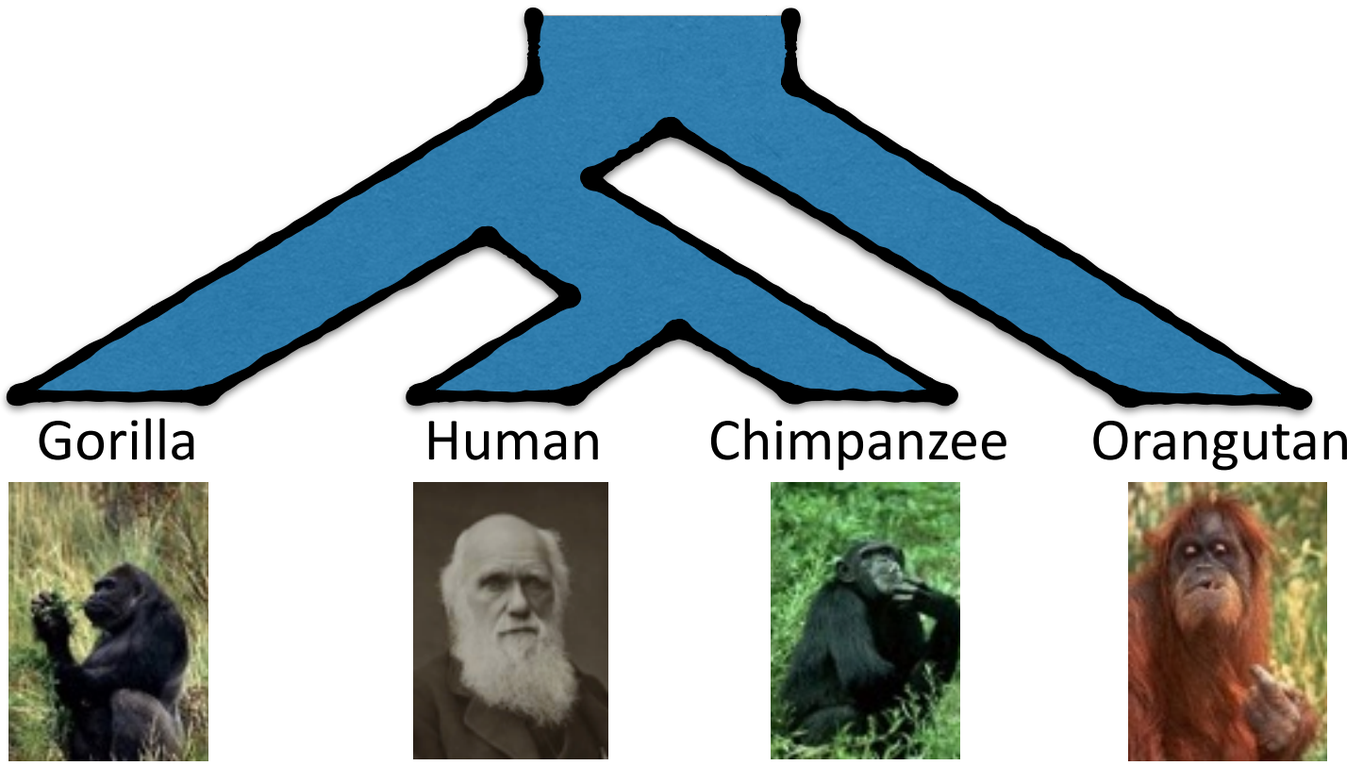
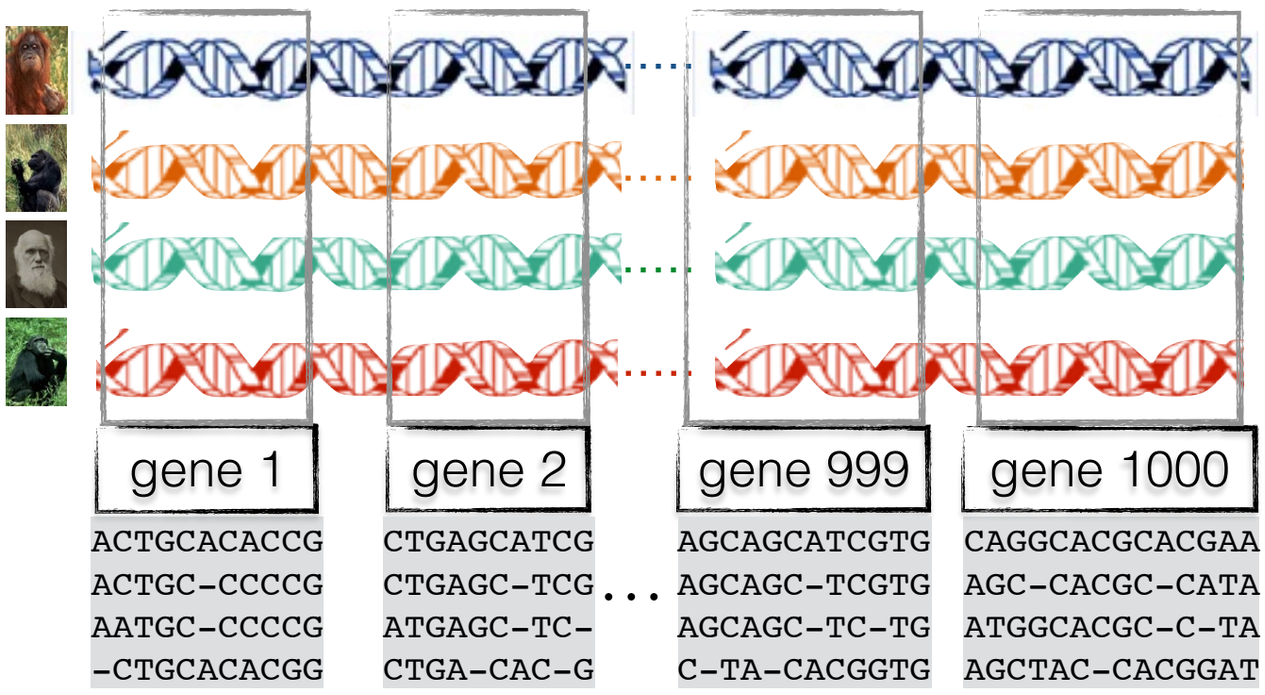
Fast coalescent-based branch support using local quartet frequencies

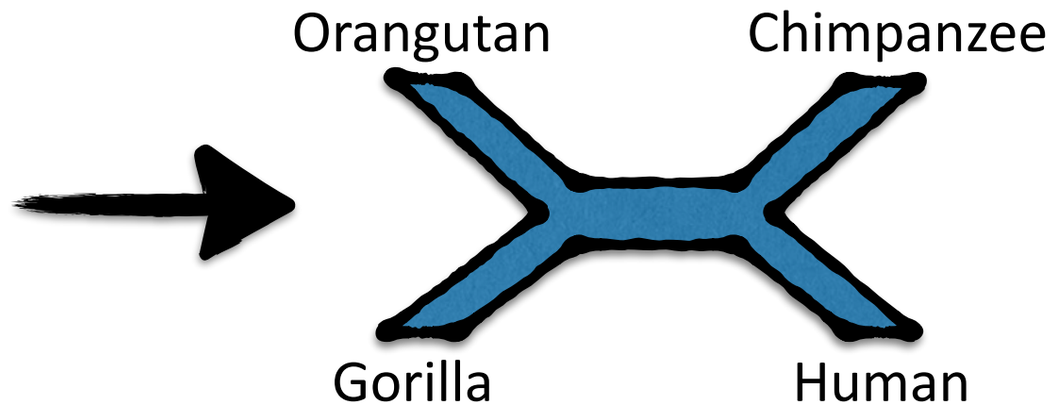
Molecular Biology and Evolution (2016)
33 (7): 1654–68

Erfan Sayyari, Siavash Mirarab
University of California, San Diego (ECE)



