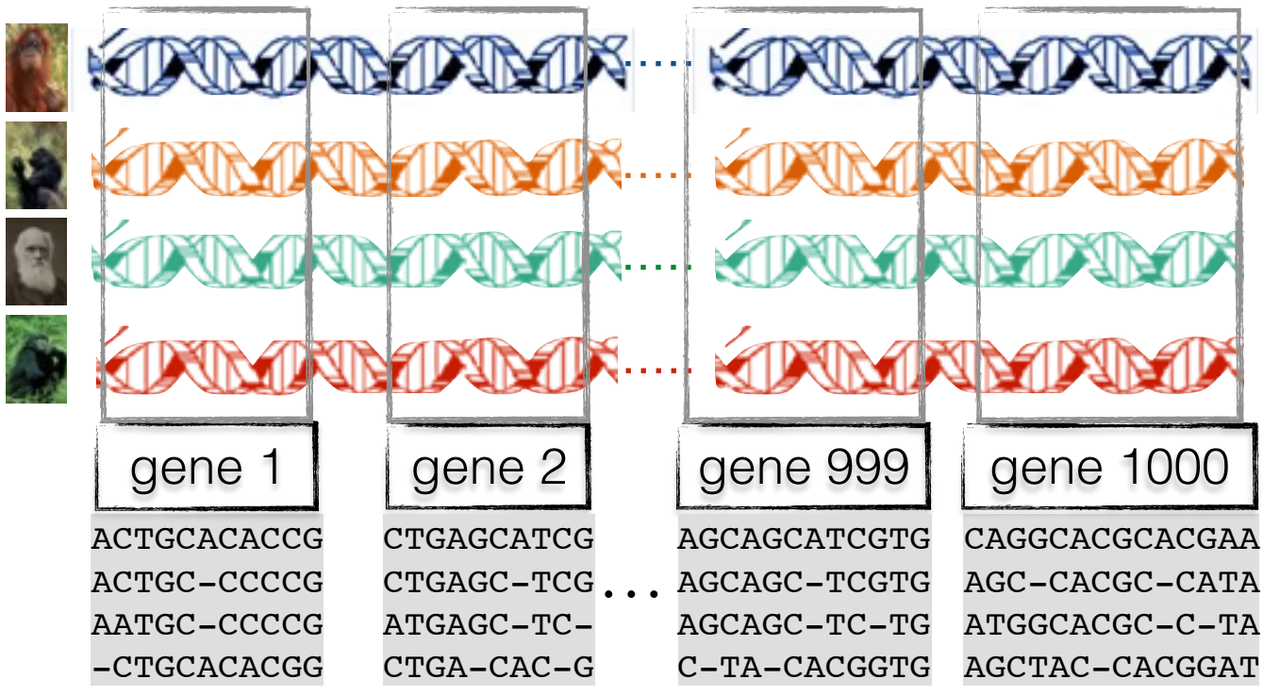
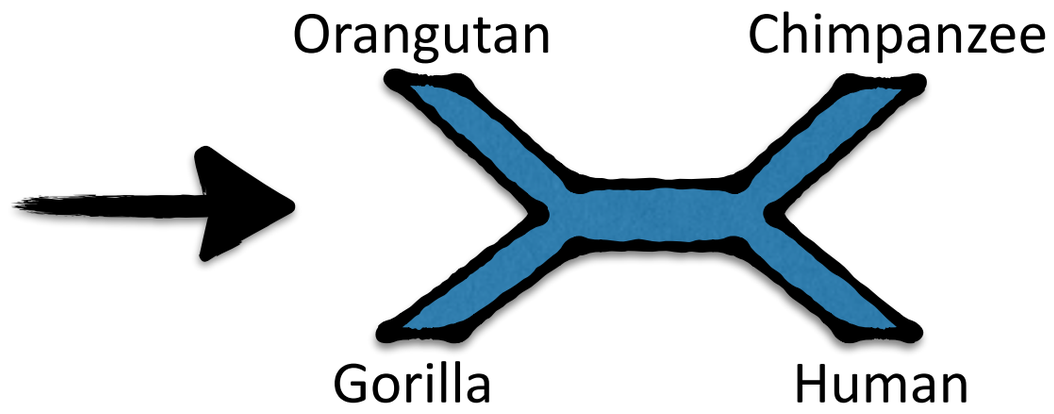
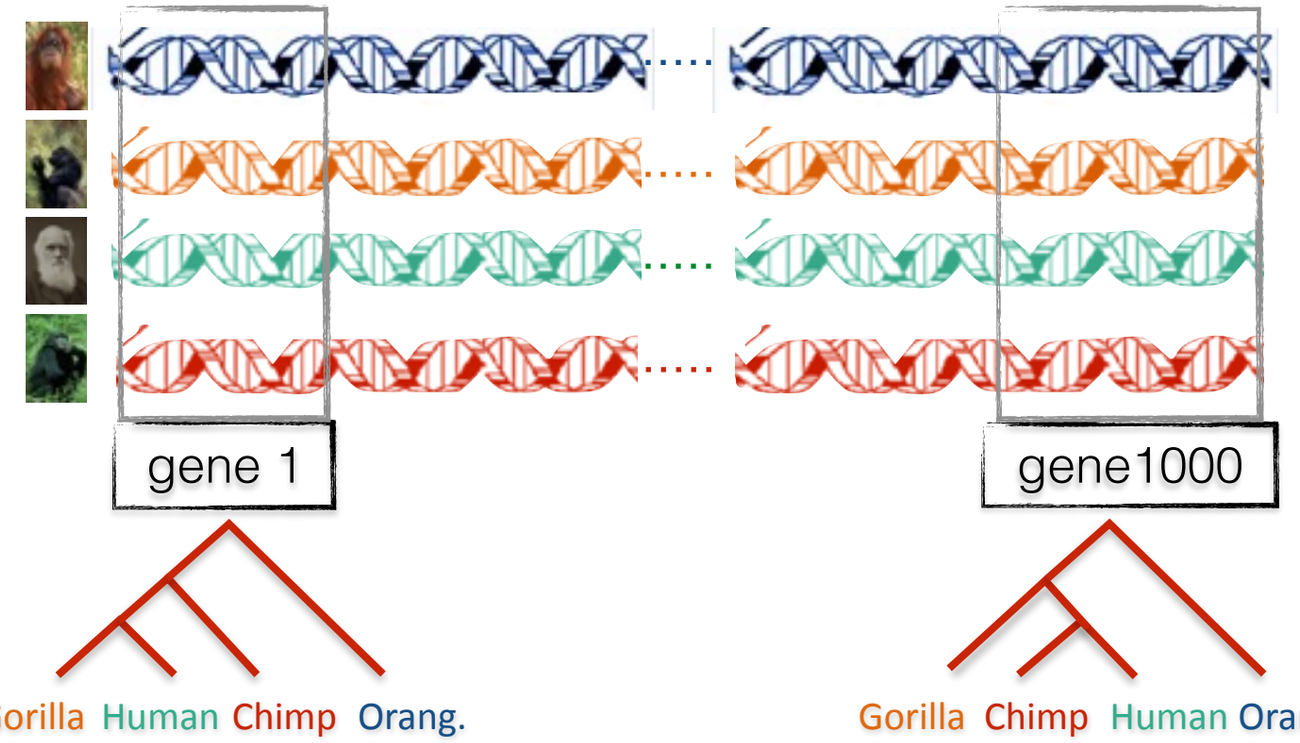
Species tree reconstruction using ASTRAL: recent advances and future directions

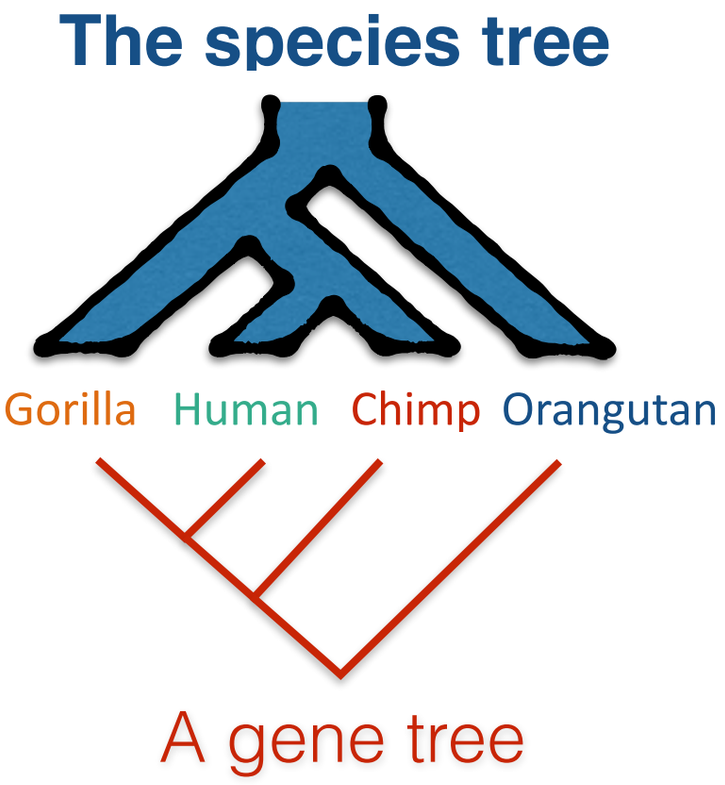
Siavash Mirarab
University of California, San Diego (ECE)





Gene tree discordance





Causes of gene tree discordance include:

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

MSC and Identifiability

- A statistical model called multi-species coalescent (MSC) is used to study ILS.

MSC and Identifiability

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

MSC and Identifiability

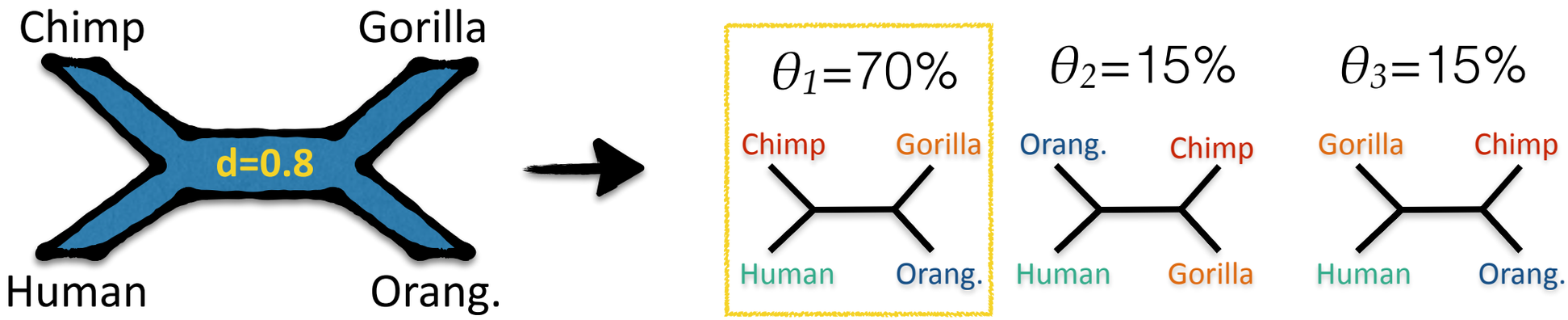
- A statistical model called multi-species coalescent (MSC) is used to study ILS.

