Tutorial on ASTRAL

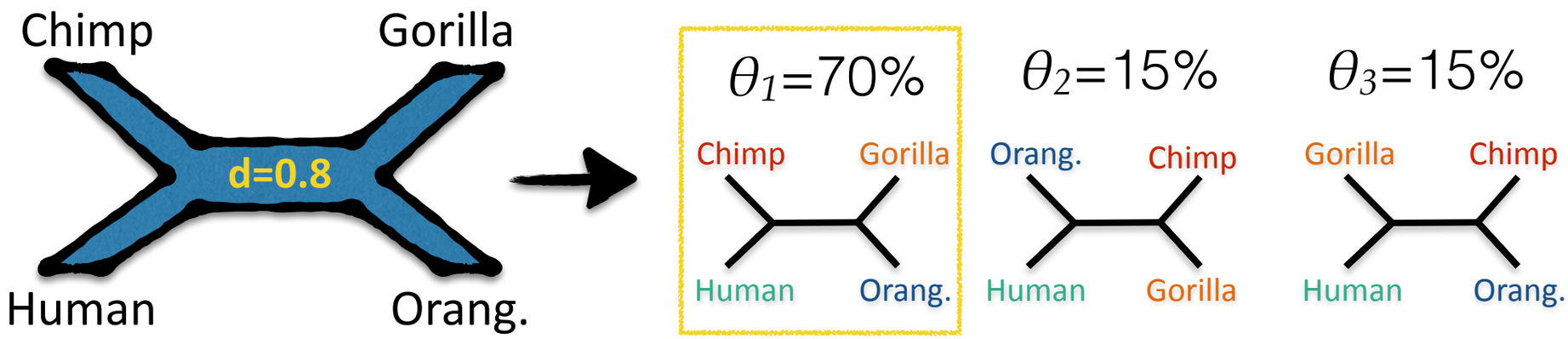
Download: https://github.com/smirarab/astral

Tutorial: https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md

Siavash Mirarab
Electrical and Computer Engineering
University of California, San Diego

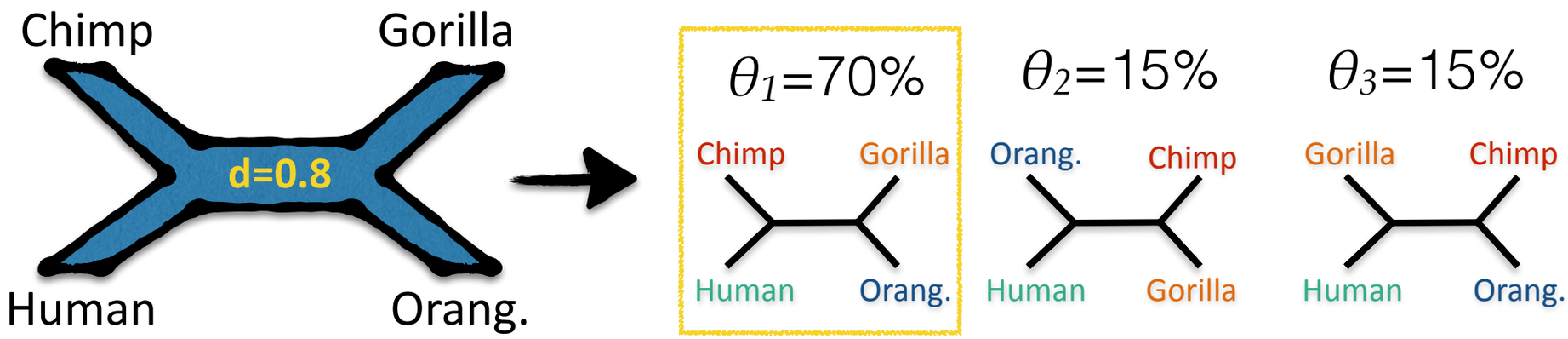
Unrooted quartets under MSC model

For a quartet (4 species), the most probable unrooted quartet tree (among the gene trees) is the unrooted species tree topology
(Allman, et al. 2010)



Unrooted quartets under MSC model

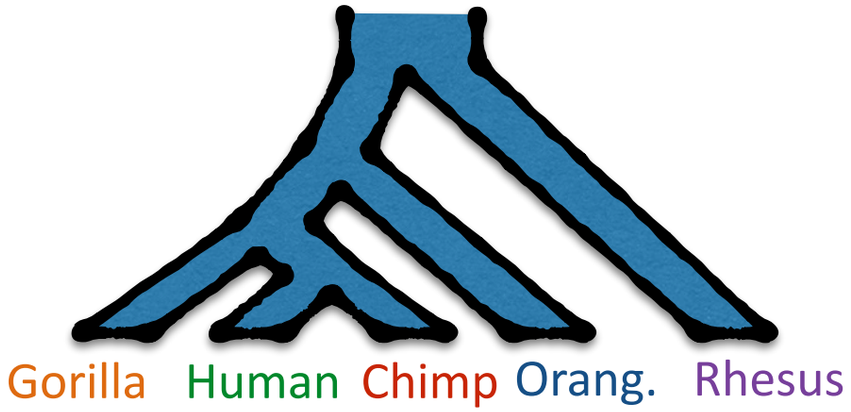
For a quartet (4 species), the most probable unrooted quartet tree (among the gene trees) is the unrooted species tree topology
(Allman, et al. 2010)



The most frequent gene tree
=
The most likely species tree

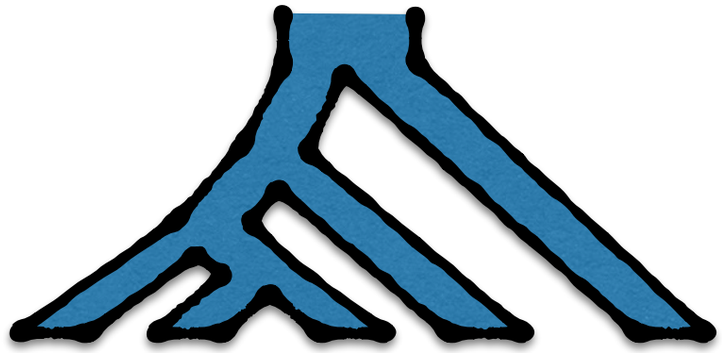
More than 4 species

For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree (called “anomaly zone”)
(Degnan, 2013)



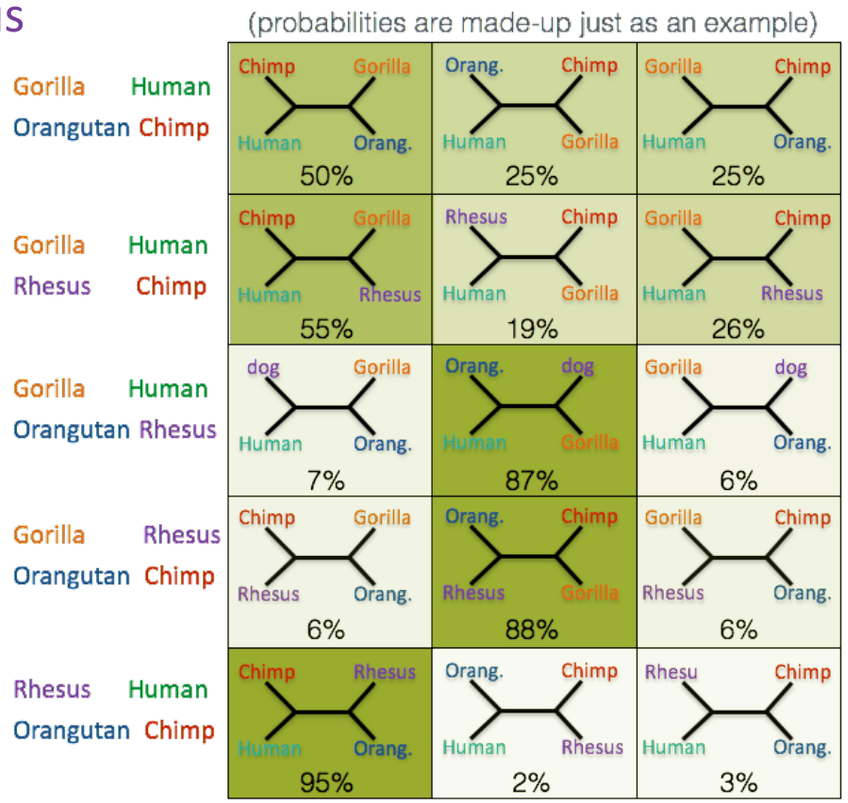
More than 4 species

For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree (called "anomaly zone")
(Degnan, 2013)



1. Break gene trees into $\binom{n}{4}$ quartets of species
2. Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

- Statistically consistent under the multi-species coalescent model with error-free input



ASTRAL versions

- ASTRAL-I (<v. 4.7.3) restricts the search space to combinations of bipartitions seen in gene trees.
  - This make it fast but it remains statistically consistent

ASTRAL versions

- ASTRAL-I (<v. 4.7.3) restricts the search space to combinations of bipartitions seen in gene trees.
  - This make it fast but it remains statistically consistent
- ASTRAL-II (<v. 5.1.0) increased the search space heuristically
  - Improved the accuracy at the expense of running time
  - Can handle polytomies in input gene trees

ASTRAL versions

- ASTRAL-I (<v. 4.7.3) restricts the search space to combinations of bipartitions seen in gene trees.
    - This make it fast but it remains statistically consistent
- ASTRAL-II (<v. 5.1.0) increased the search space heuristically
    - Improved the accuracy at the expense of running time
    - Can handle polytomies in input gene trees
- ASTRAL-III (>v. 5.1.1) changed the search space again for a better running time versus accuracy trade-off
    - Improved running time for unresolved trees; makes it feasible to remove very low support branches from gene trees.

Input/Output

- **Input:** Unrooted gene trees
  - Can have missing data
  - Can have polytomies
  - Can have multiple alleles
- **Output**: The estimated unrooted species tree
  - Will have branch lengths in coalescent units on internal branches (not super accurate)
  - Will have a measure of support called localPP

Caveats

- Assumptions for statistical consistency: a **randomly distributed** sample of **recombination-free**, **reticulation-free, error-free, orthologous** gene trees

- Practically: reduced accuracy with low accuracy gene trees (will see more)

Start of tutorial

To cover …

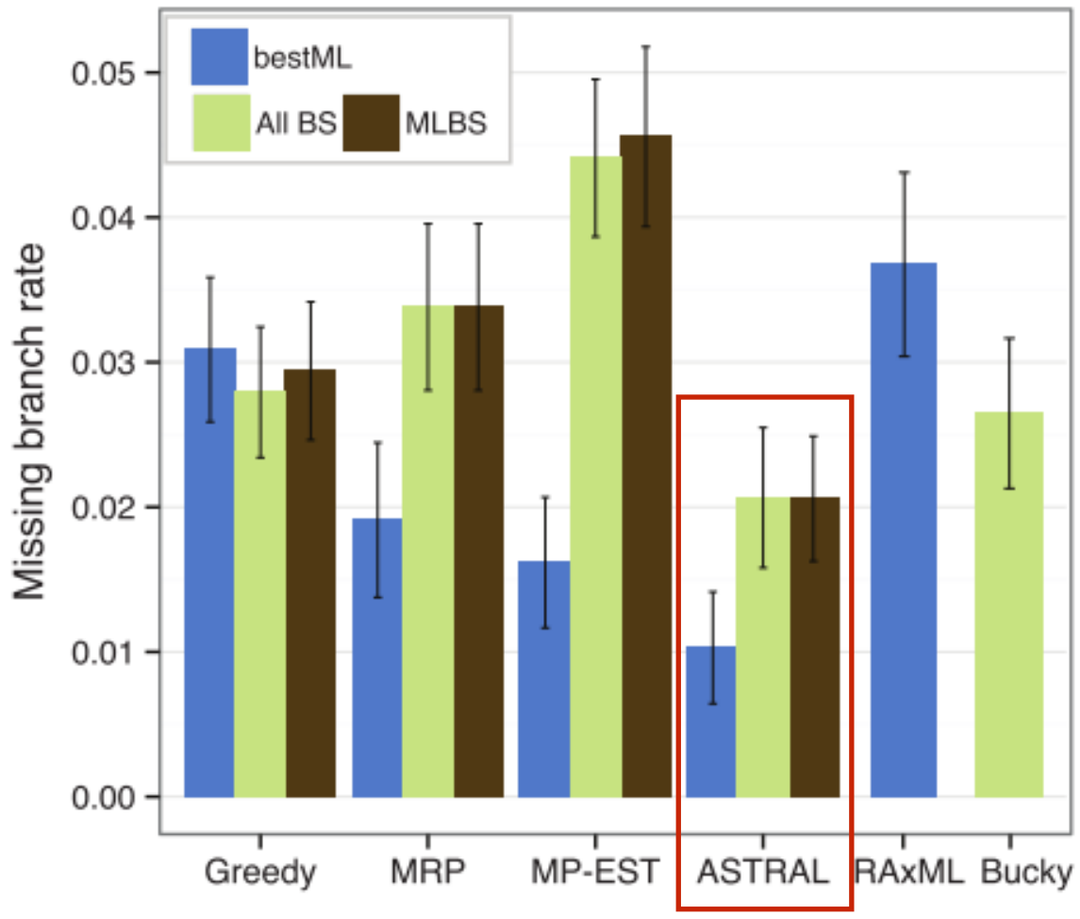
- Show installation
  - Start at https://github.com/smirarab/astral
    - A note on ASTRAL-MP
- Tutorial on https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md
  - Do a simple test run
  - Show help