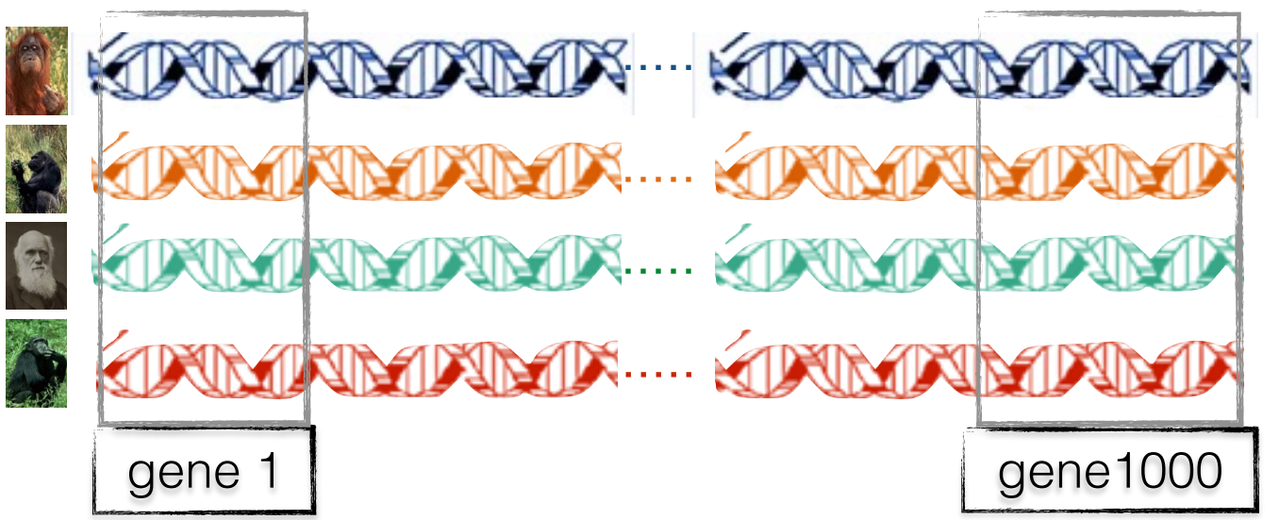
Phylogenomics



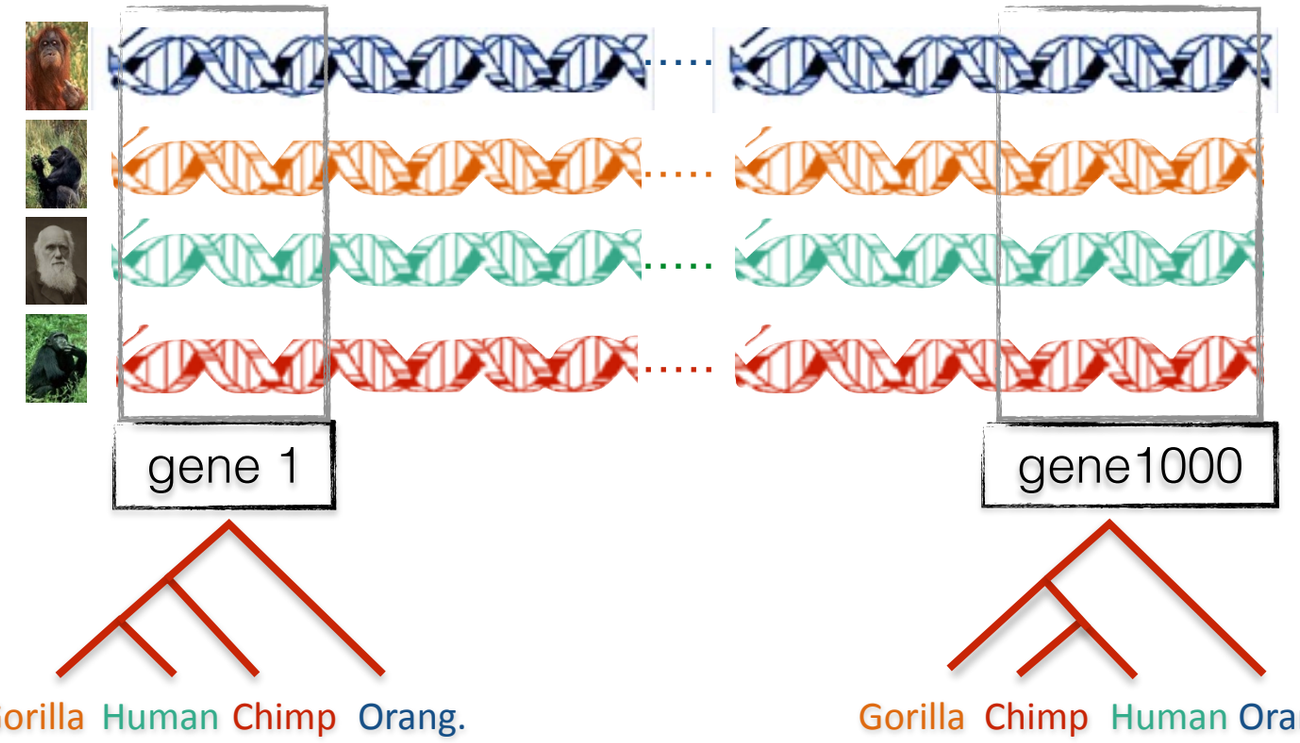


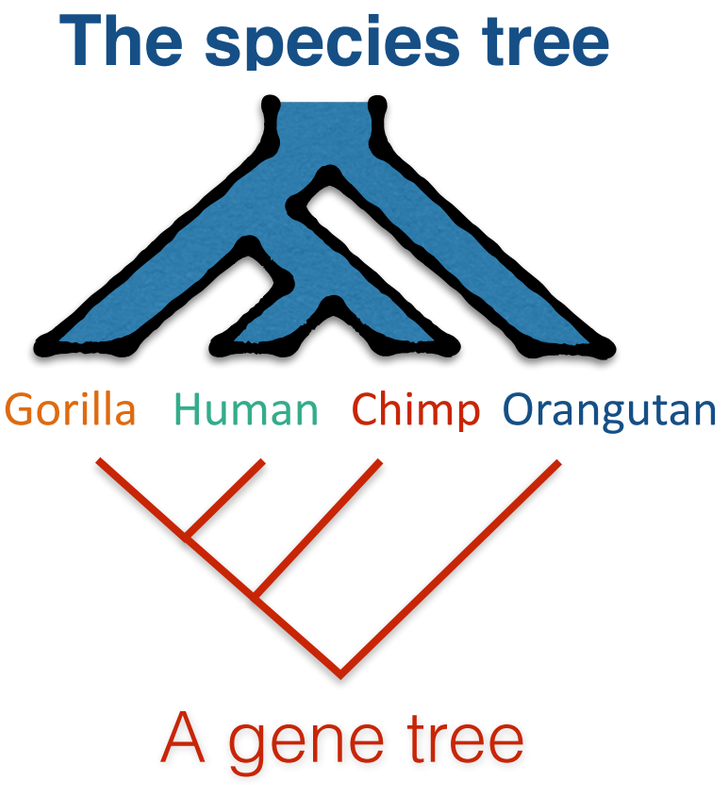
"gene" here refers to a portion of the genome
(not a functional gene)

Gene tree discordance

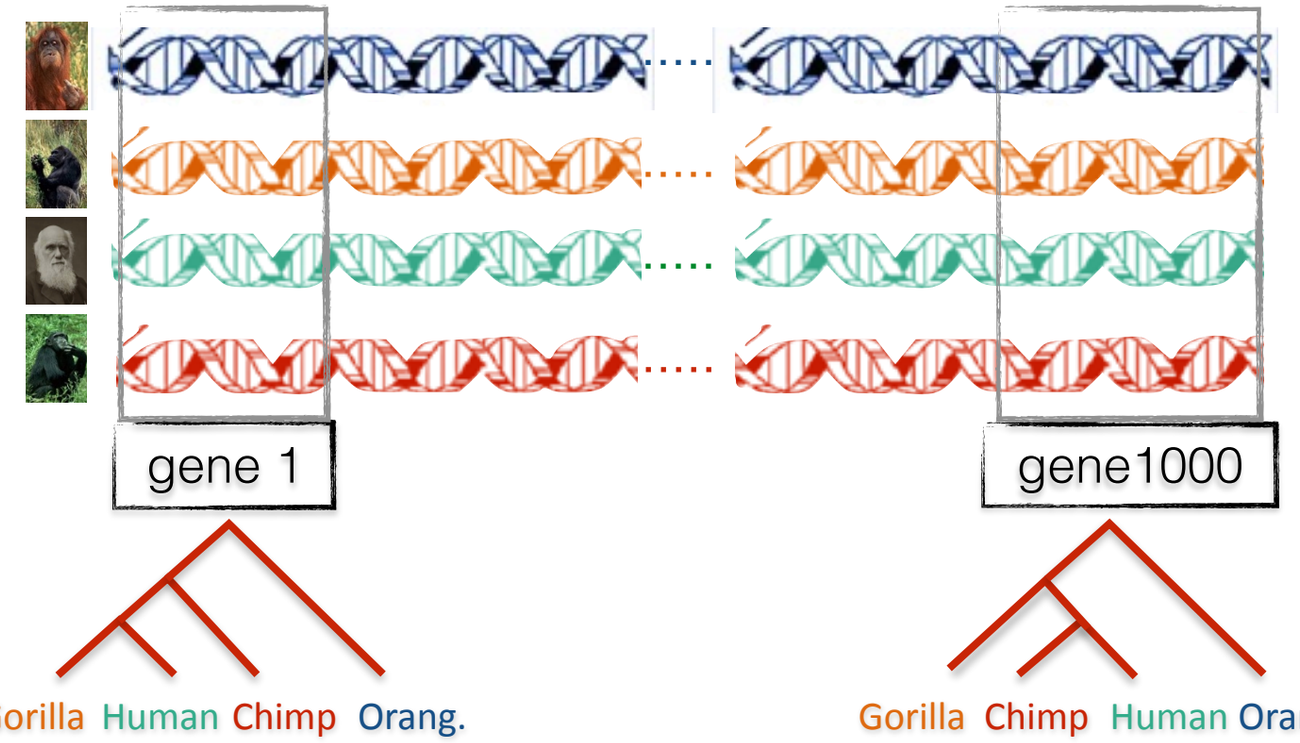


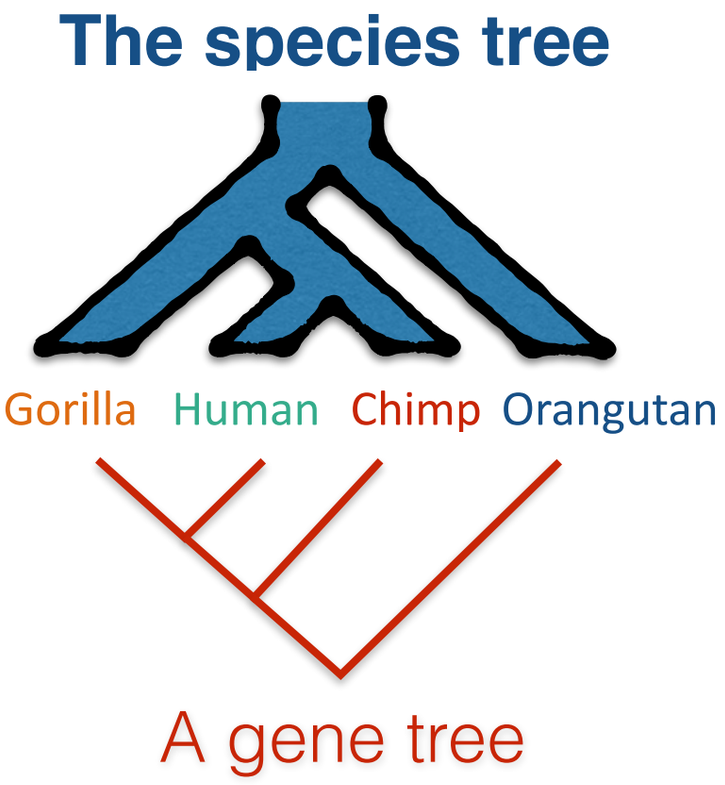
Gene tree discordance





Gene tree discordance



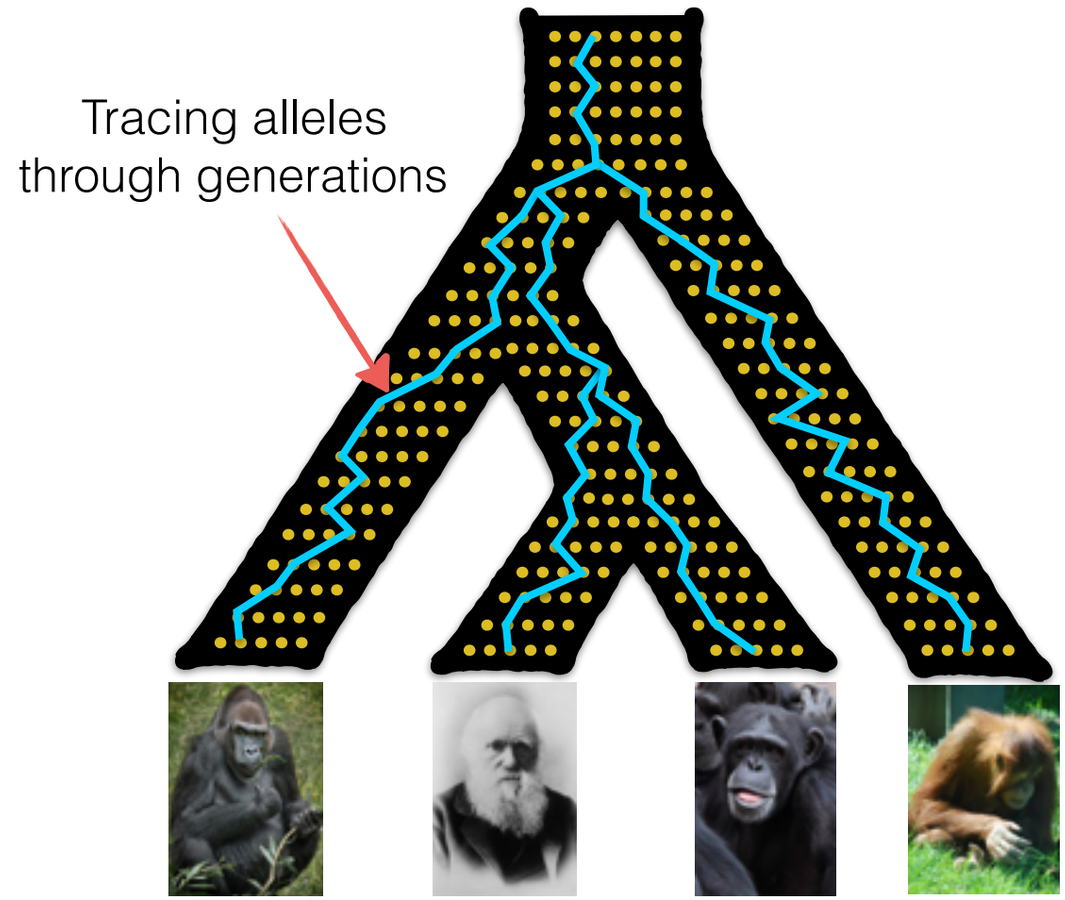


Causes of gene tree discordance include:

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

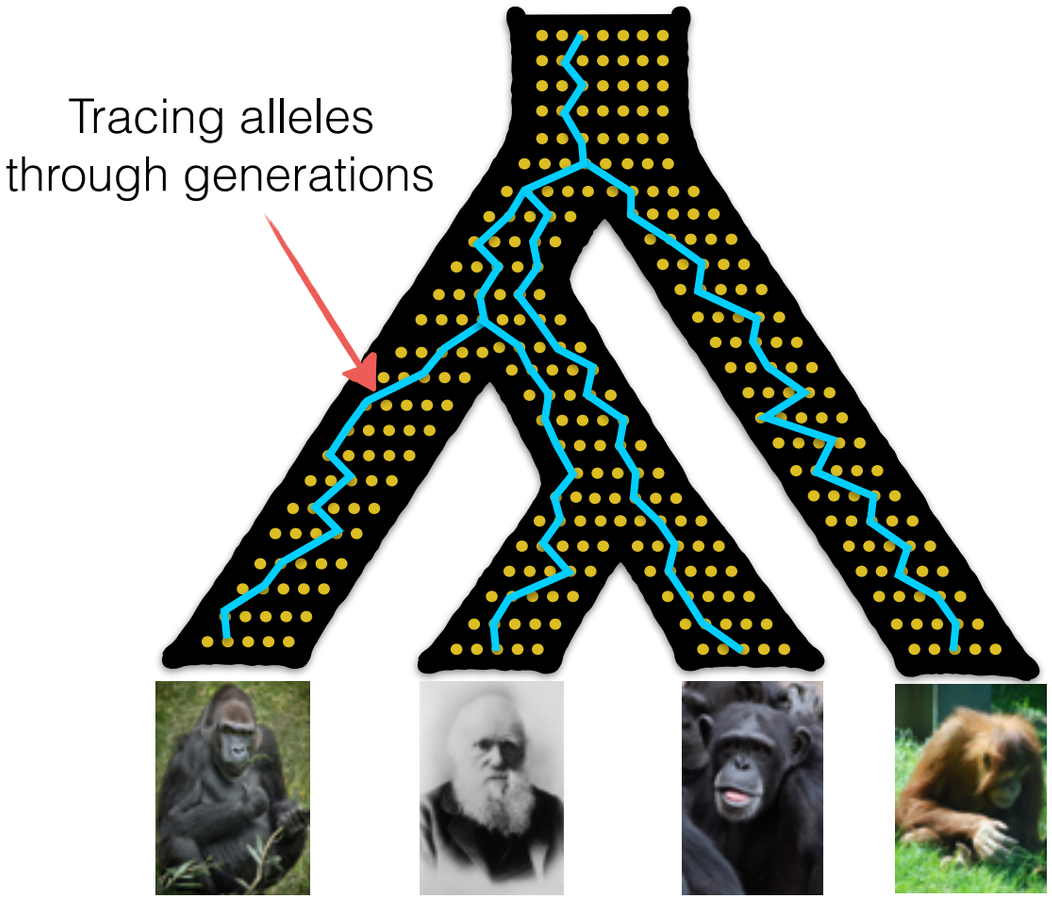
Incomplete Lineage Sorting (ILS)

