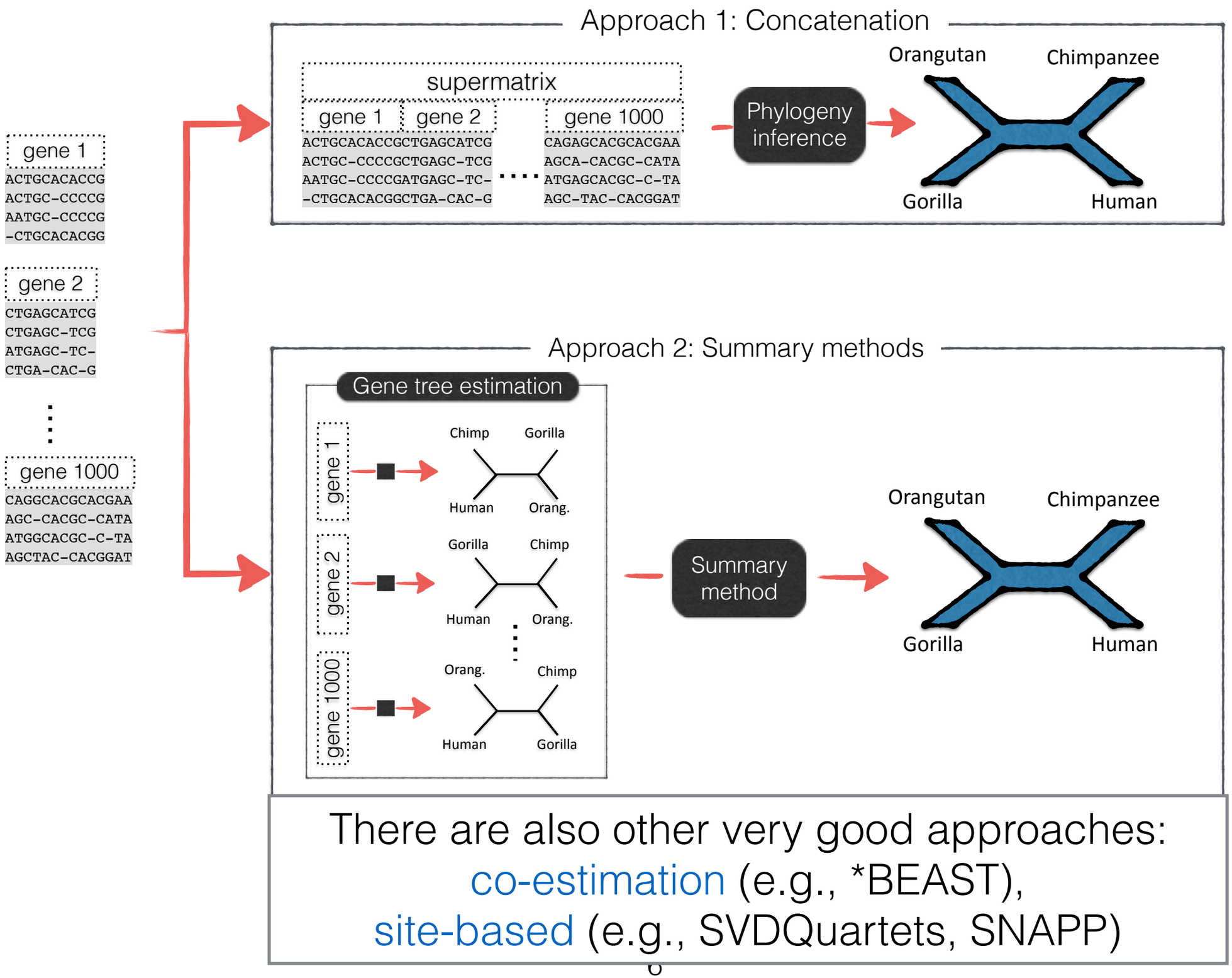
MSC and Identifiability

- A statistical model called multi-species coalescent (MSC) can generate ILS.

- Any species tree defines a unique distribution on the set of all possible gene trees

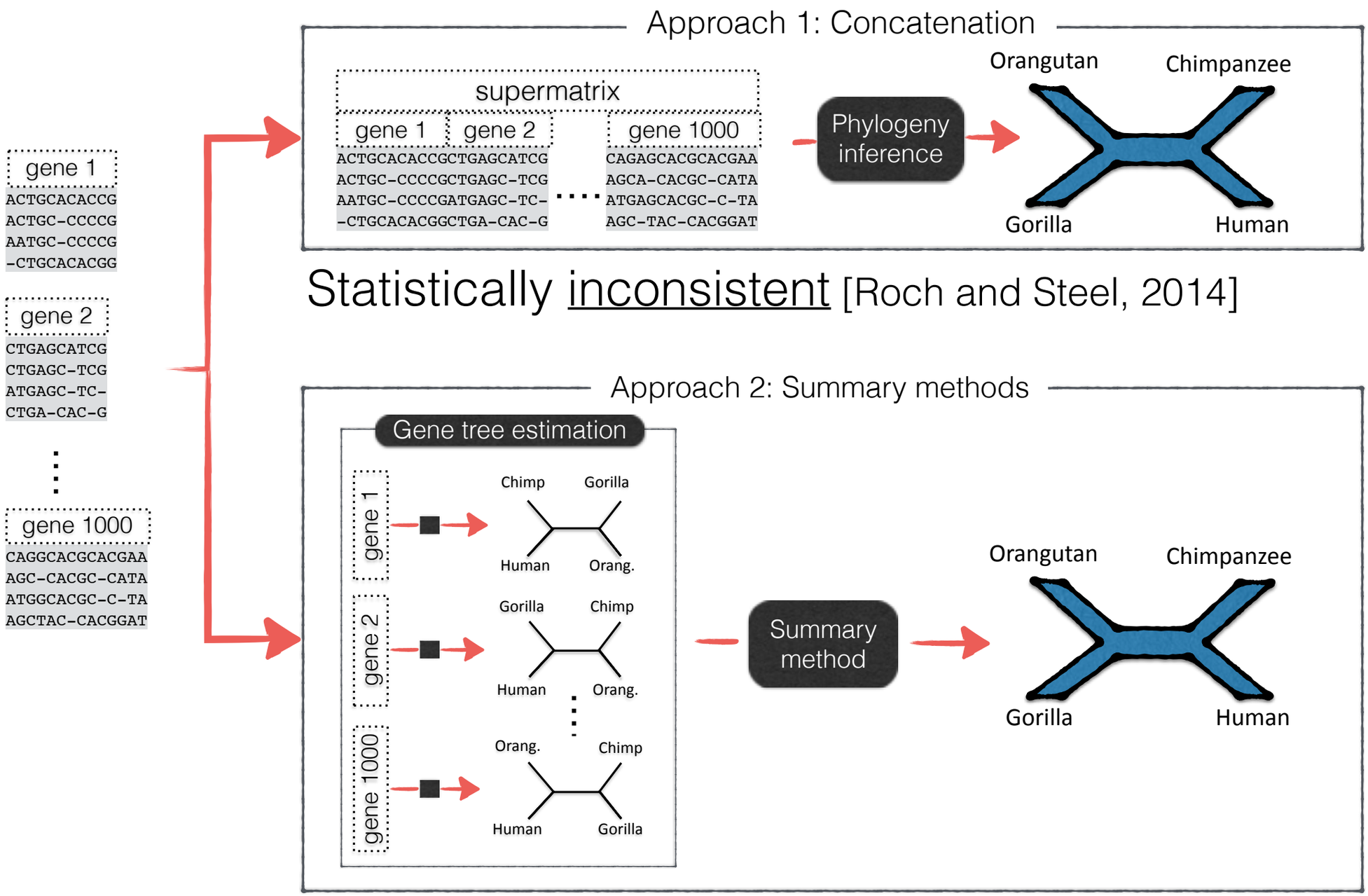
- In principle, the species tree can be identified despite high discordance from the gene tree distribution

Multi-gene tree estimation pipelines



There are also other very good approaches:
co-estimation (e.g., *BEAST),
site-based (e.g., SVDQuartets, SNAPP)

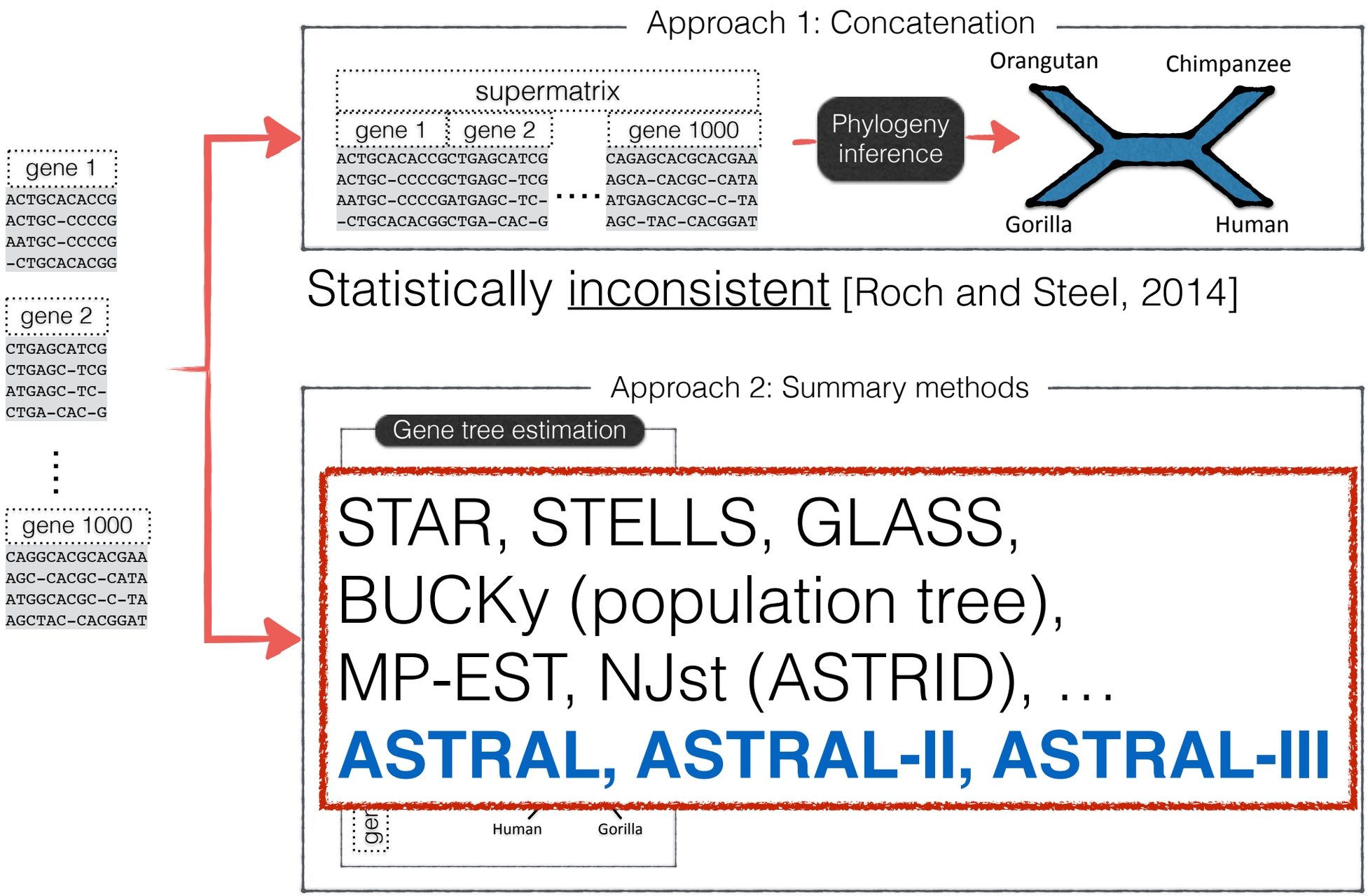
Multi-gene tree estimation pipelines



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

Can be statistically consistent given <u>true gene trees</u>

Multi-gene tree estimation pipelines

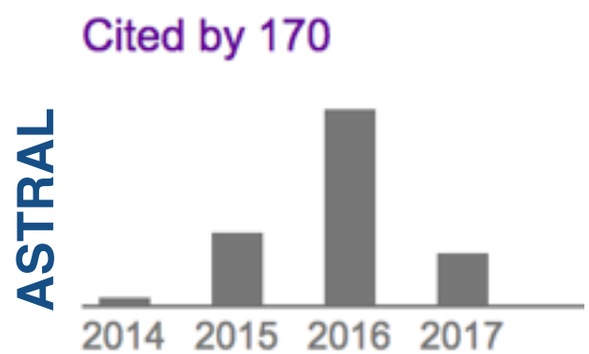


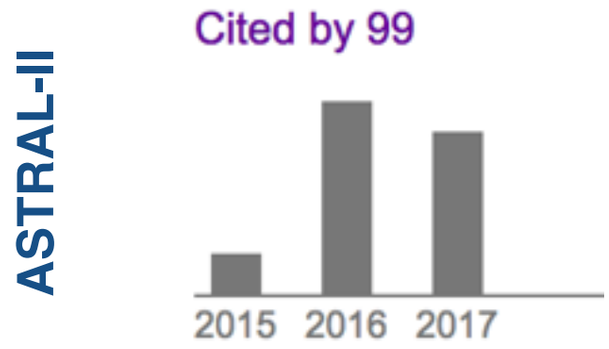
Statistically <u>inconsistent</u> [Roch and Steel, 2014]

Can be statistically consistent given <u>true gene trees</u>

ASTRAL used by the biologists

- Plants: Wickett, et al., 2014, PNAS
- Birds: Prum, et al., 2015, Nature
- Xenoturbella, Cannon et al., 2016, Nature
- Xenoturbella, Rouse et al., 2016, Nature
- Flatworms: Laumer, et al., 2015, eLife
- Shrews: Giarla, et al., 2015, Syst. Bio.
- Frogs: Yuan et al., 2016, Syst. Bio.
- Tomatoes: Pease, et al., 2016, PLoS Bio.
- Angiosperms: Huang et al., 2016, MBE
- Worms: Andrade, et al., 2015, MBE



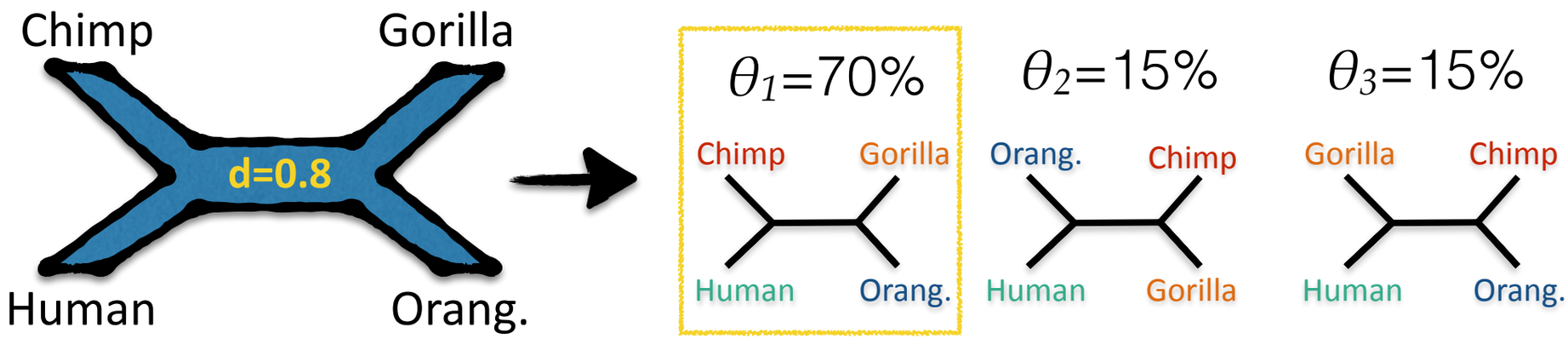


This talk: ASTRAL

- Outlines of the method
- Accuracy in simulation studies
  - With strong model violations
- The impact of
  - Fragmentary sequence data
  - Sampling multiple individuals

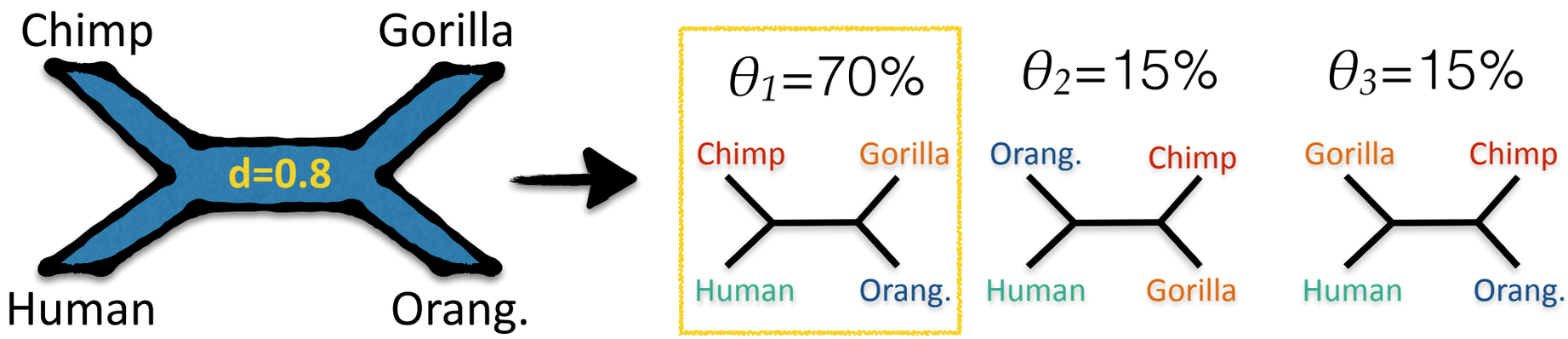
Unrooted quartets under MSC model

For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)



Unrooted quartets under MSC model

For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)



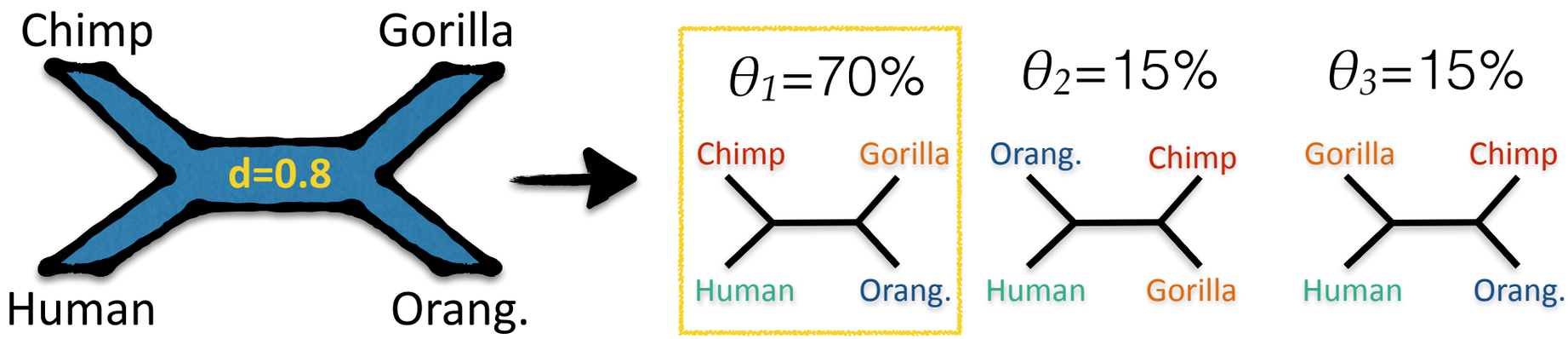
The most frequent gene tree

=

The most likely species tree

Unrooted quartets under MSC model

For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)



The most frequent gene tree

=

The most likely species tree

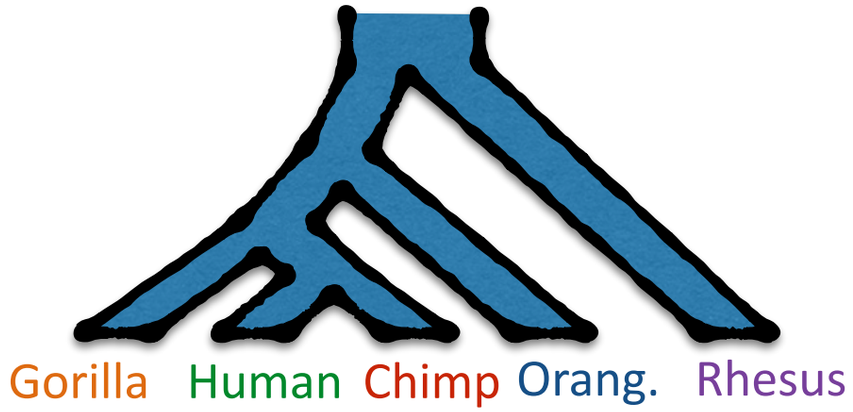
shorter branches ⇒

more discordance ⇒

a harder species tree reconstruction problem

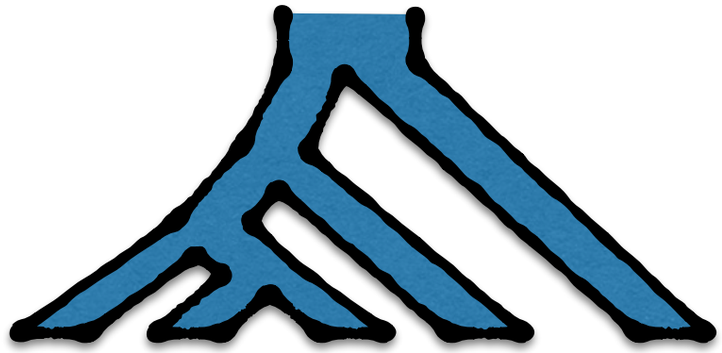
More than 4 species

For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)