- 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

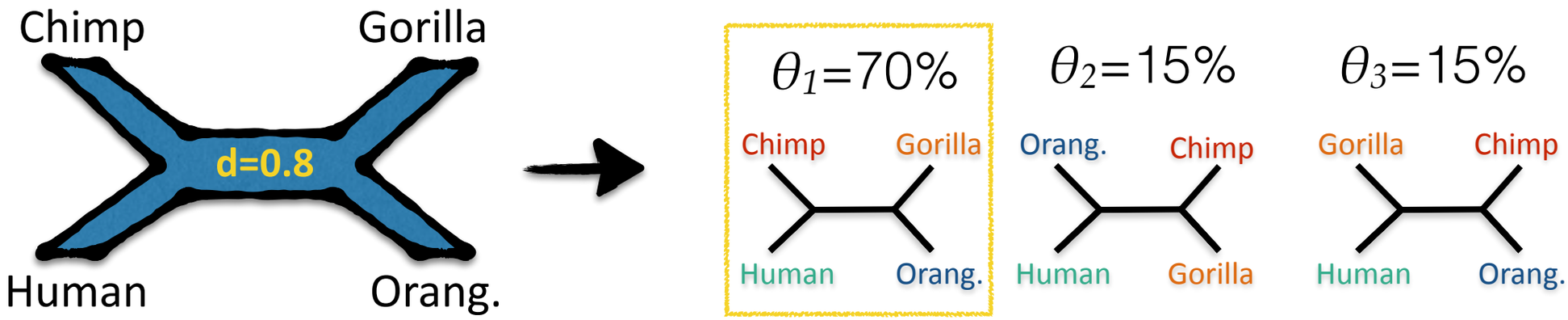
Unrooted quartets under MSC model

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)



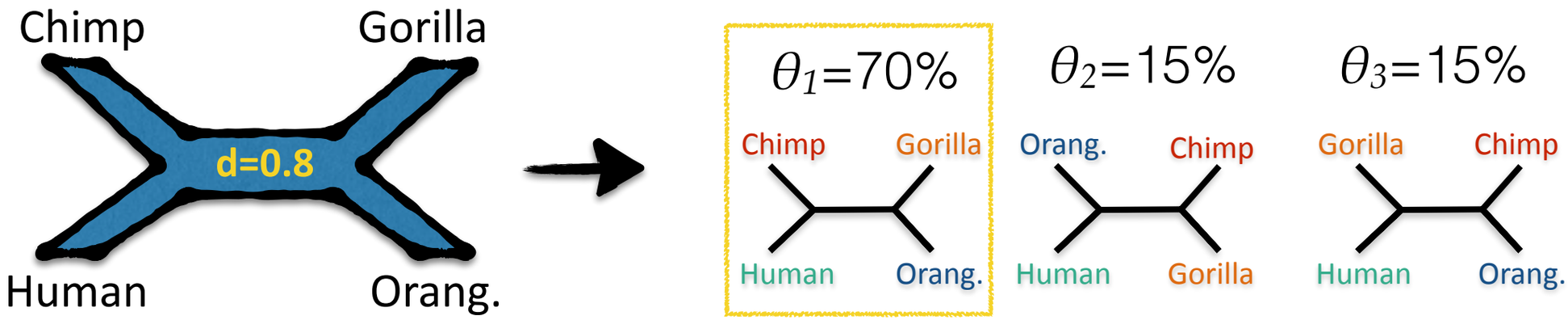
The most frequent gene tree

=

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

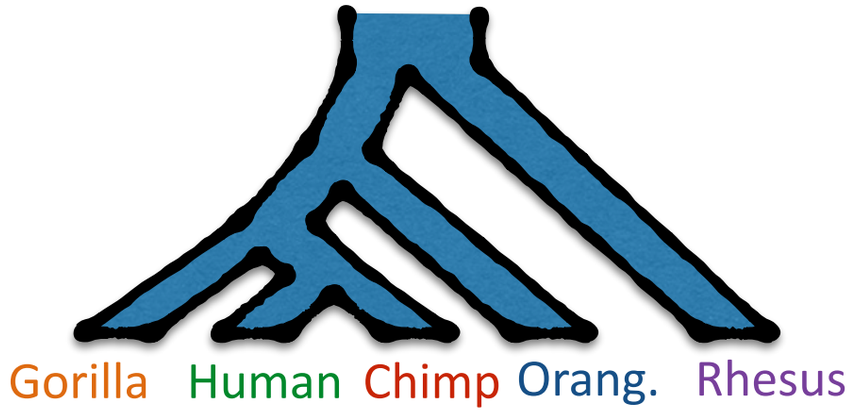
shorter branches ⇒

more discordance ⇒

a harder species tree reconstruction problem

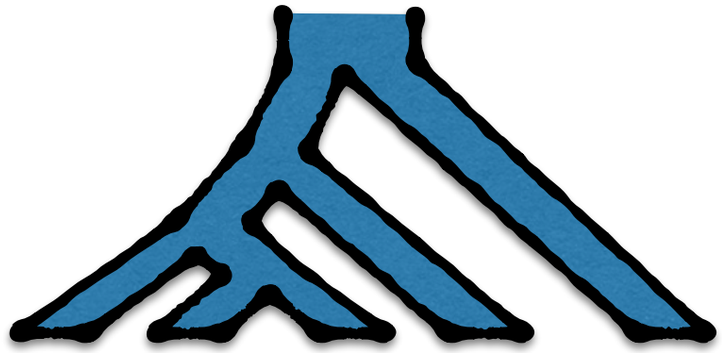
More than 4 species

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



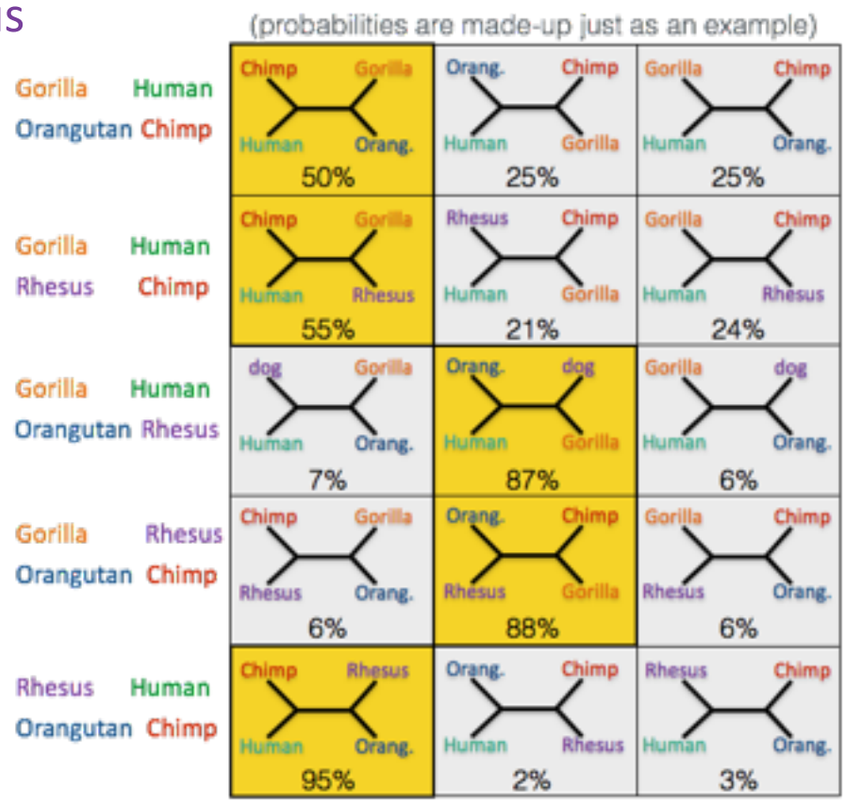
More than 4 species

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



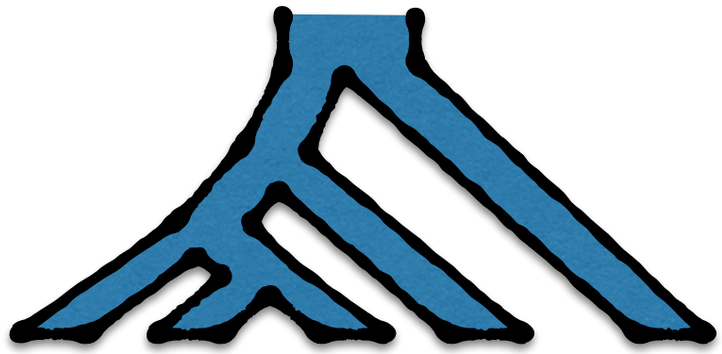
1. Break gene trees into $\binom{n}{4}$ quartets of species
2. Find the dominant tree for all quartets of taxa
3. Combine quartet trees

   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)



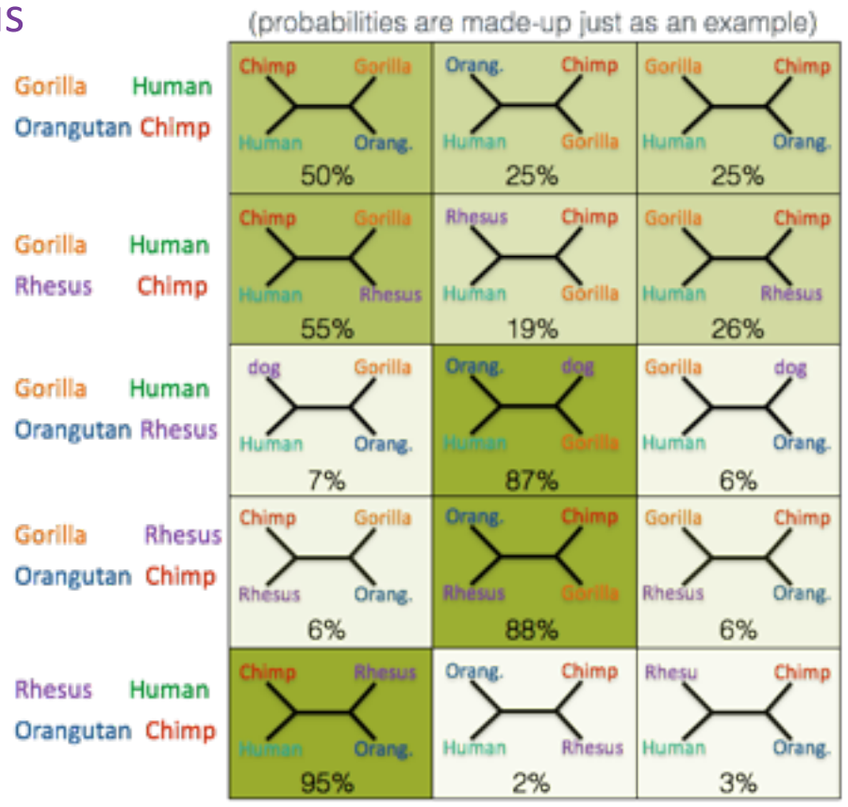
Gorilla Human Chimp Orang. Rhesus

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 (NP-hard; Lafond & Scornavaccaori):

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 (NP-hard; Lafond & Scornavaccaori):

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]

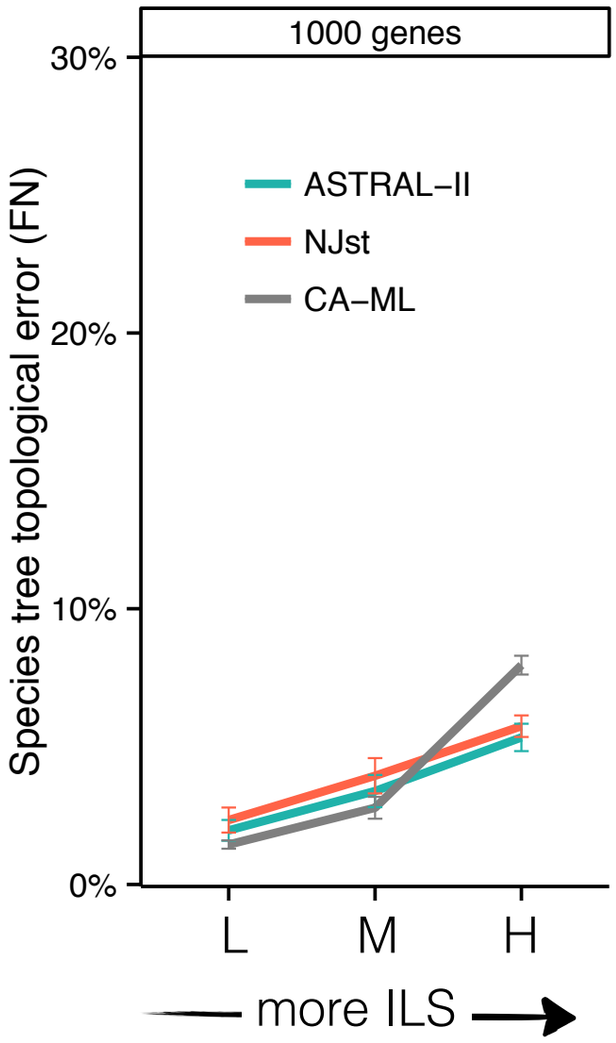
- Solve the problem exactly using dynamic programming
  - Constrains the search space to make large datasets feasible
  - The constrained version remains statistically consistent
  - Running time: polynomially increases with the number of genes and the number of species