Notes on input

bestML versus MLBS

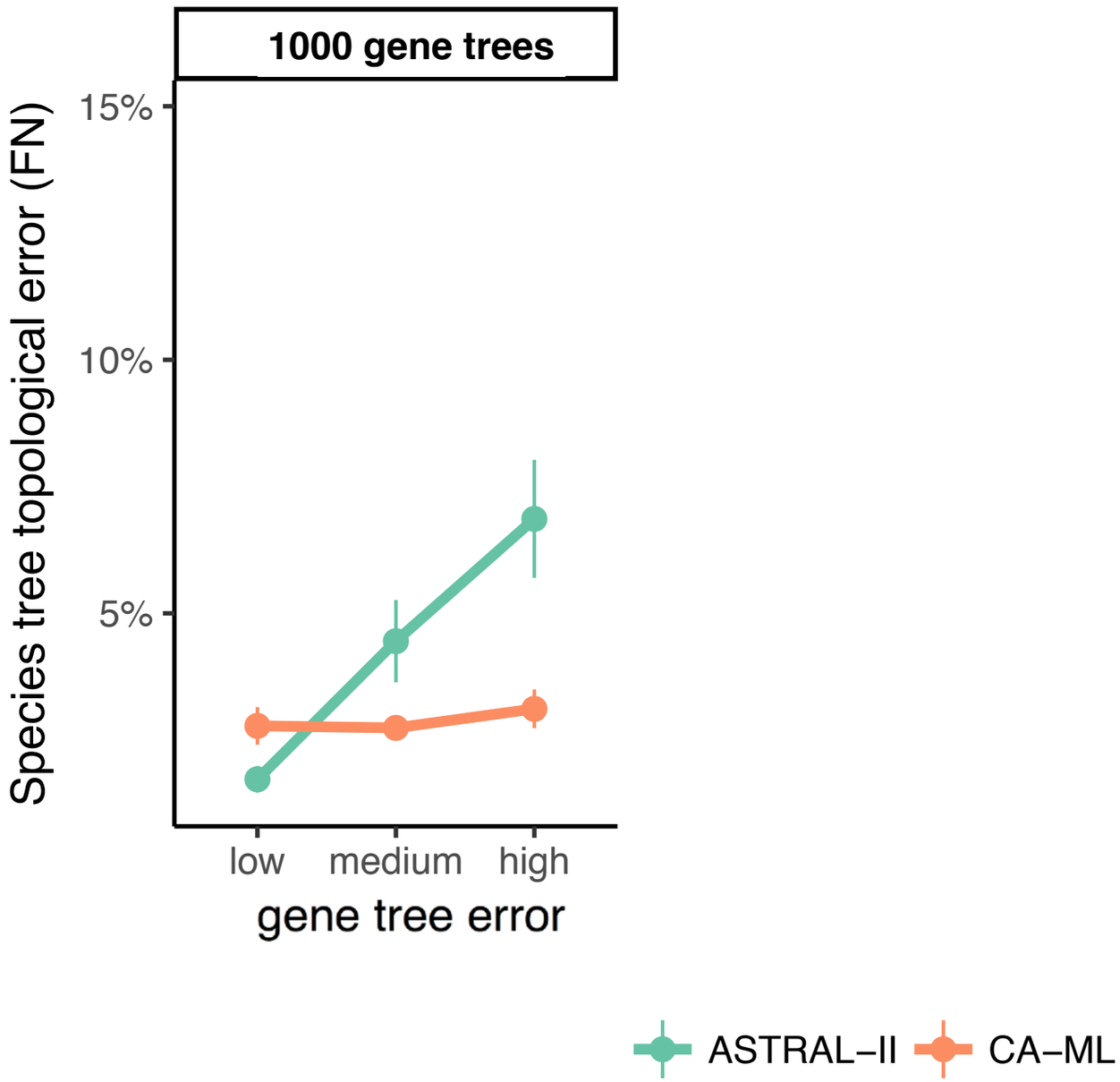
- Multi-locus Bootstrapping (MLBS) is also possible

- MLBS not suggested as it seems to degrade accuracy compared to simply using ML gene trees.



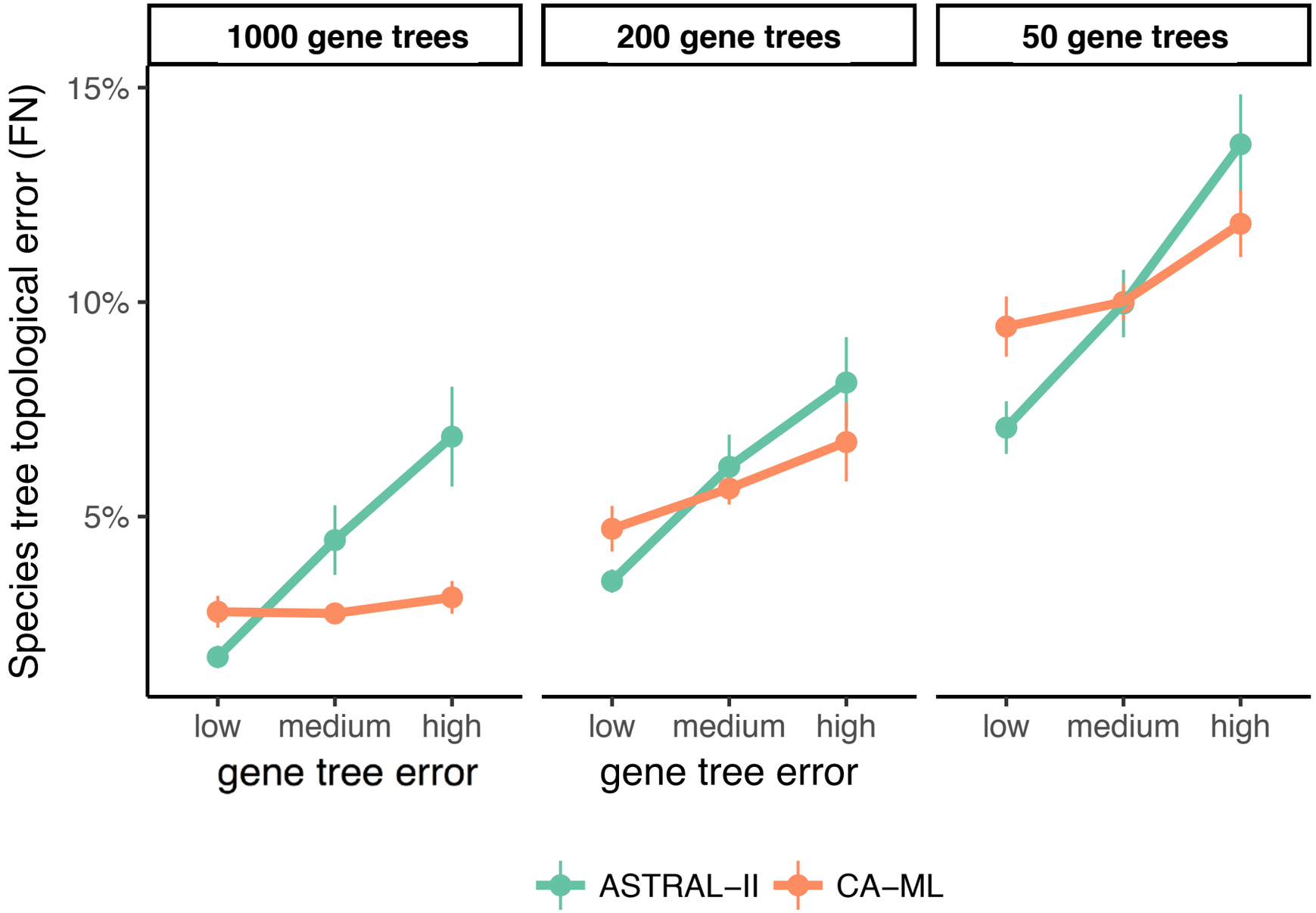
Bioinformatics, 2014, Mirarab et al.

Comparison to concatenation: depends on the level of gene tree error



Simulations, 200 species, deep medium level ILS [Mirarab and Warnow, 2016]

Comparison to concatenation: depends on the level of gene tree error

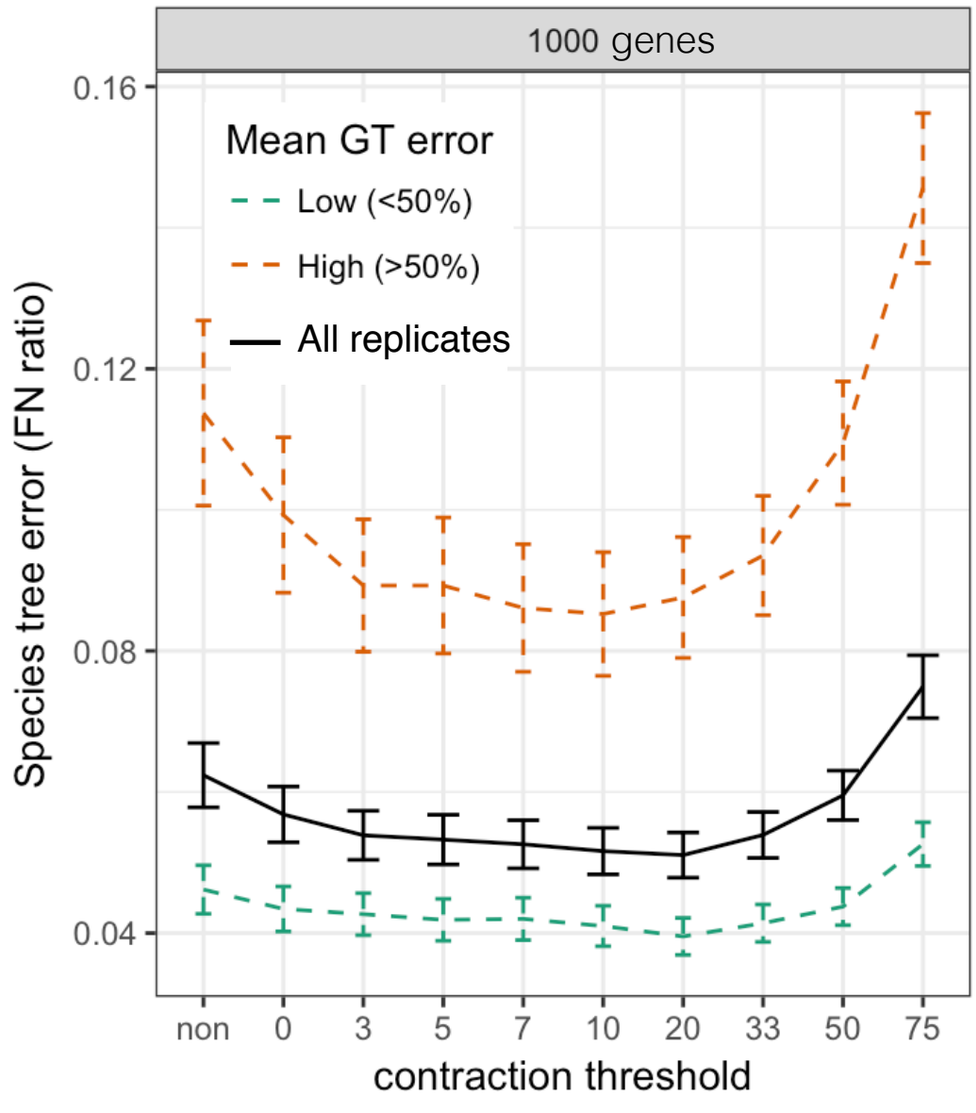


Simulations, 200 species, deep medium level ILS [Mirarab and Warnow, 2016]

Low support branches

- Gene tree support matters
- Does it help to contract branches with low support?