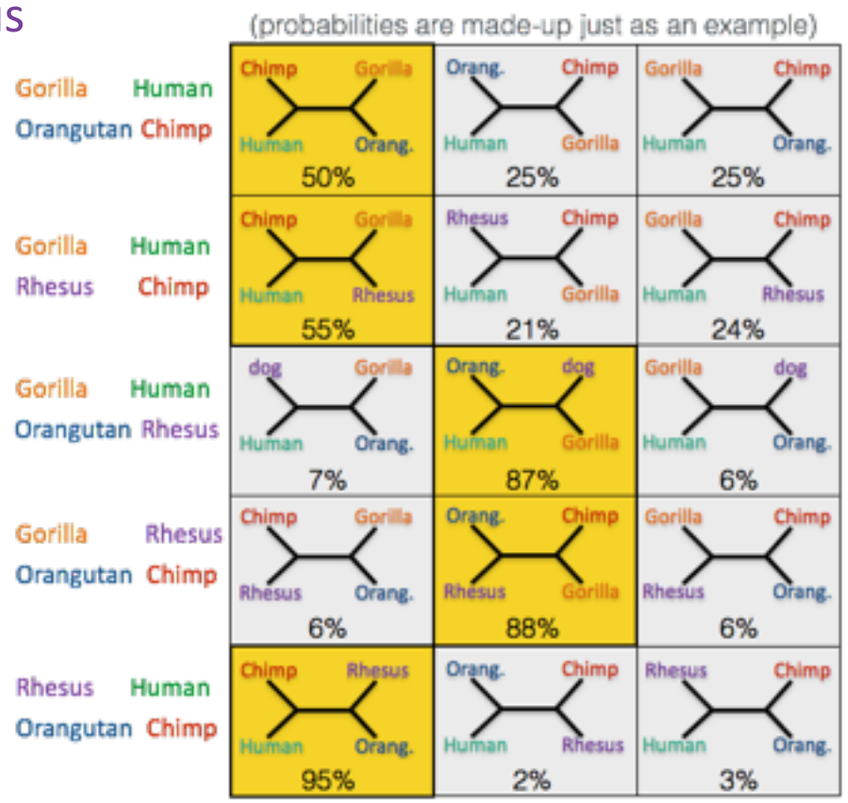
More than 4 species

For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)



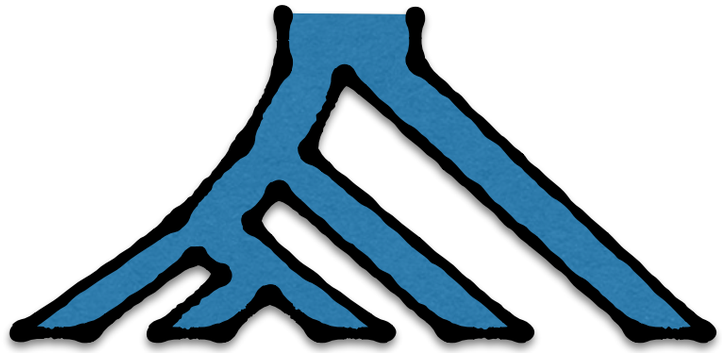
1. Break gene trees into $\binom{n}{4}$ quartets of species
2. Find the dominant tree for all quartets of taxa
3. Combine quartet trees

   Some tools (e.g.. BUCKy-p [Larget, et al., 2010])



More than 4 species

For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)



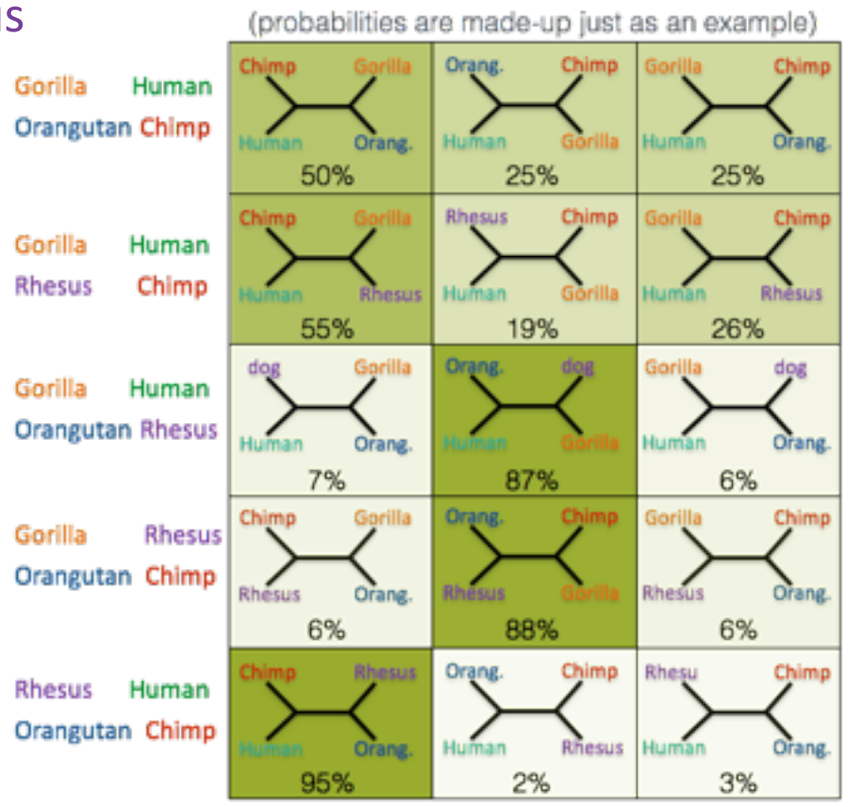
Gorilla Human Chimp Orang. Rhesus

1 s

2

3

Alternative:
weight all $3\binom{n}{4}$ quartet topologies
by their frequency
and find the optimal tree



Maximum Quartet Support Species Tree

- Optimization problem:

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

$$Score(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)|$$

the set of quartet trees induced by T

a gene tree

Maximum Quartet Support Species Tree

- Optimization problem:

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

$$Score(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)|$$

the set of quartet trees induced by T

a gene tree

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

ASTRAL-I and ASTRAL-II

[Mirarab, et al., Bioinformatics, 2014] [Mirarab and Warnow, Bioinformatics, 2015]

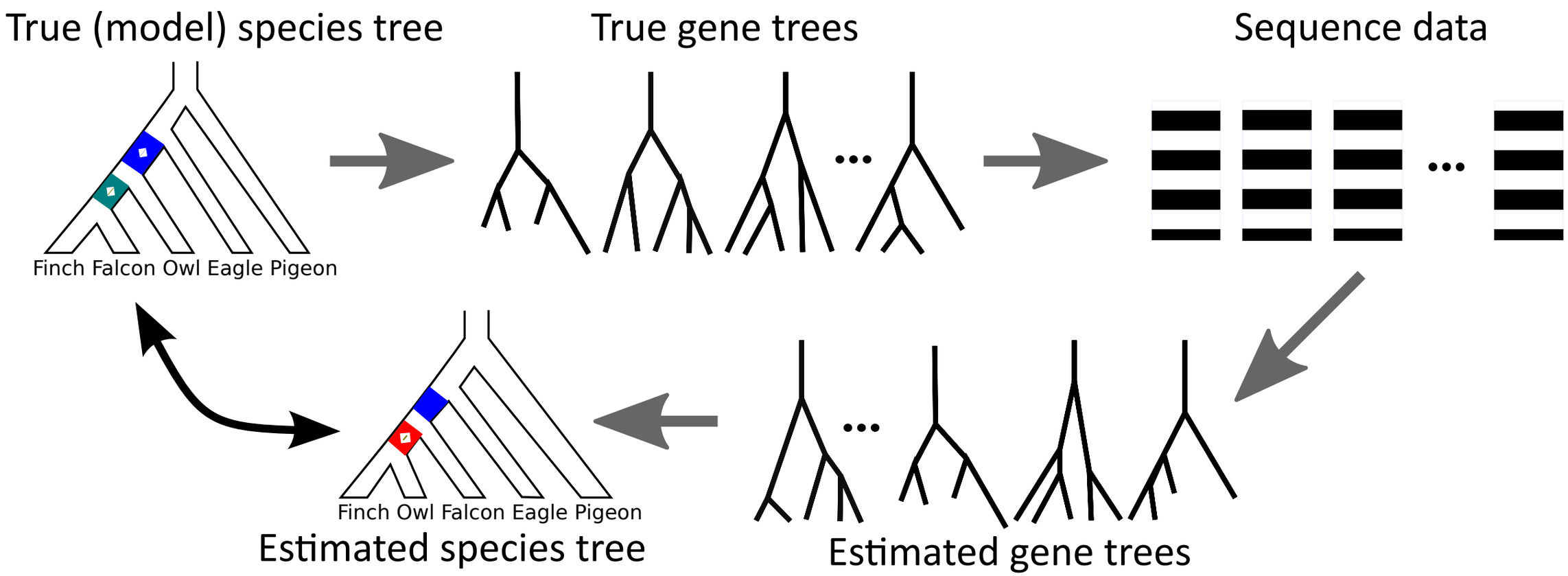
- Solve the problem exactly using dynamic programming
  - Constrains the search space to make large datasets feasible
  - The constrained version remains statistically consistent
  - Running time: polynomially increases with the number of genes and the number of species

ASTRAL-I and ASTRAL-II

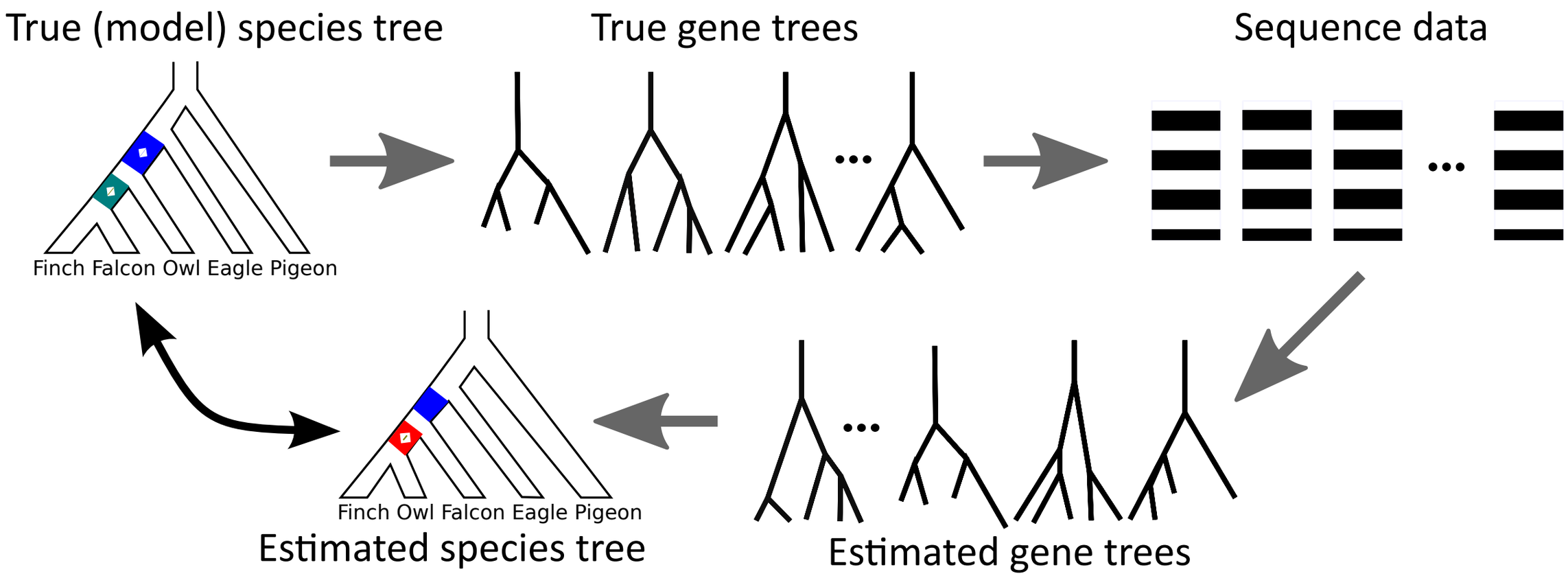
[Mirarab, et al., Bioinformatics, 2014] [Mirarab and Warnow, Bioinformatics, 2015]

- Solve the problem exactly using dynamic programming
  - Constrains the search space to make large datasets feasible
  - The constrained version remains statistically consistent
    - Running time: polynomially increases with the number of genes and the number of species
- ASTRAL-II:
  - Increased the search space
  - Improved the running time
  - Can handle polytomies (lack of resolution) in input gene trees

Simulation study

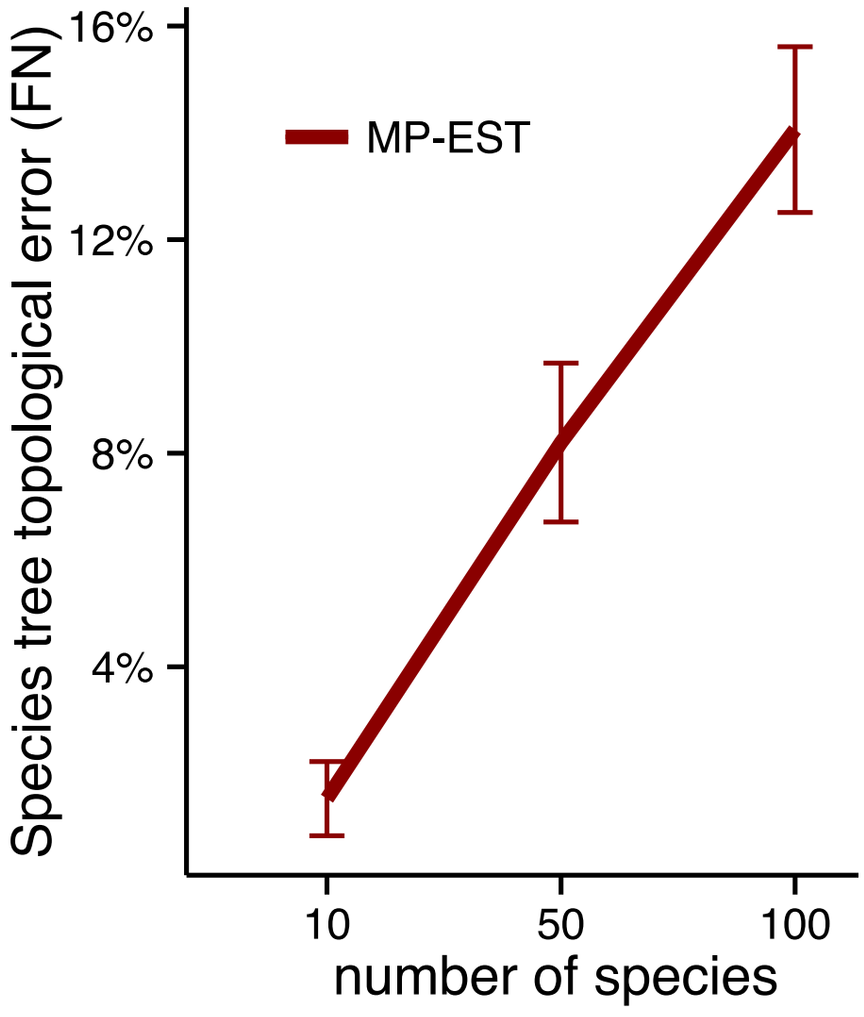


Simulation study



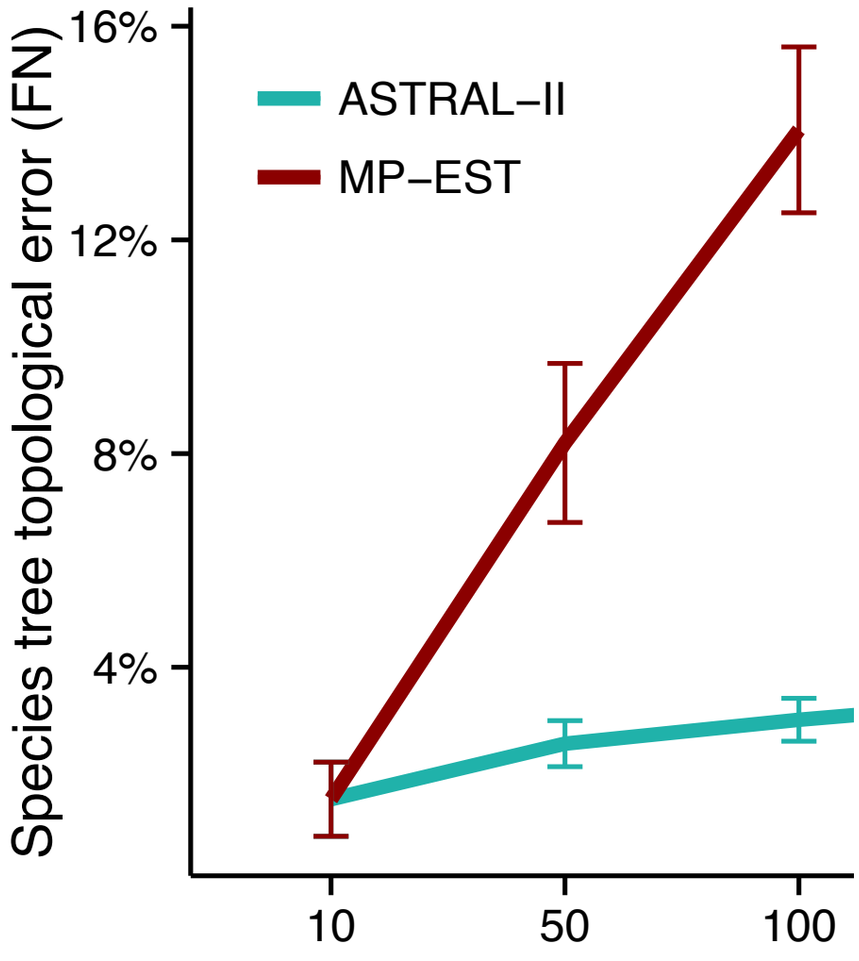
- Evaluate using the **FN rate**: the percentage of branches in the true tree that are missing from the estimated tree

Number of species impacts estimation error in the species tree



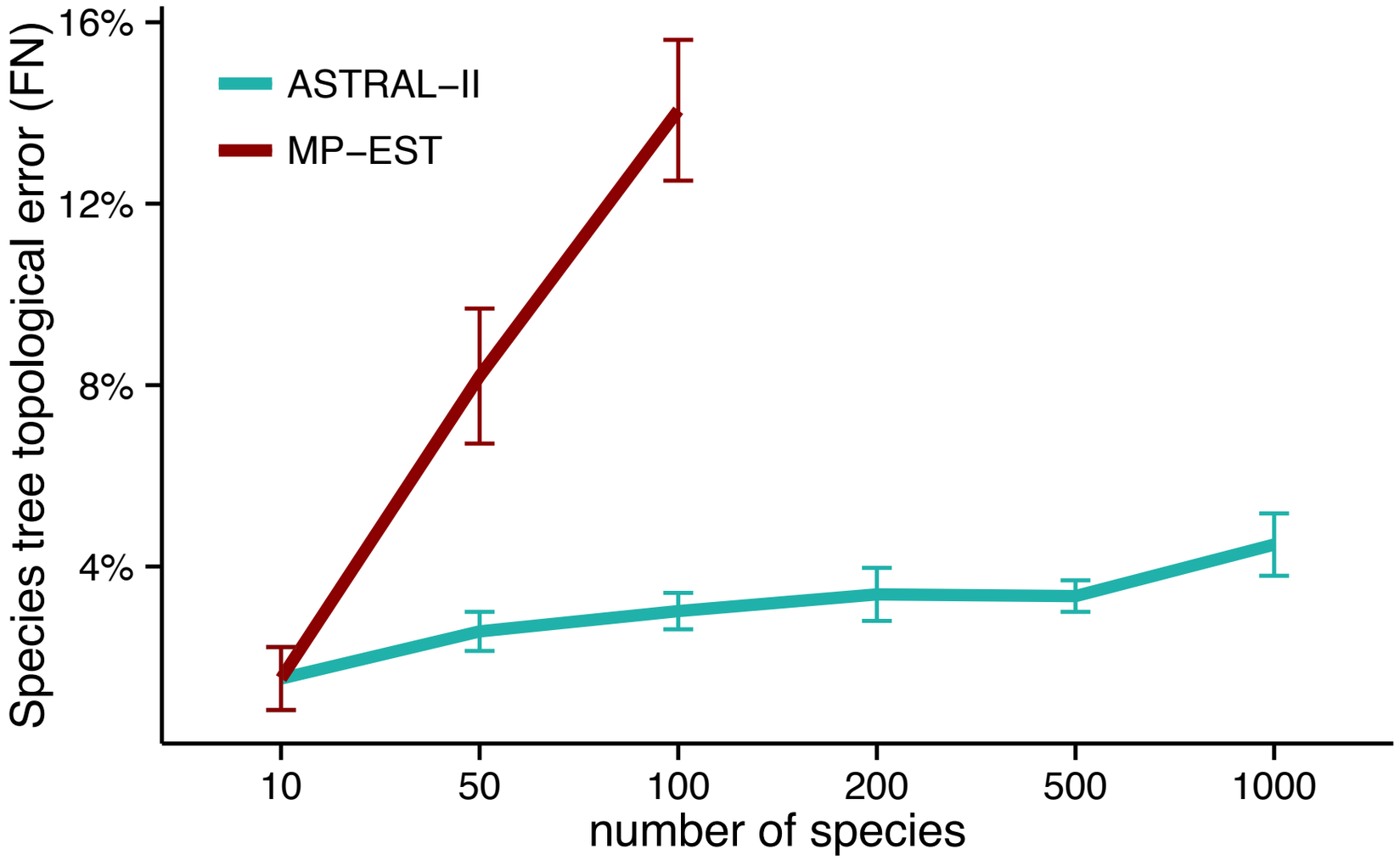
1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]