ASTRAL-I and ASTRAL-II

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

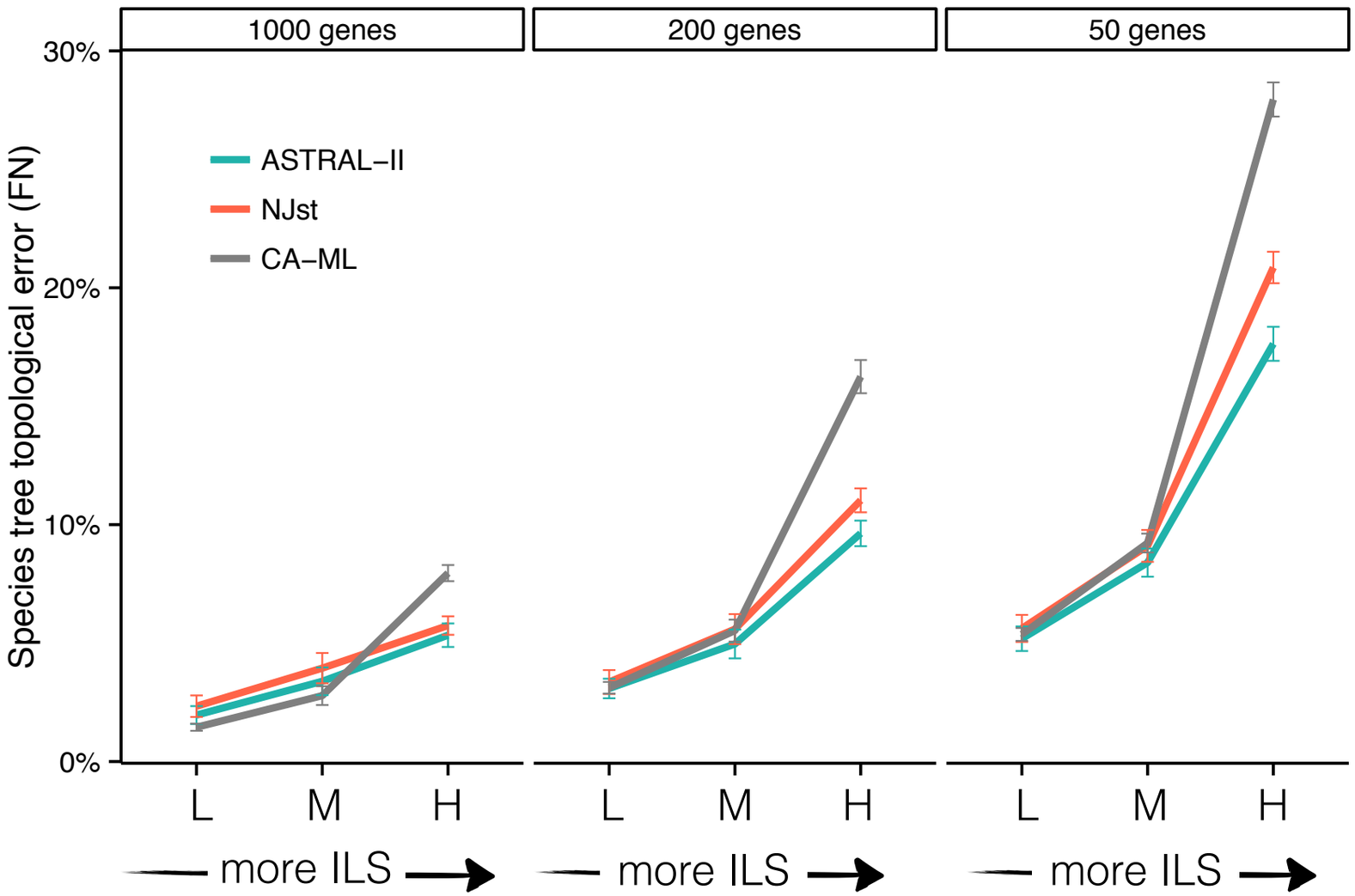
- Solve the problem exactly using dynamic programming
  - Constrains the search space to make large datasets feasible
  - The constrained version remains statistically consistent
    - Running time: polynomially increases with the number of genes and the number of species
- ASTRAL-II:
  - Increased the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

Comparison to concatenation: depends on the ILS level



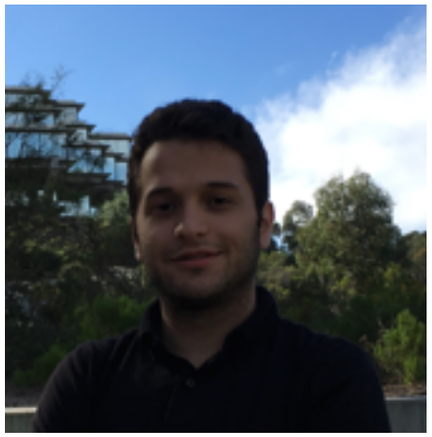
200 species, recent ILS

Comparison to concatenation: depends on the ILS level



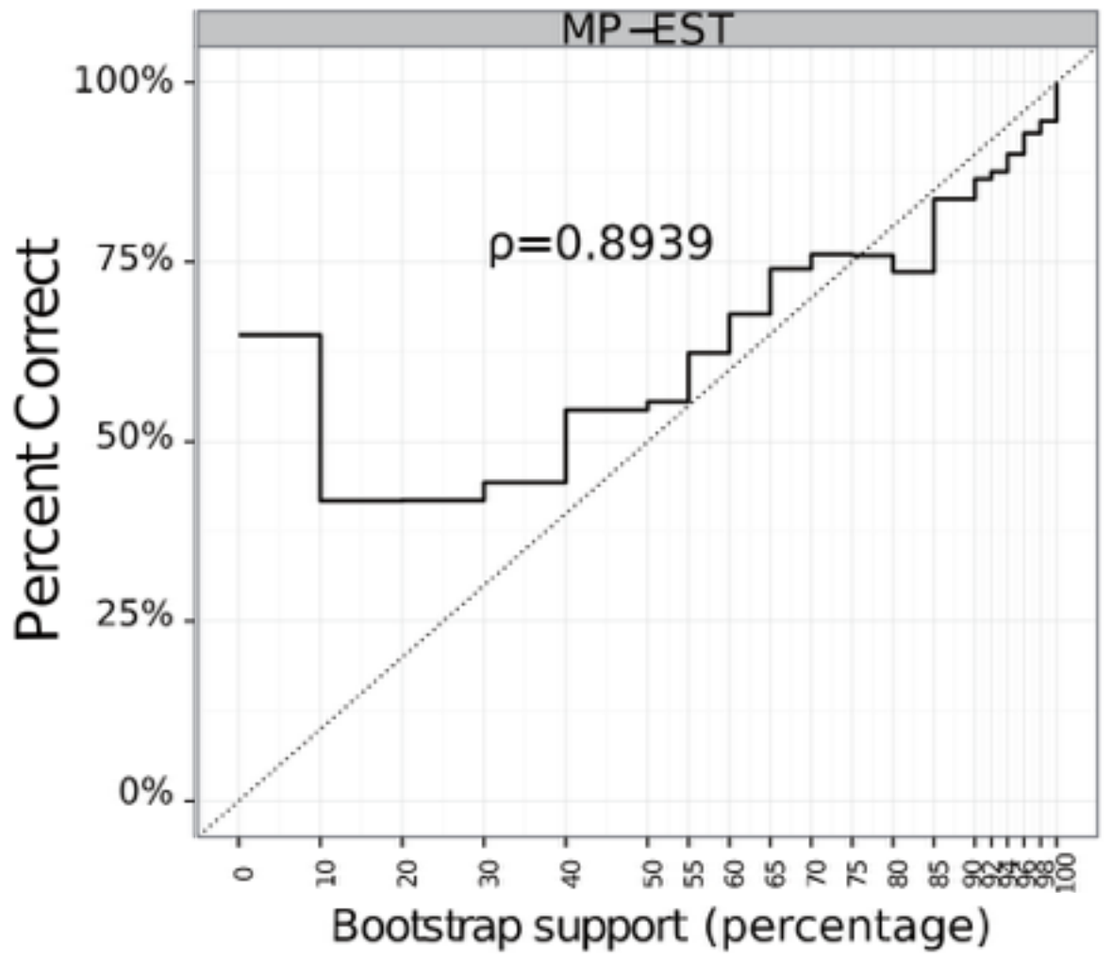
200 species, recent ILS

Branch support



Erfan Sayyari

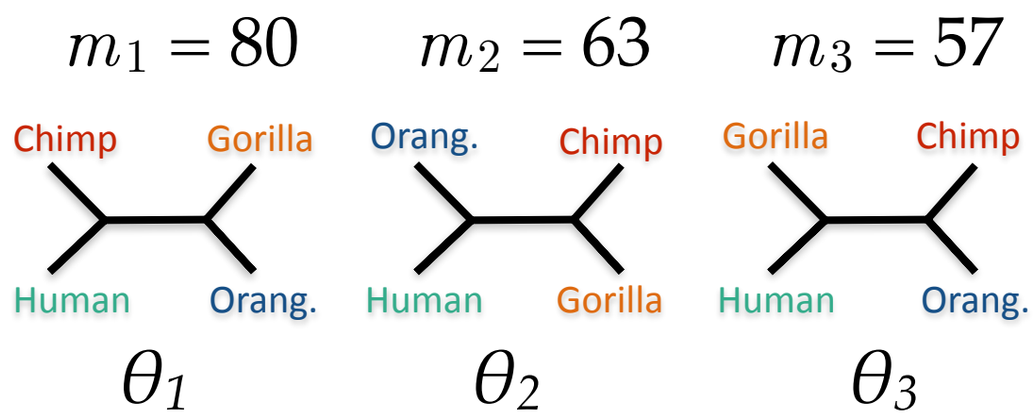
- Traditional approach:
  Multi-locus bootstrapping (MLBS)
  - Slow: requires bootstrapping all genes (e.g., $100\times m$ ML trees)
  - Inaccurate and hard to interpret
    [Mirarab et al., Sys bio, 2014; Bayzid et al., PLoS One, 2015]
- We can do better!