BMC Bioinformatics, 2018, Zhang et al.

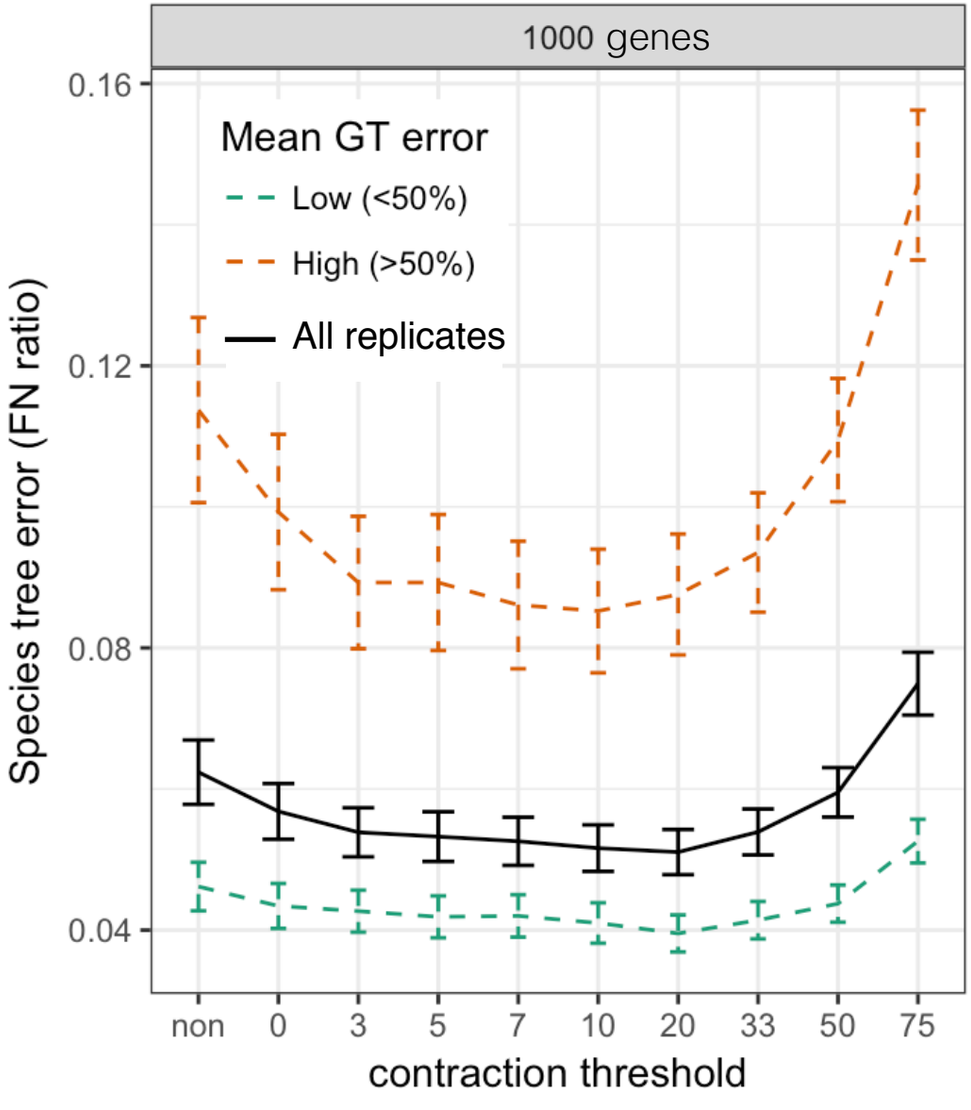


Simulations: 100 taxa, simphy,
ILS: around 46% true discordance
FastTree, support from bootstrapping

Low support branches

- Gene tree support matters
- Does it help to contract branches with low support?
- Yes, but only for very low support branches

BMC Bioinformatics, 2018, Zhang et al.

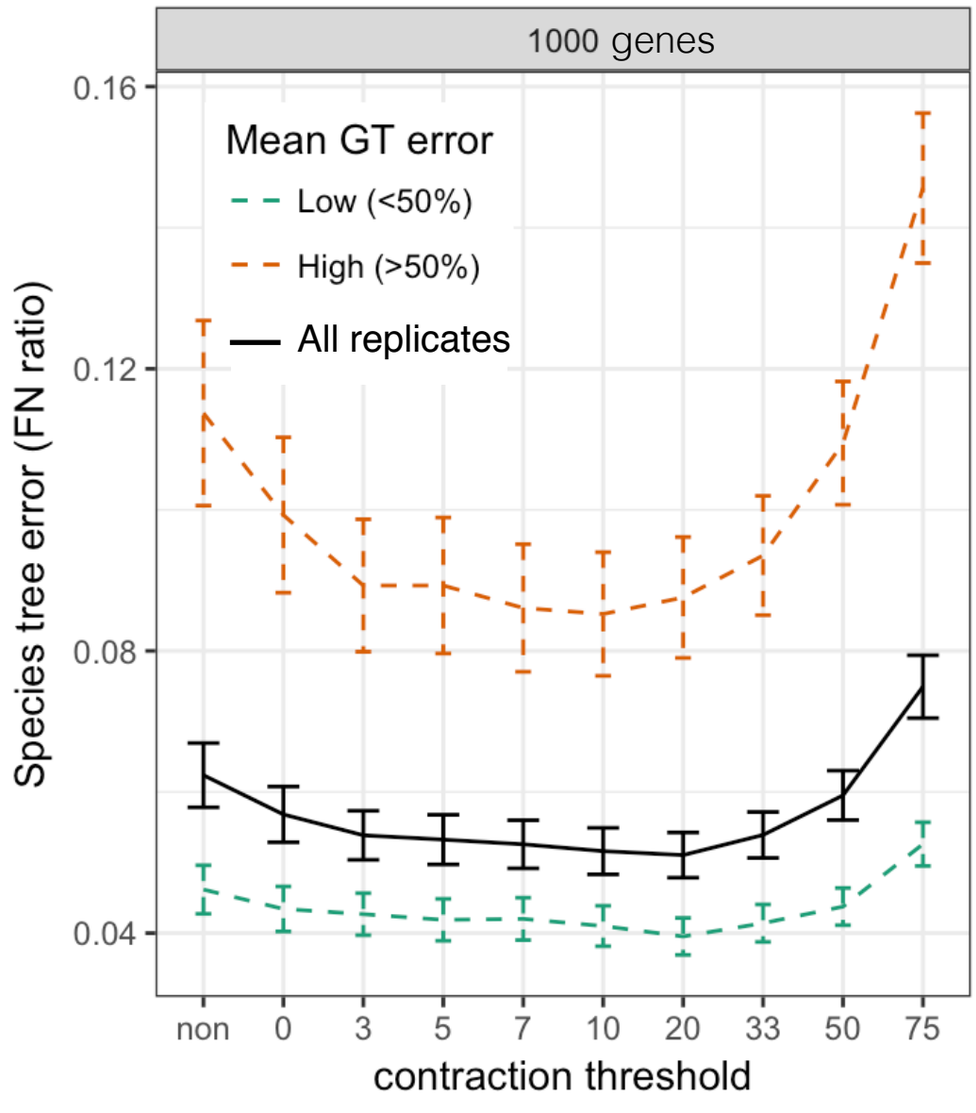


Simulations: 100 taxa, simphy,
ILS: around 46% true discordance
FastTree, support from bootstrapping

Low support branches

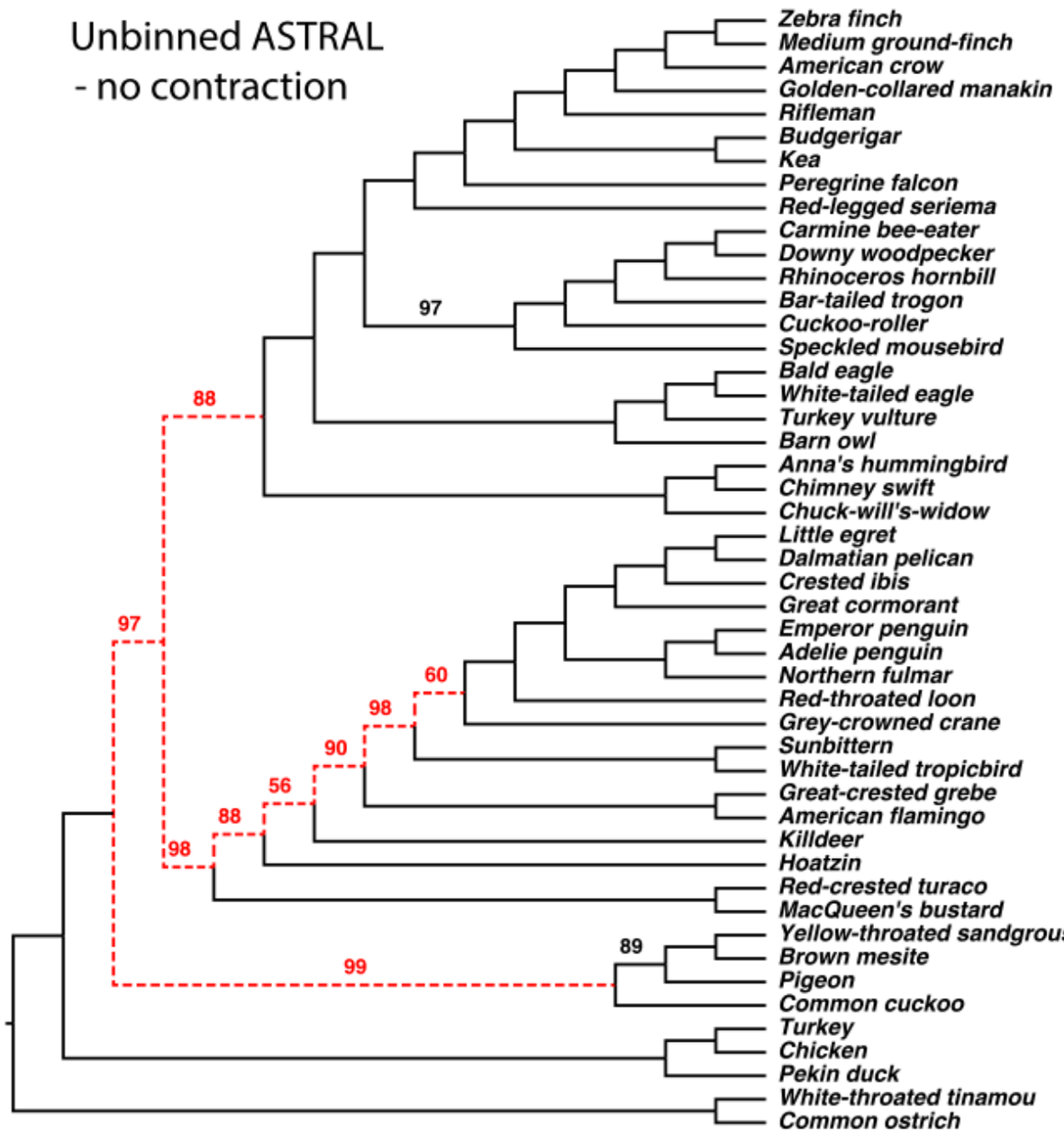
- Gene tree support matters
- Does it help to contract branches with low support?
- Yes, but only for very low support branches
- Mostly helps in the presence of low support gene trees