ASTRAL: accurate and scalable



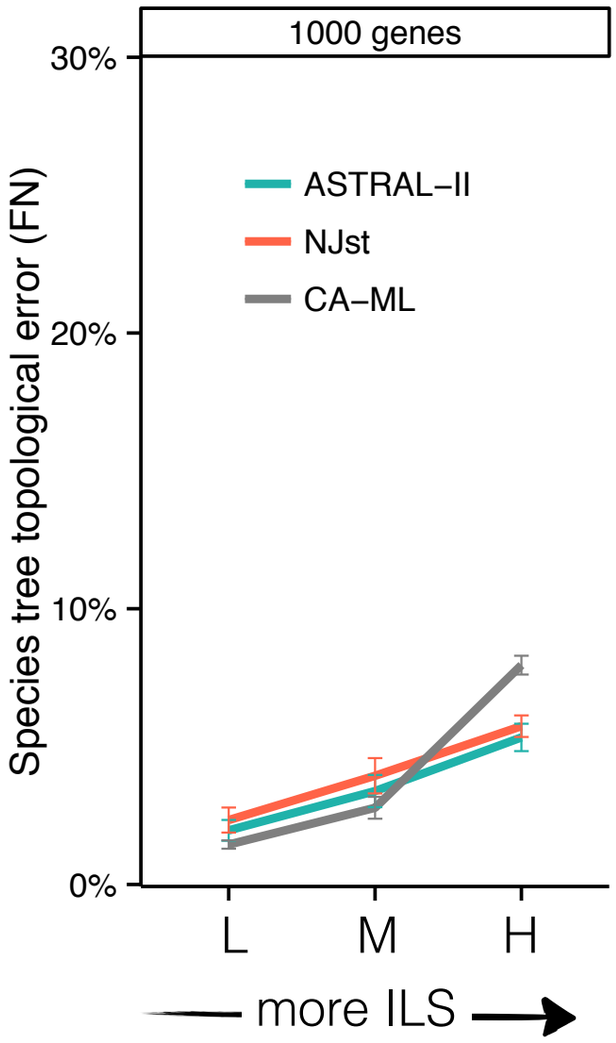
1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]

ASTRAL: accurate and scalable



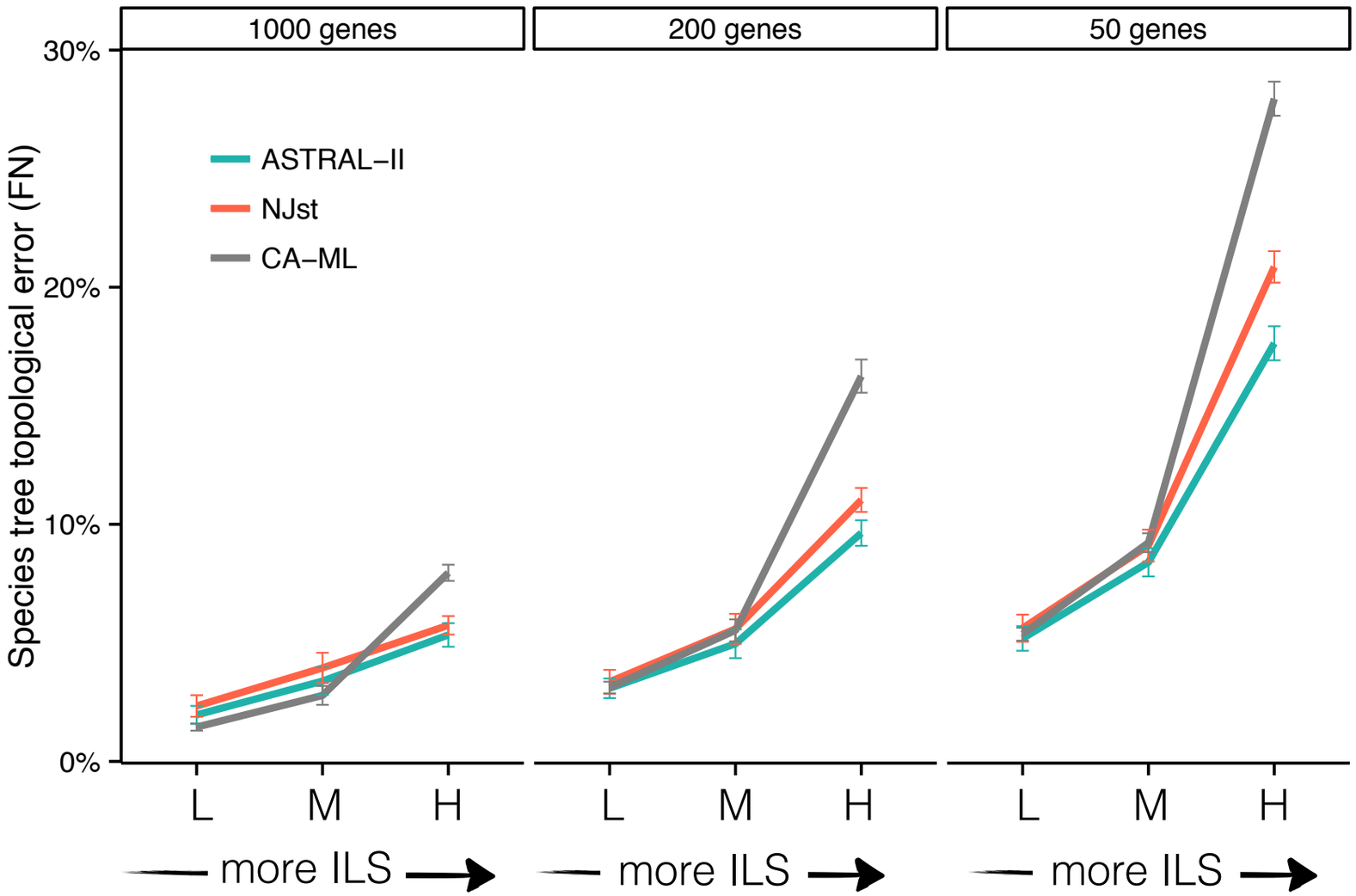
1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]

Comparison to concatenation: depends on the ILS level



200 species, recent ILS

Comparison to concatenation: depends on the ILS level

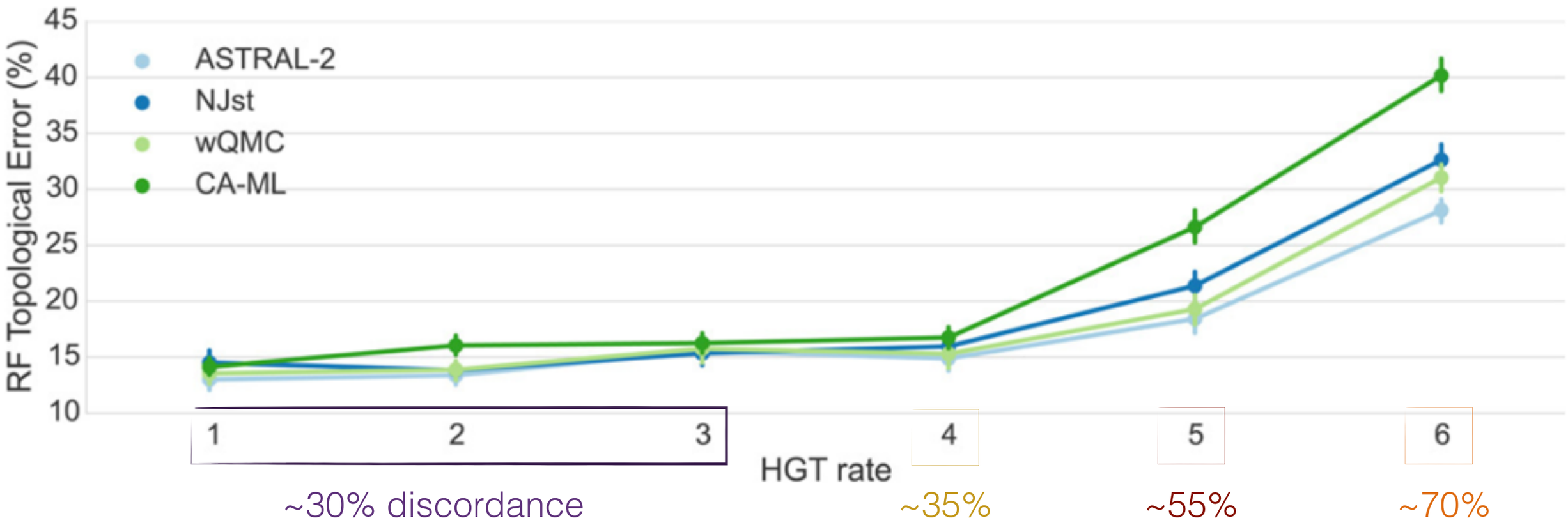


200 species, recent ILS

Horizontal Gene Transfer (HGT)

[R. Davidson et al., BMC Genomics. 16 (2015)]

Model violation: the simulated discordance is due to both ILS and randomly distributed HGT

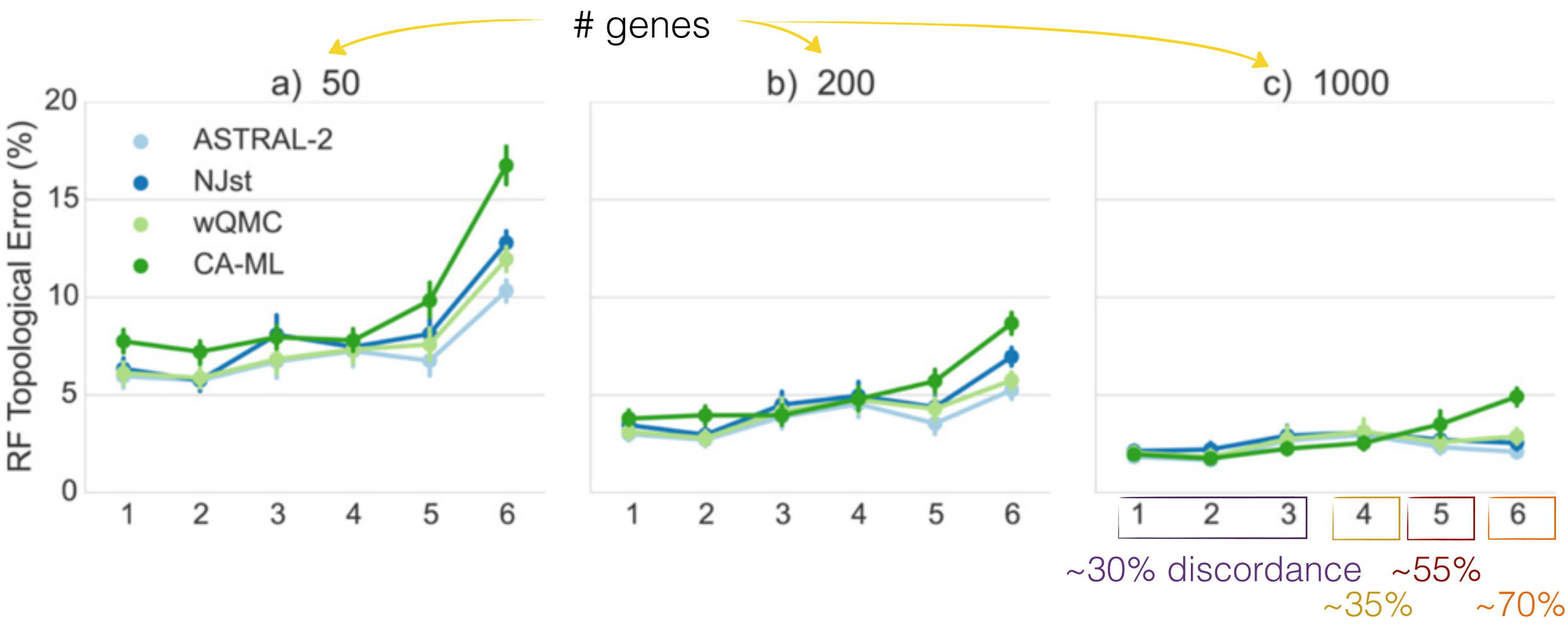


50 species, 10 genes

Horizontal Gene Transfer (HGT)

[R. Davidson et al., BMC Genomics. 16 (2015)]

Randomly distributed HGT is tolerated with enough genes



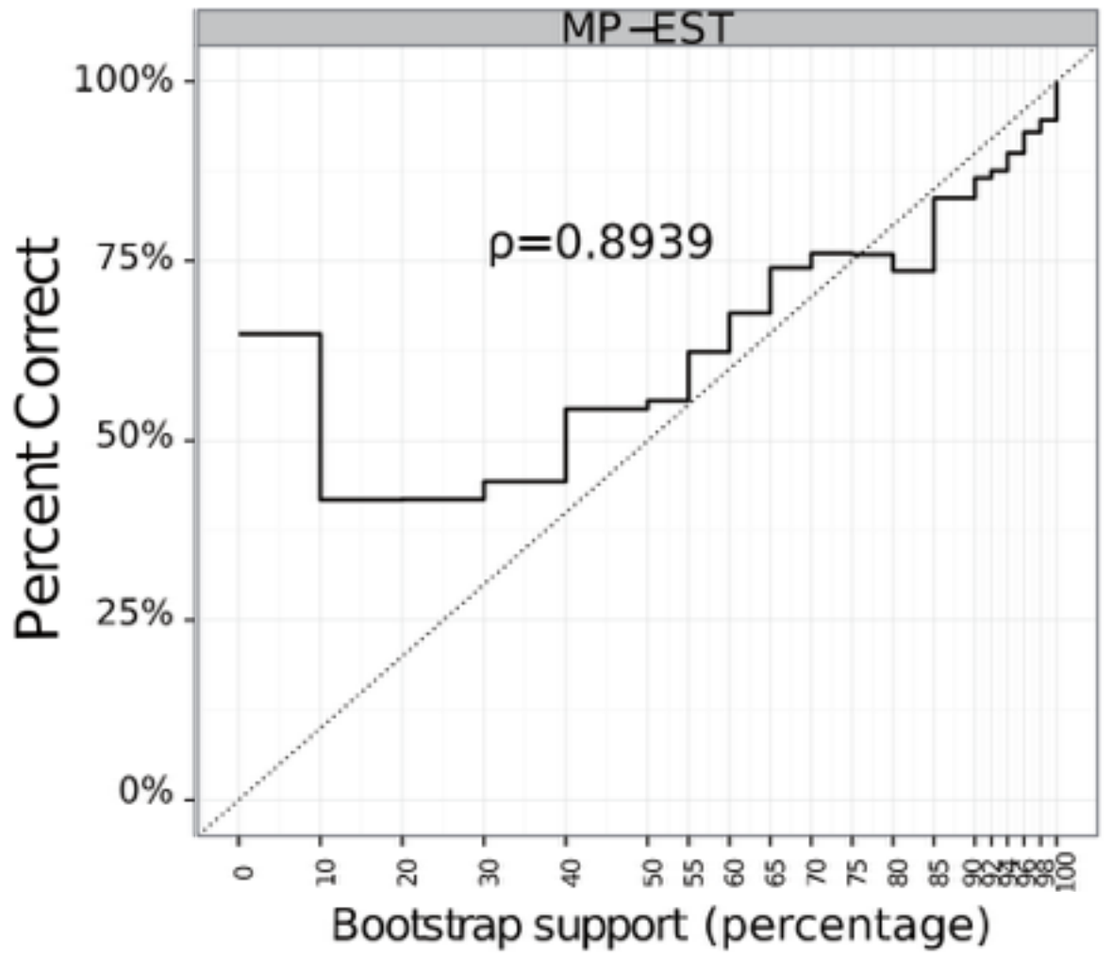
50 species, varying # genes

UCSD Work

1. Statistical support
2. ASTRAL-III
    - Better running time
    - GPU and CPU Parallelism
    - Multi-Individual datasets
3. Impact of fragmentary data

Branch support

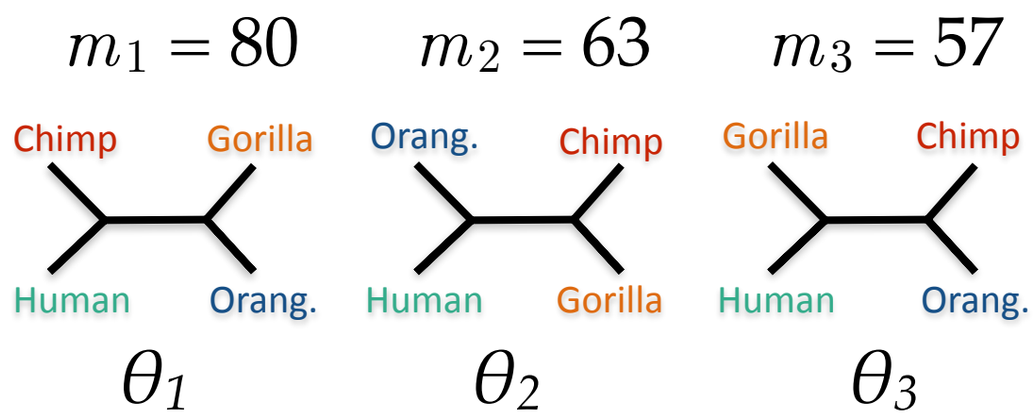
- Traditional approach:
  Multi-locus bootstrapping (MLBS)
  - Slow: requires bootstrapping all genes (e.g., $100\times m$ ML trees)
  - Inaccurate and hard to interpret
    [Mirarab et al., Sys bio, 2014; Bayzid et al., PLoS One, 2015]
- We can do better!



[Mirarab et al., Sys bio, 2014]

Local posterior probability

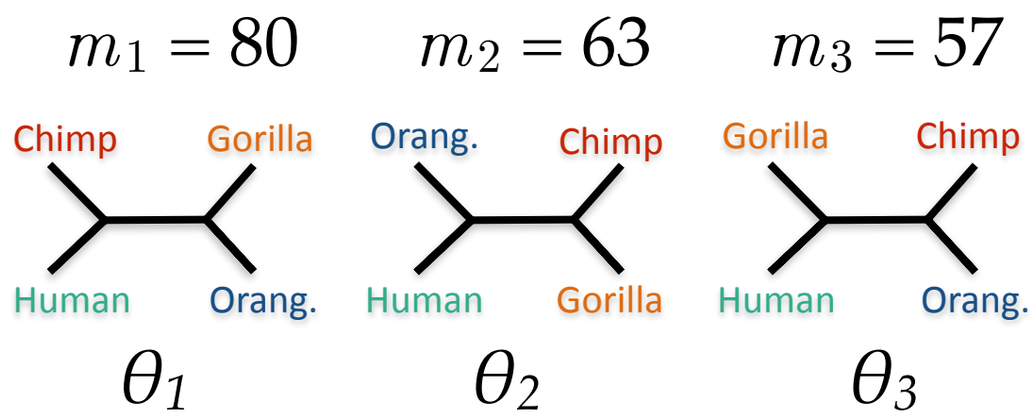
- Recall quartet frequencies follow a multinomial distribution



- $\boldsymbol{P}$ (topology seen in m_1 / m gene trees is the species tree) = $\boldsymbol{P}$ ($\theta_1 > 1/3$)