[Mirarab et al., Sys bio, 2014]

Local posterior probability

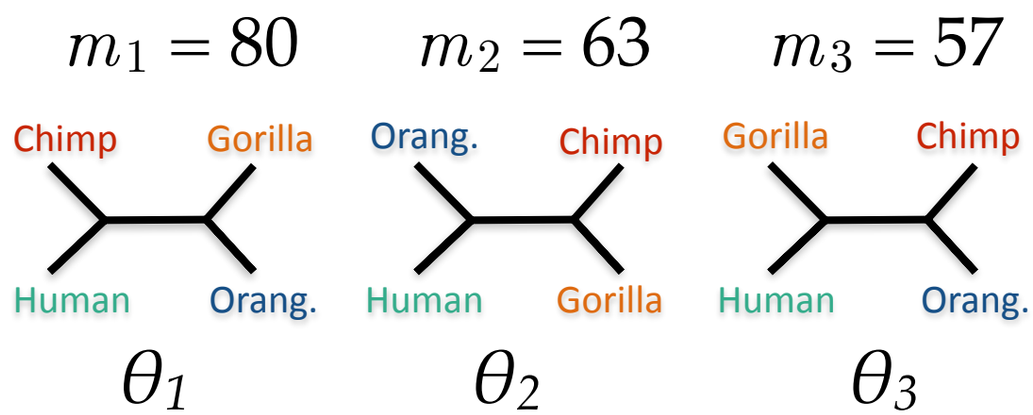
- Recall quartet frequencies follow a multinomial distribution



- $\boldsymbol{P}$ (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

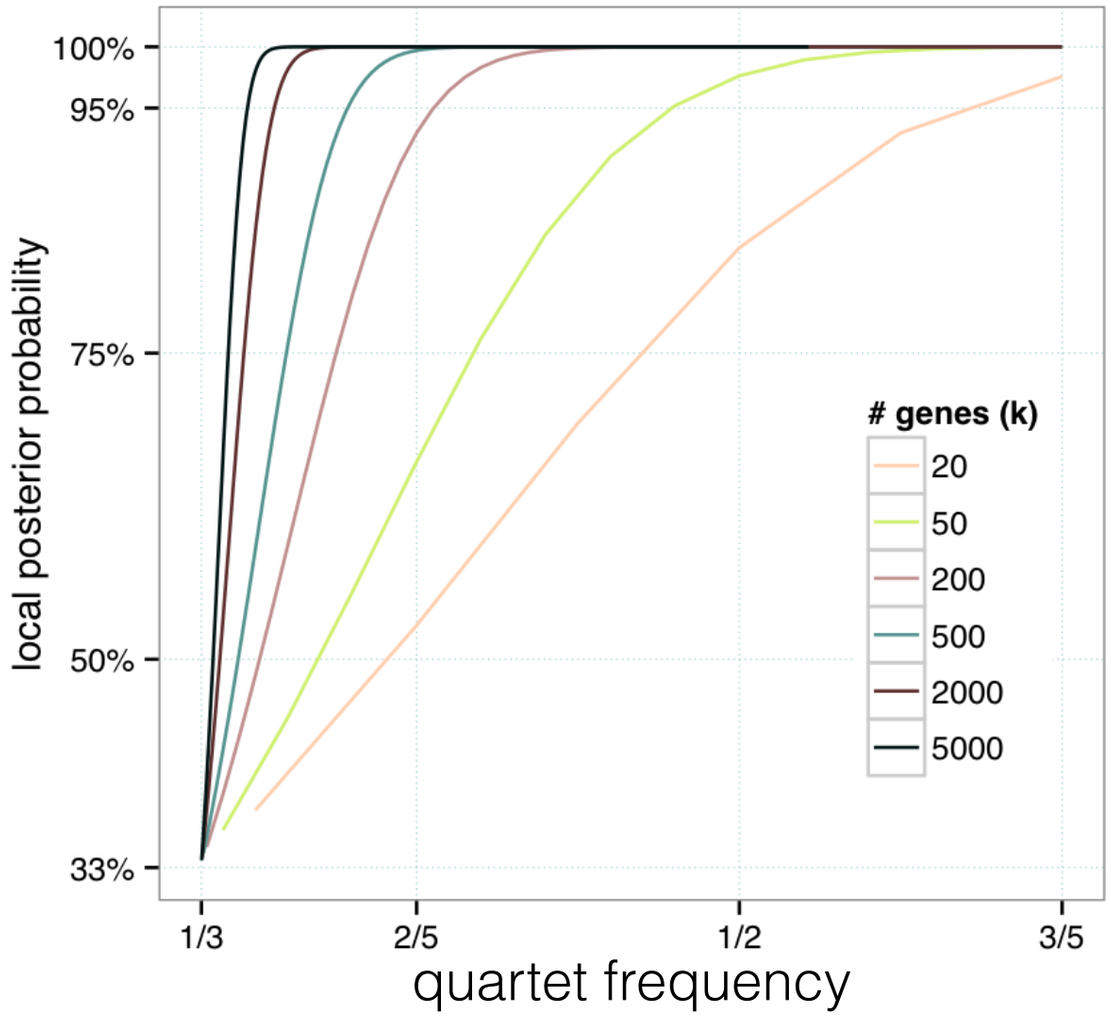


- $\boldsymbol{P}$ (topology seen in m_1 / m gene trees is the species tree) = $\boldsymbol{P}$ ($\theta_1 > 1/3$)

- Can be analytically solved

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

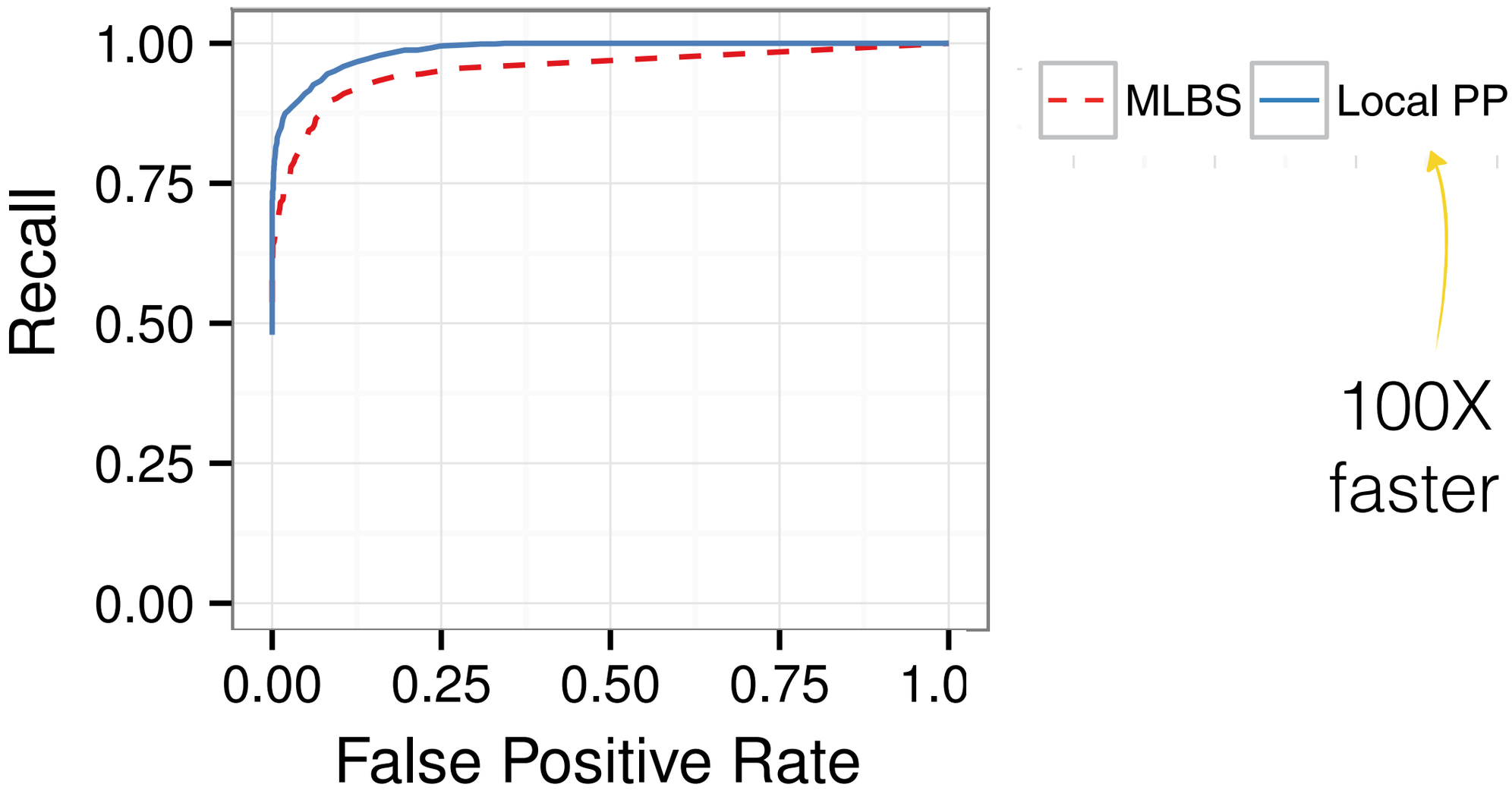
Quartet support v.s. localPP



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

Decreased discordance $\Rightarrow$ increased support

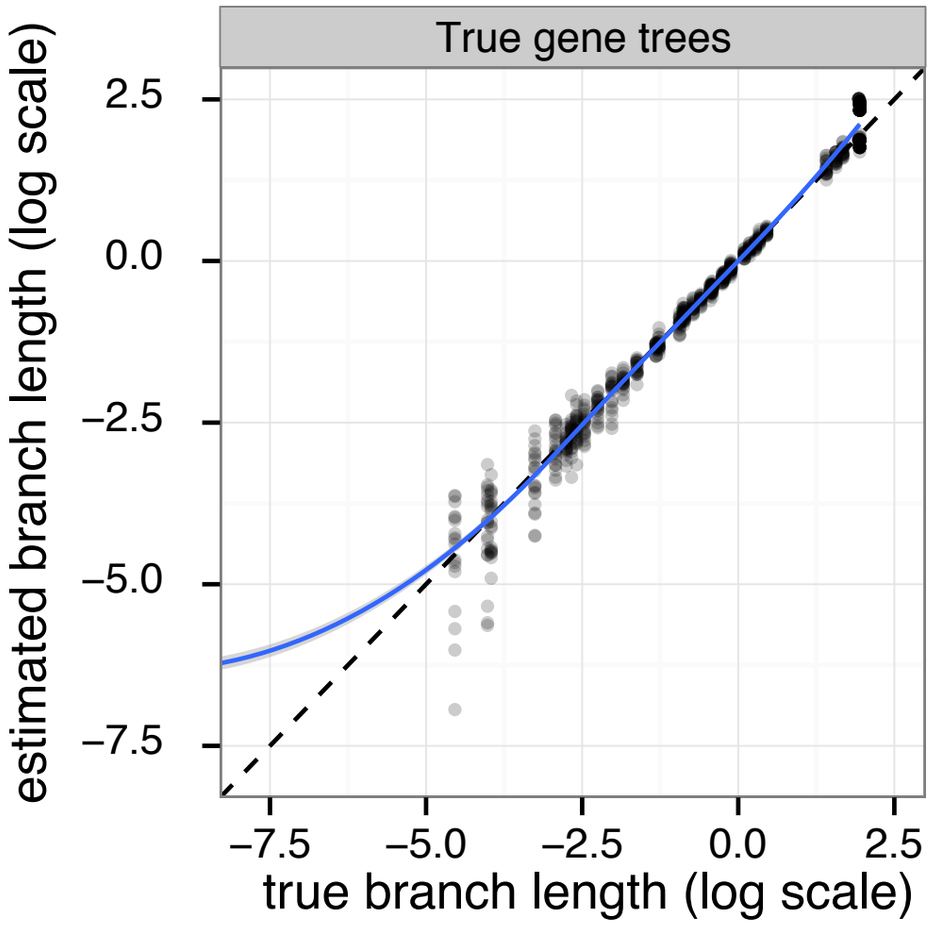
localPP is more accurate than bootstrapping



Avian simulated dataset (48 taxa, 1000 genes)