- 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

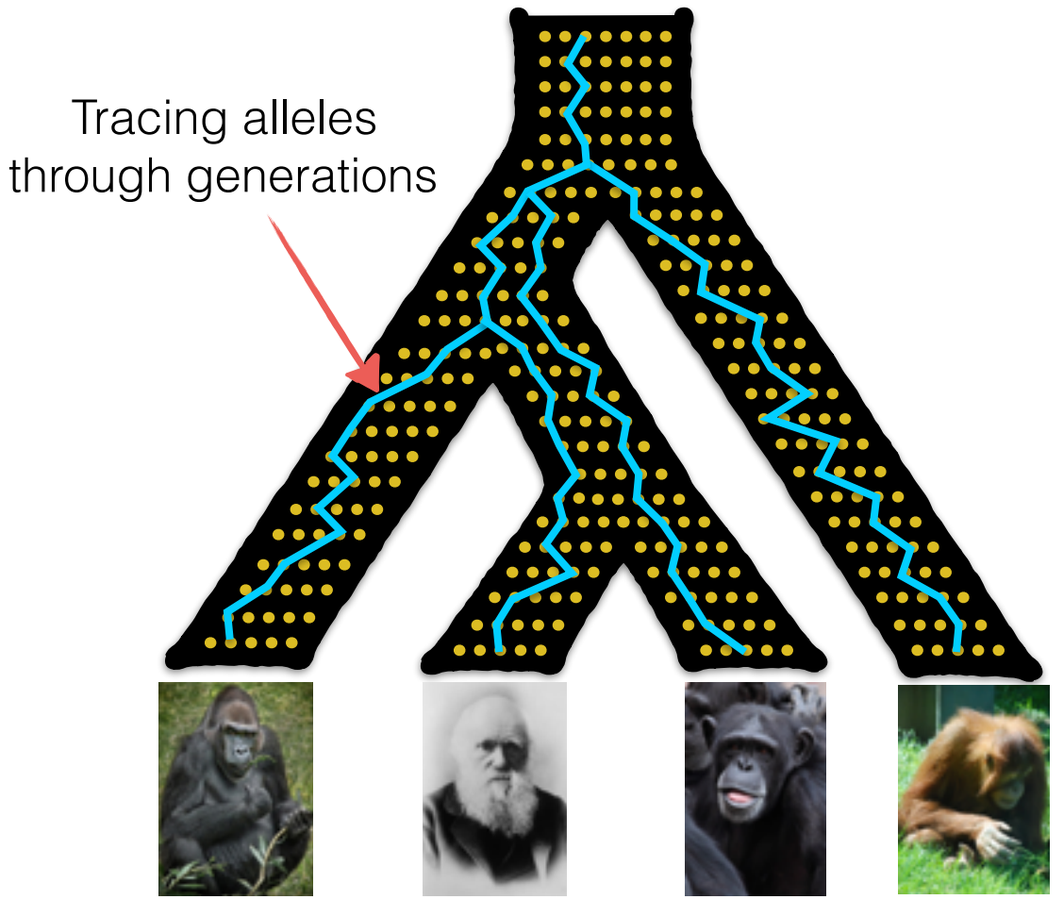


Incomplete Lineage Sorting (ILS)

- 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.

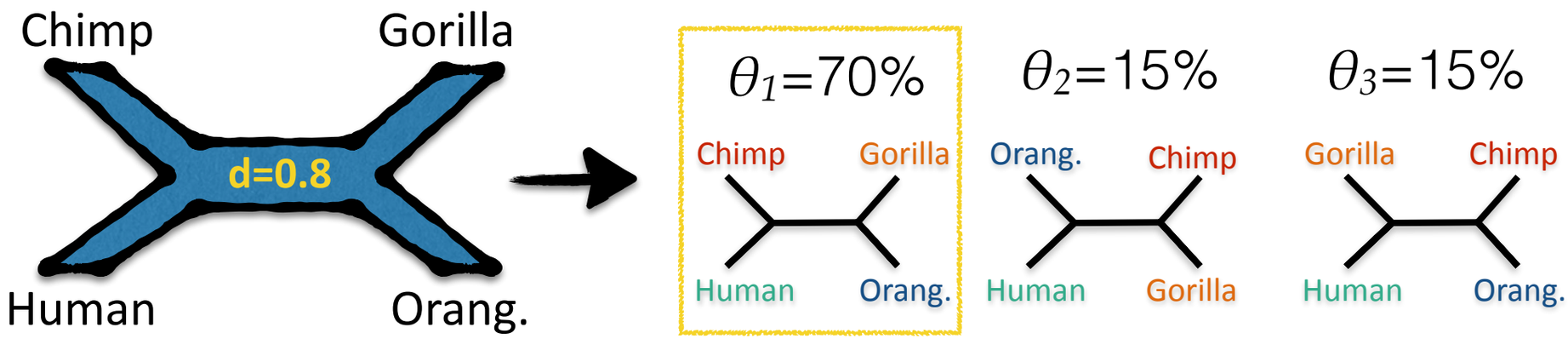
- Any species tree defines a unique distribution on the set of all possible gene trees

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

  - Likelihood calculation is not feasible.

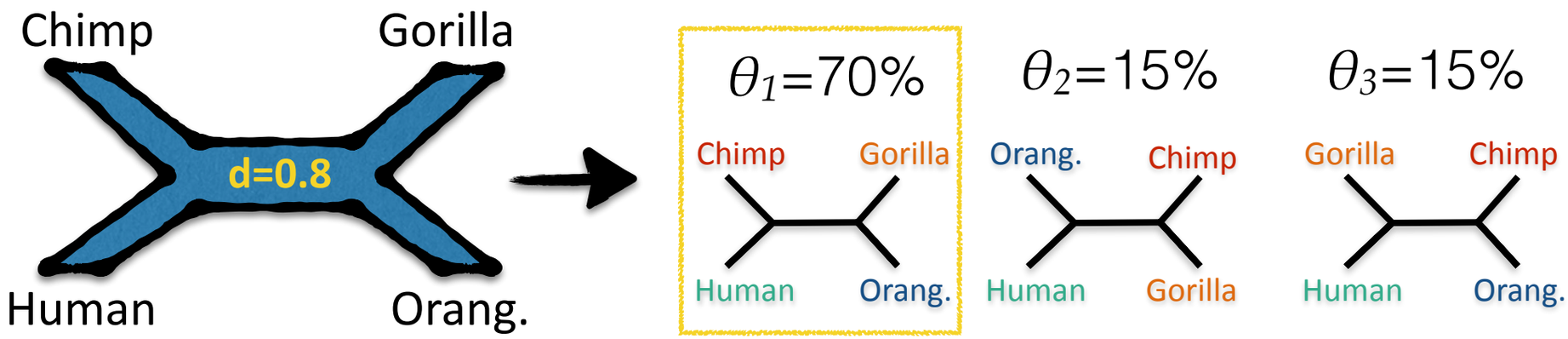
Unrooted quartets under MSC model

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

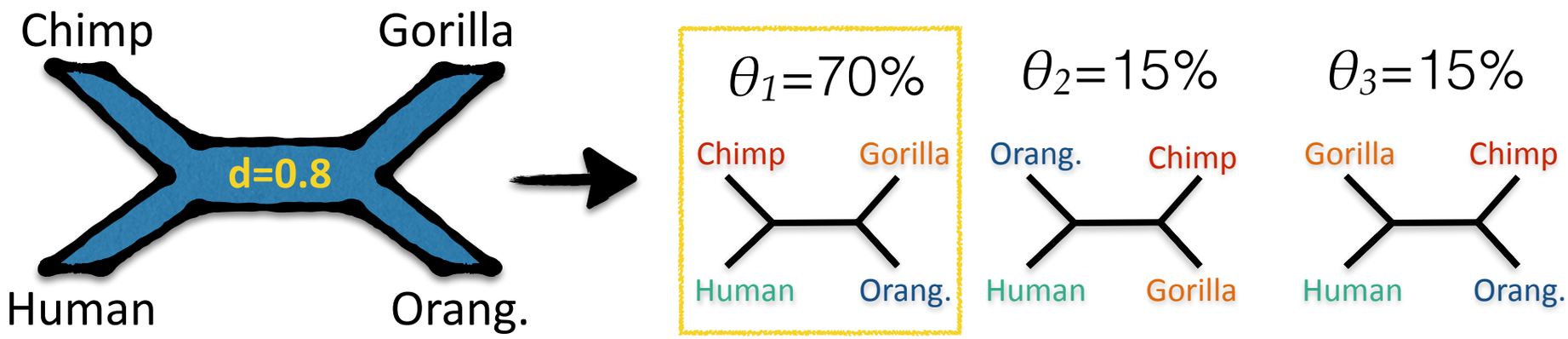
For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)



The most frequent gene tree
=
The most likely species tree

Unrooted quartets under MSC model

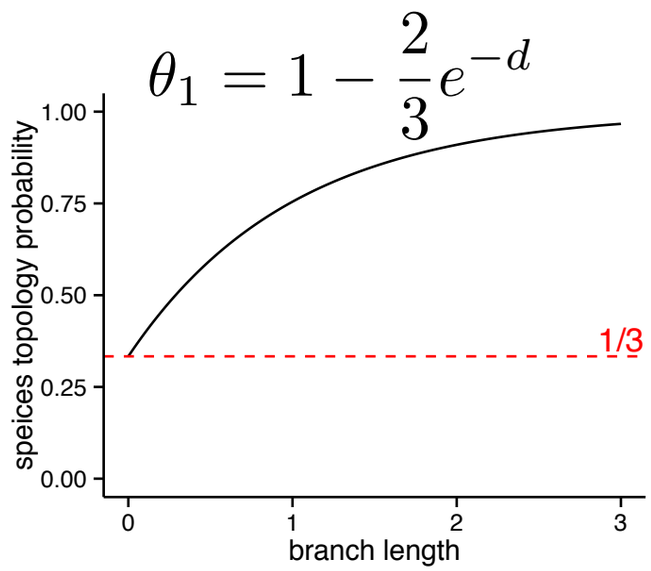
For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)