Local posterior probability

- Recall quartet frequencies follow a multinomial distribution

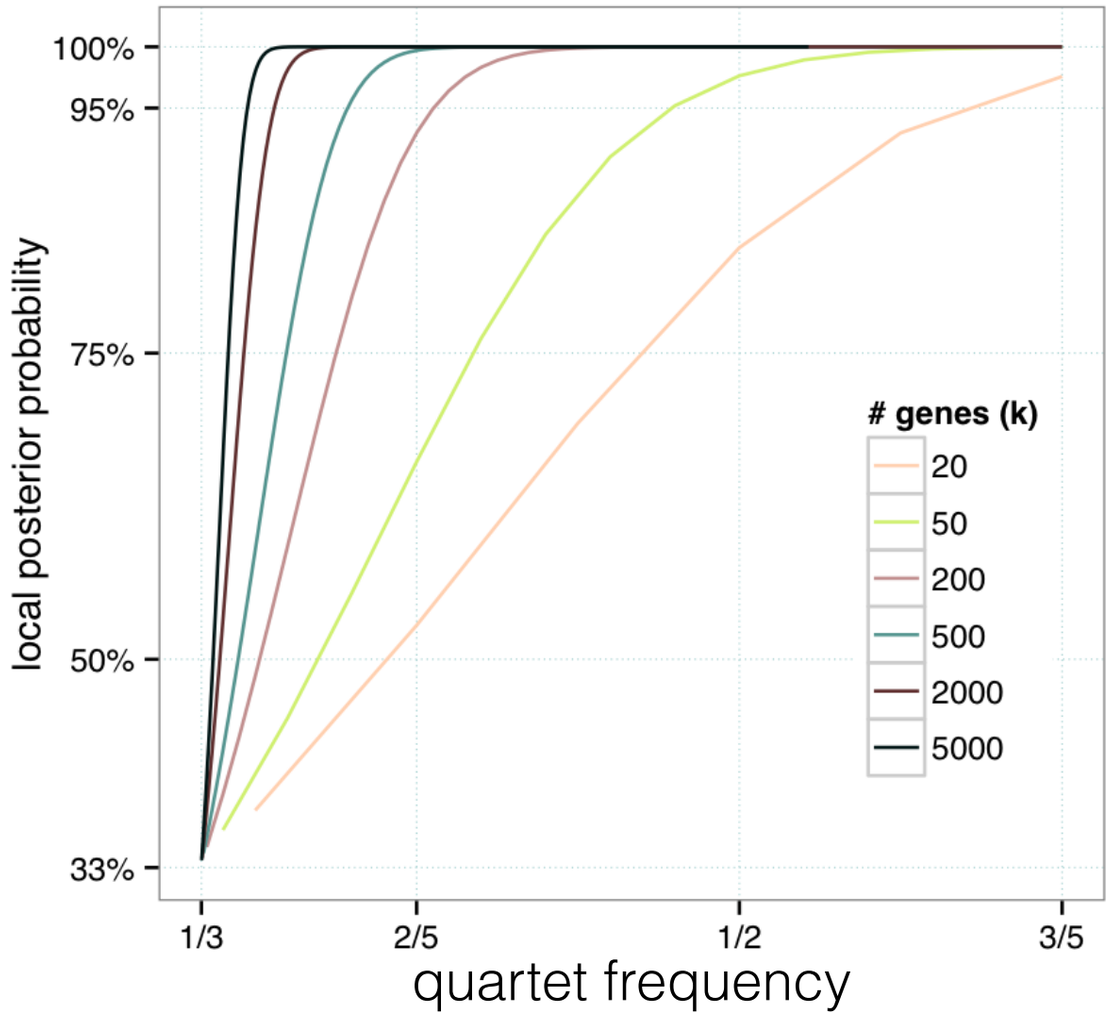


- $\boldsymbol{P}$ (topology seen in m_1 / m gene trees is the species tree) = $\boldsymbol{P}$ ($\theta_1 > 1/3$)

- Can be analytically solved

  - We implemented this idea in astral and called the measure “the local posterior probability” (localPP)

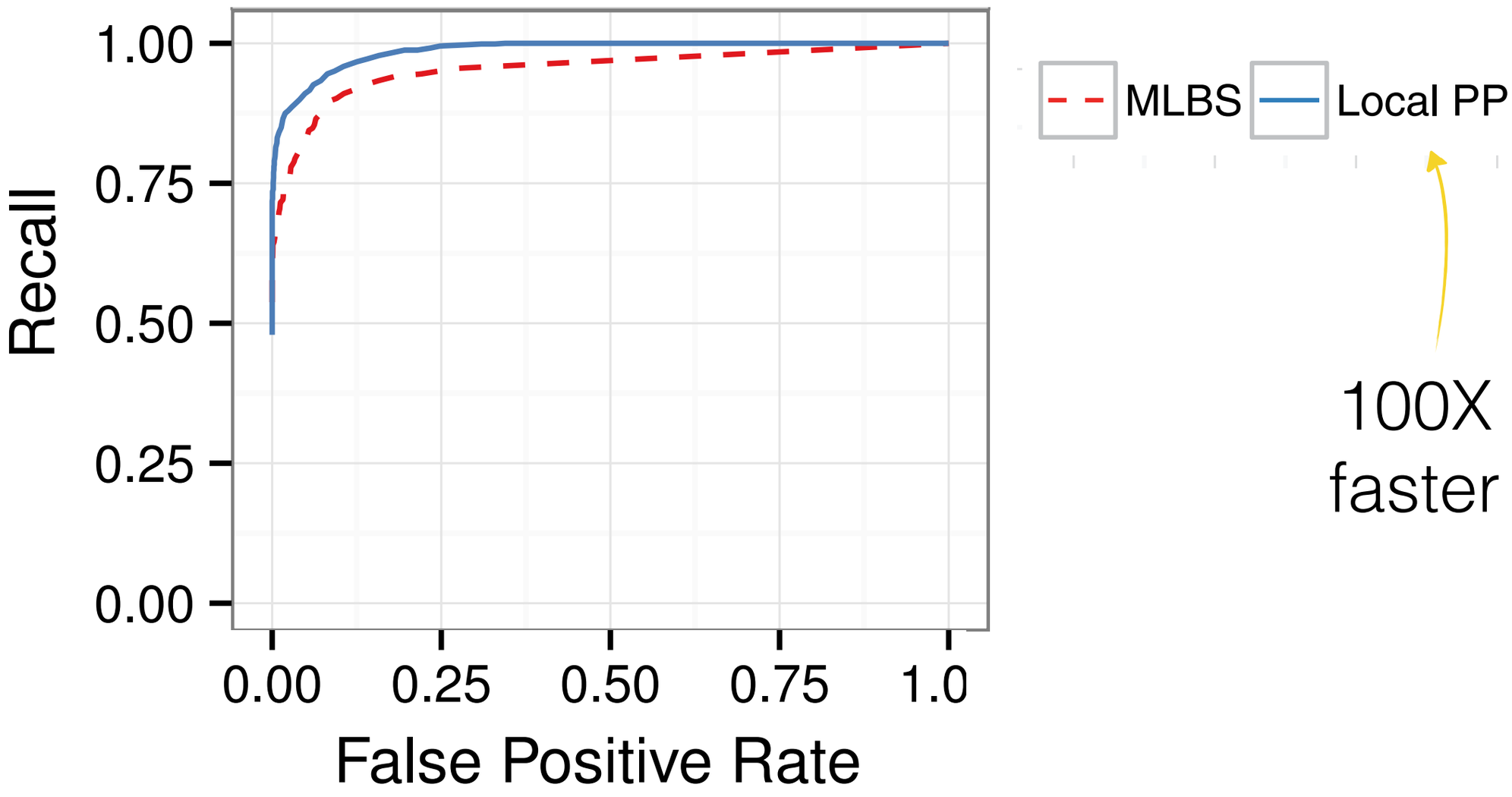
Quartet support v.s. localPP



Increased number of genes (m) $\Rightarrow$ increased support

Decreased discordance $\Rightarrow$ increased support

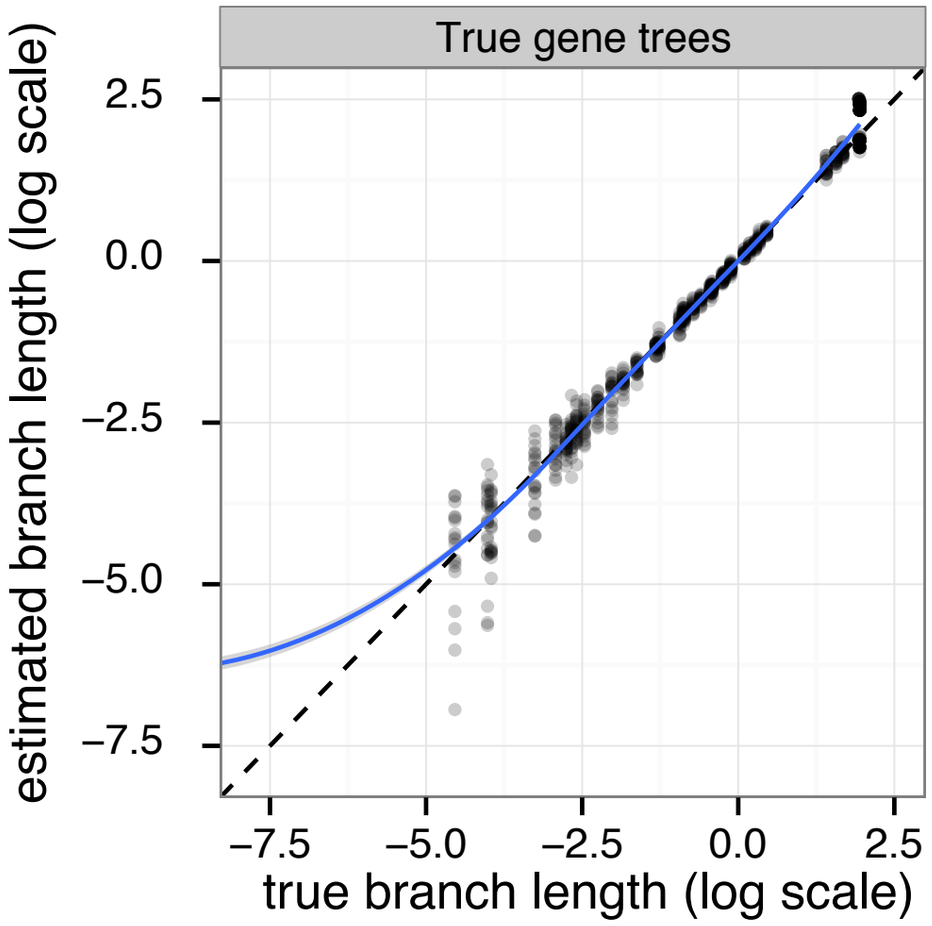
localPP is more accurate than bootstrapping



Avian simulated dataset (48 taxa, 1000 genes)

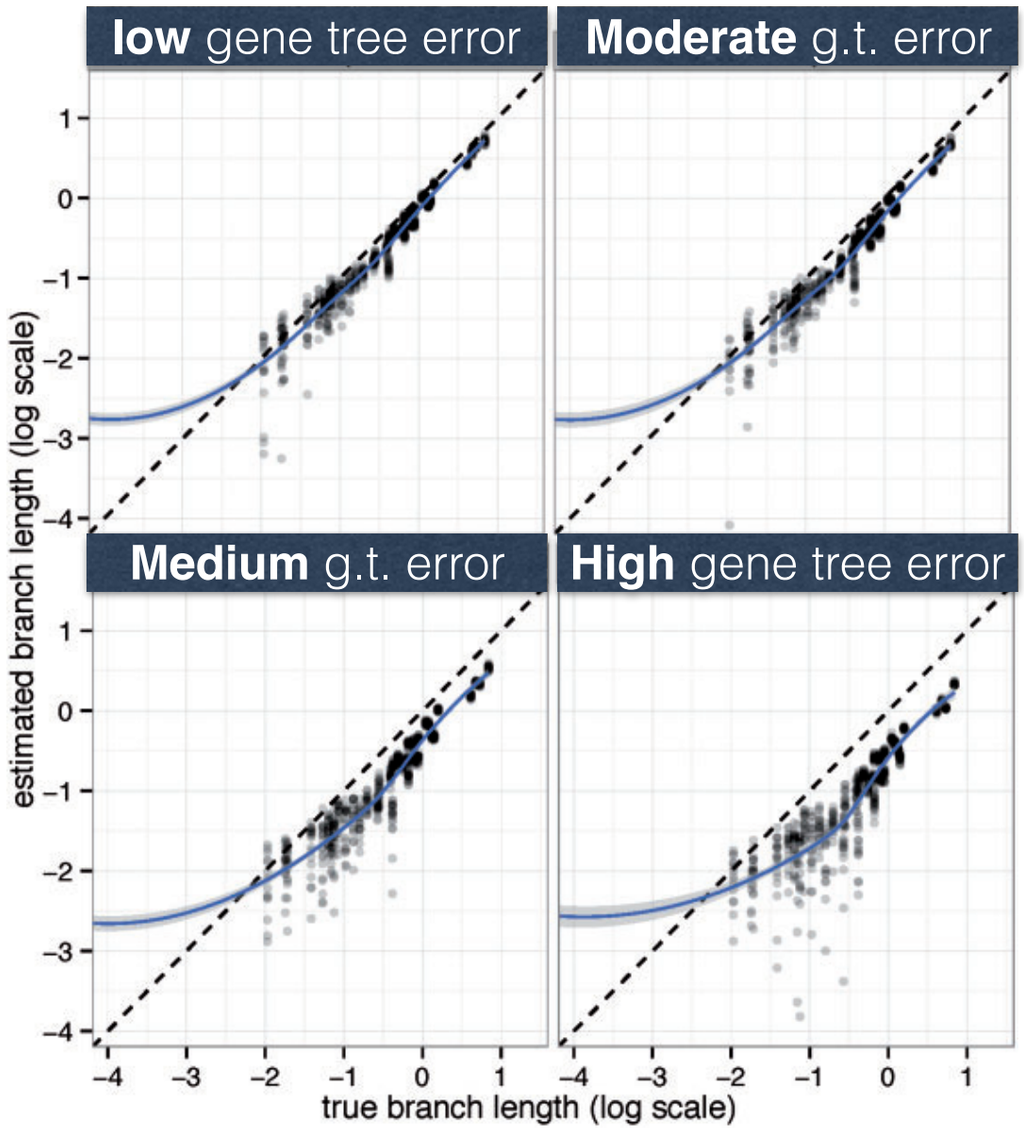
[Sayyari and Mirarab, MBE, 2016]

ASTRAL can also estimate internal branch lengths



With true gene trees, ASTRAL correctly estimates BL

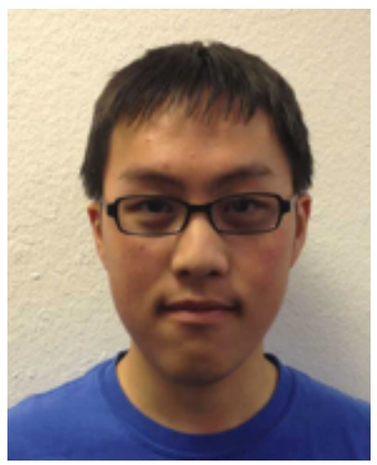
ASTRAL can also estimate internal branch lengths



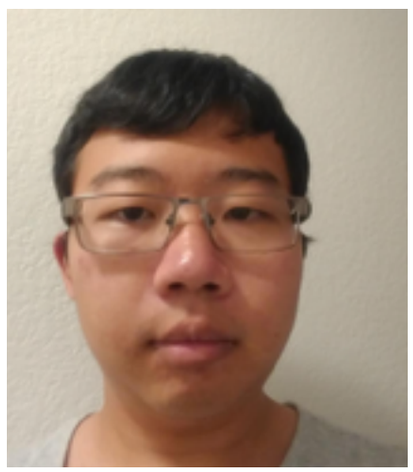
With error-prone estimated gene trees, ASTRAL underestimates BL

Running time complexity & ASTRAL-III (unpublished work)

Maryam Rabiee Hashemi



John Yin

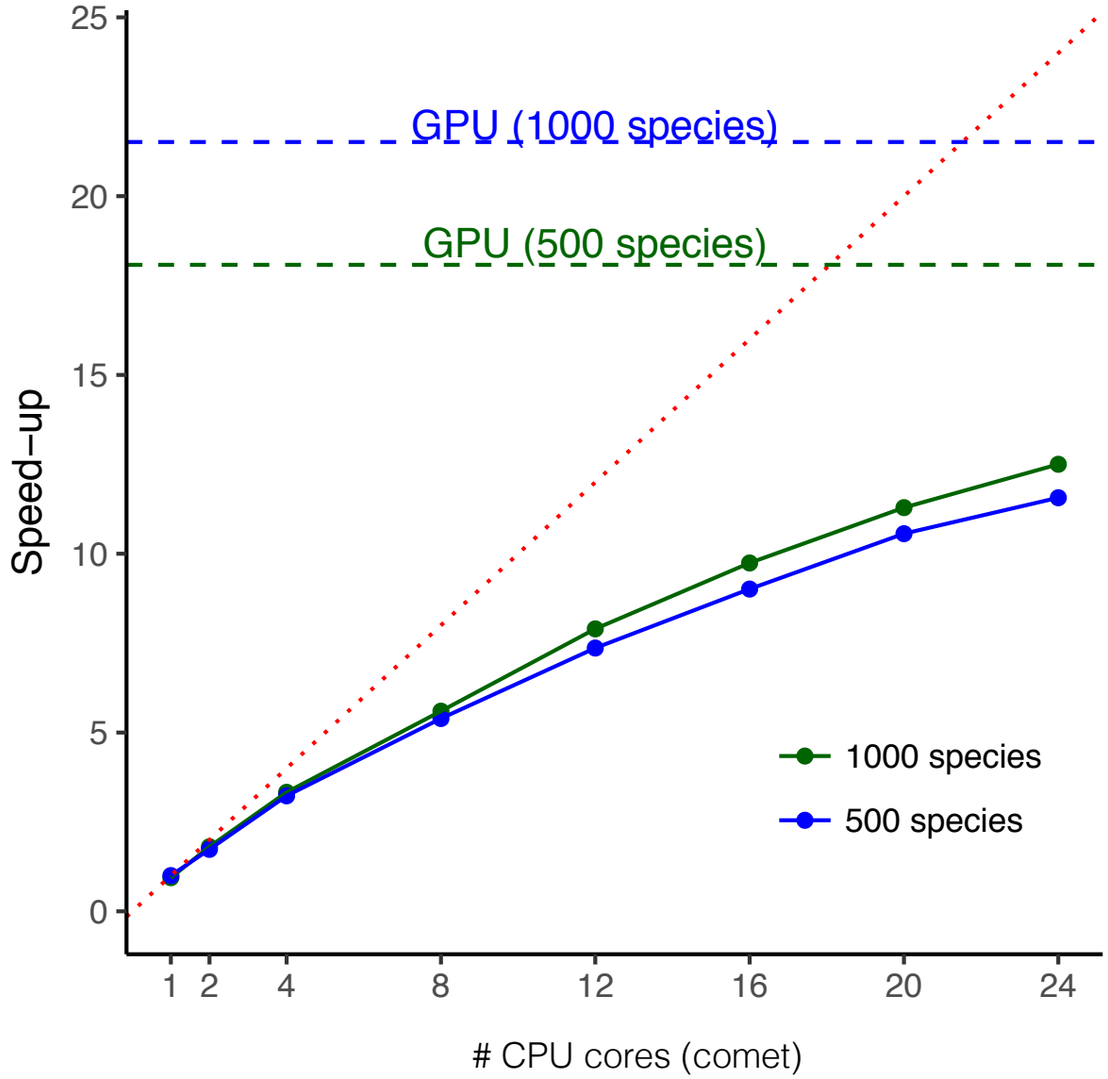


Chao Zhang

ASTRAL-III new features

- Improved running time with a single CPU
  - We have ~3-5X running time improvement, and guaranteed polynomial running time
  - ~8 hours for 1000 genes, 1000 species
- Handling datasets with multiple individuals
- GPU and CPU parallelism

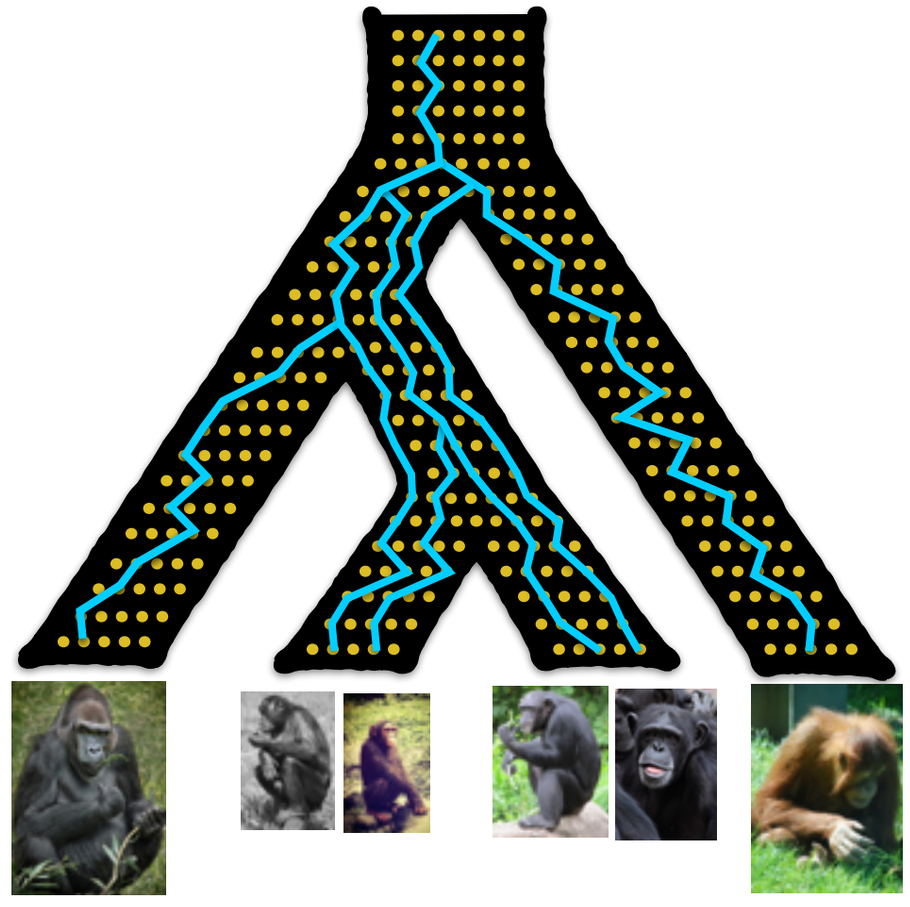
Parallelism



Can infer trees with 10,000 species & 400 genes in a day

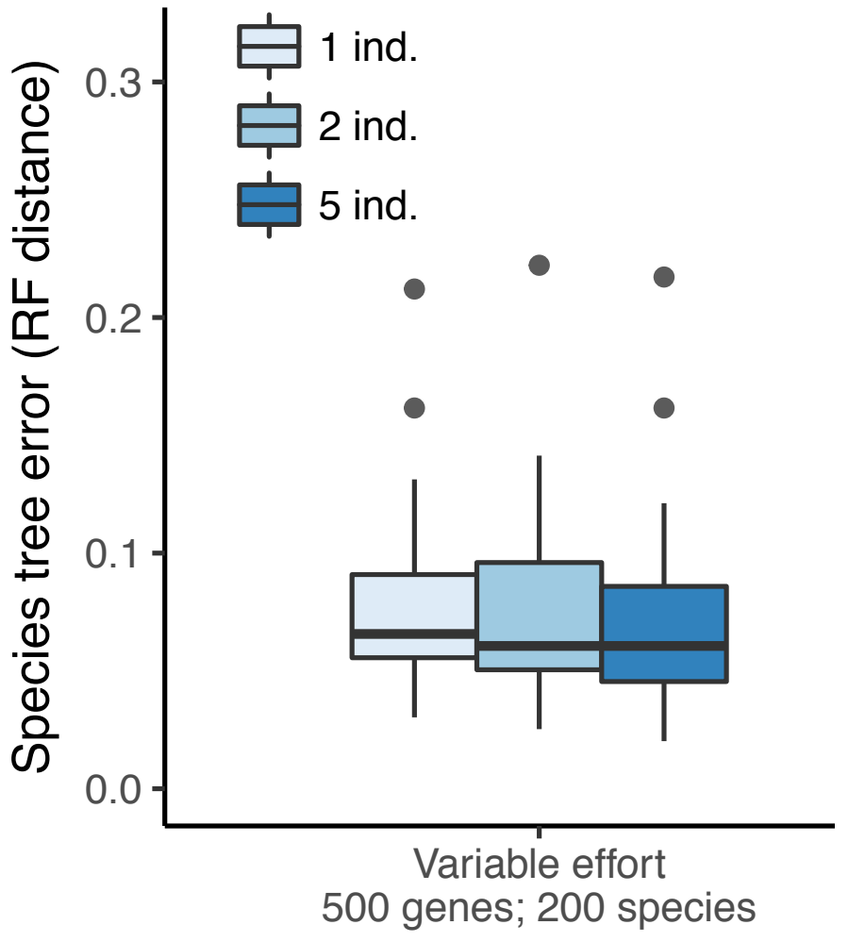
Multiple individuals

- What if we sample multiple individuals from each species?
- In recently diverged species individuals *may* have different trees for each gene
- Sampling multiple individuals *may* provide useful information
- The ASTRAL approach can be extended to these types of data



Multiple individuals helpful?