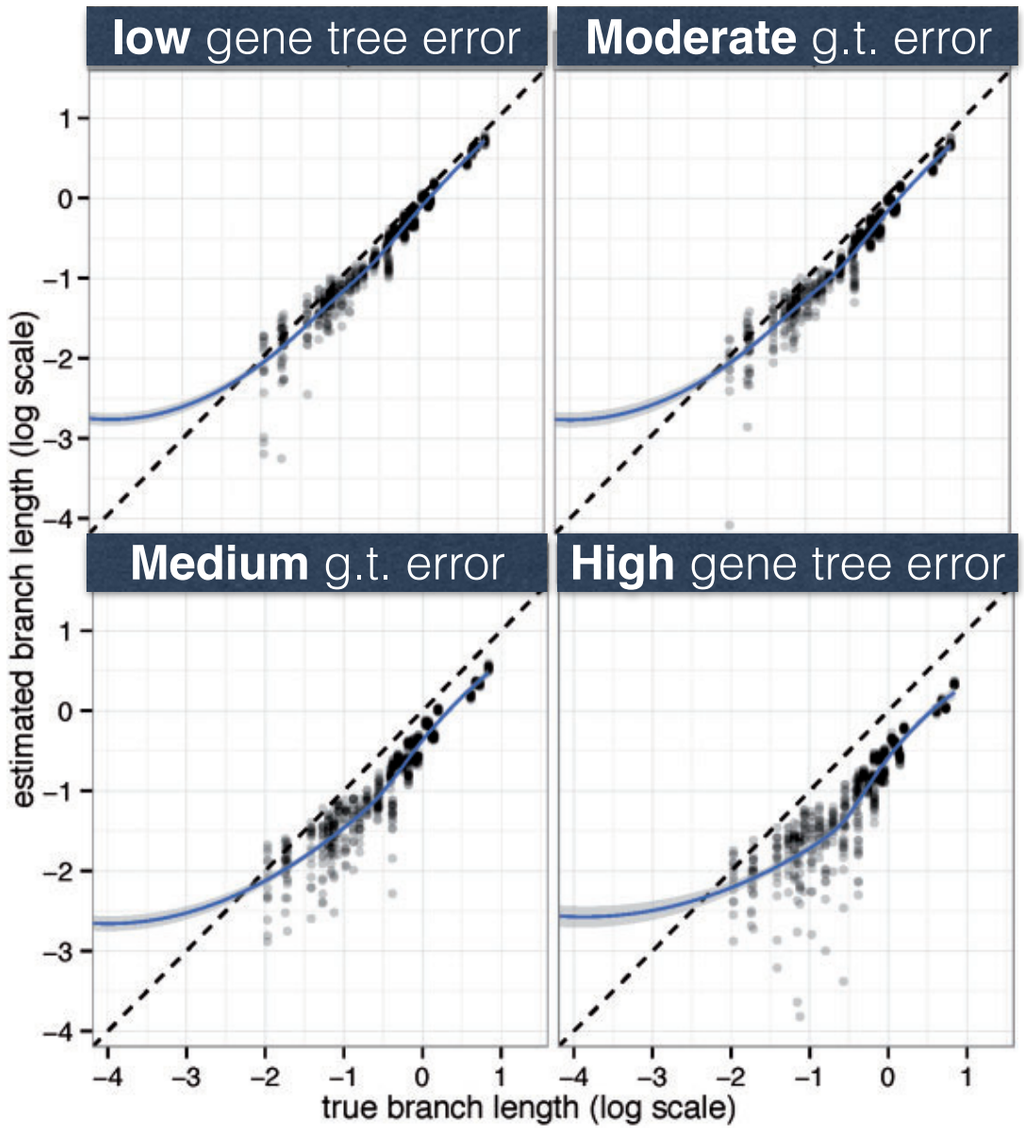
[Sayyari and Mirarab, MBE, 2016]

ASTRAL can also estimate internal branch lengths



With true gene trees, ASTRAL estimates BL with high accuracy

ASTRAL can also estimate internal branch lengths



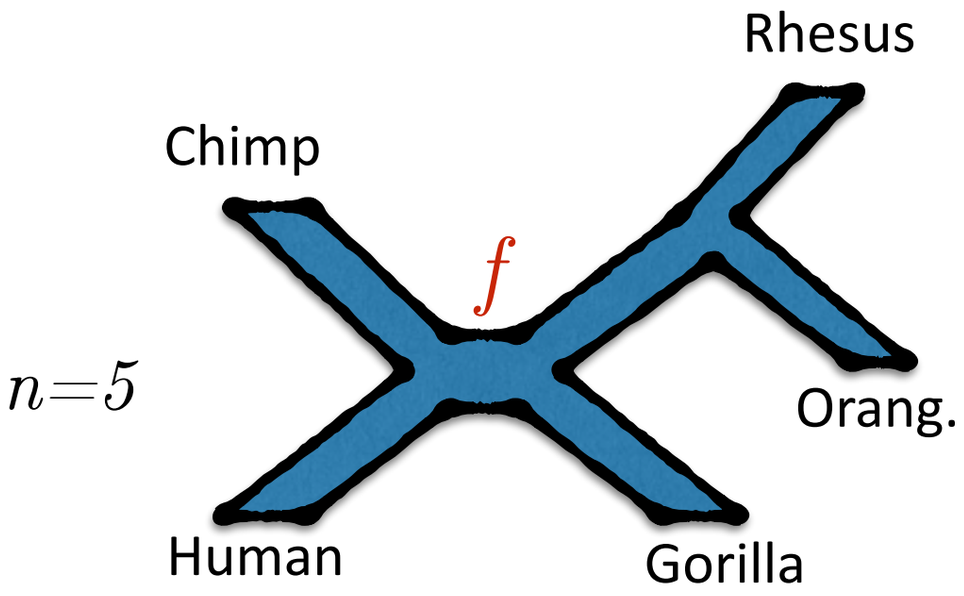
With error-prone estimated gene trees, ASTRAL underestimates BL

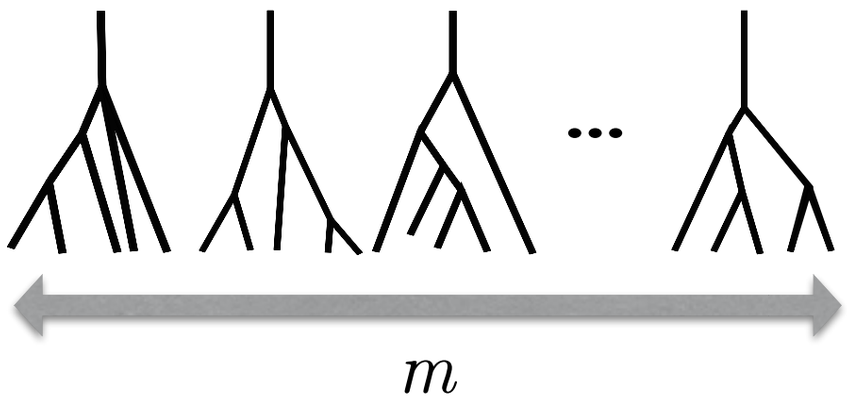
Sample complexity?

Shubhanshu Shekhar

Sebastien Roch

- **How many genes** are needed to guarantee an arbitrarily high probability of finding the true species tree?
  - f: the length of the shortest branch (difficulty)
    - Find m, as a function of f and n for probability of error ϵ





Theorem 1

Consider a model species tree with minimum branch length $f < \log(\sqrt{2})$. *Then, for any* $\epsilon > 0$, *ASTRAL (exact) returns the true species tree with probability at least* $1 - \epsilon$ *if the number of input error-free gene trees satisfies*

$$m > \frac{9}{2} \log\left(\frac{4}{\epsilon}\binom{n}{4}\right) \frac{1}{(1 - e^{-f})^2} \tag{1}$$

Theorem 2

For any $\rho \in (0, 1)$ *and* $a \in (0, 1)$, *there exist constants* f_0 *and* n_0 *such that the following holds. For all* $n \geq n_0$ *and* $f \leq f_0$, *there exists a species tree with* n *leaves and shortest branch length* f *such that when ASTRAL (exact) is used with* $m \leq \frac{a \log n}{5f^2}$ *gene trees, the event* E *that ASTRAL (exact) reconstructs the wrong tree has probability*

$$\mathbf{P}(E) \geq 1 - \rho. \tag{3}$$

Theorem 1

Consider a model species tree with minimum branch length $f < \log(\sqrt{2})$. *Then, for any* $\epsilon > 0$, *ASTRAL (exact) returns the true species tree with probability at least* $1 - \epsilon$ *if the number of input error-free gene trees satisfies*

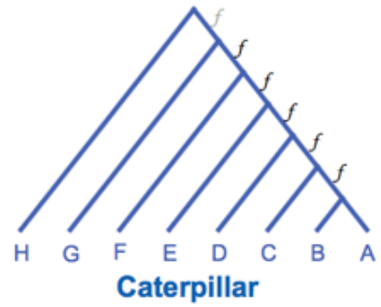
The sample complexity of the ASTRAL optimization problem (exact solution) is

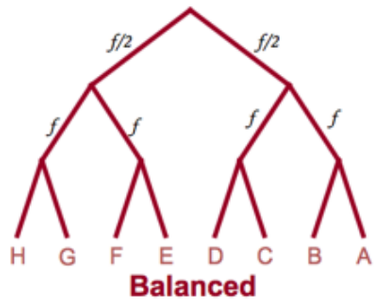
$$O(\log(n) f^{-2})$$

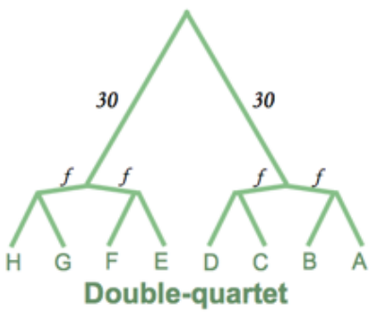
such that when ASTRAL (exact) is used with $m \leq \frac{a \log n}{5f^2}$ *gene trees, the event* E *that ASTRAL (exact) reconstructs the wrong tree has probability*

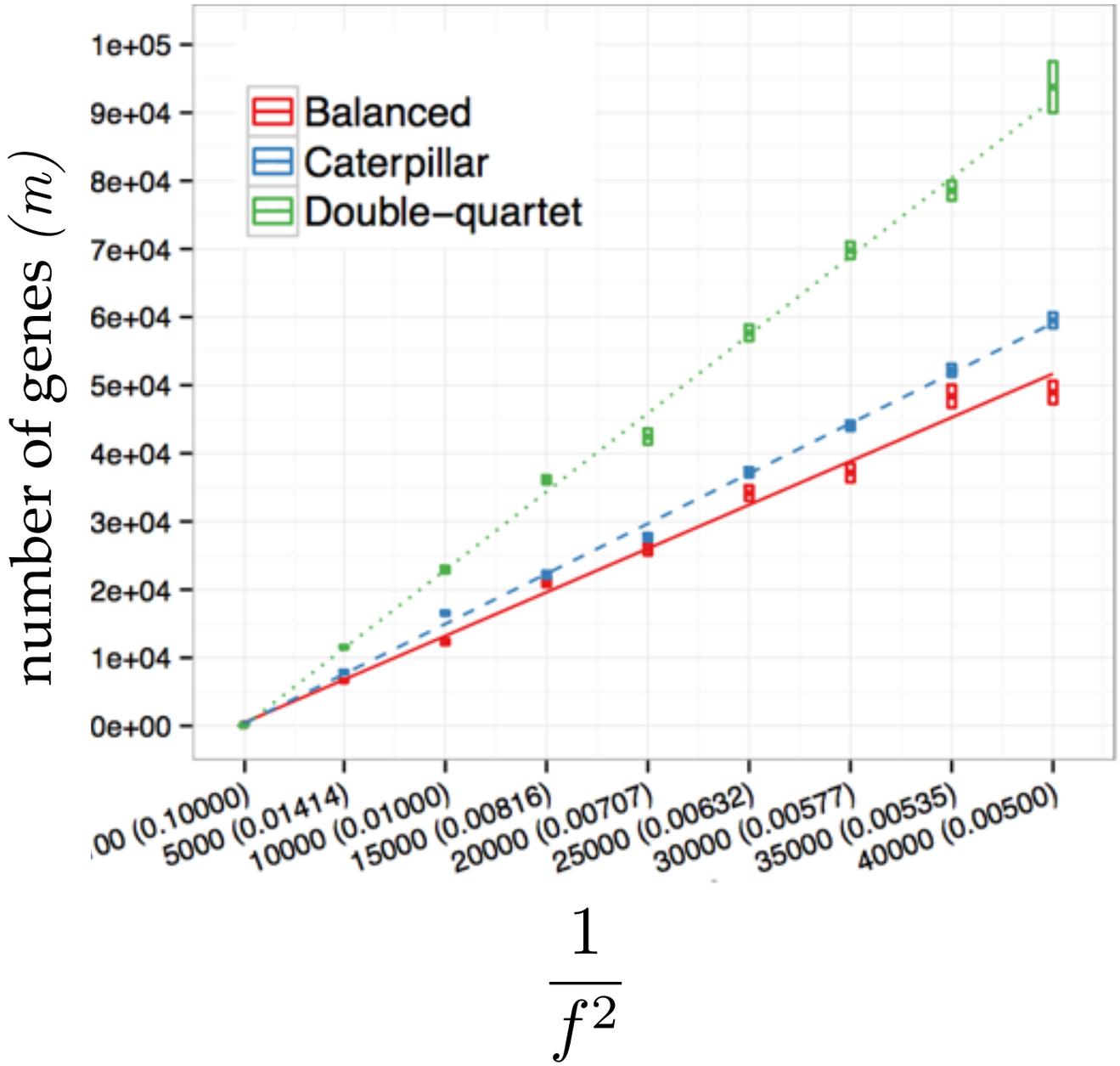
$$\mathbf{P}(E) \geq 1 - \rho. \tag{3}$$