The most frequent gene tree

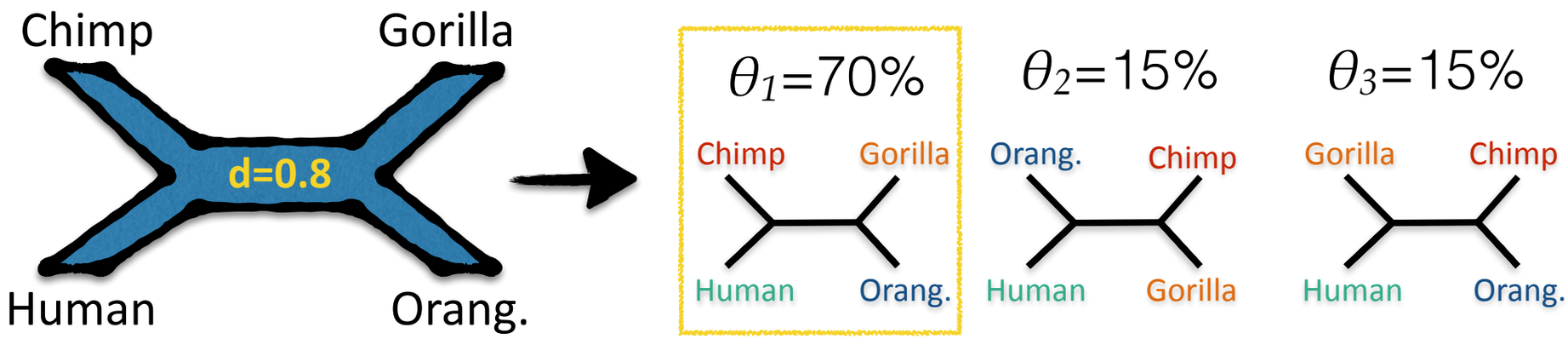
=

The most likely species tree

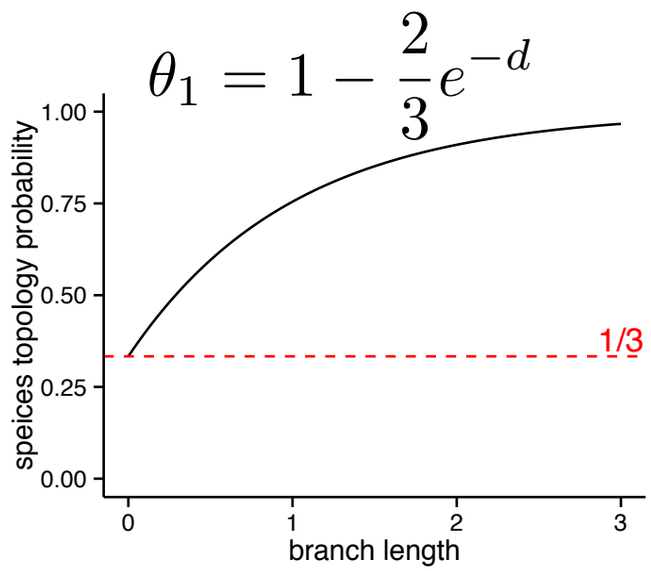


Unrooted quartets under MSC model

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



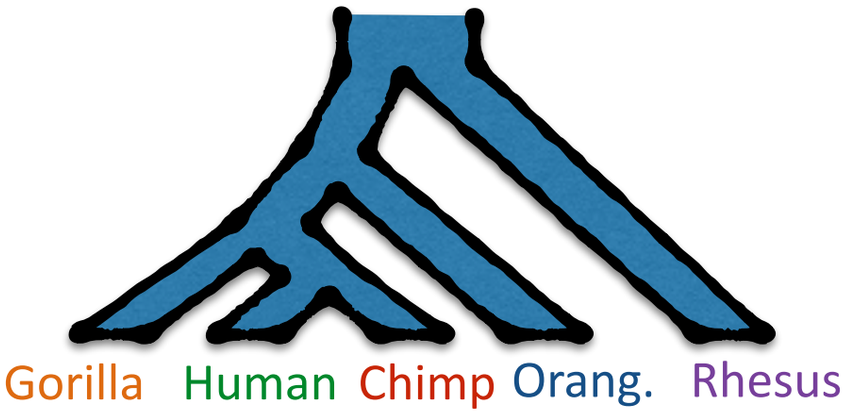
shorter branches ⇒

more discordance ⇒

a harder species tree reconstruction problem

Species tree inference for >4 species

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

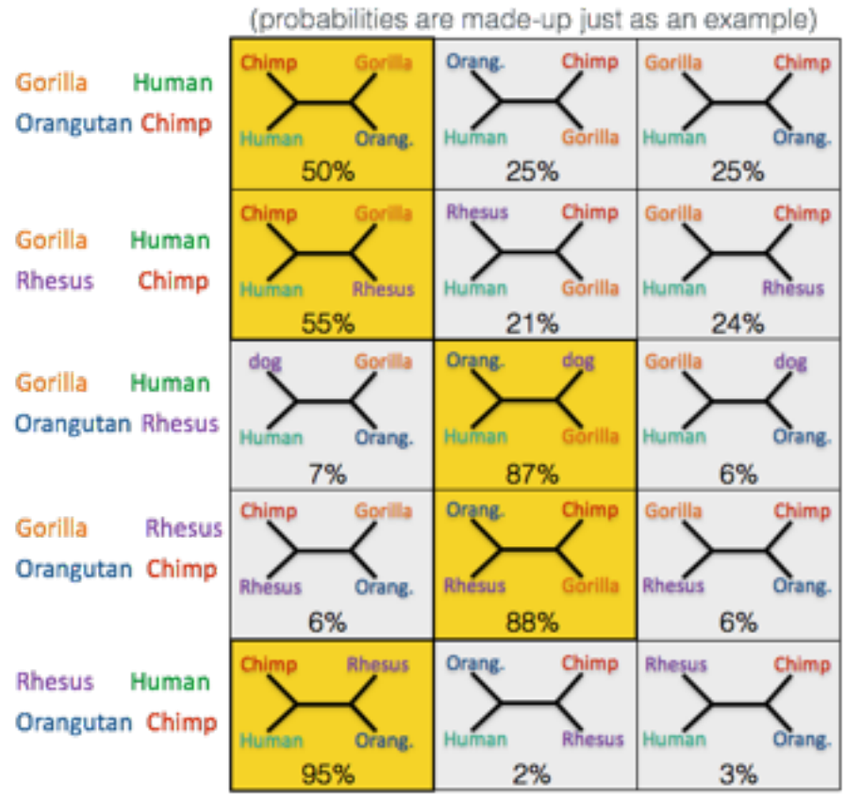


Species tree inference for >4 species

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])



Species tree inference for >4 species

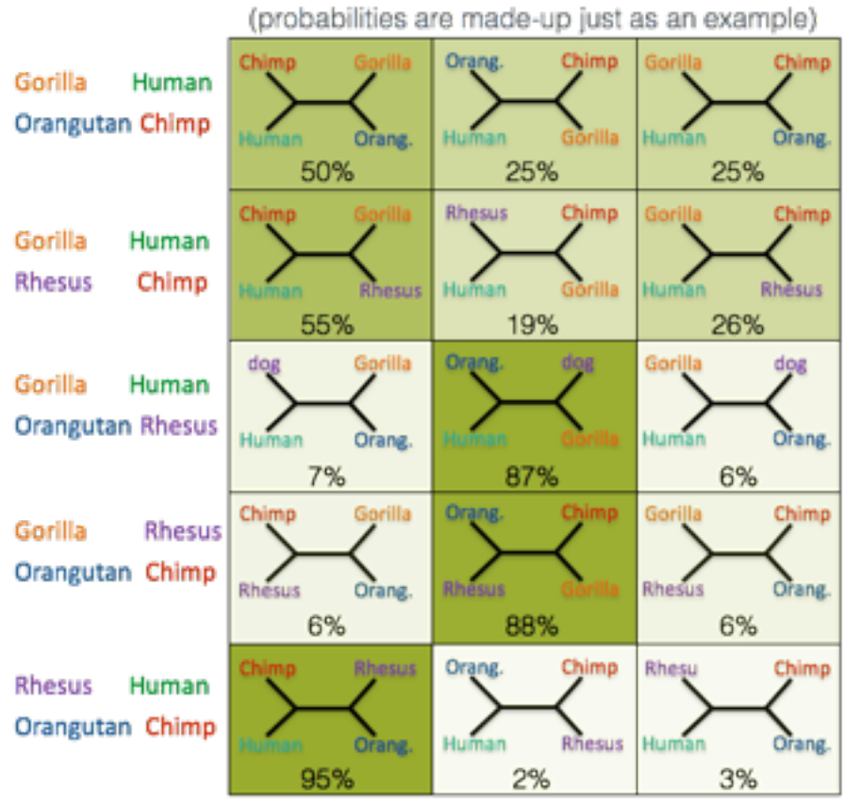
For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)

ASTRAL:

weight all $3\binom{n}{4}$ quartet topologies by their frequency in gene trees

&

find the optimal species tree using dynamic programming

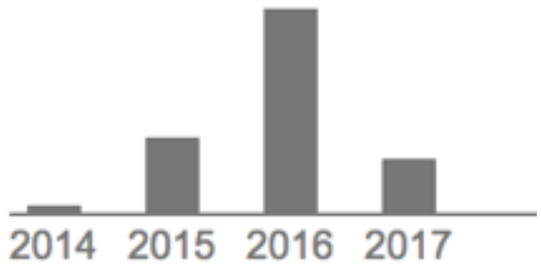


ASTRAL used by biologists

- 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

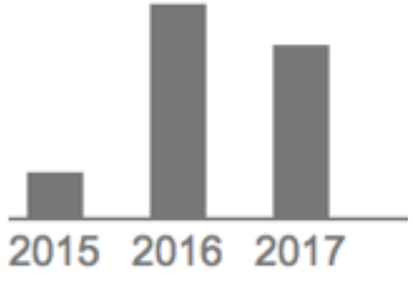
ASTRALI:
[Mirarab et al., 2014, Bioinformatics]

Cited by 170



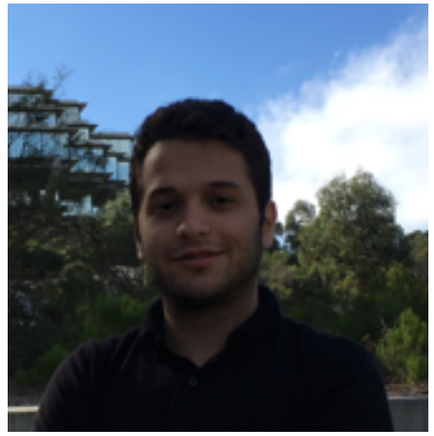
ASTRAL-II:
[Mirarab and Warnow, 2015, Bioinformatic]

Cited by 98



Going beyond the topology

[Sayyari and Mirarab, Molecular Biology & Evolution, 2016]

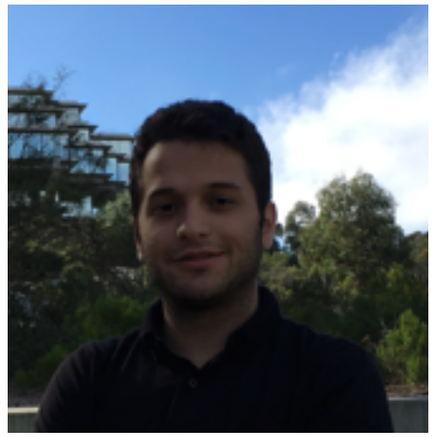


Erfan Sayyari

- **Branch length (BL):**
  - ASTRAL did not estimate branch length
  - We added branch length estimation
    - in coalescent units (#generations/population size)
    - only for internal branches

Going beyond the topology

[Sayyari and Mirarab, Molecular Biology & Evolution, 2016]



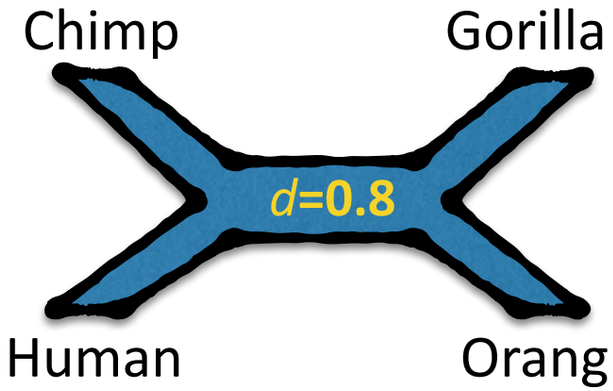
Erfan Sayyari

- **Branch length (BL):**
  - ASTRAL did not estimate branch length
  - We added branch length estimation
    - in coalescent units (#generations/population size)
    - only for internal branches
- **Branch support:** how reliable is a branch?
  - ASTRAL relied on bootstrapping
  - We added a “native” Bayesian support

Branch Length

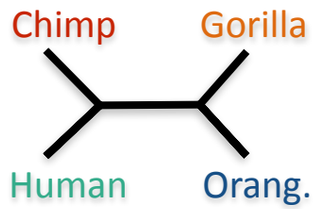
[Sayyari and Mirarab, MBE, 2016]