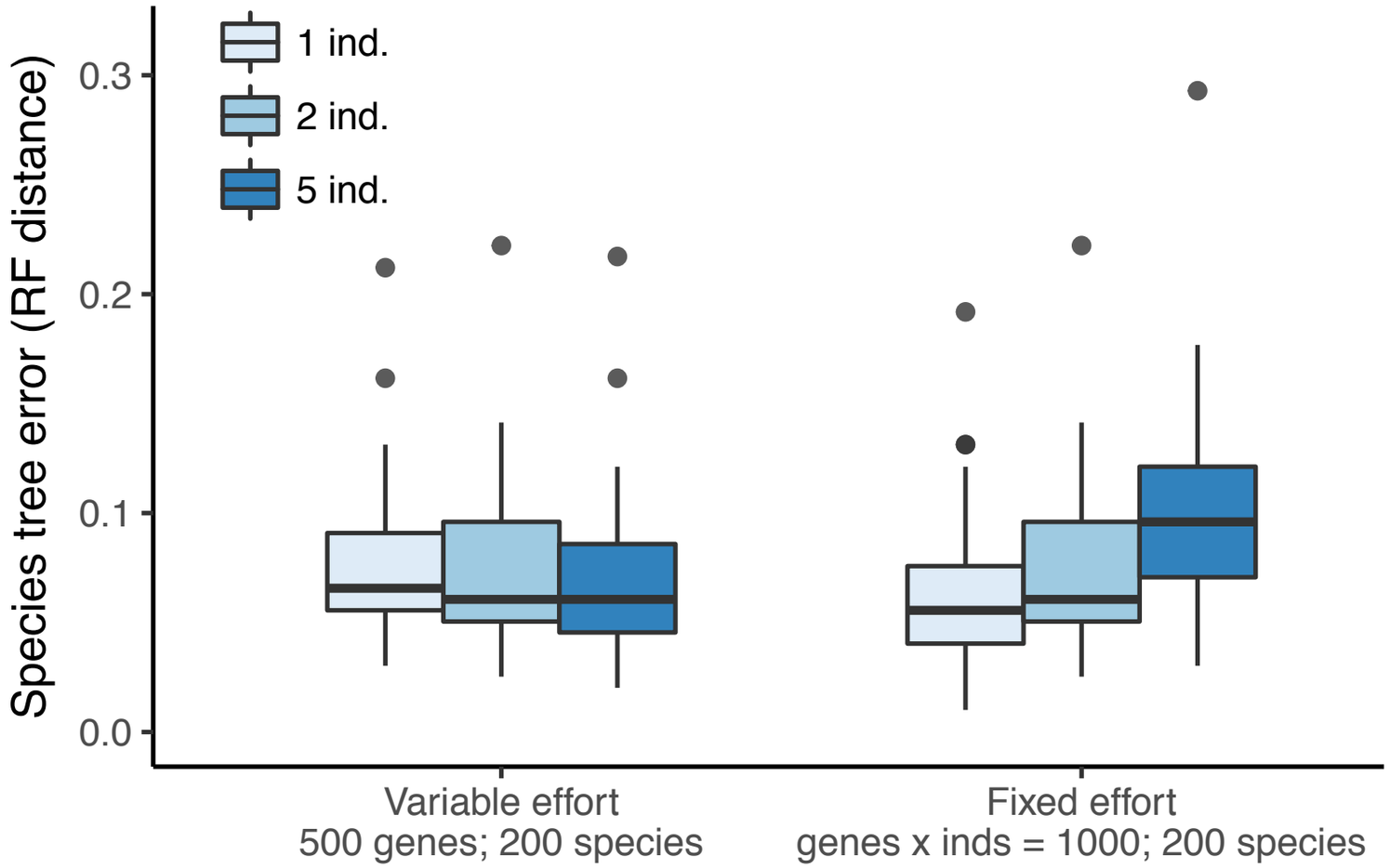
Simulations with very sallow trees (500K generations in total)



Yes, it marginally helps accuracy

Multiple individuals helpful?

Simulations with very sallow trees (500K generations in total)

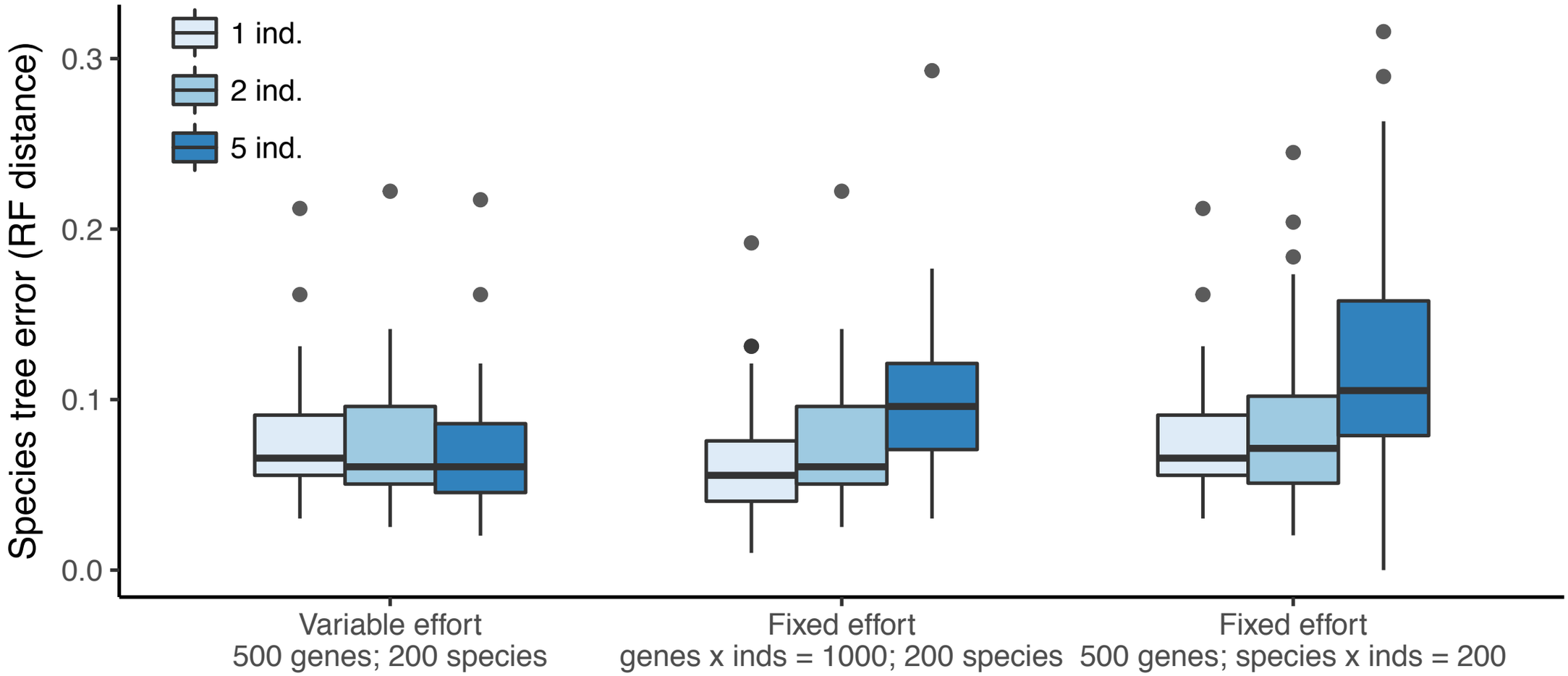


Yes, it marginally helps accuracy

But **not** if sequencing **effort** is kept **fixed**

Multiple individuals helpful?

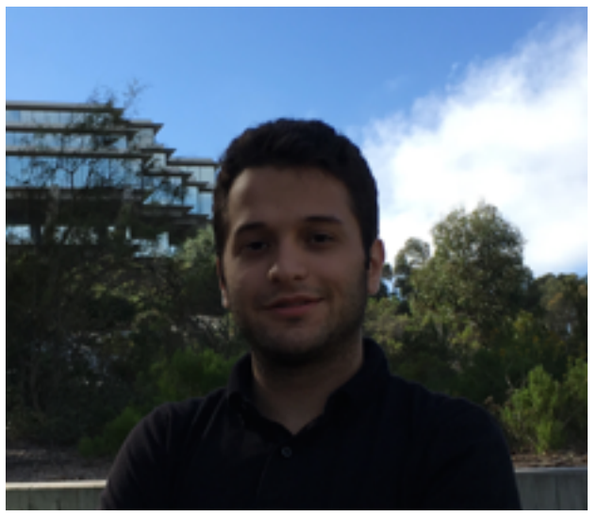
Simulations with very sallow trees (500K generations in total)



Yes, it marginally helps accuracy

But **not** if sequencing **effort** is kept **fixed**

Impact of fragmentary data

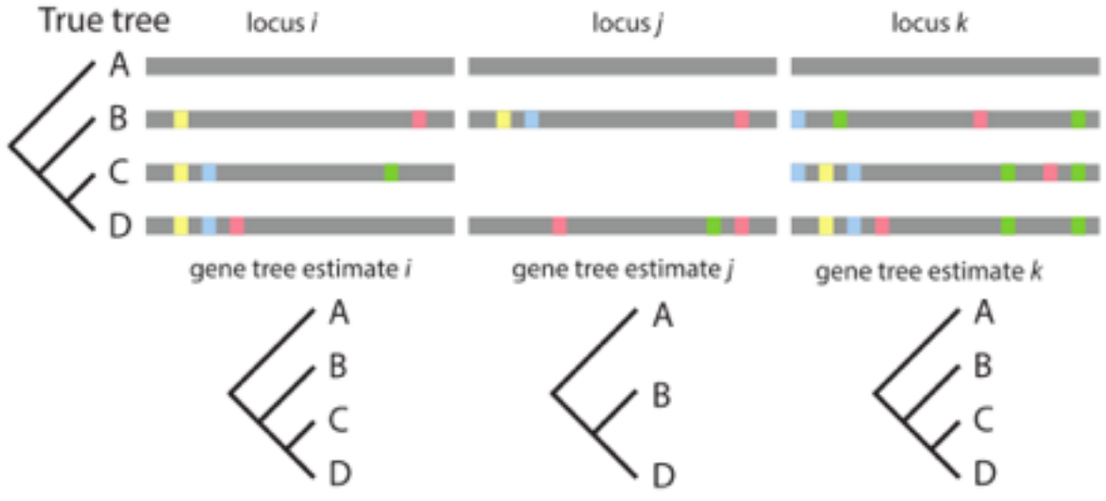


Erfan Sayyari

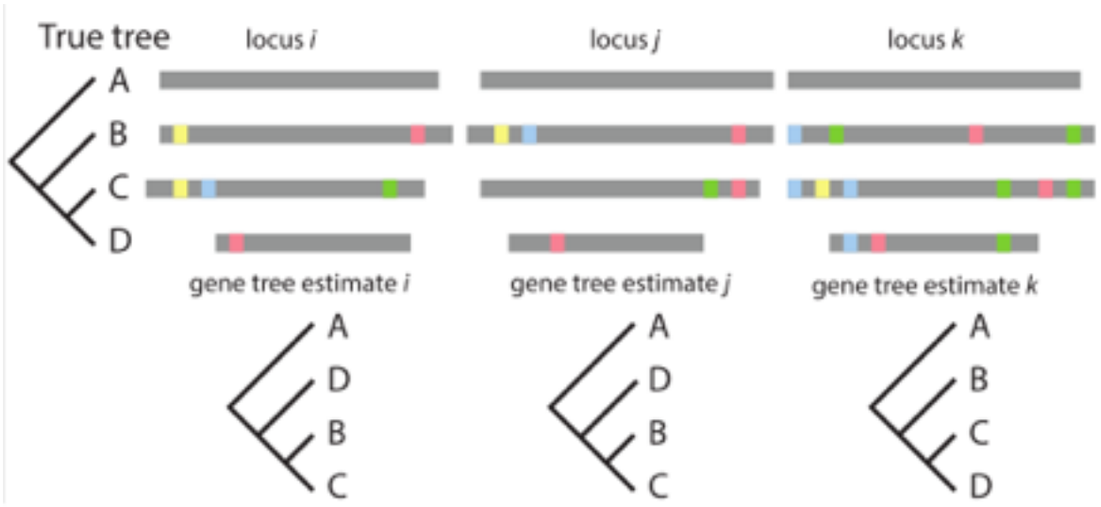
James B. Whitfield

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Two types of missing data



- Missing genes: a taxon misses the gene fully (taxon occupancy)



- Fragmentary data: the species is partially sequenced for some genes

Figure from Hosner et al, 2016

Does missing data matter?

- Missing genes:
  - Evidence suggests summary methods are often robust

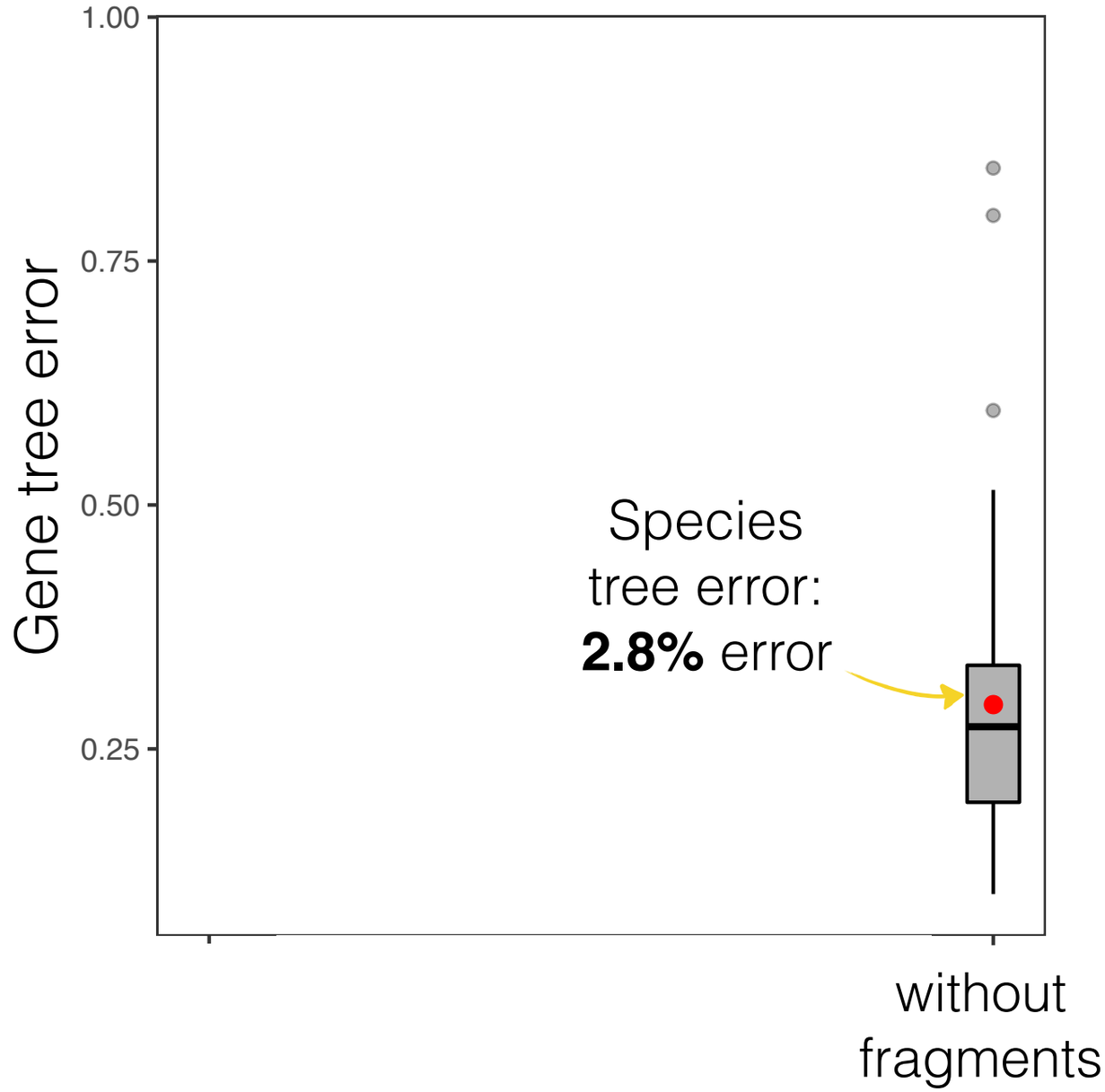
Does missing data matter?