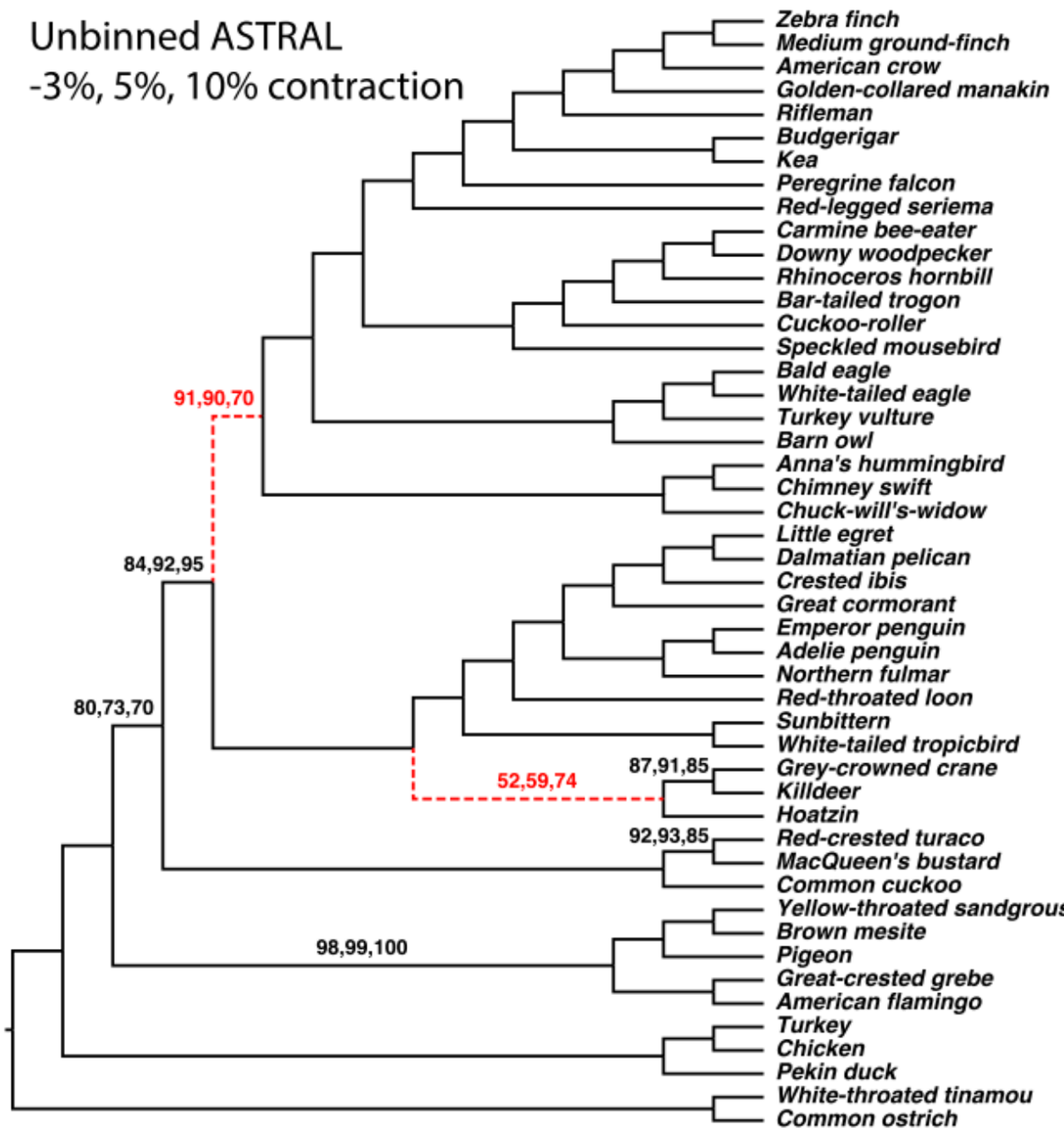
BMC Bioinformatics, 2018, Zhang et al.



Simulations: 100 taxa, simphy,
ILS: around 46% true discordance
FastTree, support from bootstrapping

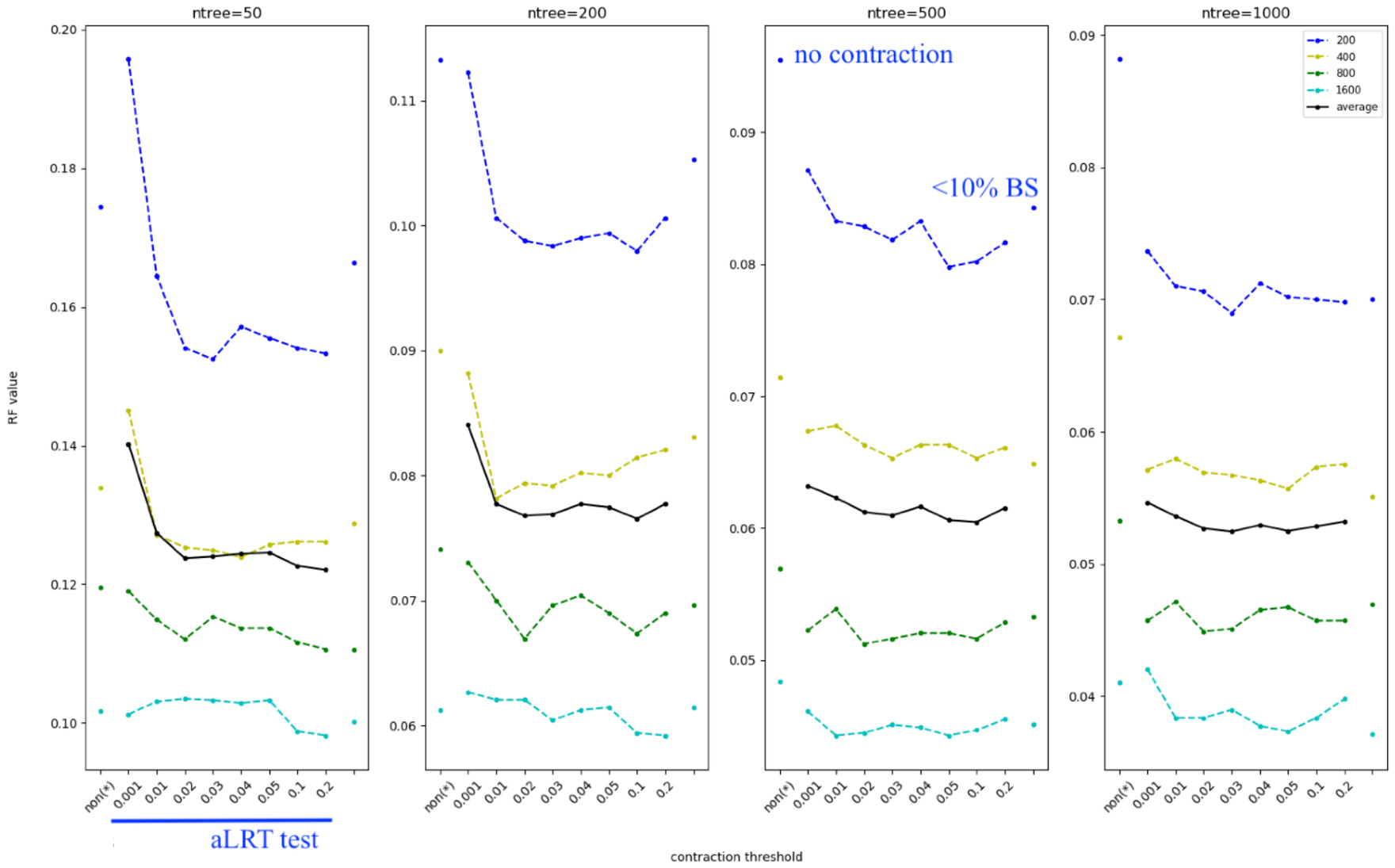
ASTRAL-III on all 14,446 unbinned gene trees





Zhang, Chao, et al. BMC Bioinformatics 19, no. S6 (2018): 153. https://doi.org/10.1186/s12859-018-2129-y

Using aLRT tests



PhyML_mac -m GTR -o n -d nt -i ...phylip -u somffile -b -2 -n 1000

- Back to tutorial on https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md

```
nw_ed  1KP-genetrees.tre 'i & b<=10' o > 1KP-genetrees-BS10.tre
```

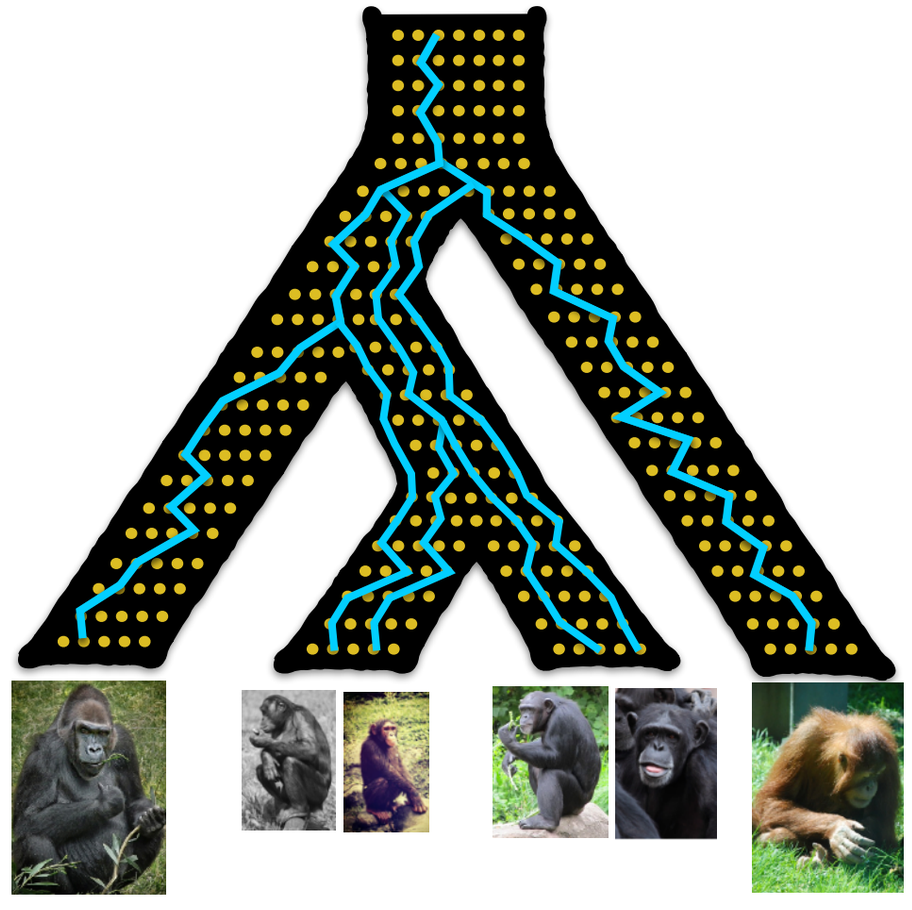
```
java -jar astral.5.7.3.jar -i test_data/1KP-genetrees-BS10.tre
-o test_data/1kp-BS10.tre 2> test_data/1kp-bs10.log
```

Multiple individuals

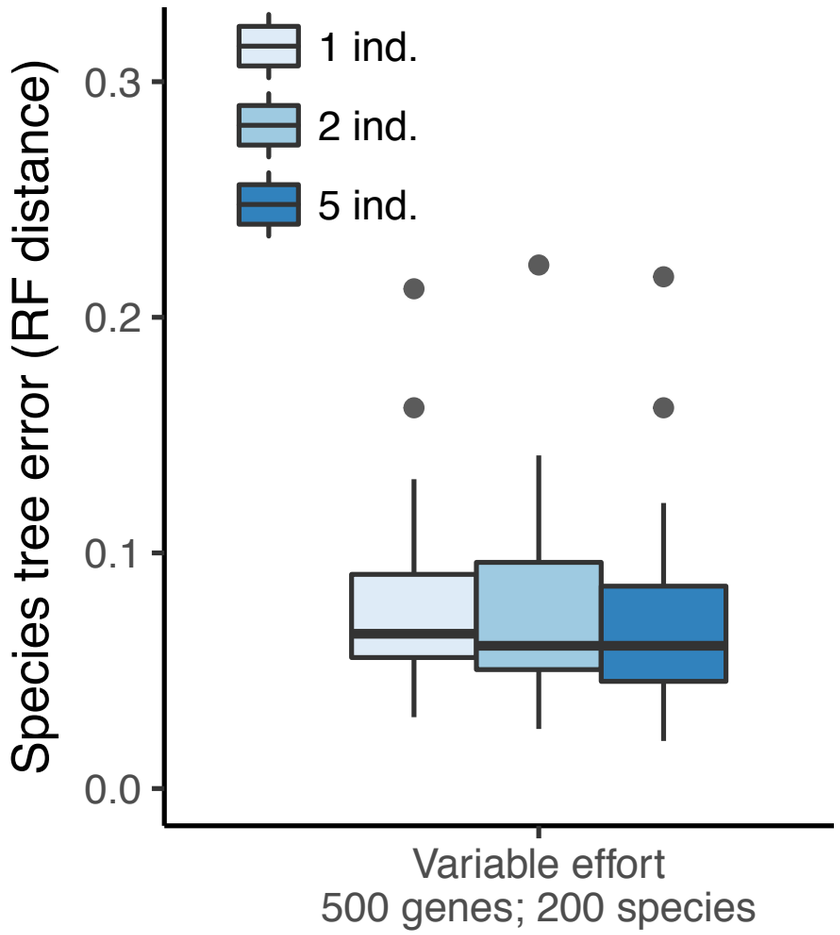
- What if we sample multiple individuals from each species?