- Simply a function of the level of discordance

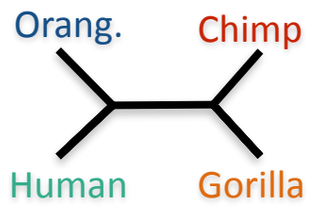


$$\theta_1 = 1 - \frac{2}{3}e^{-d}$$

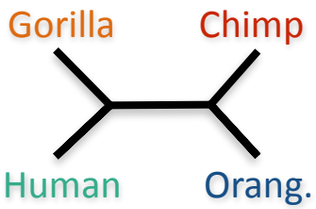
θ_1=70%



θ_2=15%



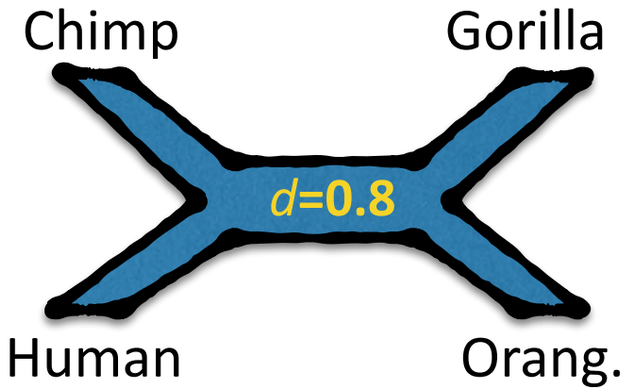
θ_3=15%



Branch Length

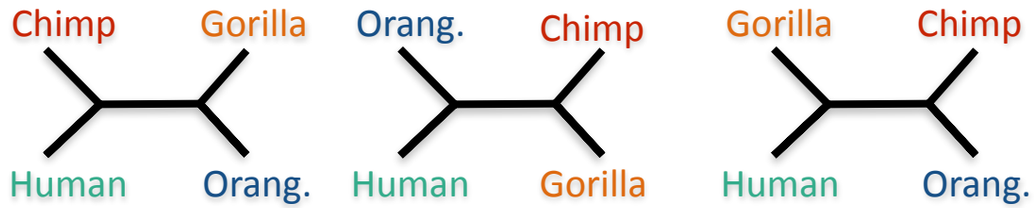
[Sayyari and Mirarab, MBE, 2016]

- Simply a function of the level of discordance
- A single quartet ($n=4$): reverse the discordance formula to get the ML estimate



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

θ_1=70% θ_2=15% θ_3=15%



$$-\ln\frac{3}{2}(1-\hat{\theta}_1)$$

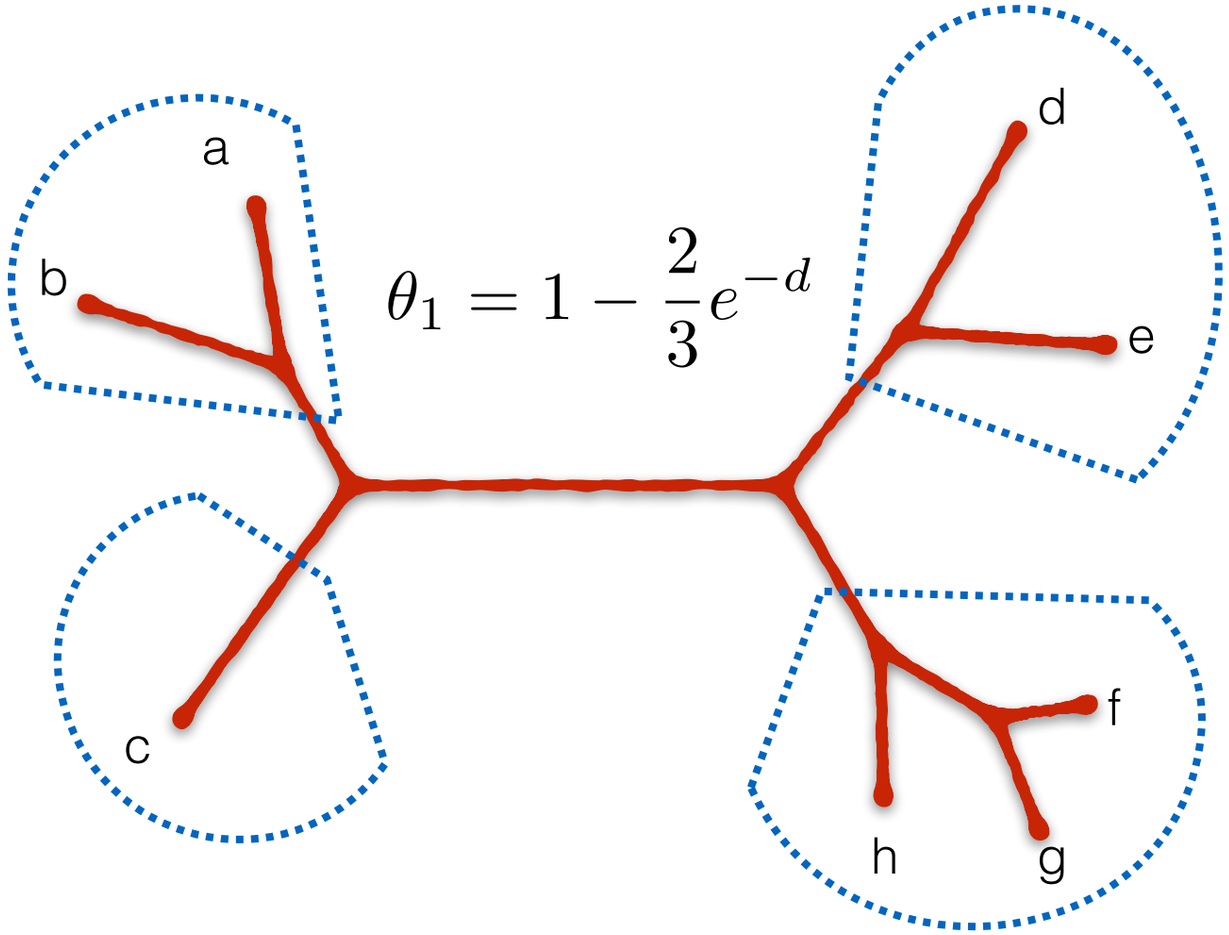
$m_1 = 132$ $m_2 = 32$ $m_3 = 36$

$\hat{\theta}_1$=66% $\hat{\theta}_2$=16% $\hat{\theta}_3$=18%

$\hat{d}$=0.67

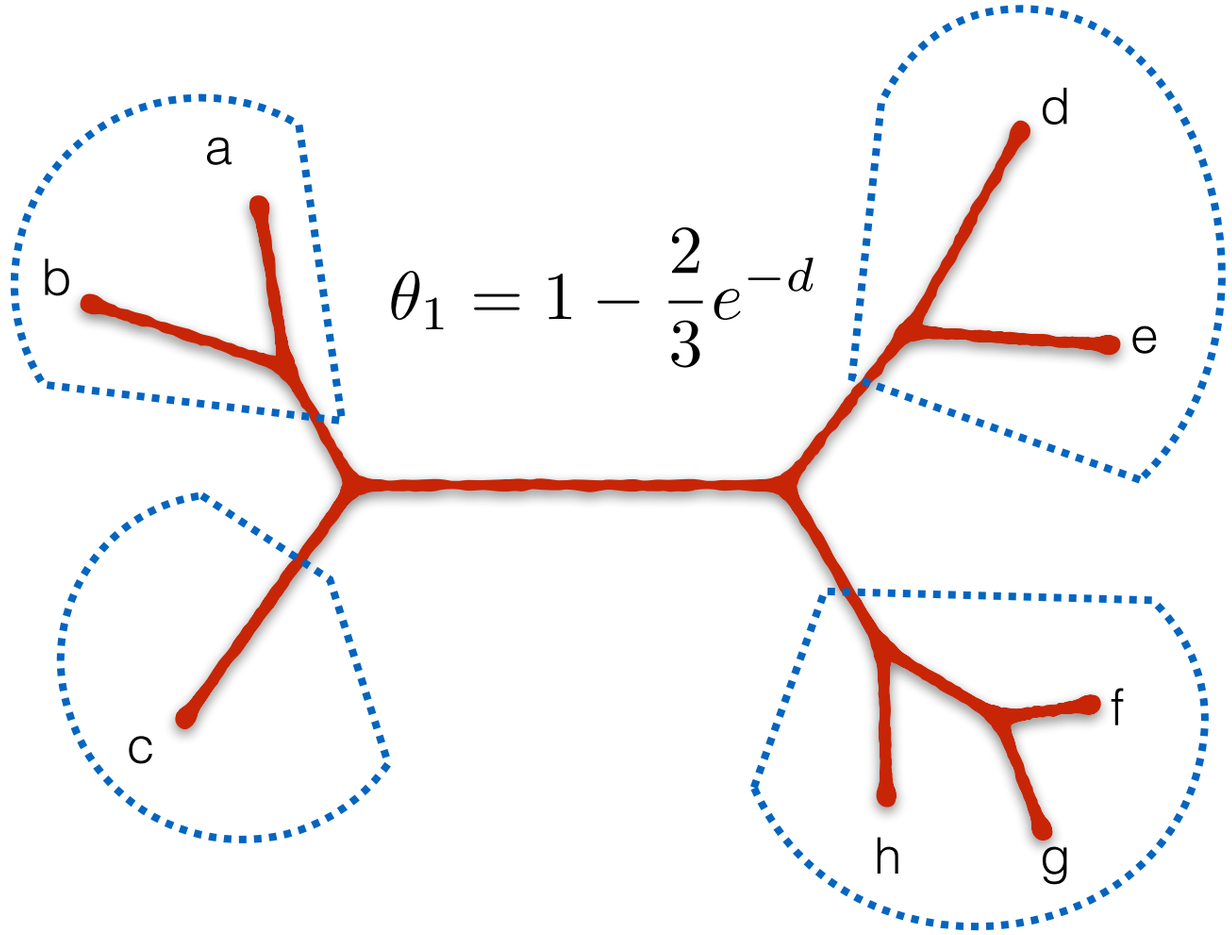
Branch length for $n > 4$

- Simply average all quartet frequencies "around" that branch
- Justified given some assumptions

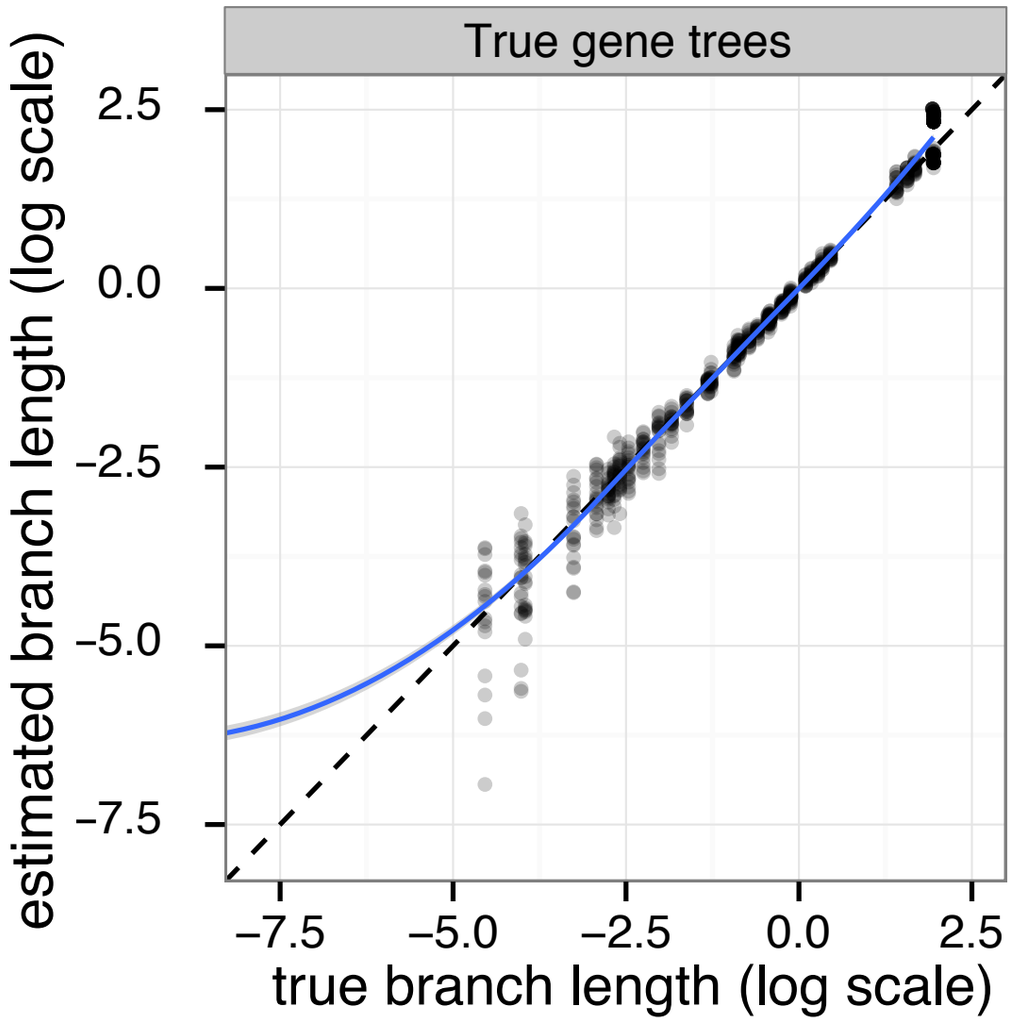


Branch length for $n > 4$

- Simply average all quartet frequencies "around" that branch
  - Justified given some assumptions
- Can be done efficiently in $\Theta(n^2 m)$ for all branches for n species and m genes