- Missing genes:
  - Evidence suggests summary methods are often robust
- Fragmentary data:
  - Less extensively studied
  - Hosner et al. indicated:
    - Fragments matter
    - They suggest removing genes with many fragments

Our study

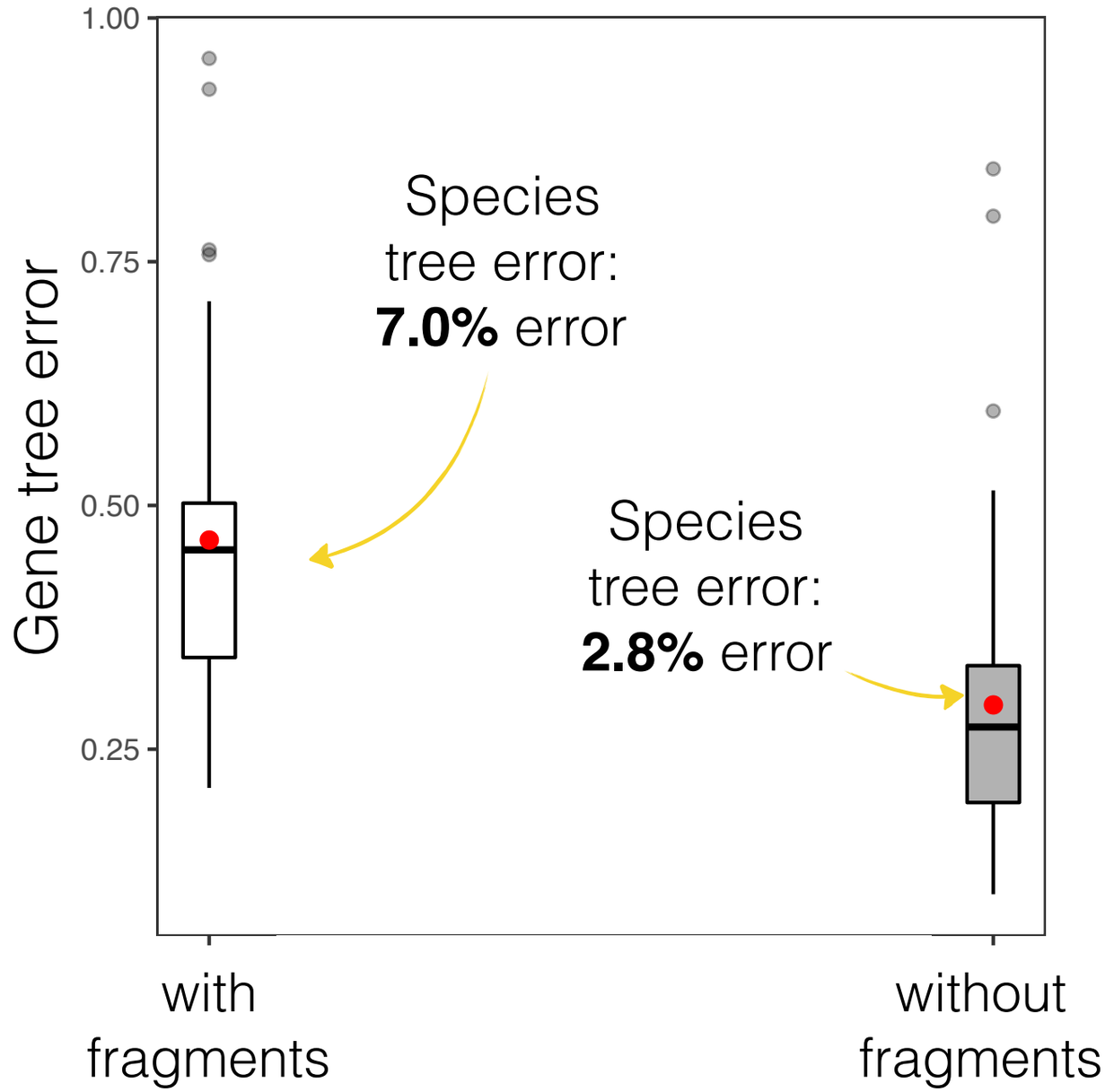
- Empirical data:
  - An insect dataset of 144 species and 1478 genes
  - 600 million years of evolution
  - Transcriptomics, so plenty of missing data of both types
- Simulated data:
  - 100 taxon, medium ILS, 1000 genes
  - Added fragmentation with patterns similar to the insect data

Fragmentary data hurt gene trees and species trees



Fragmentary data hurt gene trees and species trees

Fragmentary data dramatically increase gene tree and the species tree error



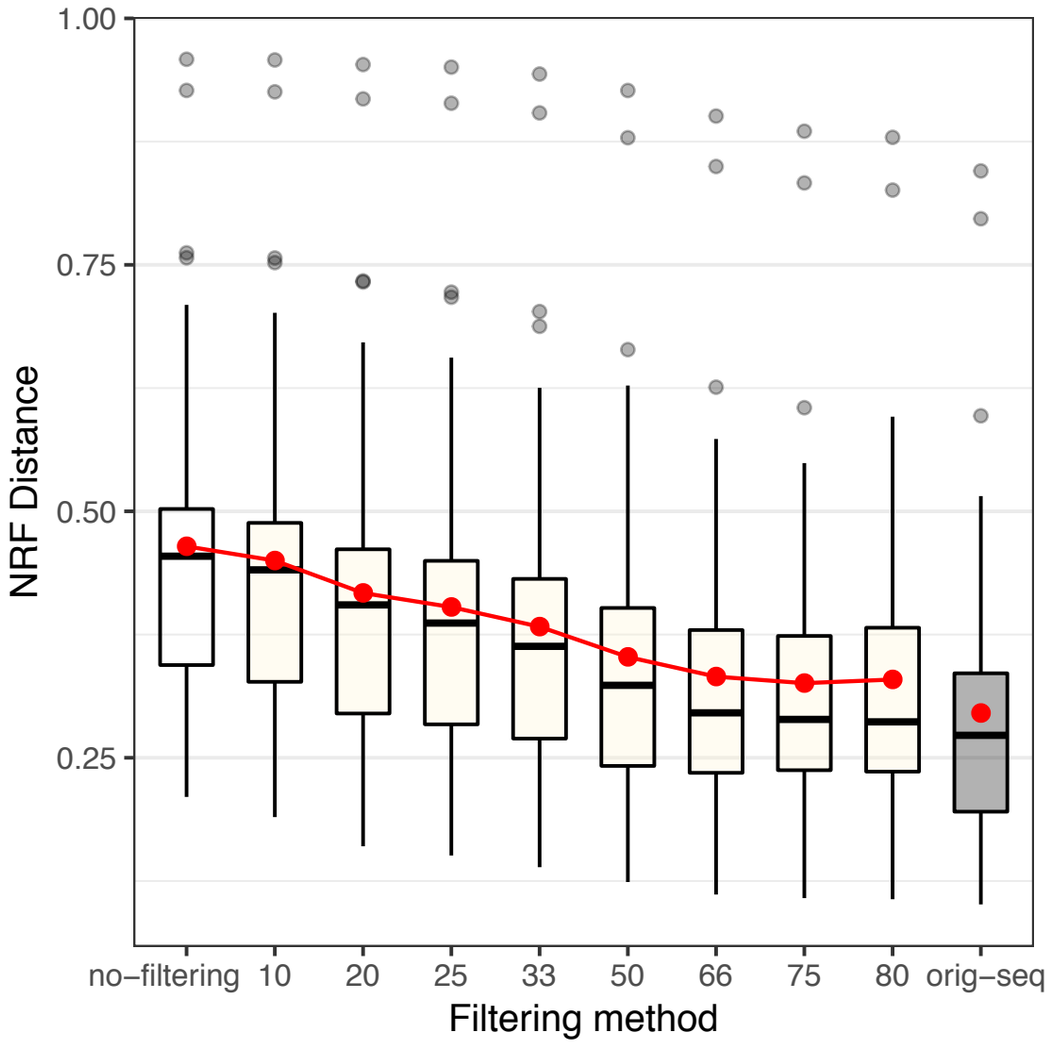
How do we solve this?

- Hosner et al. suggested removing entire genes.
  - This is too conservative in our opinion.

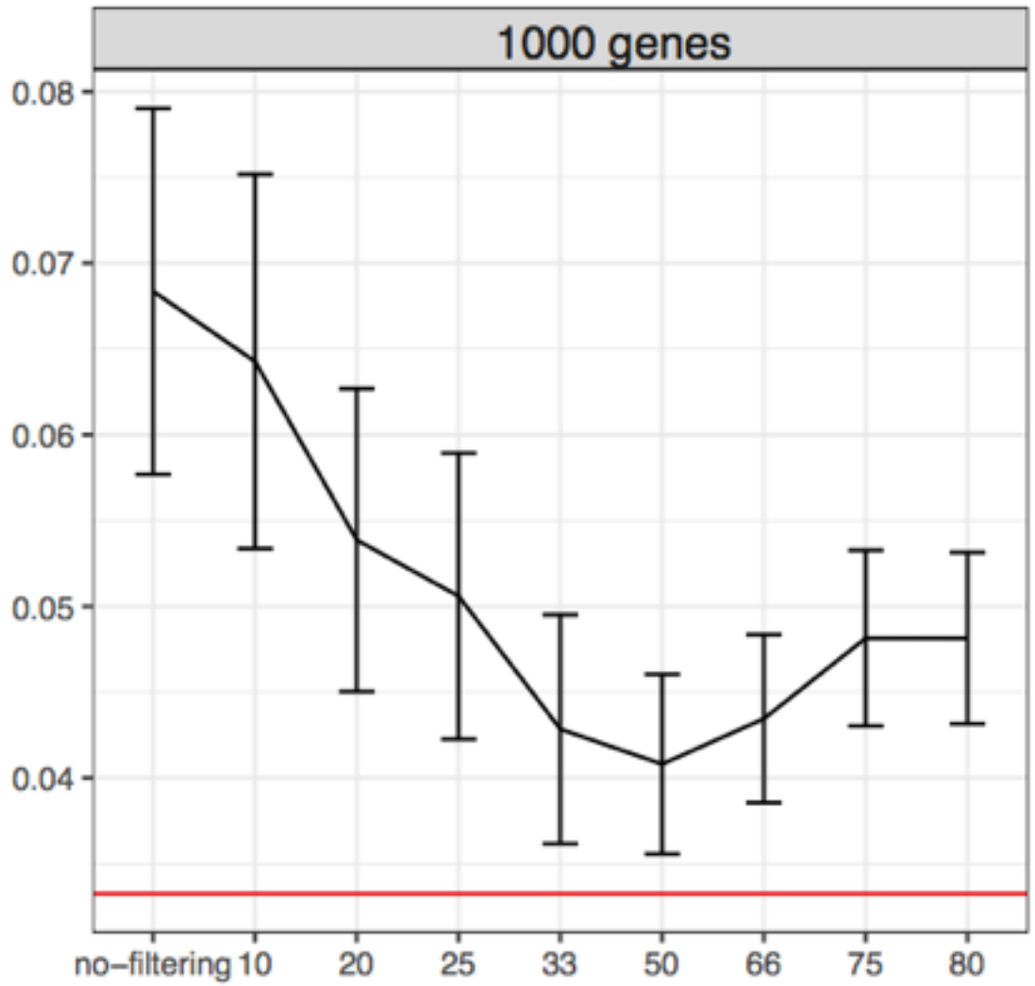
How do we solve this?

- Hosner et al. suggested removing entire genes.
  - This is too conservative in our opinion.
- We simply remove the fragmentary sequence from the gene and keep the rest of the gene
  - What is fragmentary?
    - Anything that has at most X% of sites
    - We study different X values

Filtering helps to some level (simulations)



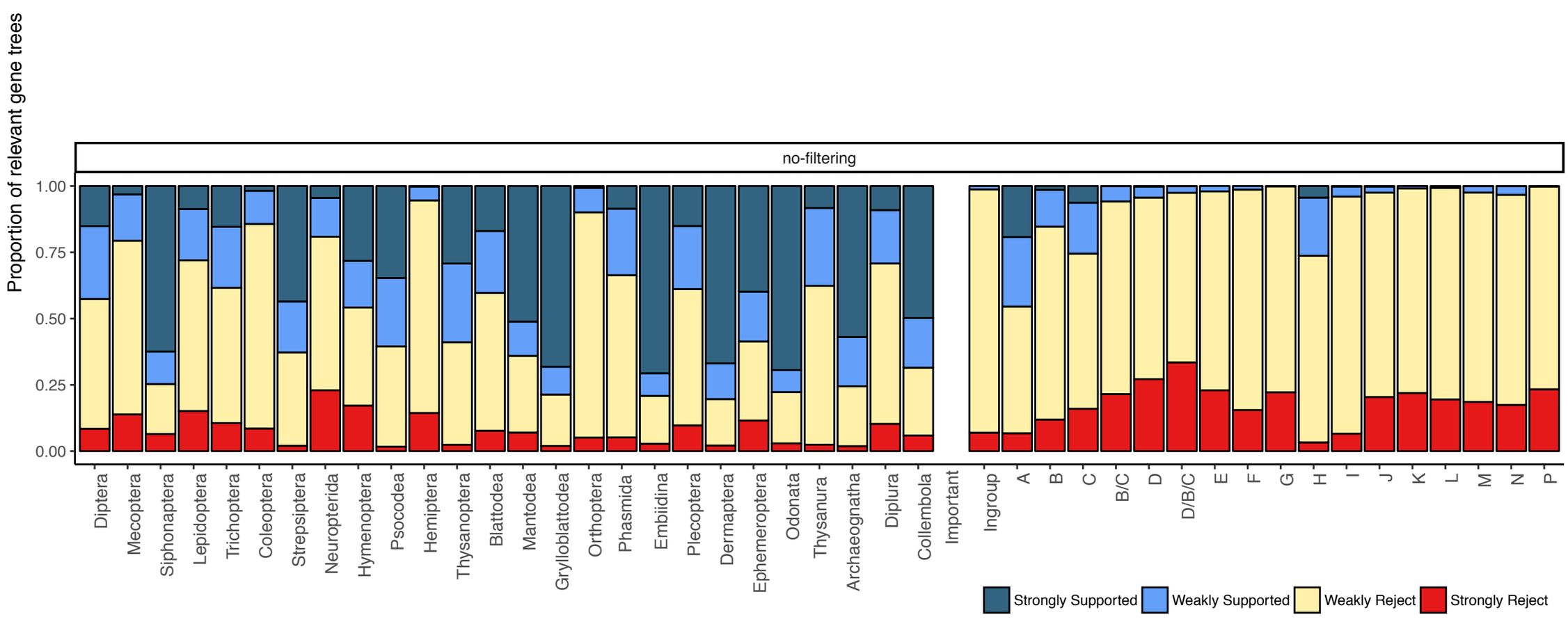
Gene tree error



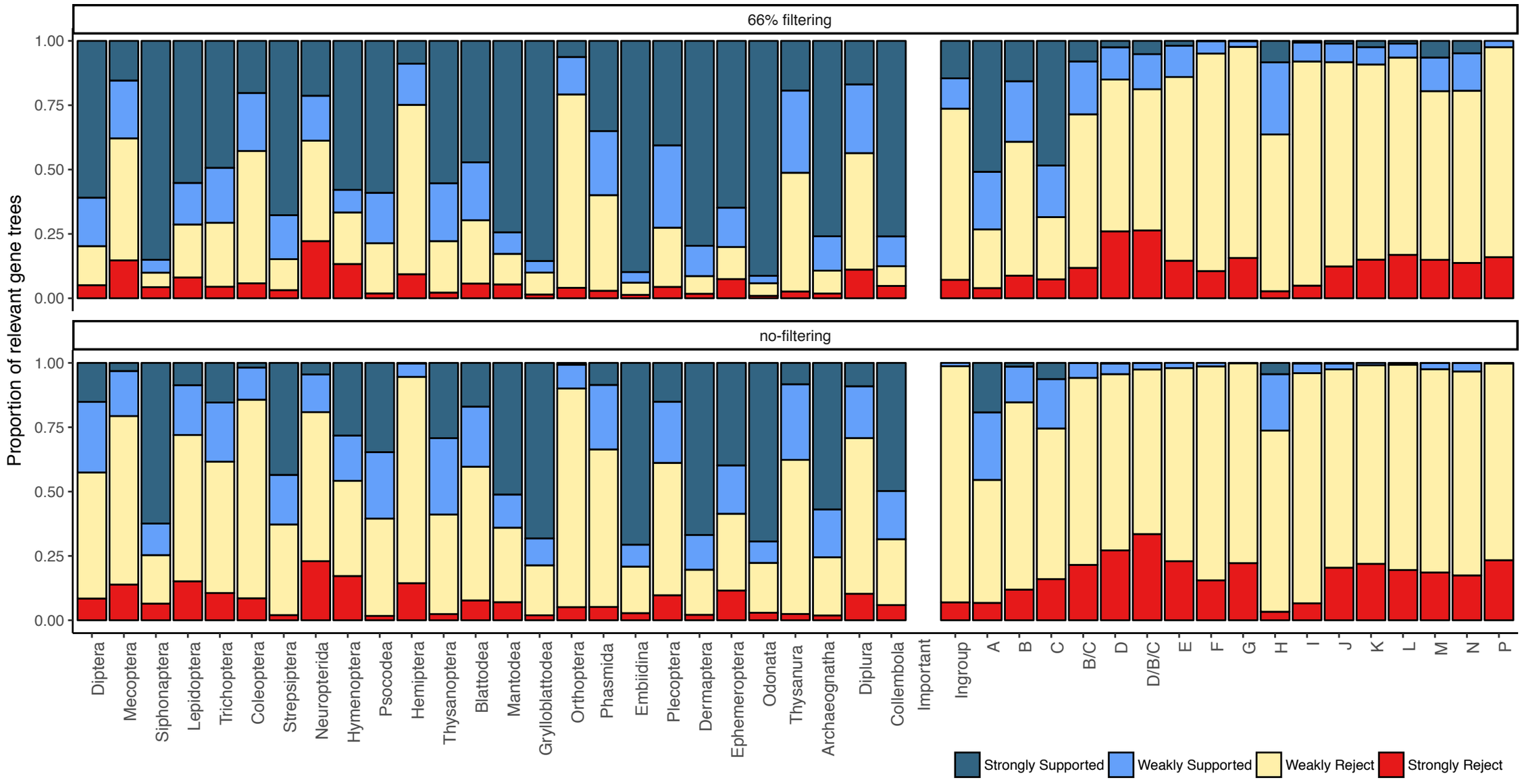
Species tree error

How about real data?

Gene trees seem to improve (less gene tree discordance)

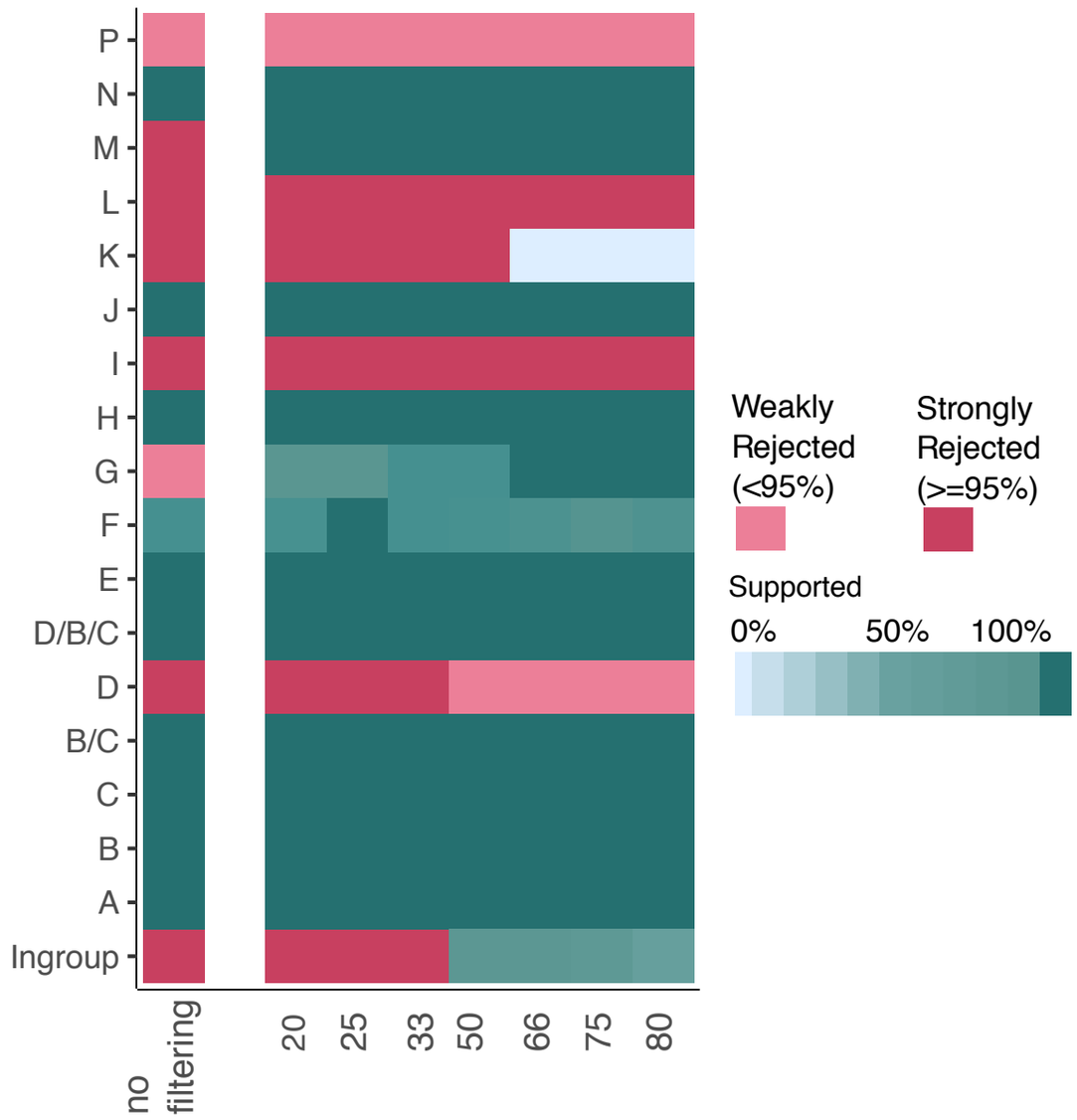


Gene trees seem to improve (less gene tree discordance)



The species tree also improves

- ASTRAL trees differed with concatenation initially
- After filtering, ASTRAL and concatenation were very similar
  - Differences on controversial nodes



ASTRAL …

- Reconstructs species trees from gene trees with both
  - high accuracy
  - scalability to large datasets
    - robust to *some* levels of model violations
- Like any other summary method, remains sensitive to gene tree estimation error
  - try best to obtain highly accurate gene trees
  - removing fragmentary data seems to help

Future of ASTRAL

- Can it be changed to use characters directly? possible but slow for binary characters …
- More broadly, can alternative scalable methods be developed for better gene tree estimation?

Future of ASTRAL

- Can it be changed to use characters directly? possible but slow for binary characters …
  - More broadly, can alternative scalable methods be developed for better gene tree estimation?

- ASTRAL scales to 10K leaves. We have 90K bacterial genomes.

  - Can we scale further?