- In recently diverged species individuals *can* be non-monophyletic in gene trees

- Sampling multiple individuals may provide extra signal

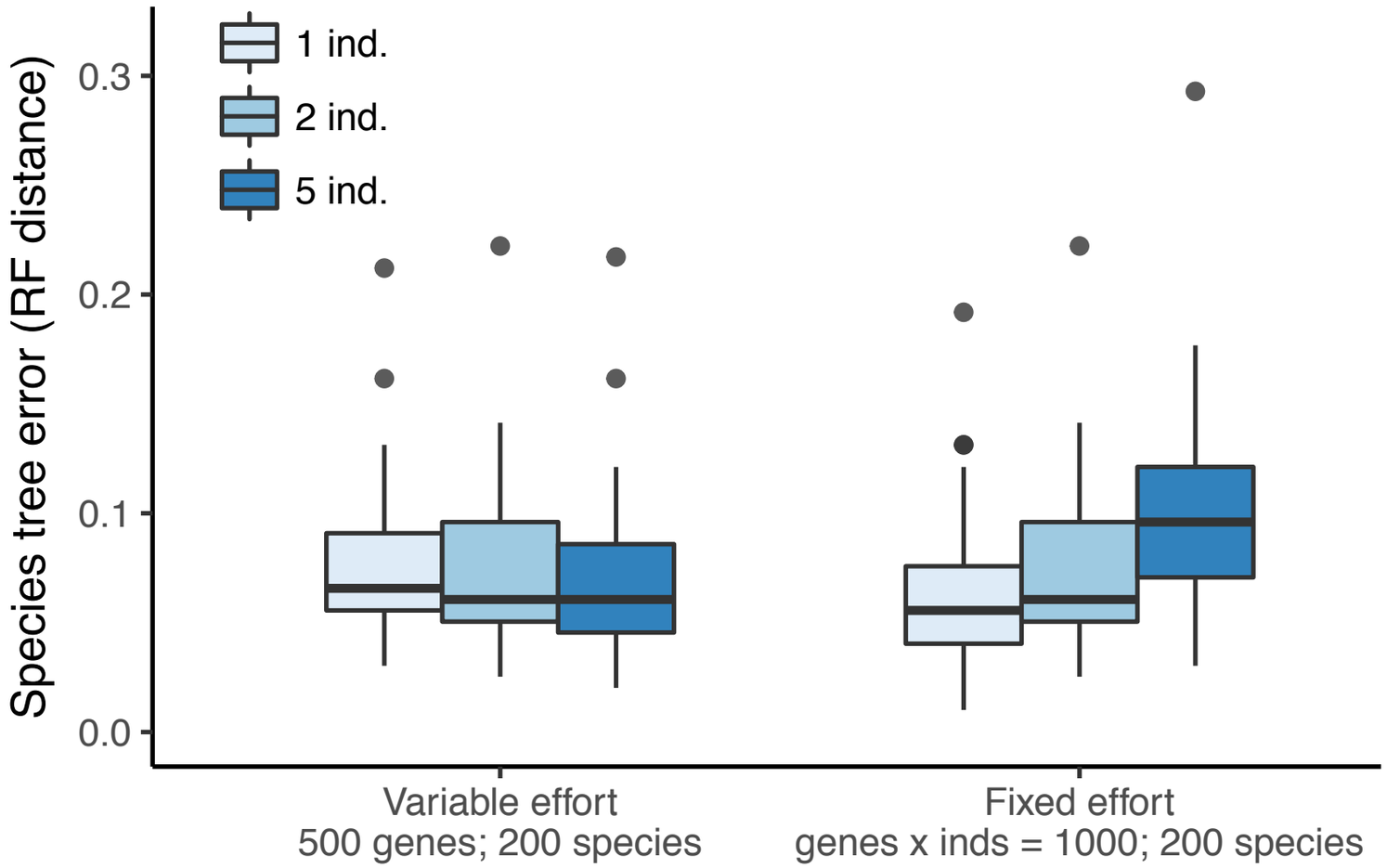


Multiple individuals helpful?



Yes, it marginally helps accuracy

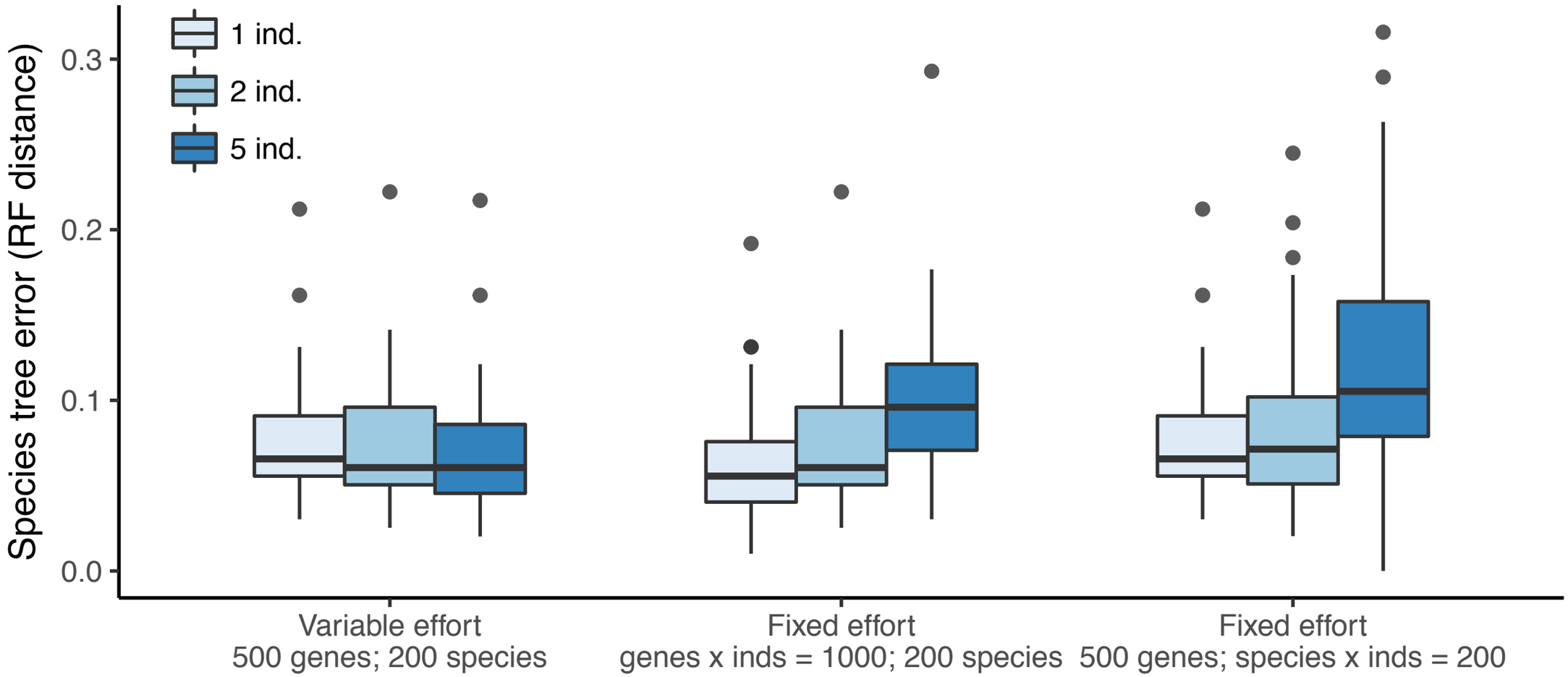
Multiple individuals helpful?



Yes, it marginally helps accuracy

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

Multiple individuals helpful?



Yes, it marginally helps accuracy

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

- Back to tutorial on https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md

```
java -jar astral.5.7.3.jar -i test_data/easy-multi.tre -a
test_data/mapping-10.txt
```

Sorry, the test files are not on the website right now!

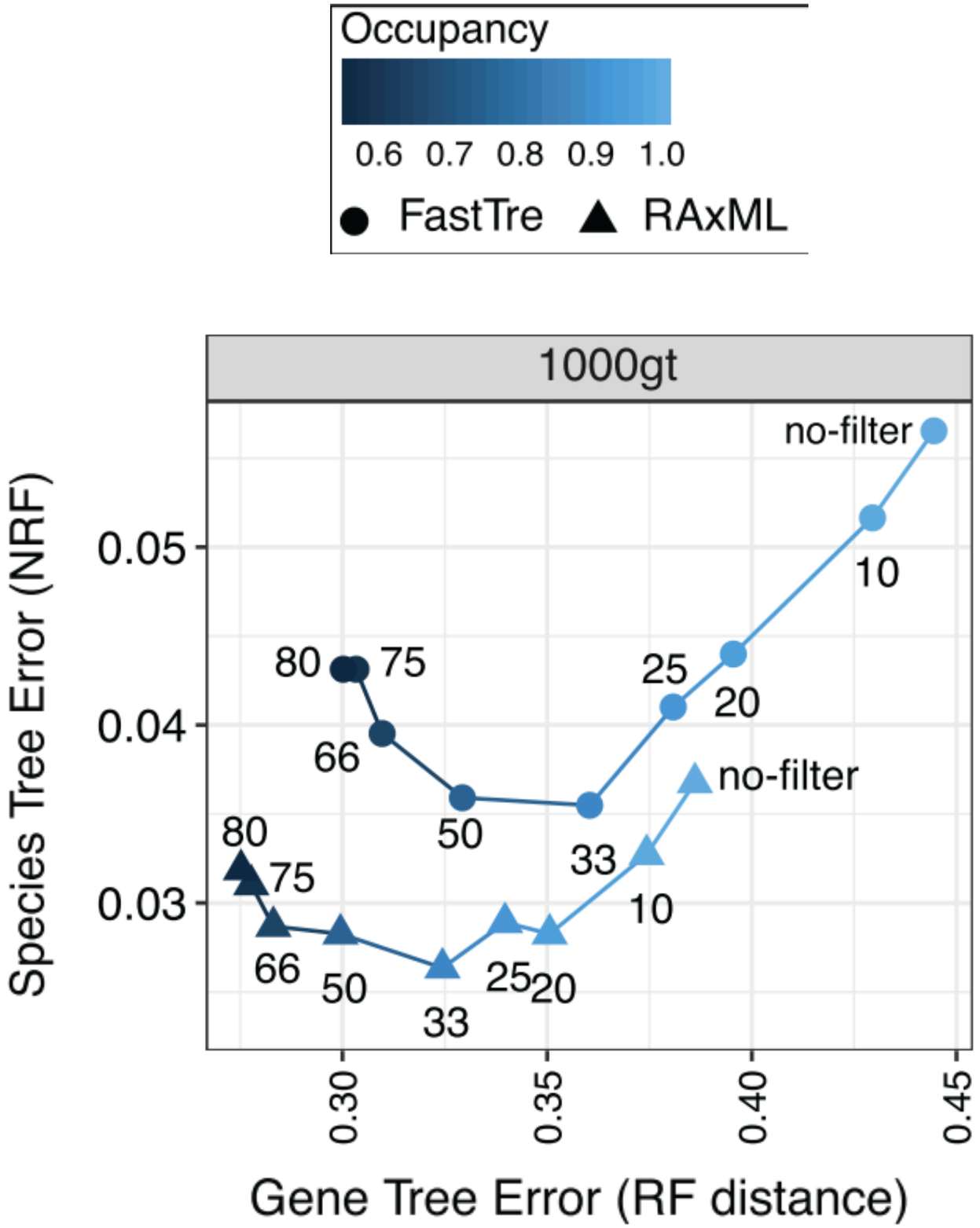
Should you filter?