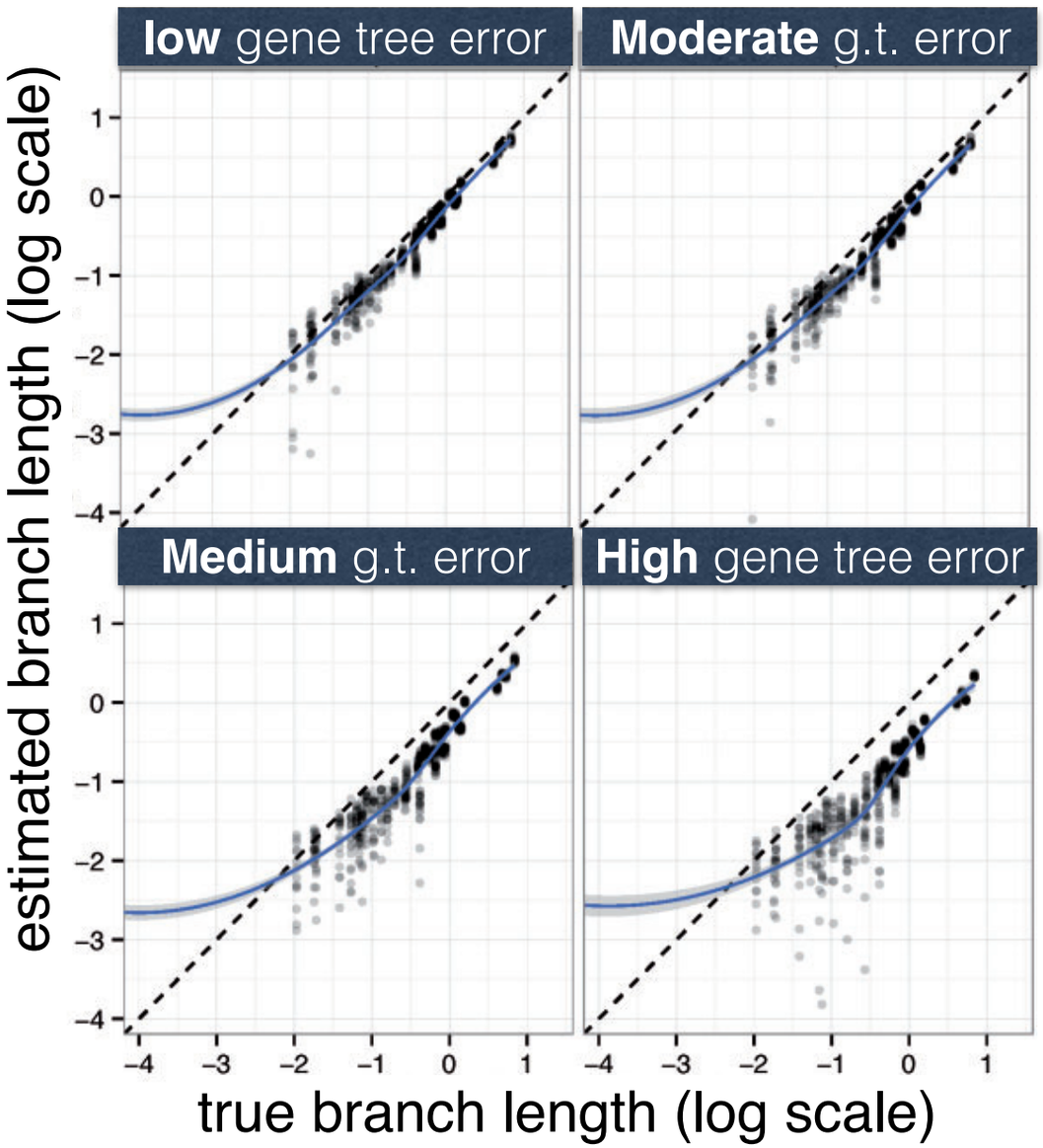


Branch length accuracy



With true gene trees, ASTRAL correctly estimates BL

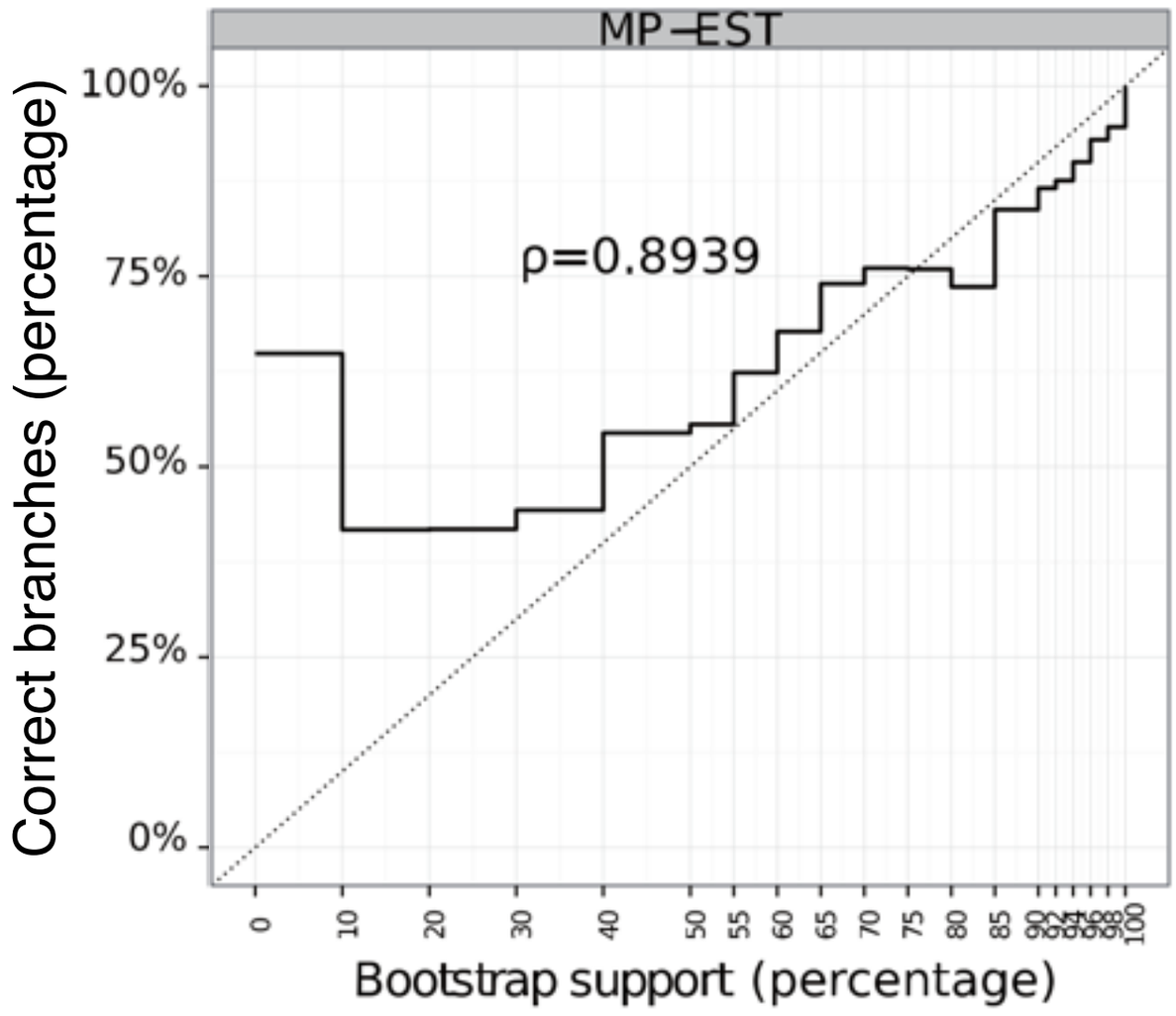
Branch length accuracy



With error-prone estimated gene trees, ASTRAL underestimates BL

Branch support (common practice)

- Multi-locus bootstrapping (MLBS)
- Slow: requires bootstrapping all genes (e.g., 100m ML trees)
- Inaccurate and hard to interpret
  [Mirarab et al., Sys bio, 2014; Bayzid et al., PLoS One, 2015]



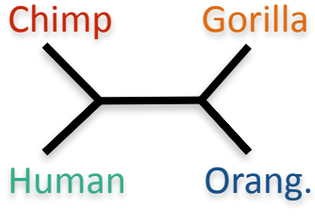
[Mirarab et al., Sys bio, 2014]

Branch support idea: $n=4$

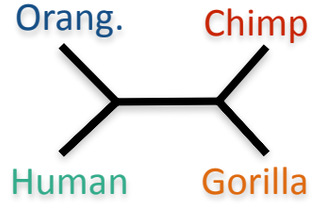
- Recall quartet frequencies follow a multinomial distribution

$m = 200$

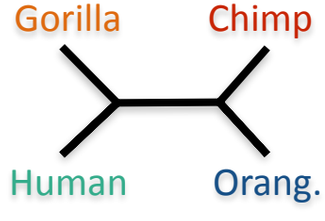
$m_1 = 80$



$m_2 = 63$



$m_3 = 57$



θ_1 θ_2 θ_3

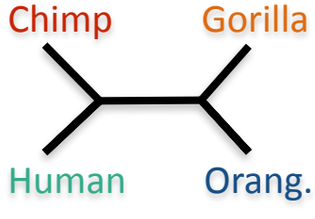
- $\boldsymbol{P}$ (topology seen in m_1 / m gene trees is the species tree) =
  $\boldsymbol{P}$ ($\theta_1 > 1/3$) =
  $\boldsymbol{P}$ (a 3-sided coin tossed m times is biased towards
  the side that shows up m_1 times)

Branch support idea: $n=4$

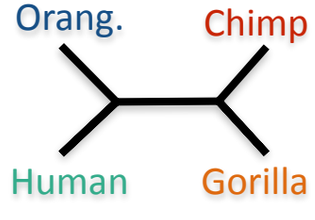
- Recall quartet frequencies follow a multinomial distribution

$m = 200$

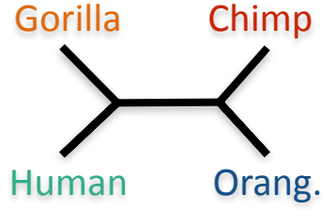
$m_1 = 80$



$m_2 = 63$



$m_3 = 57$



θ_1 θ_2 θ_3

- $\boldsymbol{P}$ (topology seen in m_1 / m gene trees is the species tree) =
  $\boldsymbol{P}$ ($\theta_1 > 1/3$) =
  $\boldsymbol{P}$ (a 3-sided coin tossed m times is biased towards
  the side that shows up m_1 times)

- Can be analytically solved

Posterior

$$\Gamma t^{z_1}\left(\frac{1-t}{2}\right)^{n-z_1}$$

Prior: Yule process become conjugate

$$P\left(\theta_1 > \frac{1}{3}|\bar{Z}=\bar{z}\right) = \frac{\int_{\frac{1}{3}}^{1} P(\bar{Z}=\bar{z}|\theta_1=t) f_{\theta_1}(t)\mathrm{d}t}{P(\bar{Z}=\bar{z})}$$

$$\sum_{j=1}^{3}\int_{\frac{1}{3}}^{1} P(\bar{Z}=\bar{z}|\theta_j=t) f_{\theta_j}(t)\mathrm{d}t$$

- Fast to calculate
- Depends on the frequency of not just the first topology, but also the frequency of second and third topologies