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Tandy Warnow

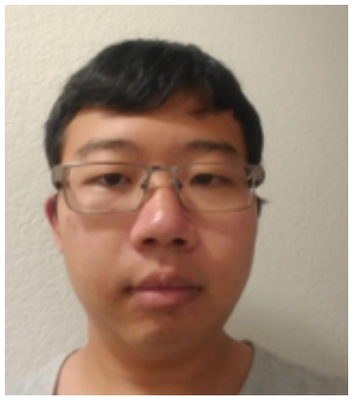
James B. Whitfield

S.M. Bayzid

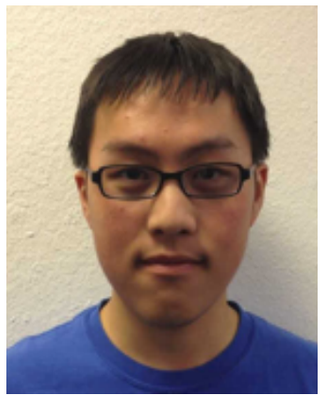
Théo Zimmermann

UC San Diego

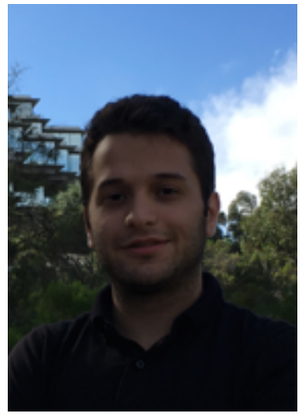
Maryam Rabiee Hashemi



Chao Zhang



John Yin



Erfan Sayyari

Shubhanshu Shekhar

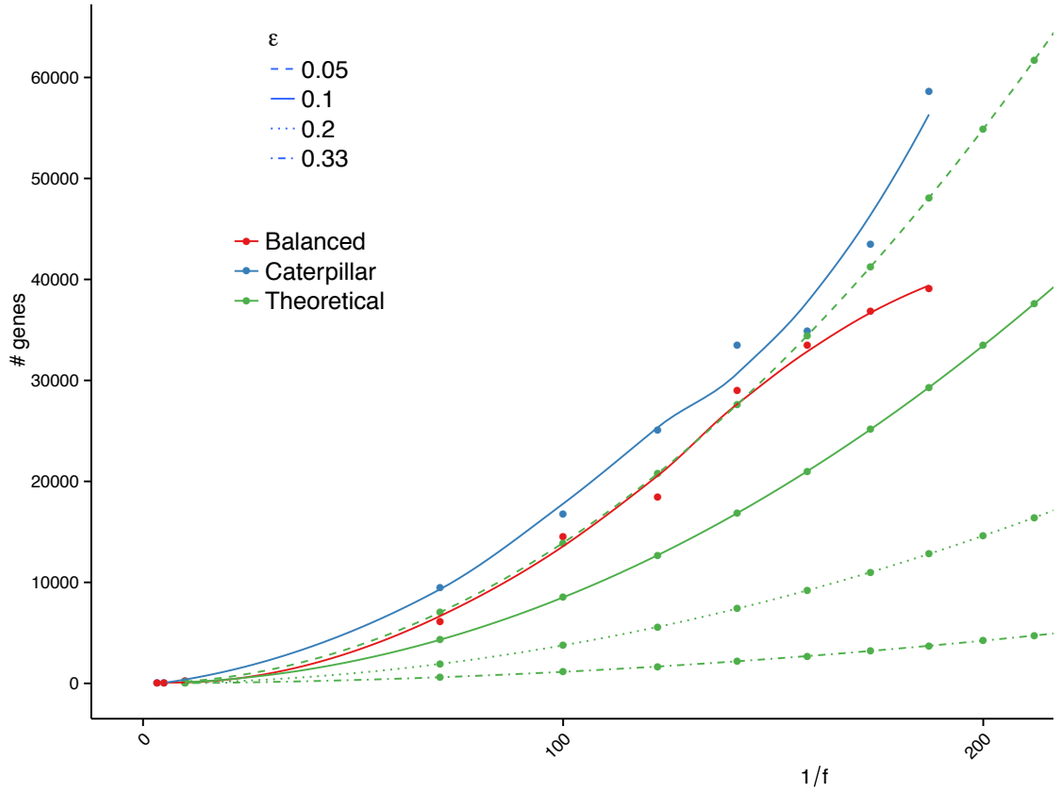
NSF

HHMI
HOWARD HUGHES MEDICAL INSTITUTE

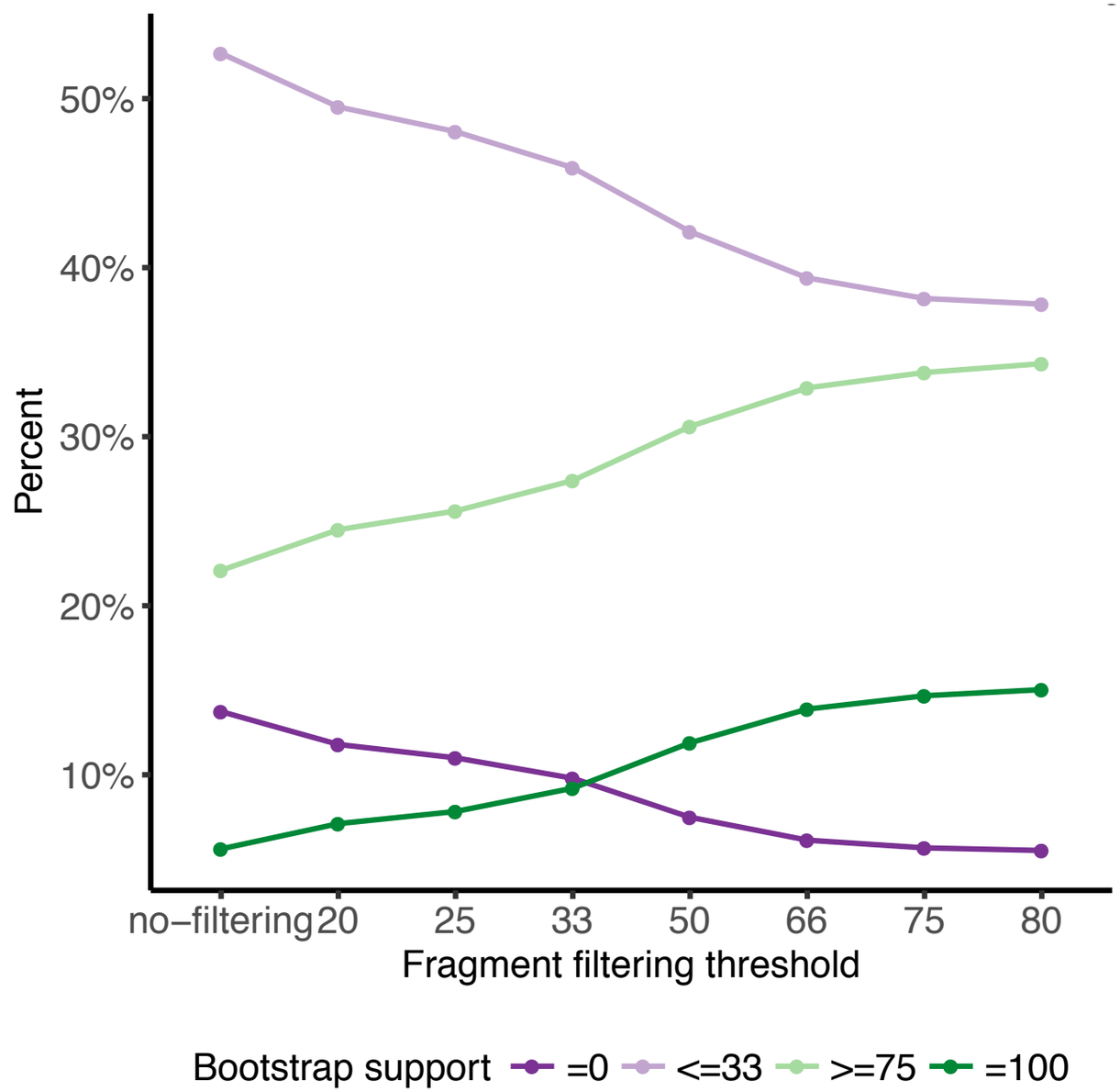
Theoretical sample complexity results

How many genes are enough to reconstruct the tree?

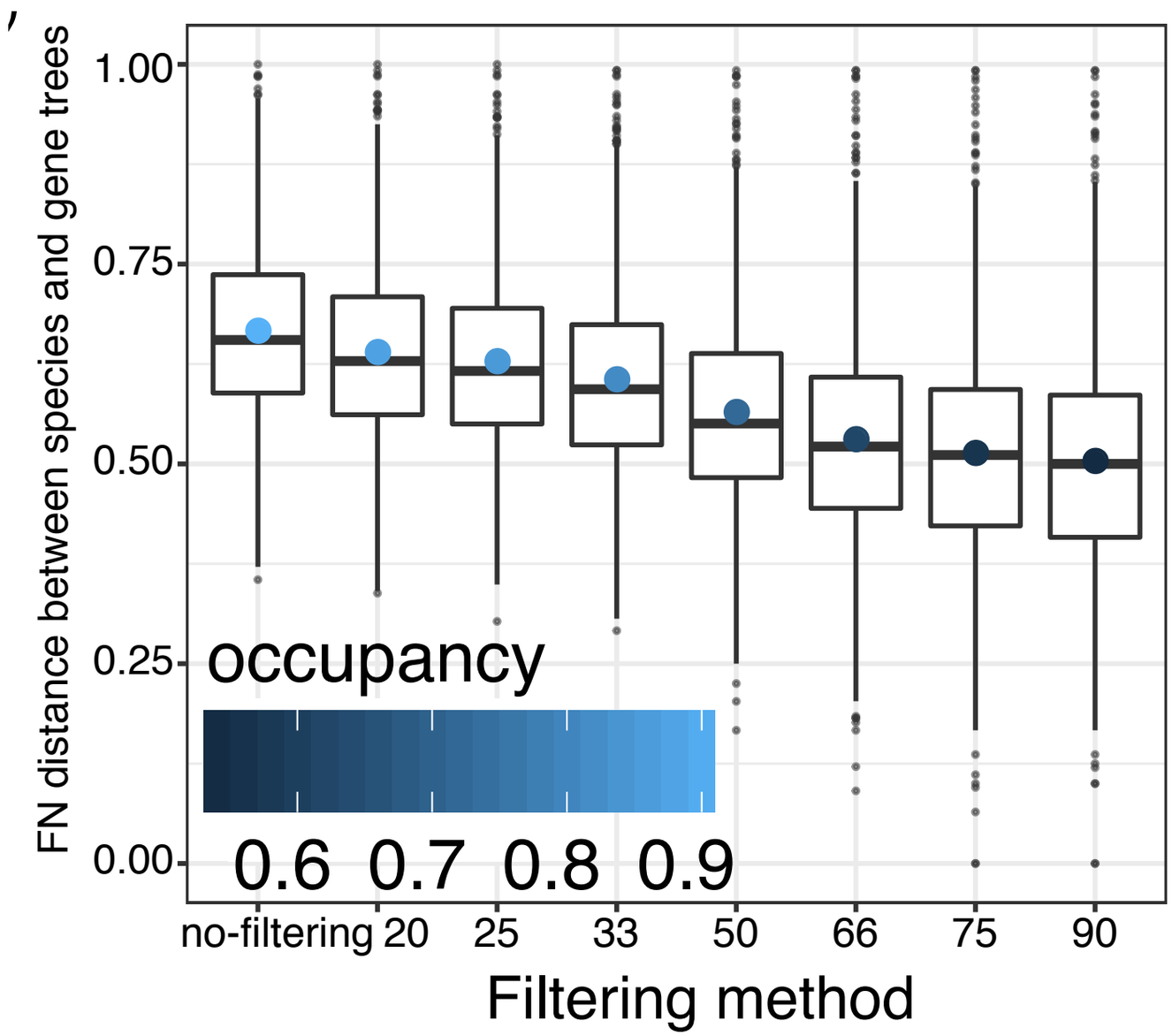
$$m \geq \frac{9}{2} \log \left(\frac{4\binom{n}{4}}{\epsilon} \right) \frac{c}{\alpha^2 f^2}$$



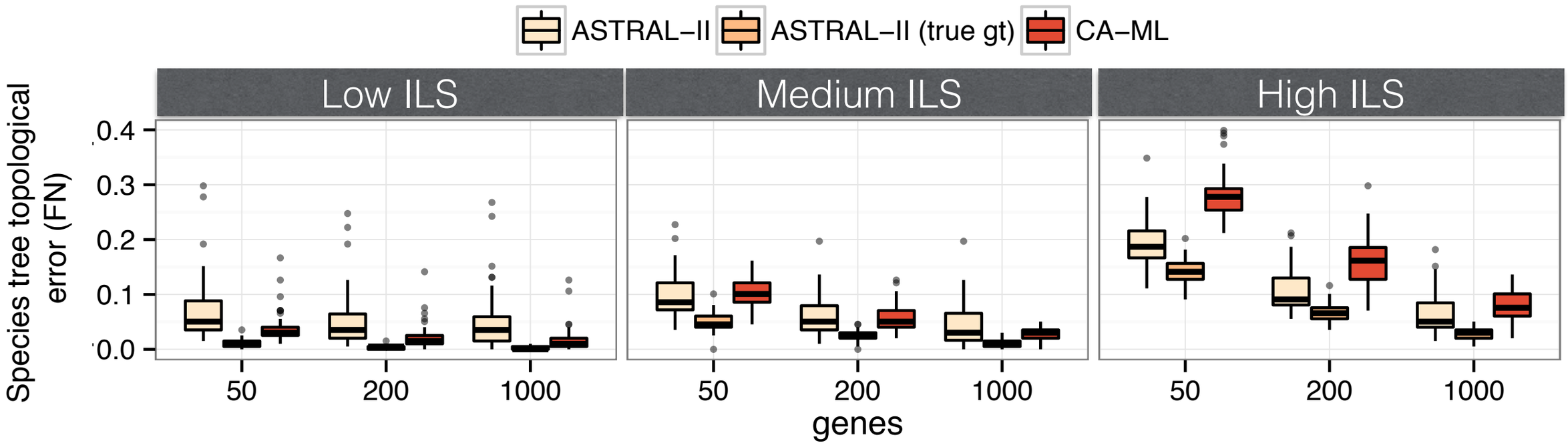
Gene trees seem to improve (bootstrap)



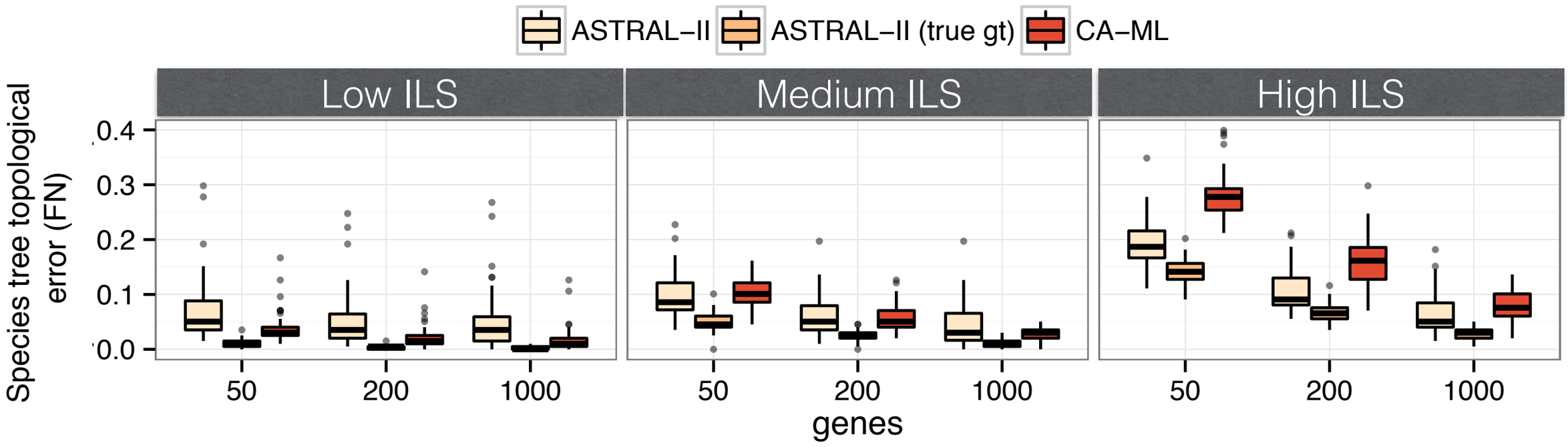
Gene trees seem to improve (less gene tree discordance)



Impact of gene tree error (using true gene trees)

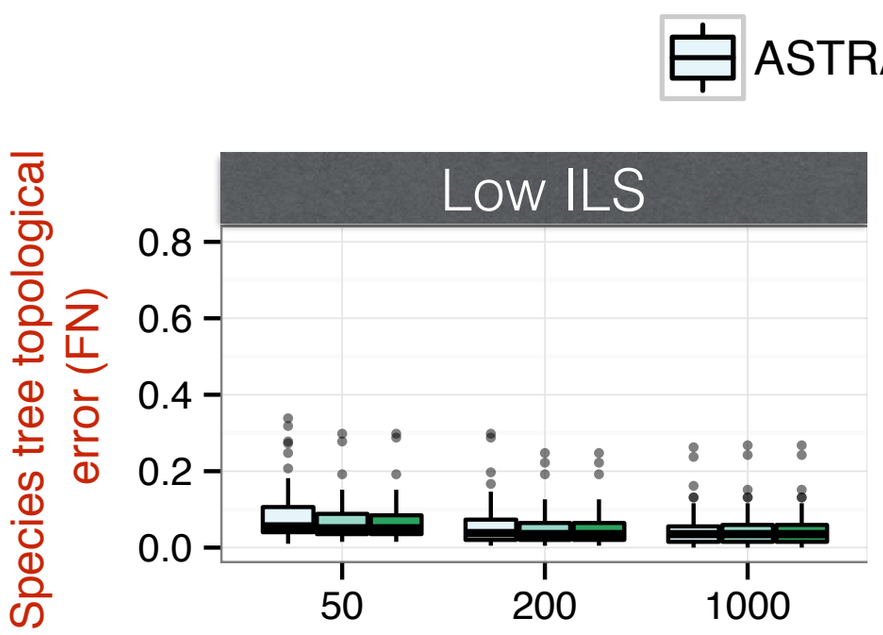


Impact of gene tree error (using true gene trees)



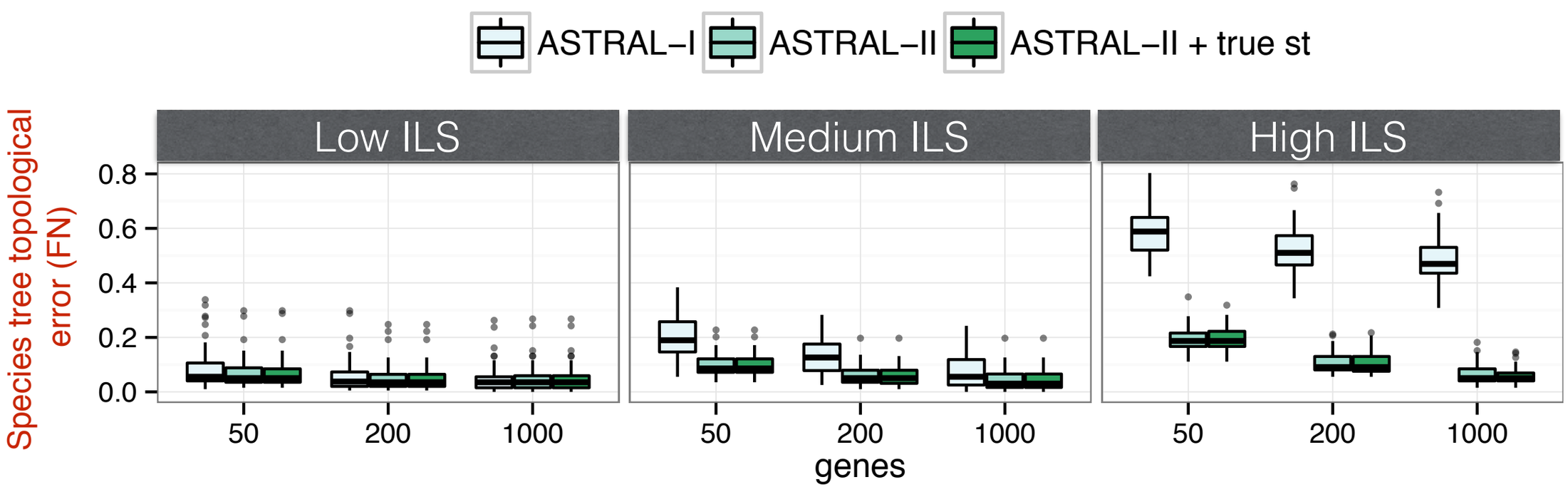
- When we divide our 50 replicates into low, medium, or high gene tree estimation error, ASTRAL tends to be better with low error