Simulations match theory

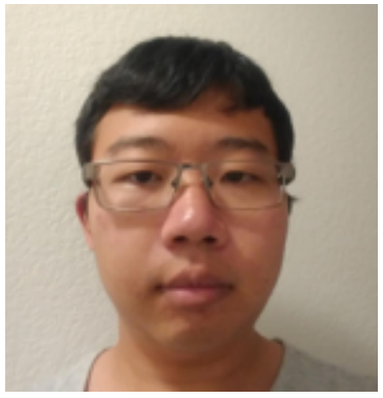








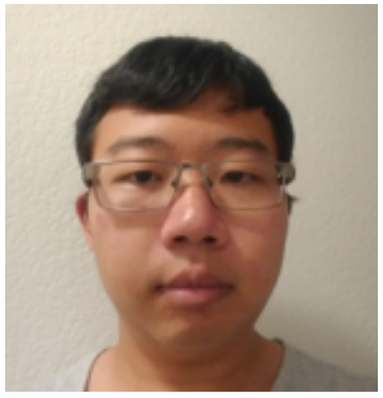
ASTRAL-III



Chao Zhang

- Improved running time (to be released in July)
  - Can handle polytomies without slowing down
  - Can exploit similarities between gene trees to reduce running time

ASTRAL-III

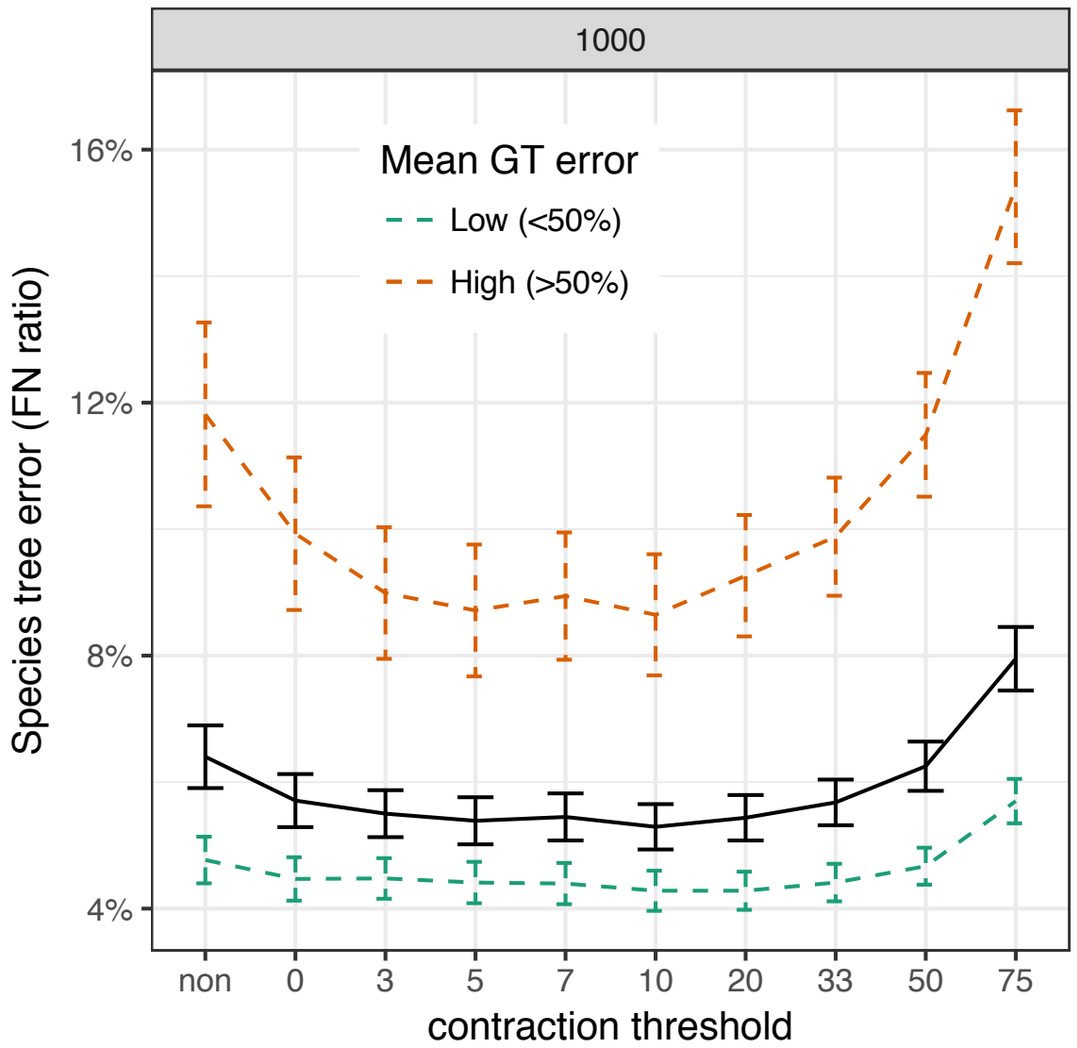


Chao Zhang

- Improved running time (to be released in July)
  - Can handle polytomies without slowing down
  - Can exploit similarities between gene trees to reduce running time
- To be released in August
  - Handling datasets with multiple individuals per species
  - GPU and CPU parallelism

Low support branches

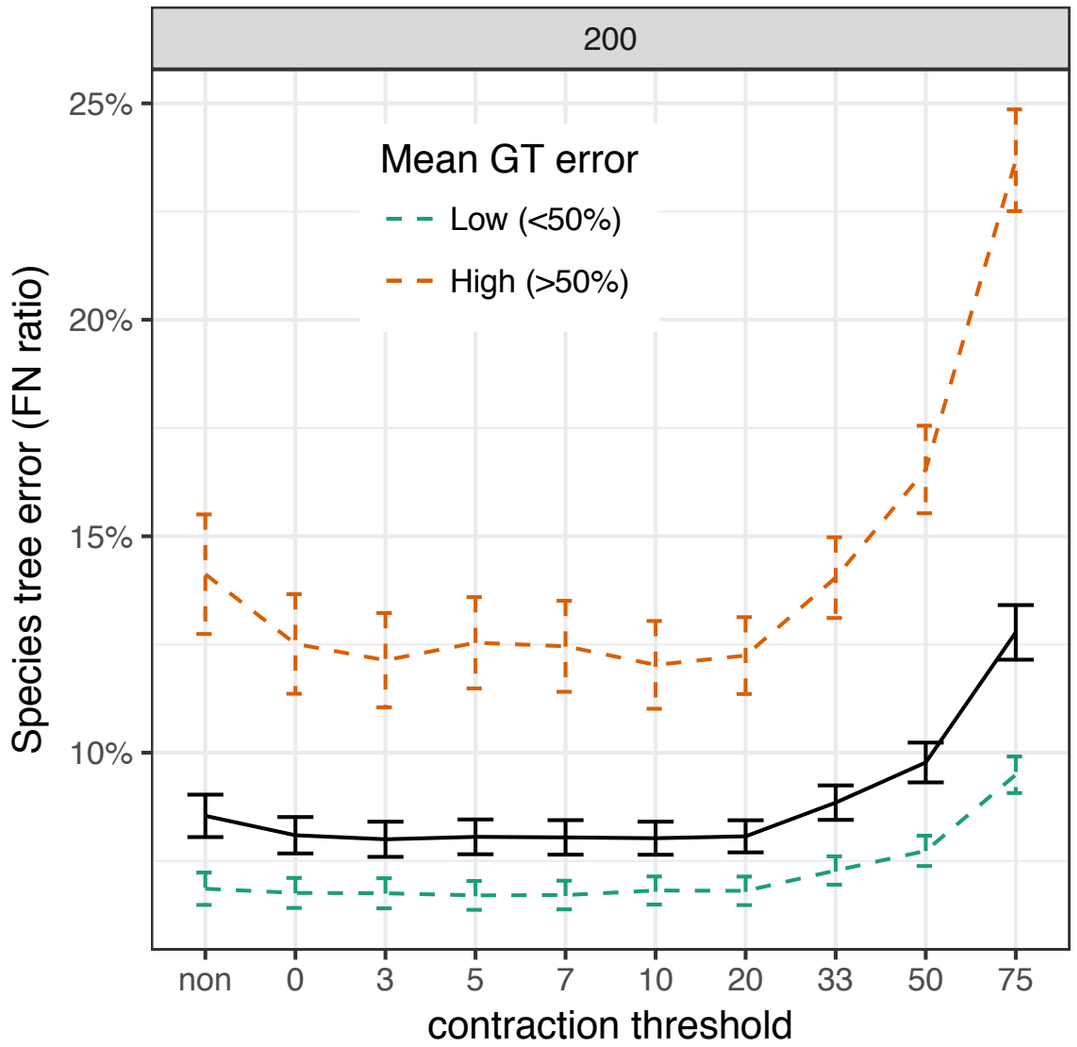
- Does it help to contract branches with low support?
- Yes, but only for **very low** supports
- Theoretical justifications not clear



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

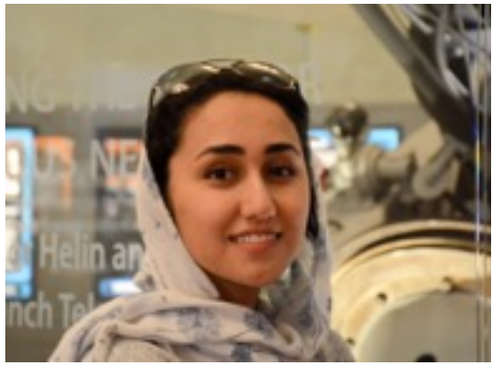
Low support branches

- Does it help to contract branches with low support?
- Yes, but only for **very low** supports
- Theoretical justifications not clear