- Filtering genes based on missing species from specific genes?
    - Generally not beneficial (see Molloy and Warnow, Systematic Biology 2018)
- Filtering genes based on gene tree estimation error?
    - Depends on conditions (see Molloy and Warnow, Systematic Biology 2018)
- Filtering fragmentary sequences from genes while keeping the gene?
    - Often beneficial (see Sayyari, Whitfield, and Mirarab, MBE 2018)

Filtering fragments

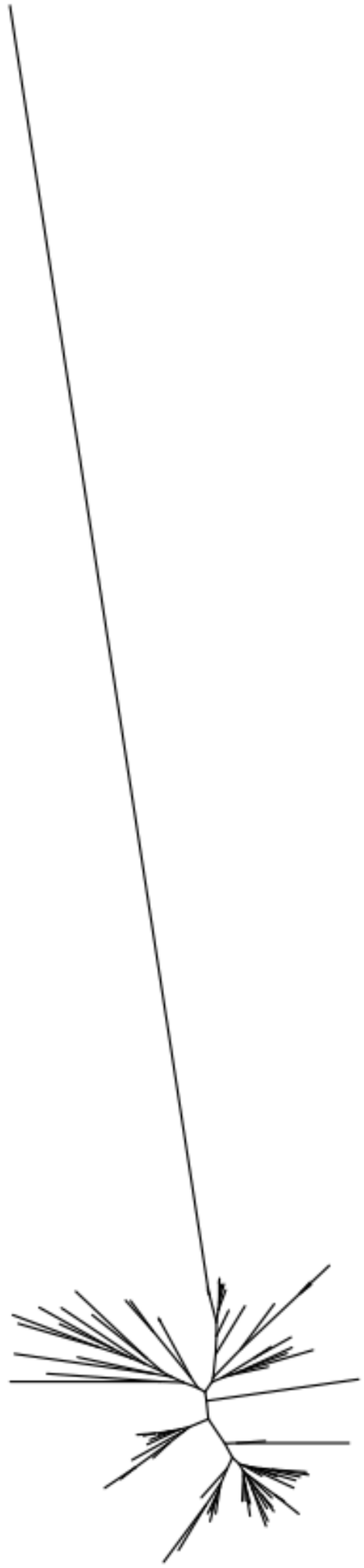
- Added fragmentation to simulated data with patterns similar to the Misof. et. al. insect (transcriptome data)

- Filtering simply removes fragmentary data from genes but keeps the gene



TreeShrink

- Removes super long branches from gene trees
- Is automatic
- Improves gene trees



Notes on output

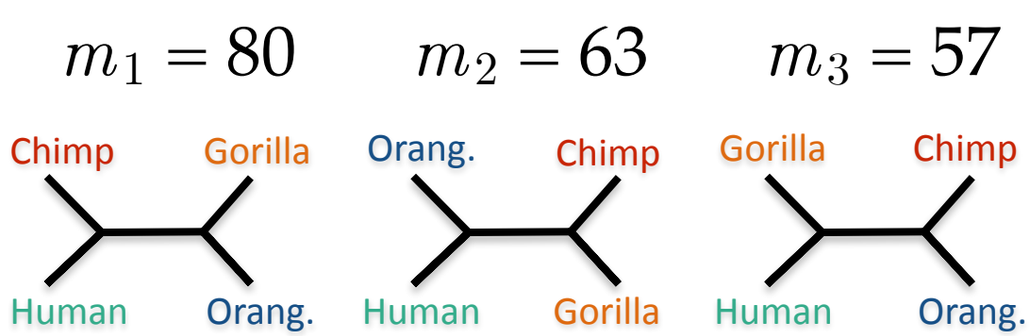
- Back to tutorial on https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md

- Open the tree
    - Comment on rooting

- Examine log information

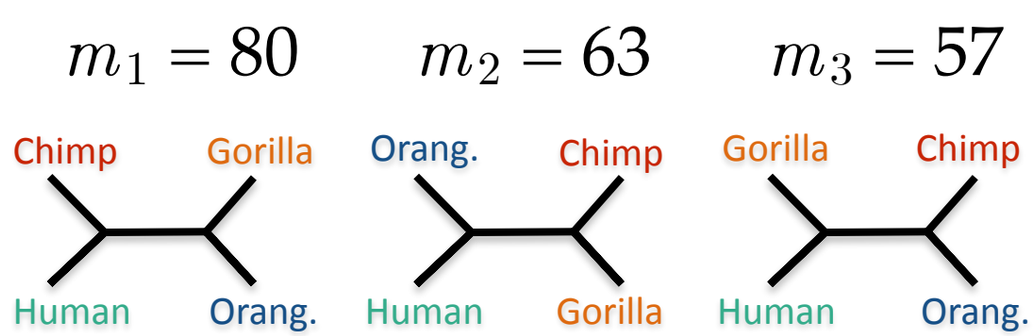
Local posterior probability

- Recall quartet frequencies follow a multinomial distribution



Local posterior probability

- Recall quartet frequencies follow a multinomial distribution

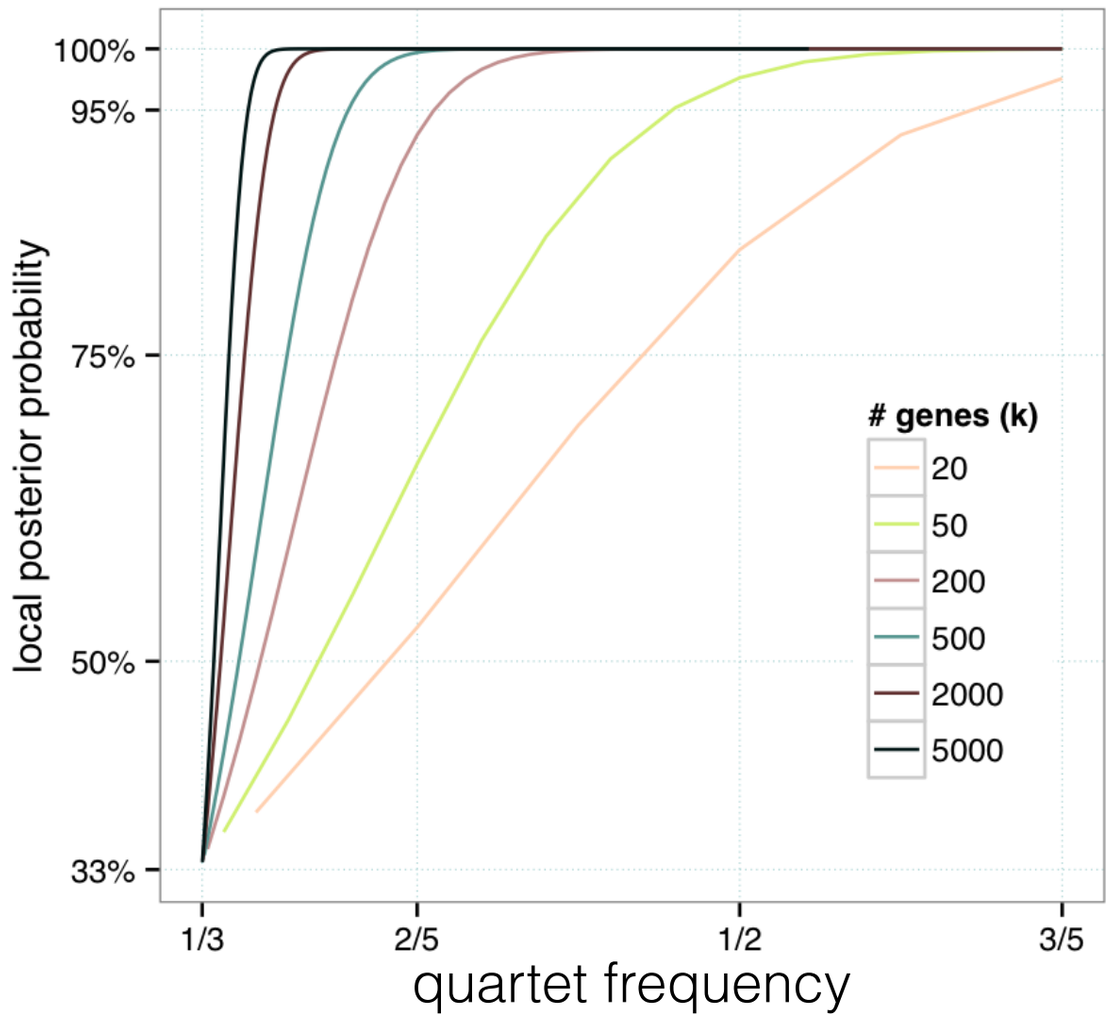


- $\boldsymbol{P}$**(**gene tree seen m_1/m times = species tree**)** = $\boldsymbol{P}$**(**$\theta_1 > 1/3$**)**

  - Can be solved analytically. The resulting measure is called "the local posterior probability" (localPP)