ASTRAL-I versus ASTRAL-II



200 species, deep ILS

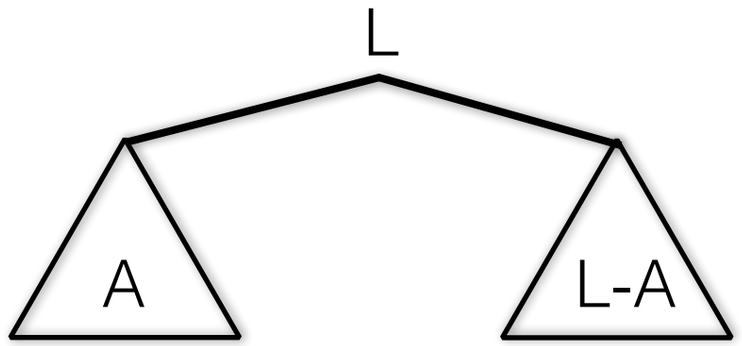
ASTRAL-I versus ASTRAL-II



200 species, deep ILS

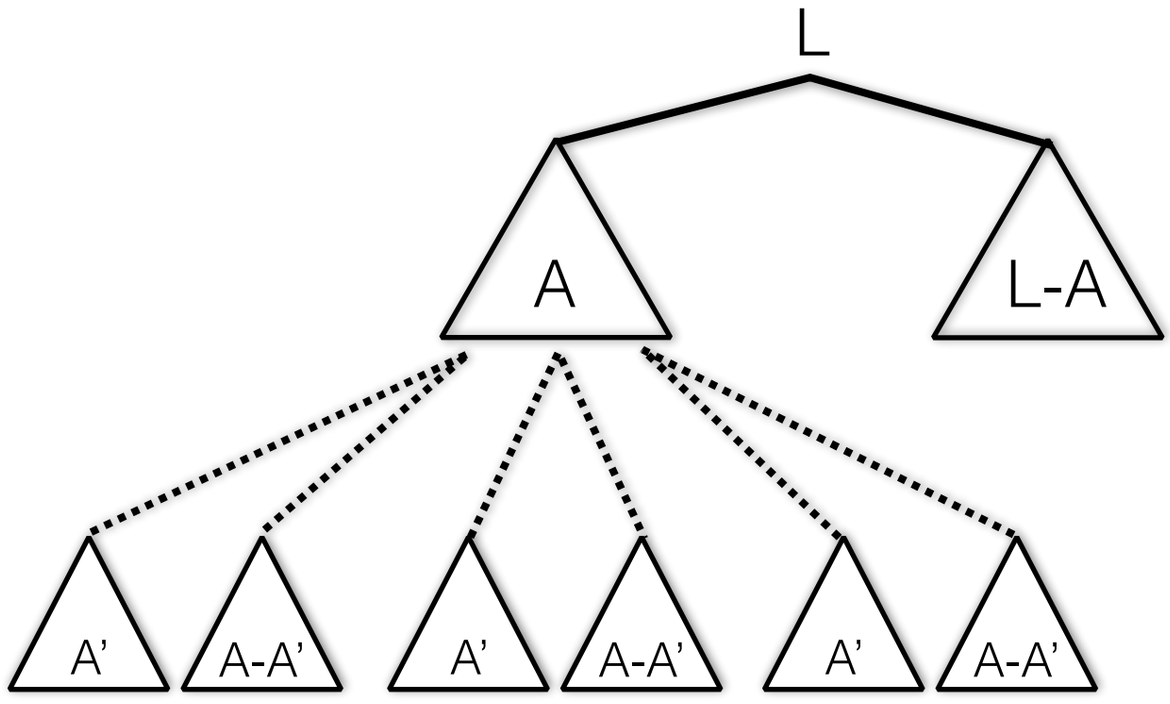
Dynamic programming

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



Dynamic programming

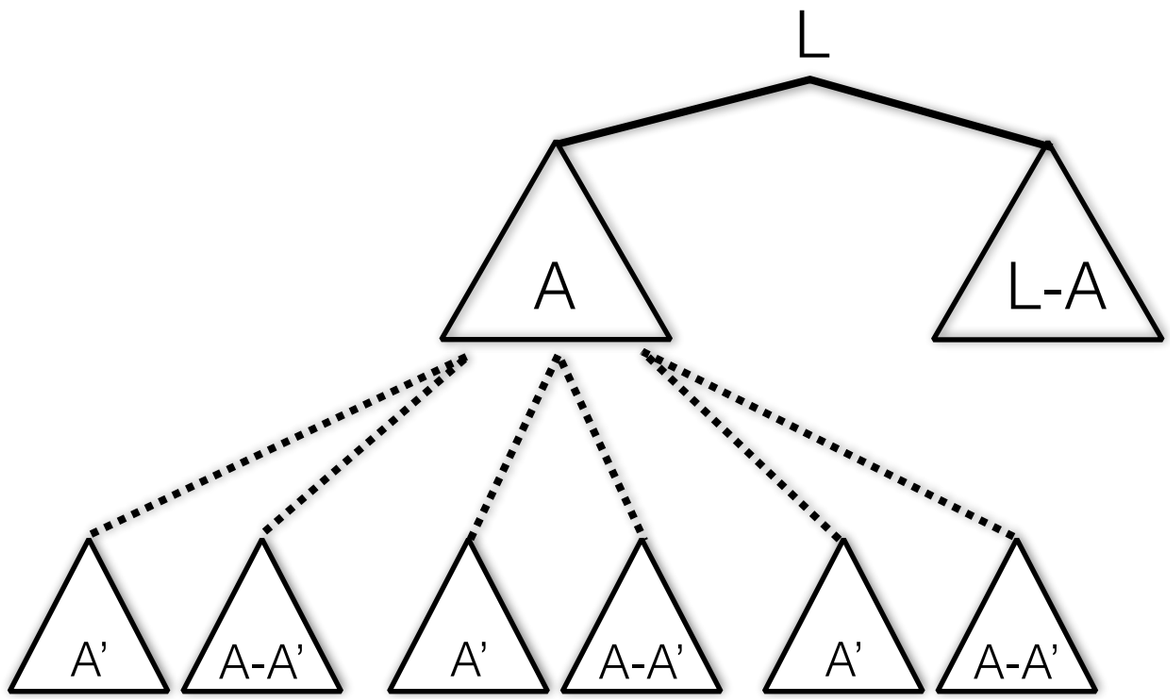
$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$

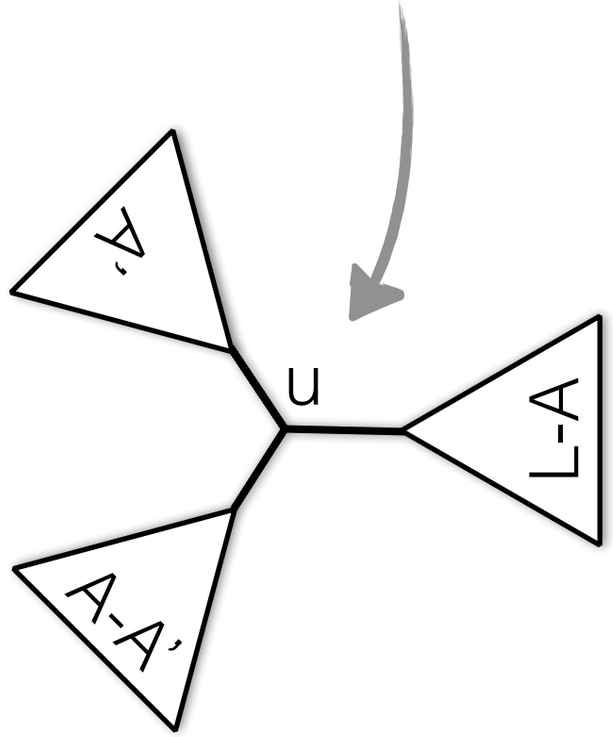


- Recursively break subsets of species into smaller subsets

Dynamic programming

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



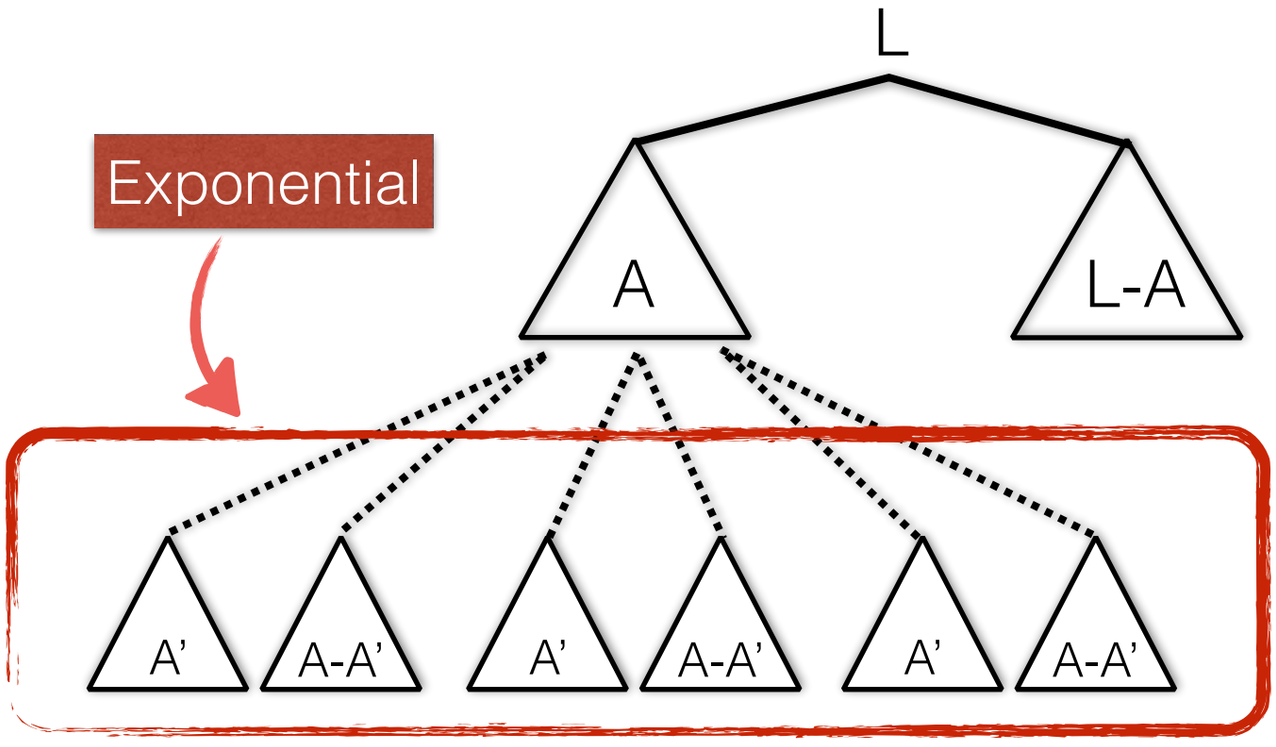


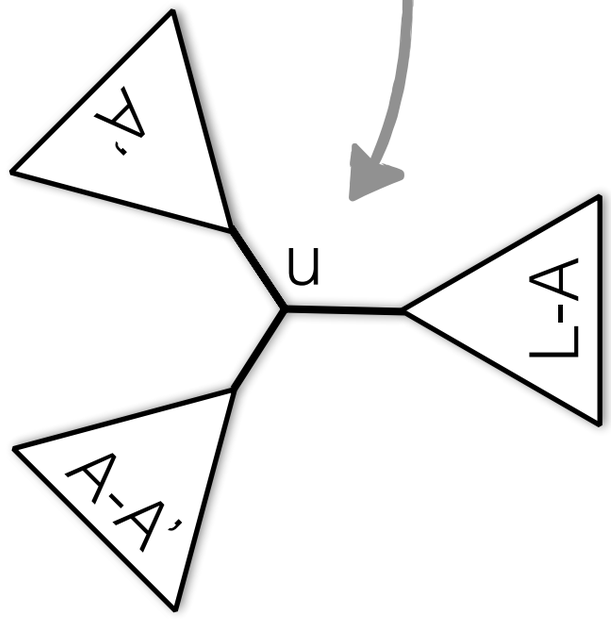
- Recursively break subsets of species into smaller subsets

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

Dynamic programming

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



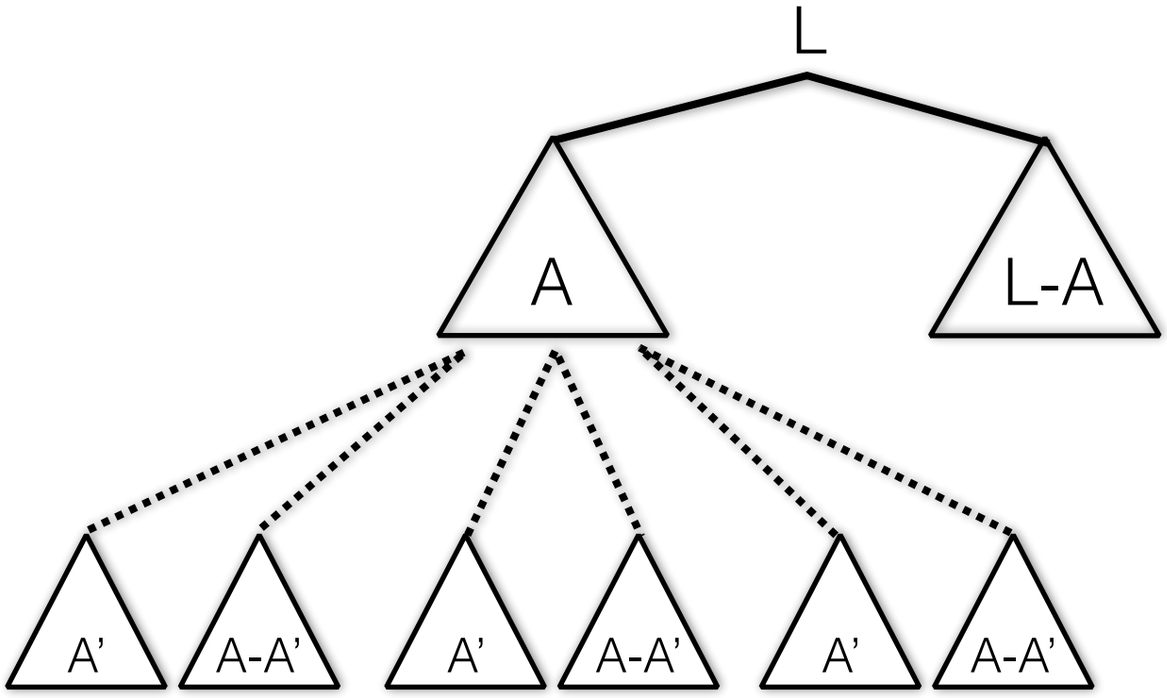


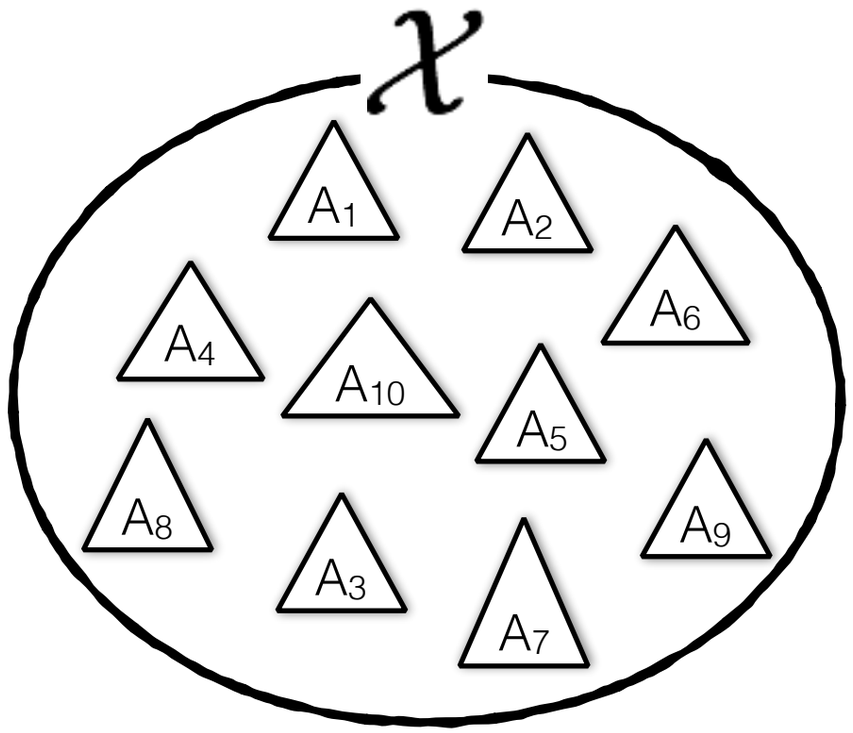
- Recursively break subsets of species into smaller subsets

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

Constrained version

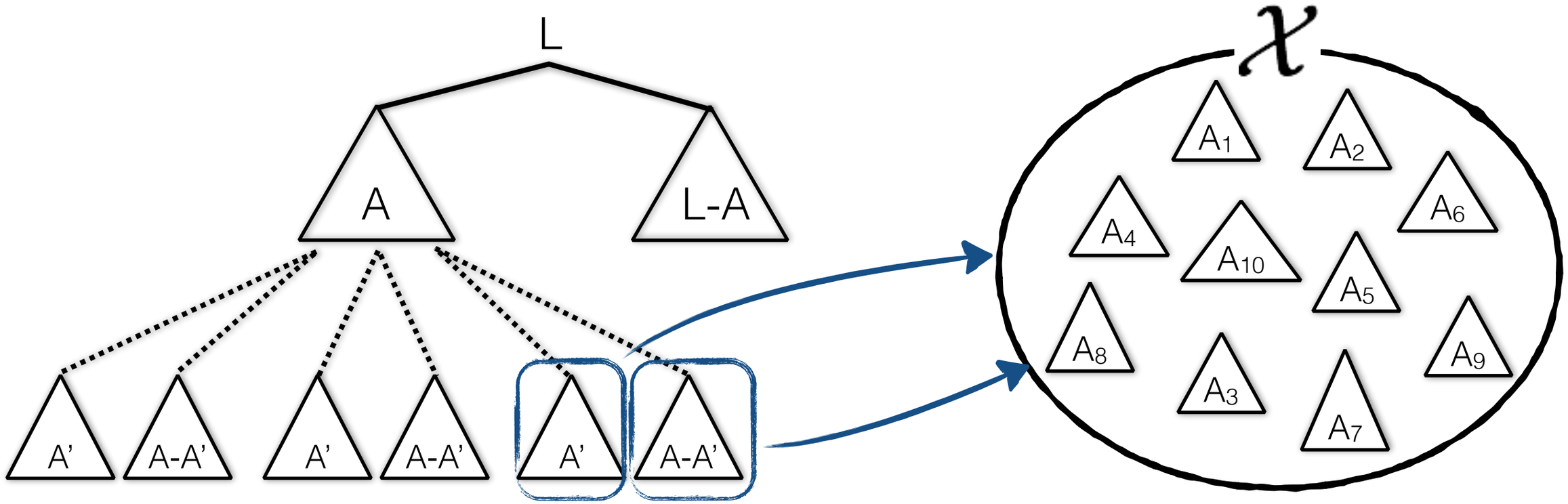
$$S(\mathcal{A}) = \max_{\{\mathcal{A}', \mathcal{A}-\mathcal{A}'\} \subset \mathcal{X}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$





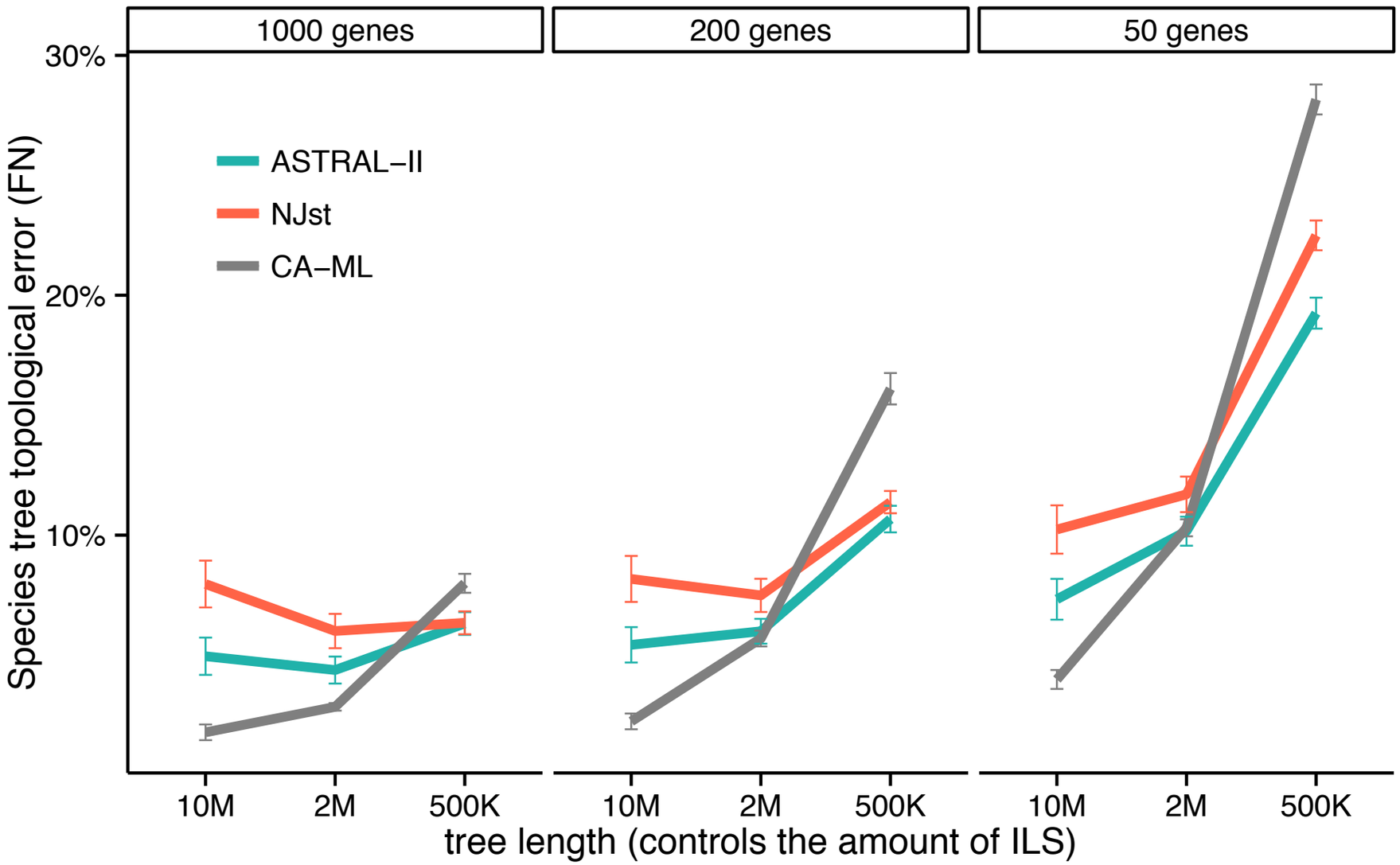
Constrained version

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



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

Deep ILS



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