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

Multiple individuals

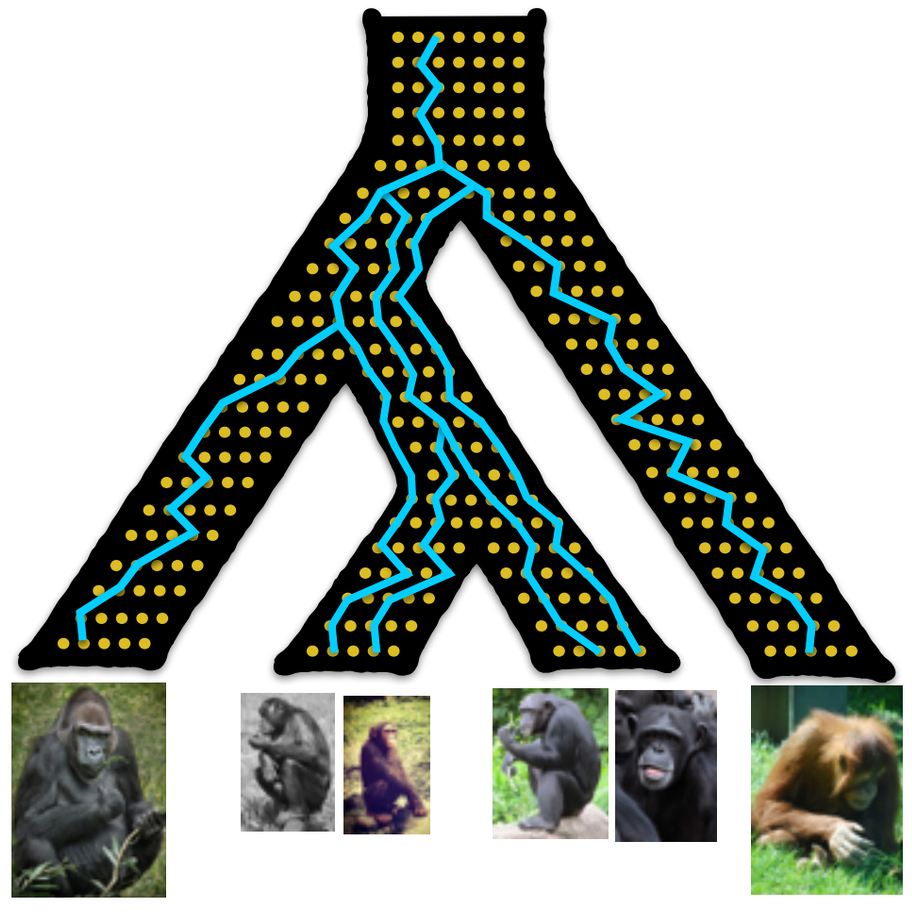


Maryam Rabiee Hashemi

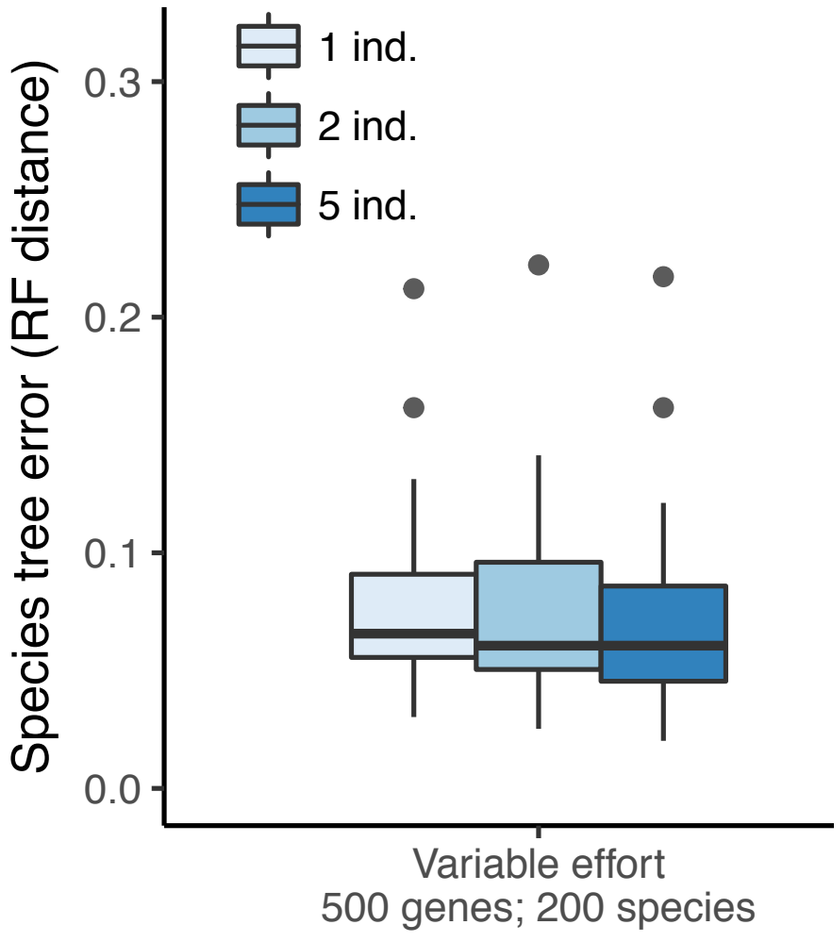
- What if we sample multiple individuals from each species?

- In recently diverged species individuals *may* have different trees for each gene

- Sampling multiple individuals may provide extra signal

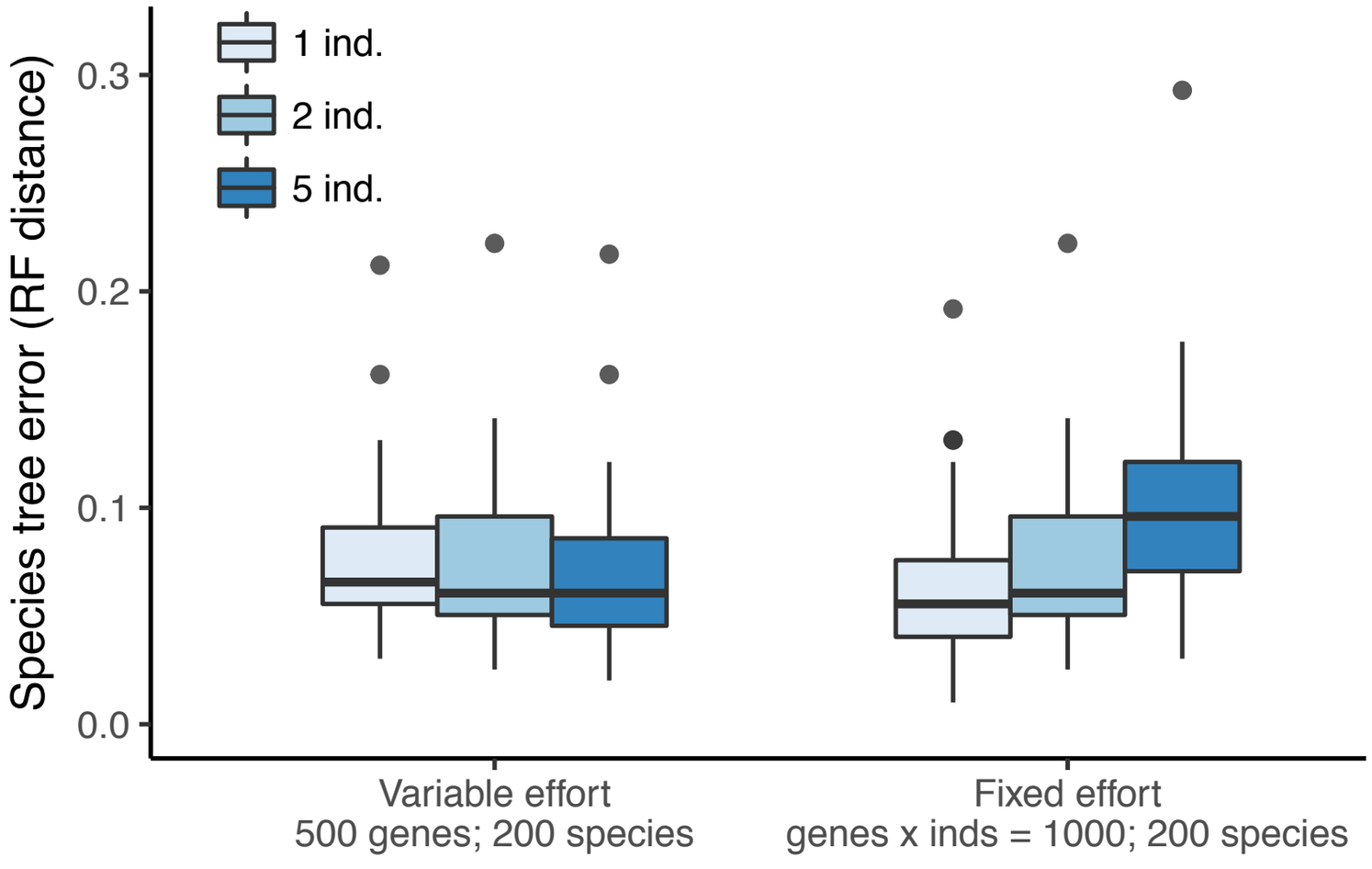


Multiple individuals helpful?



Yes, it marginally helps accuracy

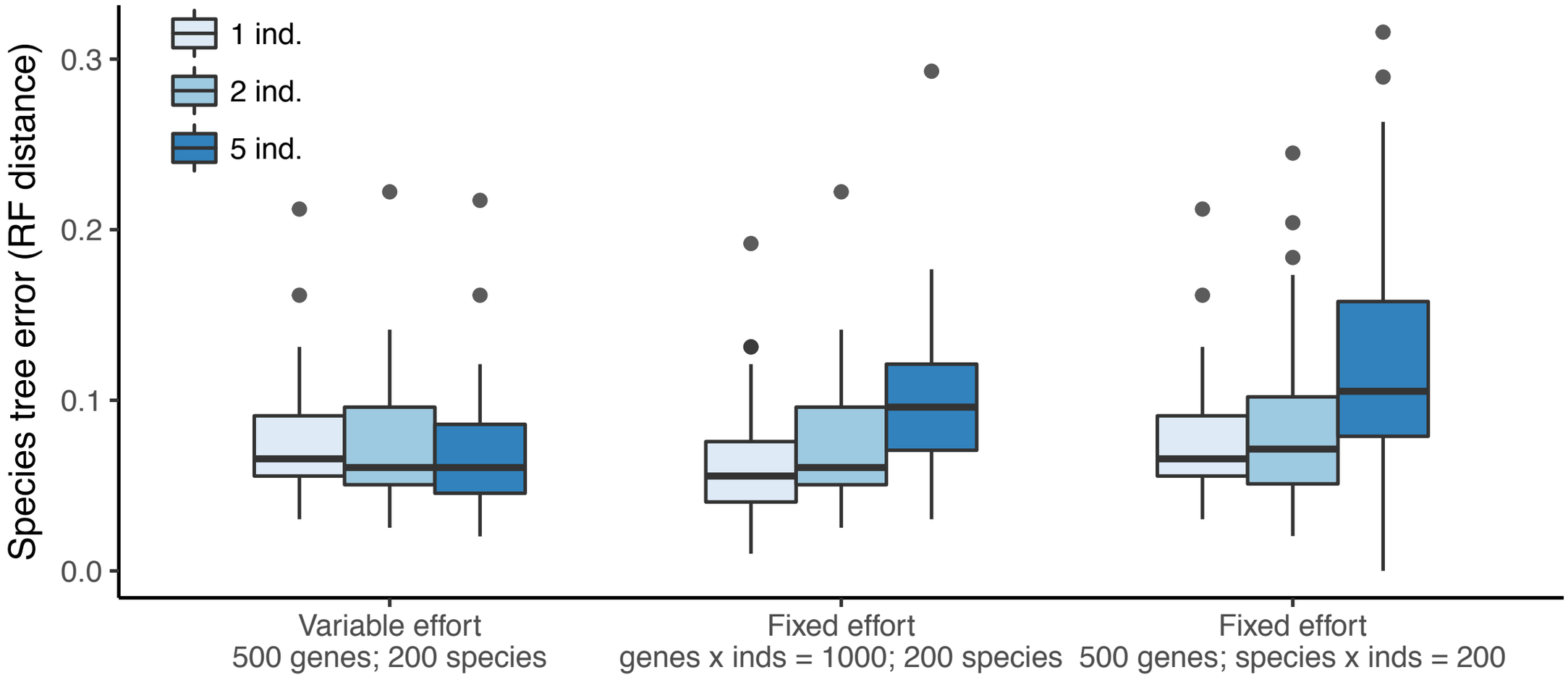
Multiple individuals helpful?