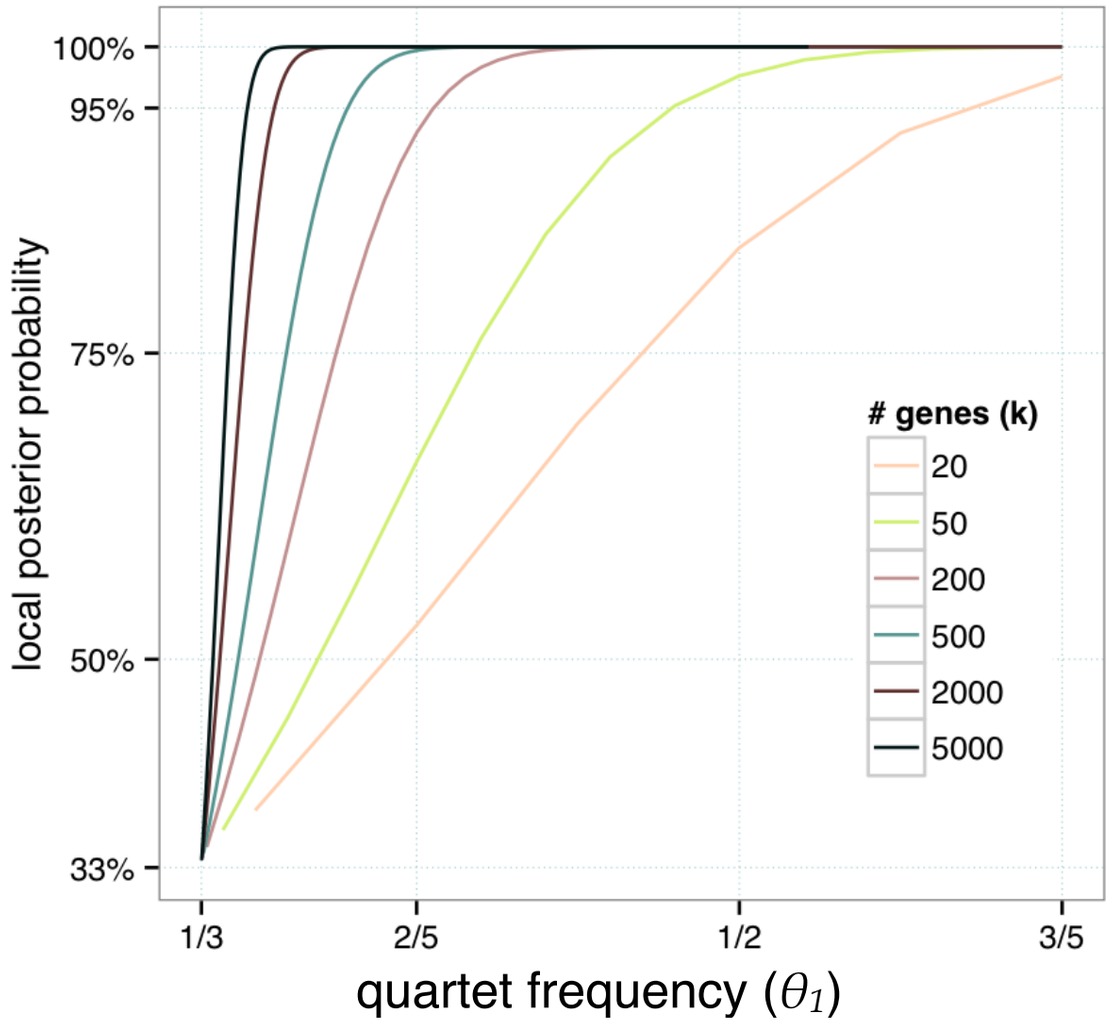
Conjugate prior

- All three topologies have equally prior

$$Pr(\theta_1 > \frac{1}{3}) = Pr(\theta_2 > \frac{1}{3}) = Pr(\theta_3 > \frac{1}{3}) = \frac{1}{3}$$

- The species tree generated through a birth-only (Yule) process with rate λ
  - Turns out to be the conjugate prior
  - (default) λ =0.5 → uniformly distributed branch lengths

Quartet support v.s. posterior

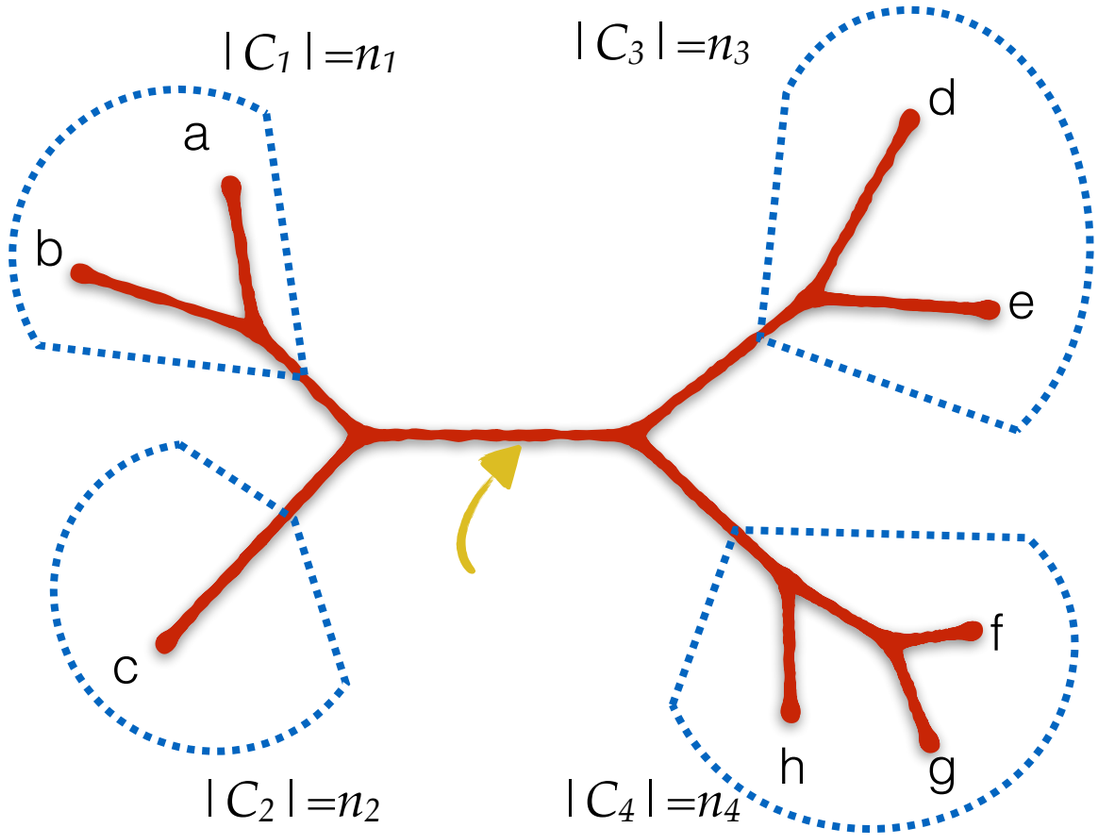


Increased number of genes (m) ⇒ increased support

Decreased discordance ⇒ increased support

How about $n>4$?

- Locality Assumption: All four clusters around a branch are correct
  - Treat branches independently



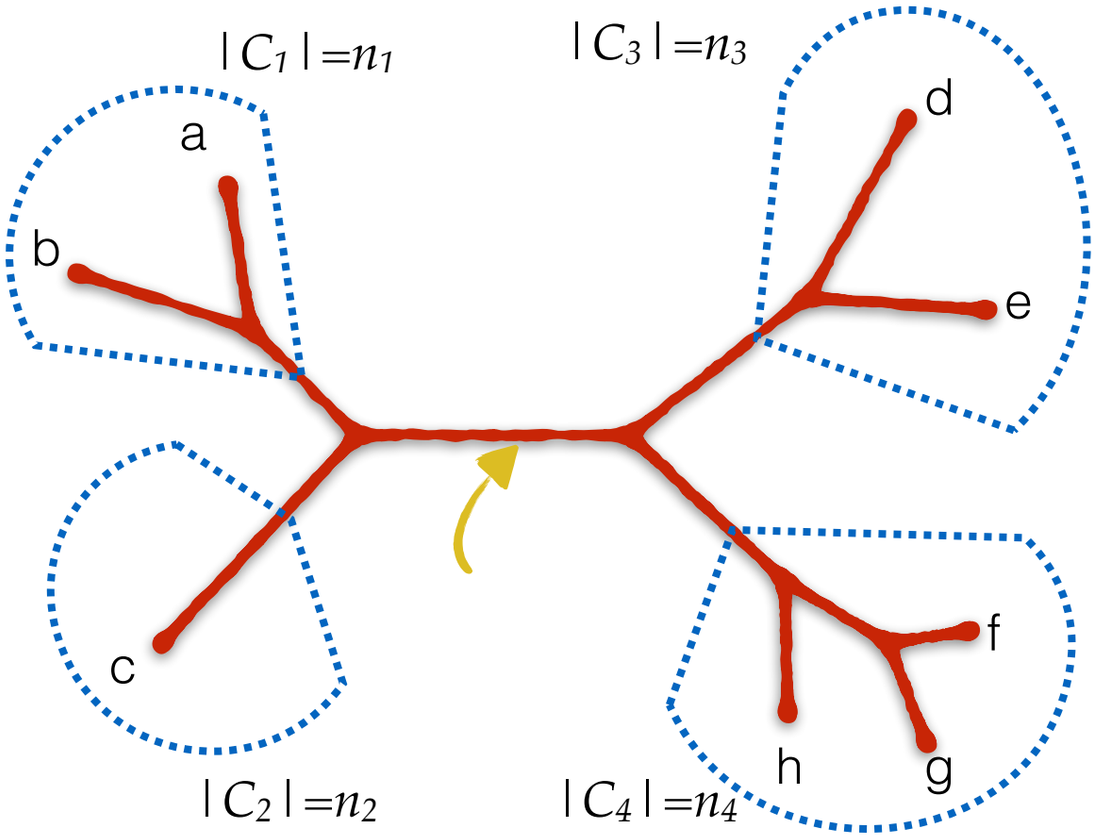
$$k = n_1 \times n_2 \times n_3 \times n_4$$

How about $n>4$?

- Locality Assumption: All four clusters around a branch are correct
  - Treat branches independently
- k quartets around a branch?
  - Independence assumption is too liberal ($m \times k$ tosses of the coin)
  - Fully dependent assumption:

    — all quartets give noisy estimates of a single hidden true frequency

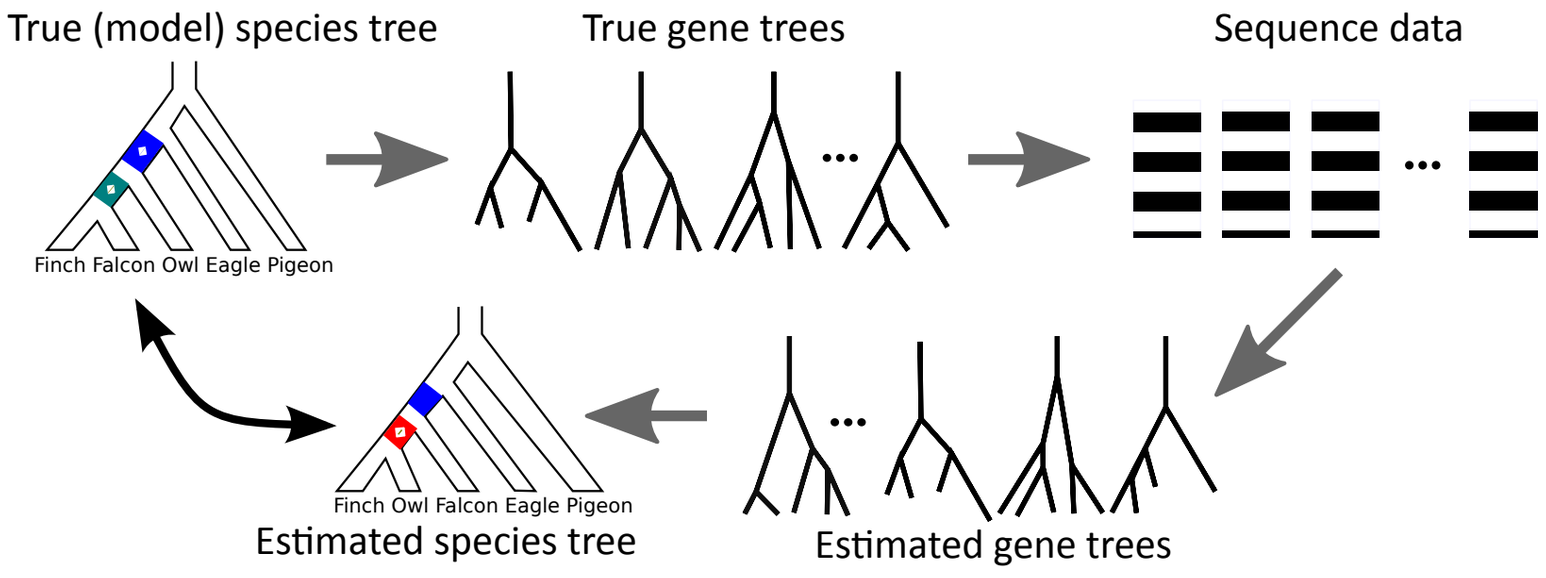
    — Simply **average** their frequencies