  - Handle $n > 4$ by averaging quartet scores

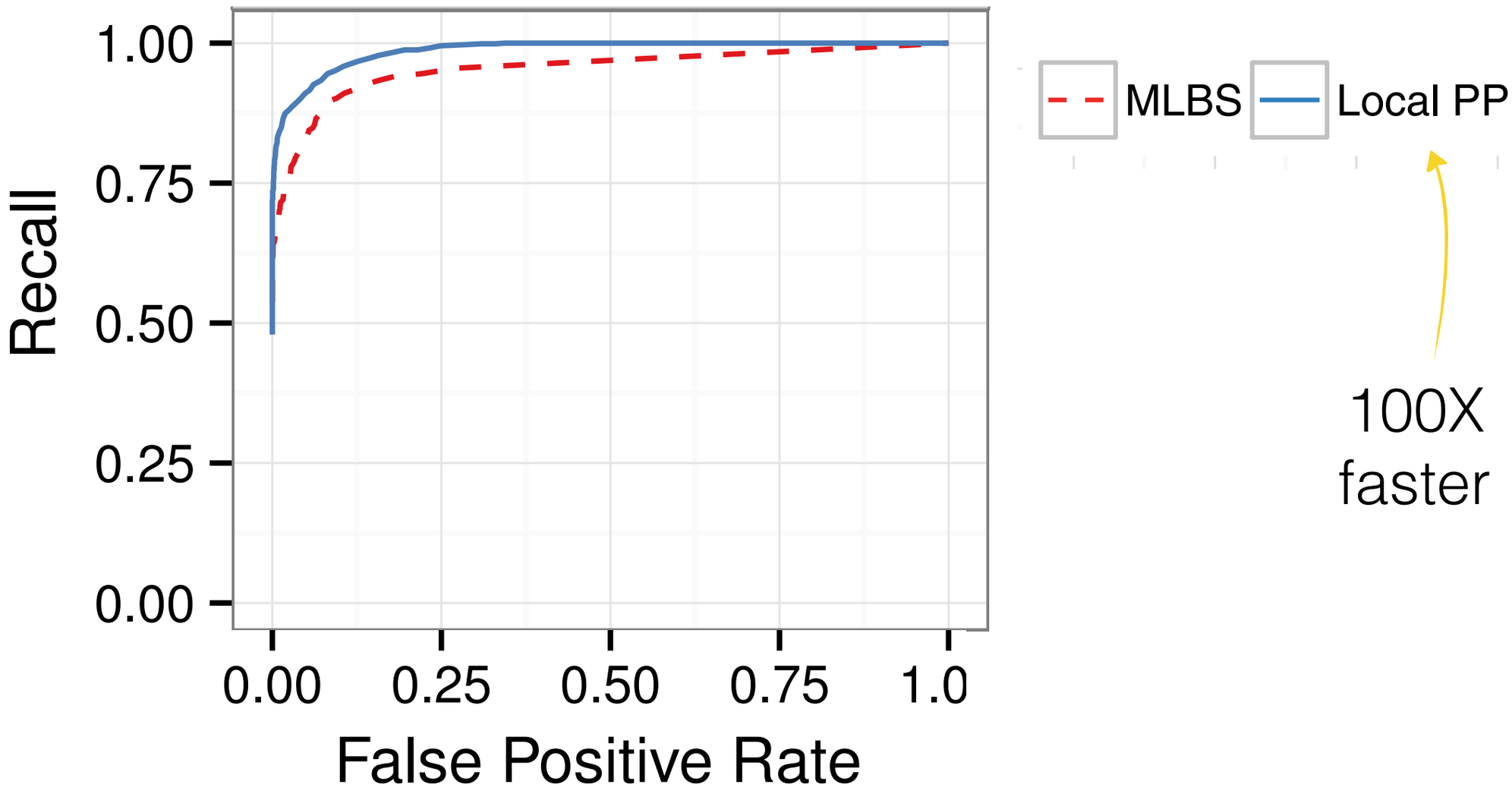
Quartet support vs. localPP



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

Decreased discordance $\Rightarrow$ increased support

localPP is more accurate than bootstrapping



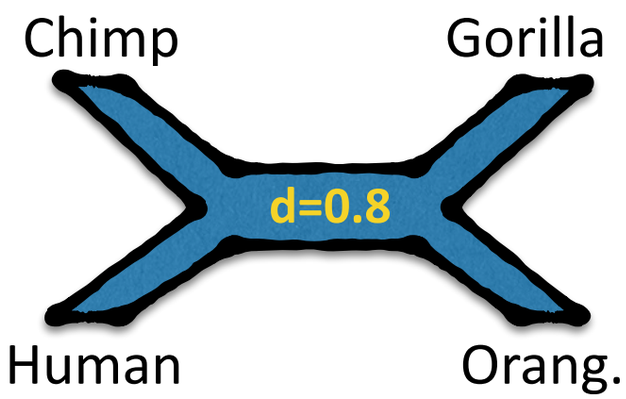
Avian simulated dataset (48 taxa, 1000 genes)

[Sayyari and Mirarab, MBE, 2016]

Branch Length

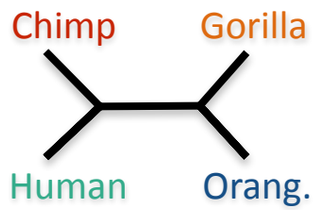
[Sayyari and Mirarab, MBE, 2016]

- The level of discordance is a function of coalescent unit branch length

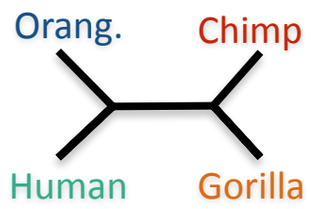


$$\theta_1 = 1 - \frac{2}{3}e^{-d}$$

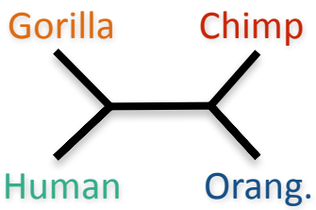
θ_1=70%



θ_2=15%



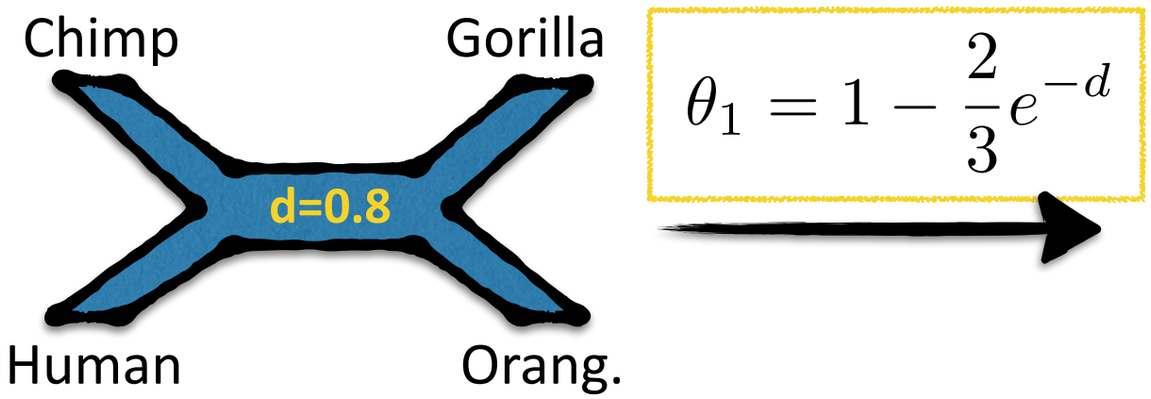
θ_3=15%



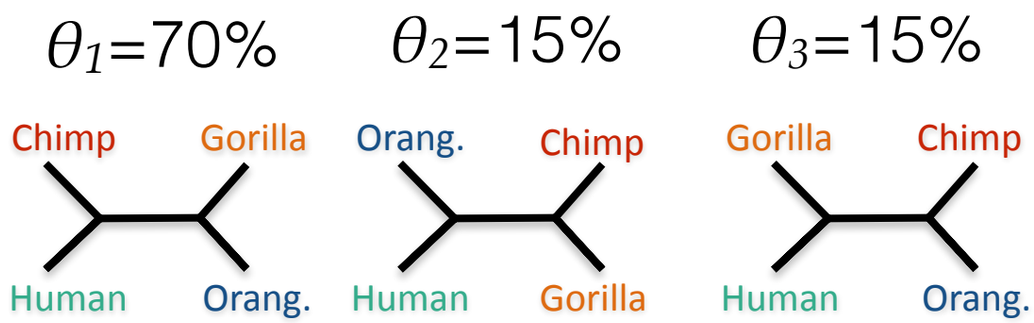
Branch Length

[Sayyari and Mirarab, MBE, 2016]

- The level of discordance is a function of coalescent unit branch length
- A single quartet ($n=4$): just reverse the discordance formula to get the ML estimate



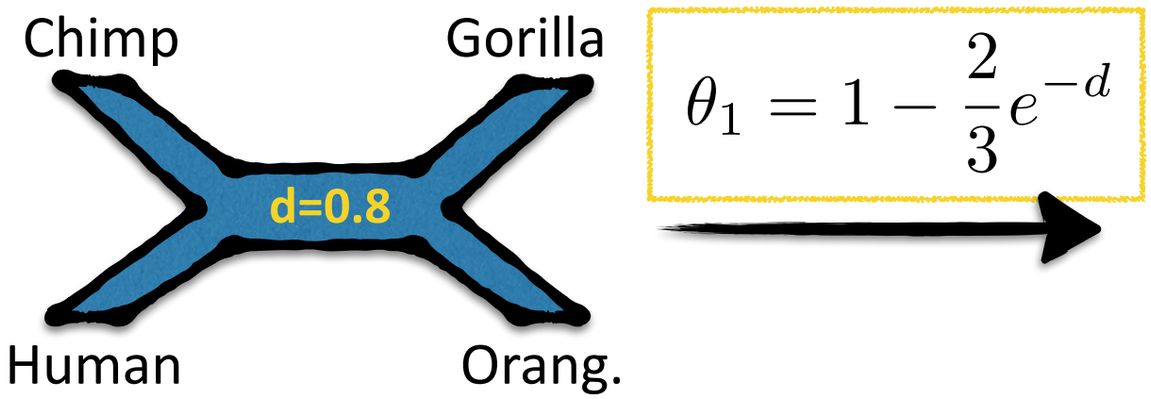
$$\theta_1 = 1 - \frac{2}{3}e^{-d}$$



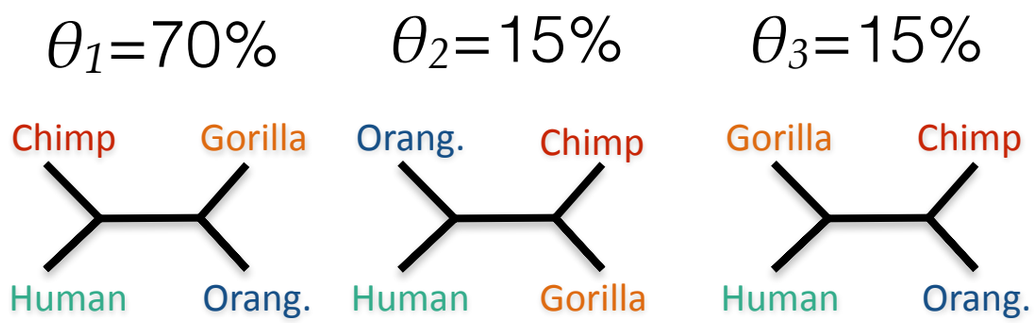
Branch Length

[Sayyari and Mirarab, MBE, 2016]

- The level of discordance is a function of coalescent unit branch length
- A single quartet ($n=4$): just reverse the discordance formula to get the ML estimate
- $n>4$: average frequencies around a branch