Yes, it marginally helps accuracy

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

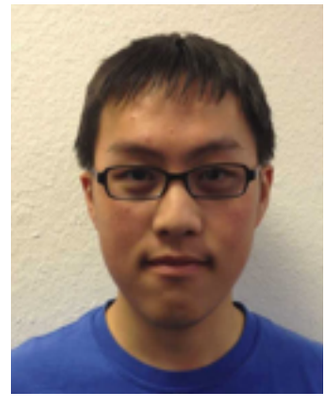
Multiple individuals helpful?



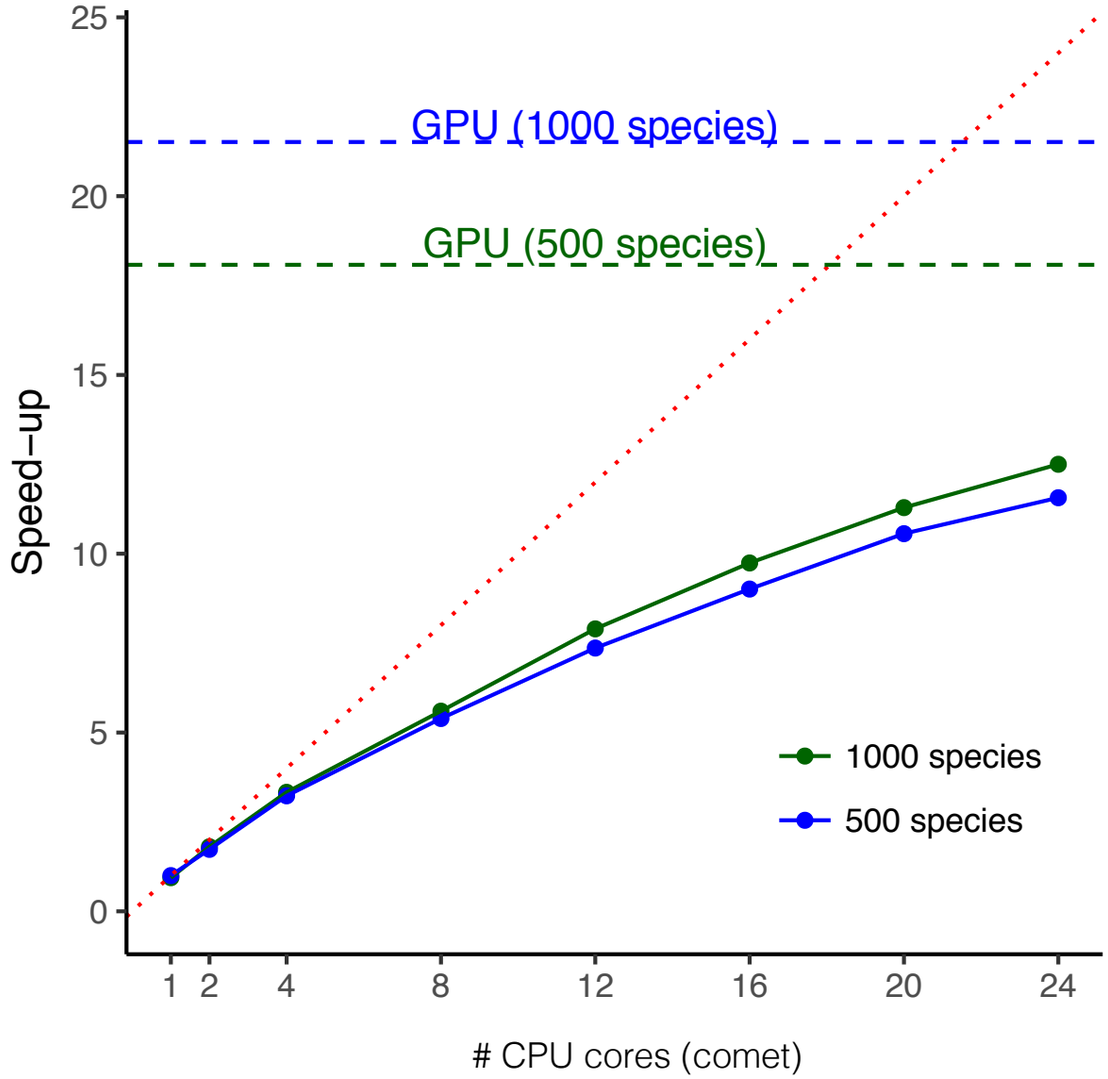
Yes, it marginally helps accuracy

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

Parallelism



John Yin

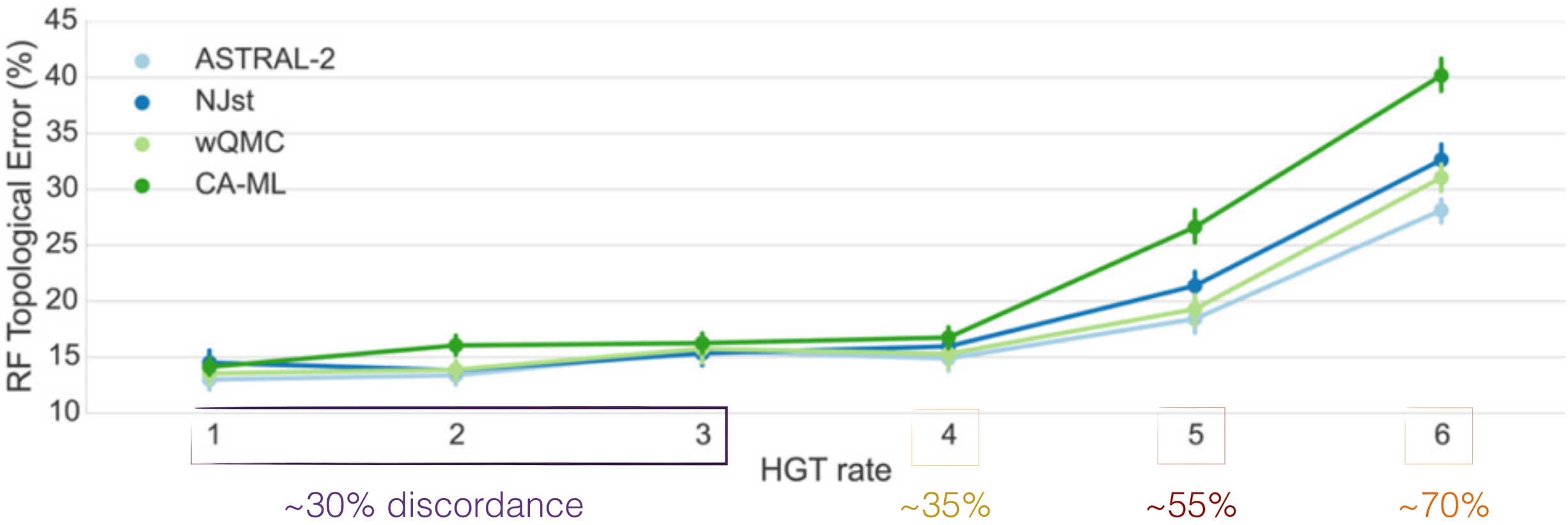


Can infer trees with 10,000 species & 400 genes in a day

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

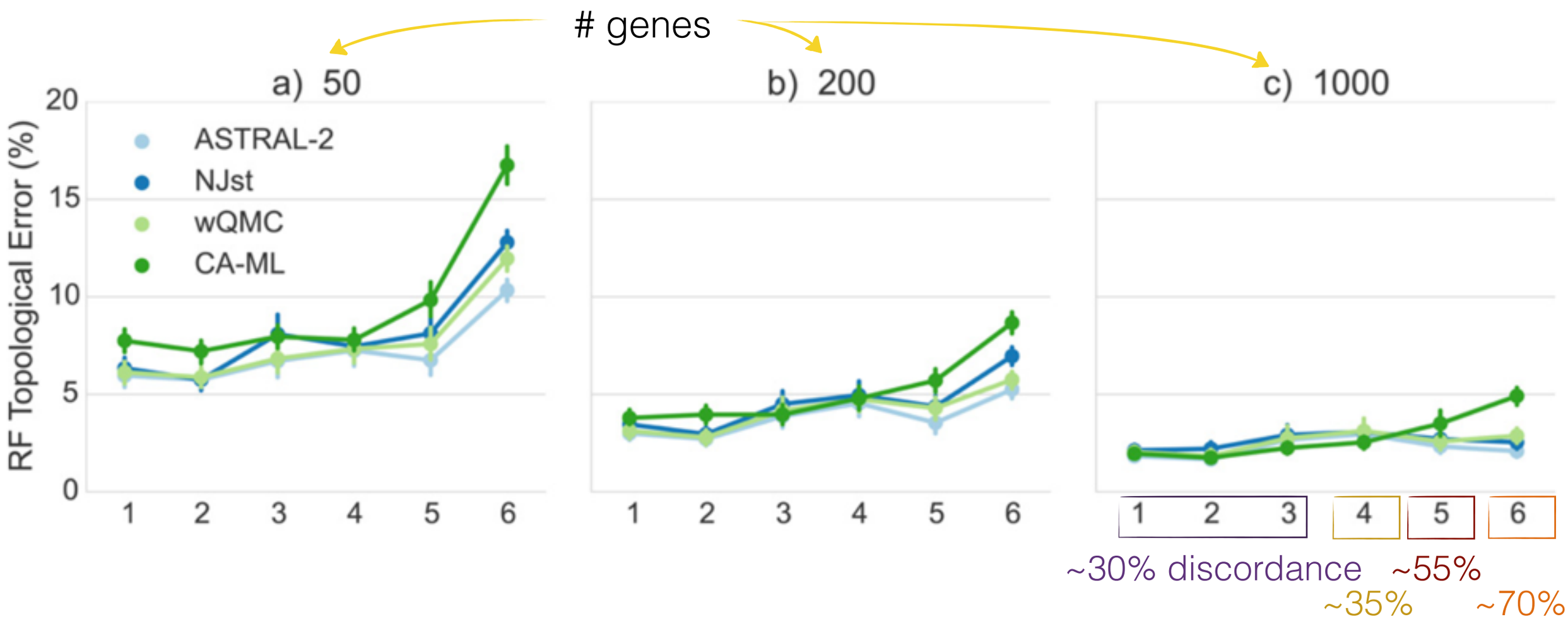


50 species, 10 genes

Horizontal Gene Transfer (HGT)

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

Randomly distributed HGT is tolerated with enough genes



50 species, varying # genes

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

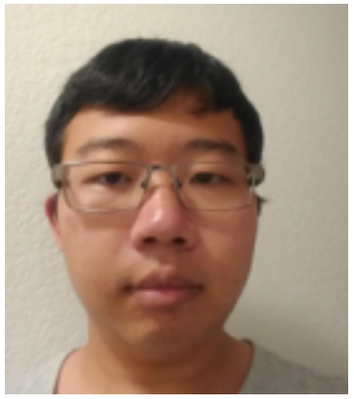
Tandy Warnow

S.M. Bayzid

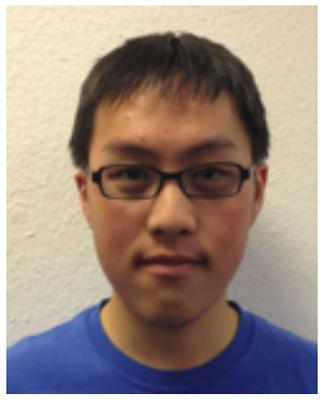
Théo Zimmermann

UC San Diego

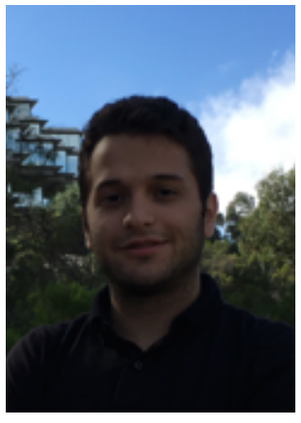
Maryam Rabiee Hashemi



Chao Zhang



John Yin



Erfan Sayyari

Shubhanshu Shekhar

NSF

HHMI
HOWARD HUGHES MEDICAL INSTITUTE