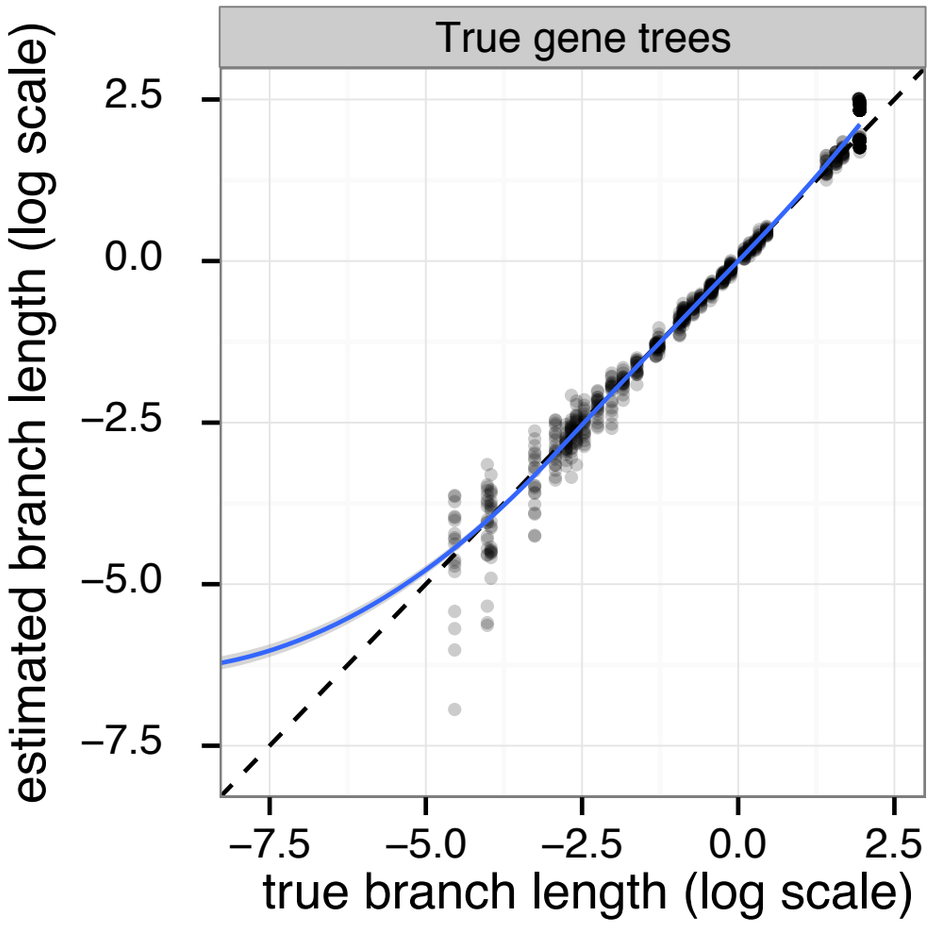


$$\theta_1 = 1 - \frac{2}{3}e^{-d}$$



Branch lengths accuracy

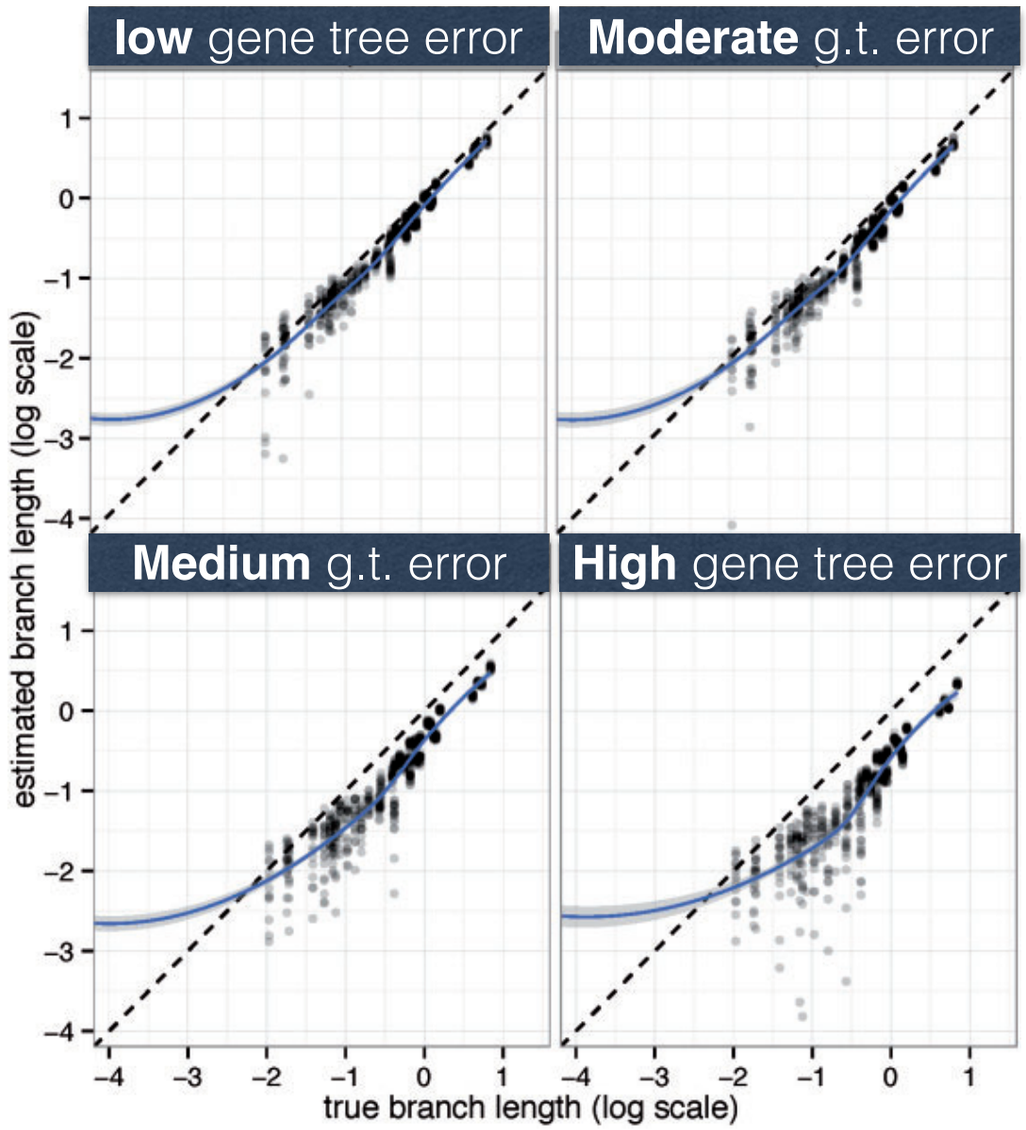
[Sayyari and Mirarab, MBE, 2016]



With true gene trees, ASTRAL correctly estimates BL

Branch lengths accuracy

[Sayyari and Mirarab, MBE, 2016]



With error-prone estimated gene trees, ASTRAL underestimates BL

Caveats

- Branch length:
  - Only for internal branches unless you have multiple individuals for a species
    - In coalescent units, so the *true* value is still a function of population size and generation time in addition to actual time
- Local PP:
  - Empirically better than BS support but based on many assumptions

- Back to tutorial on https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md

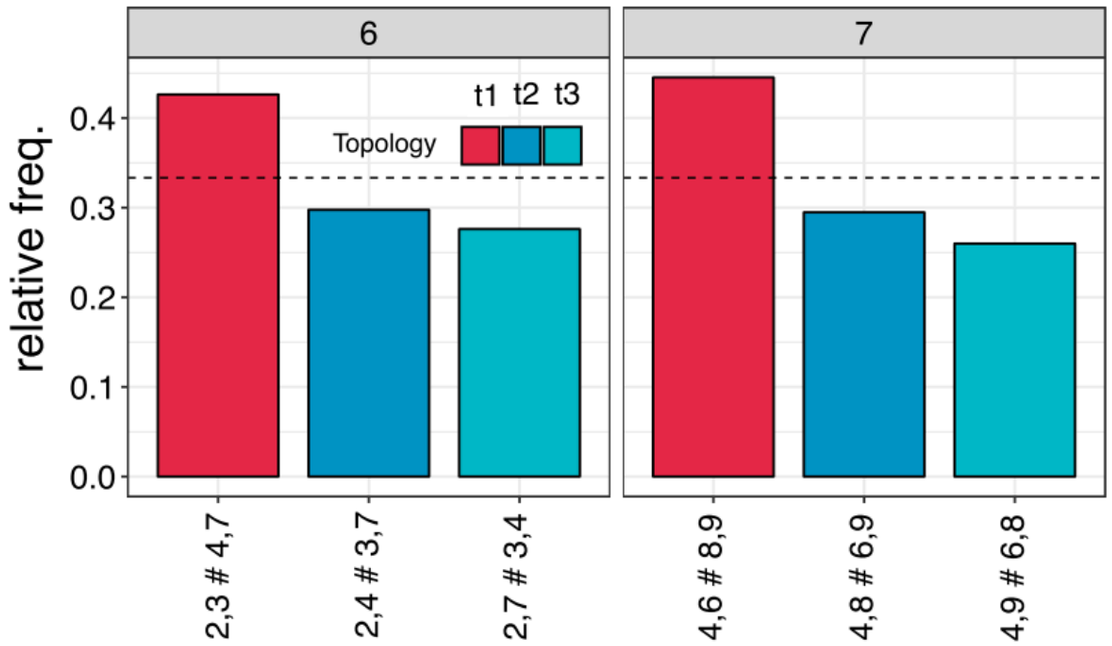
- Score an existing tree

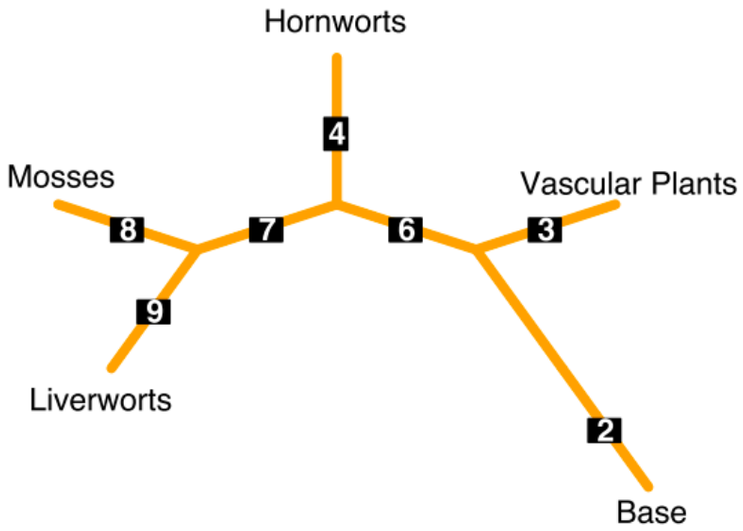
    - Extensive branch annotation

- Polytomy test (`-t 10`)

    - Sayyari, Erfan, and Siavash Mirarab. “Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies.” Genes 9, no. 3 (February 28, 2018): 132. https://doi.org/10.3390/genes9030132 .

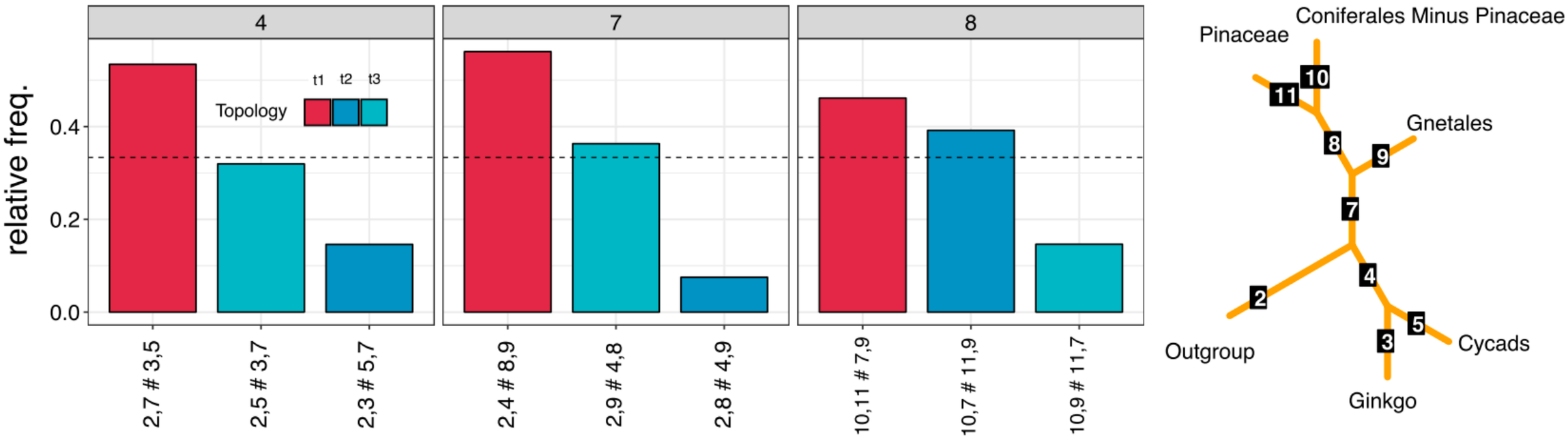
Discovista: visualizing discordance





- https://github.com/esayyari/DiscoVista
- Sayyari, et. al. “DiscoVista: Interpretable Visualizations of Gene Tree Discordance.” MPE 122 (2018): 110–15.

Discovista: visualizing discordance



- https://github.com/esayyari/DiscoVista
- Sayyari, et. al. “DiscoVista: Interpretable Visualizations of Gene Tree Discordance.” MPE 122 (2018): 110–15.

ASTRAL-MP

- A separate branch on GitHub from the main branch https://github.com/smirarab/ASTRAL/tree/MP

- Try Installation:
  https://github.com/smirarab/ASTRAL/tree/MP#installation

  - Compiling from the code may or may not be needed. Make sure you do the test below (AVX2)

```
java -D"java.library.path=lib/" -jar native_library_tester.jar
```

  - Explore -C and -T options

Some ASTRAL-related papers

- ASTRAL performs well under ILS and HGT (Davidson, Vachaspati, Mirarab, and Warnow). BMC Genomics (2015).
- It can handle multiple alleles per species (Rabiee, Sayyari, and Mirarab). MPE (2019)
- It computes branch length and branch support and can test for polytomies using quartet frequencies (Sayyari and Mirarab). MBE (2016) and Genes (2018)
- Visualizing quartet discordance using DiscoVista (Sayyari, Whitfield, and Mirarab) MPE (2018)
- Fragmentary sequences can negatively impact ASTRAL trees (Sayyari, Whitfield, and Mirarab). MBE (2017)
- Filtering loci is not beneficial! (Molloy and Warnow), Systematic Biology (2018)
- ASTRAL consistent under models of missing data (Nute, Molloy, Chou, and Warnow). BMC Genomics (2018)
- How many genes does ASTRAL need? (Shekhar, Roch, and Mirarab). Transactions on Computational Biology and Bioinformatics (2017)
- Using ASTRAL as a supertree method (Vachaspati and Warnow). Bioinformatics (2017)

I'll stop here

Time left?

— A-Pro (https://github.com/chaoszhang/A-pro)

— MLBS (https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md#multi-locus-bootstrapping)