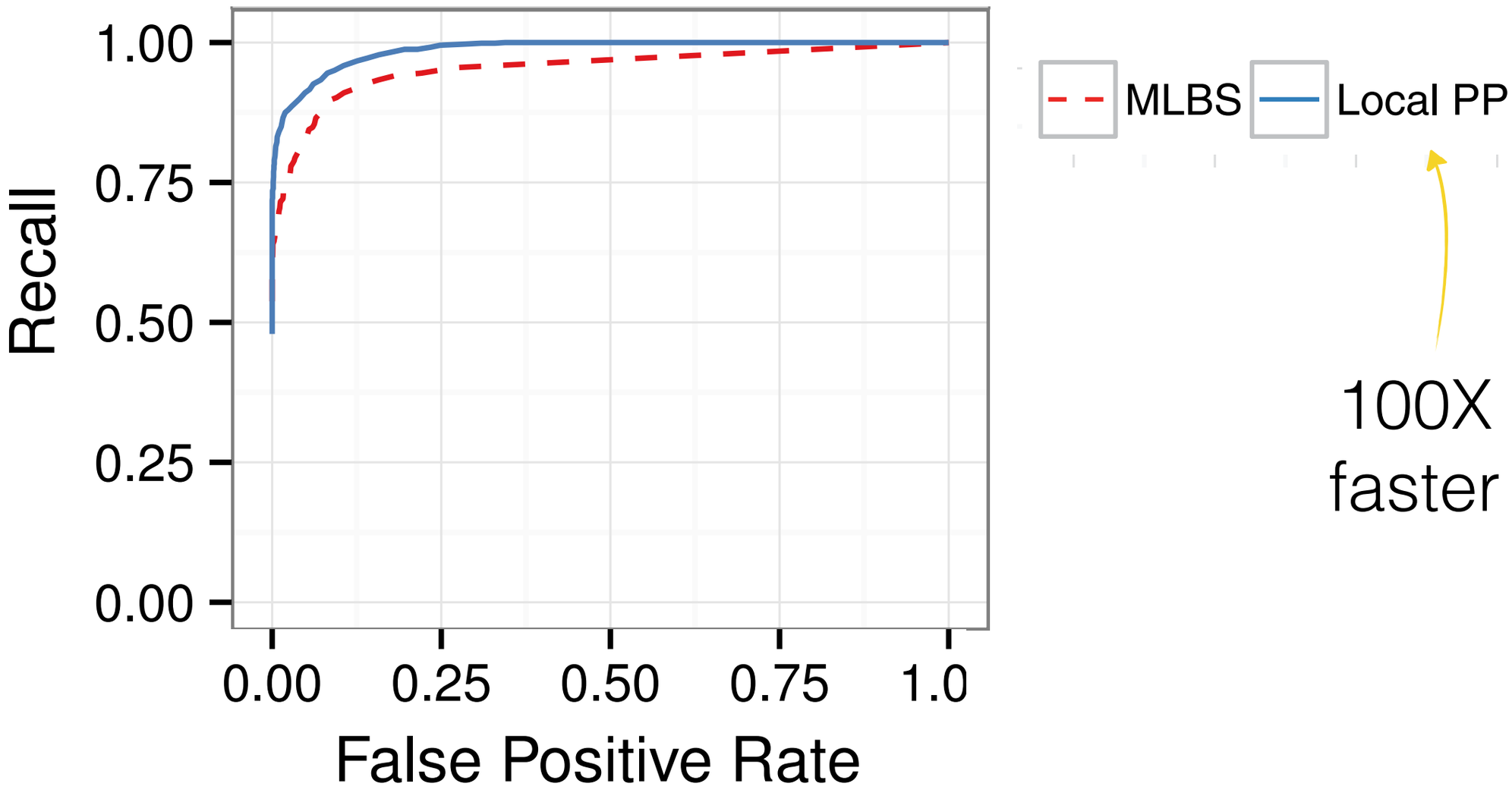
$$k = n_1 \times n_2 \times n_3 \times n_4$$

Simulation studies

- Our simulations **violate** our assumptions
  - Estimated gene trees instead of true gene trees
  - Estimated species trees: the locality assumption can be violated
- Measuring the support accuracy: the number of false positive and false negatives above various thresholds of support



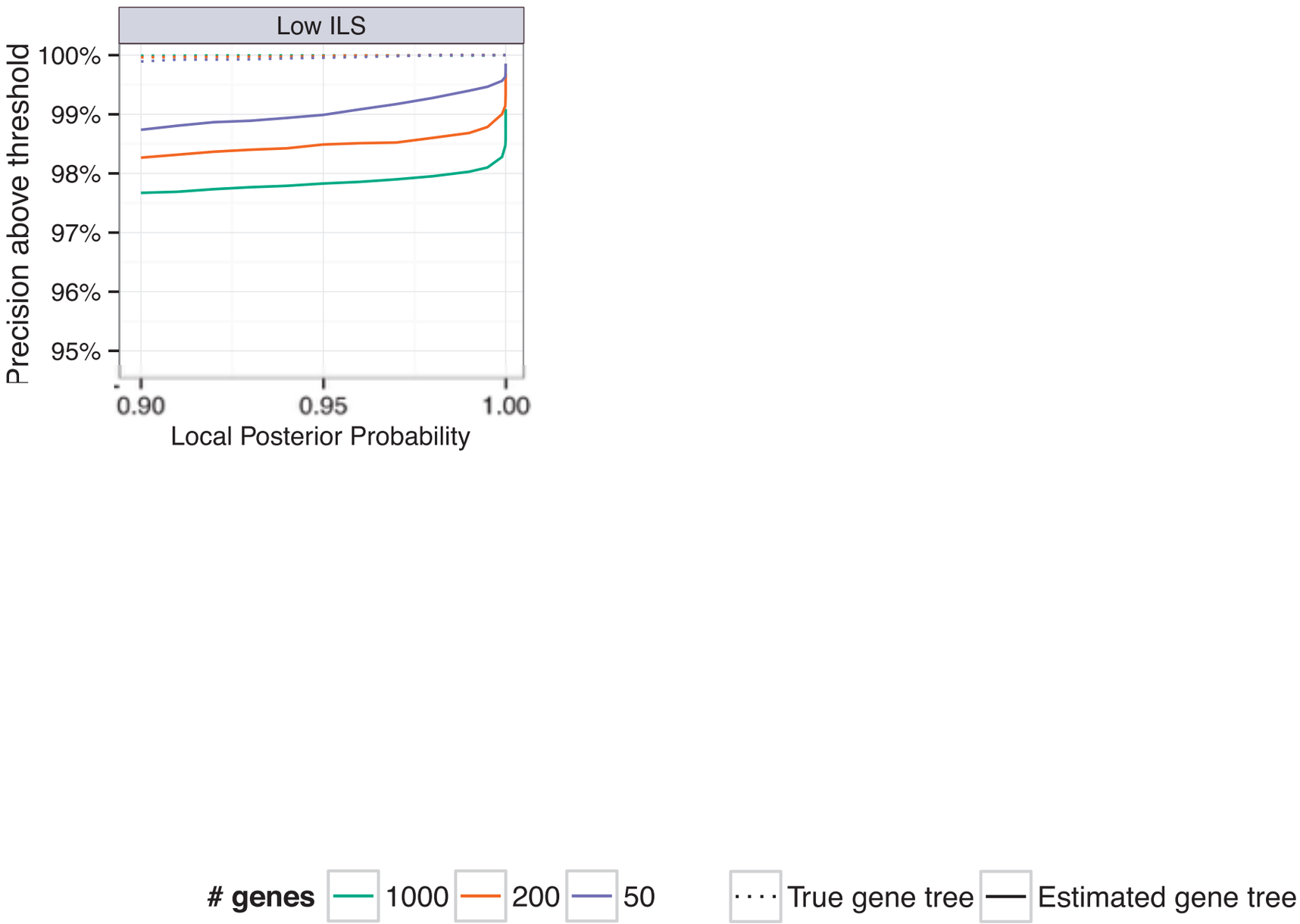
localPP is more accurate than bootstrapping



Avian simulated dataset (48 taxa, 1000 genes)

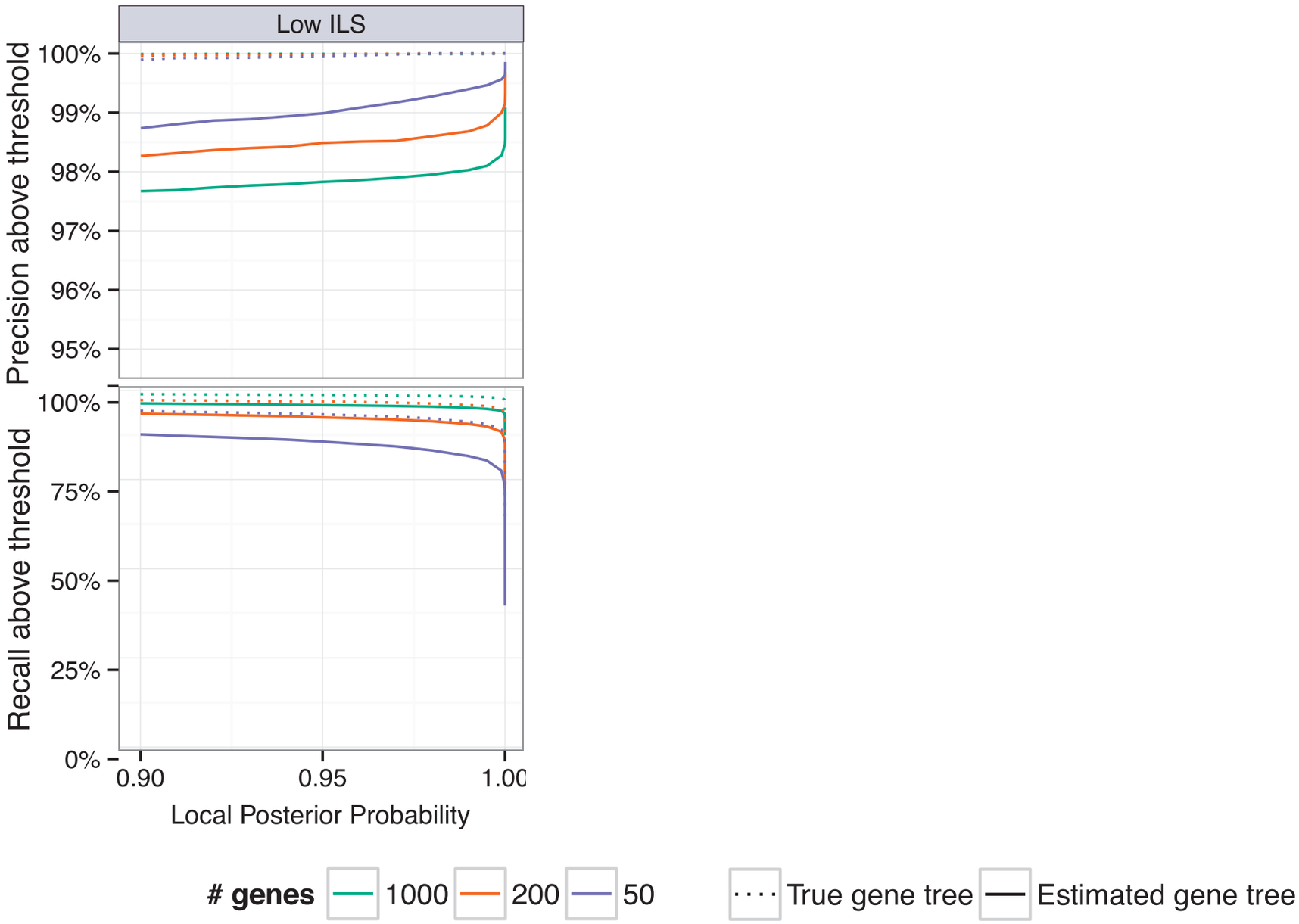
[Sayyari and Mirarab, MBE, 2016]

High precision and recall at high support



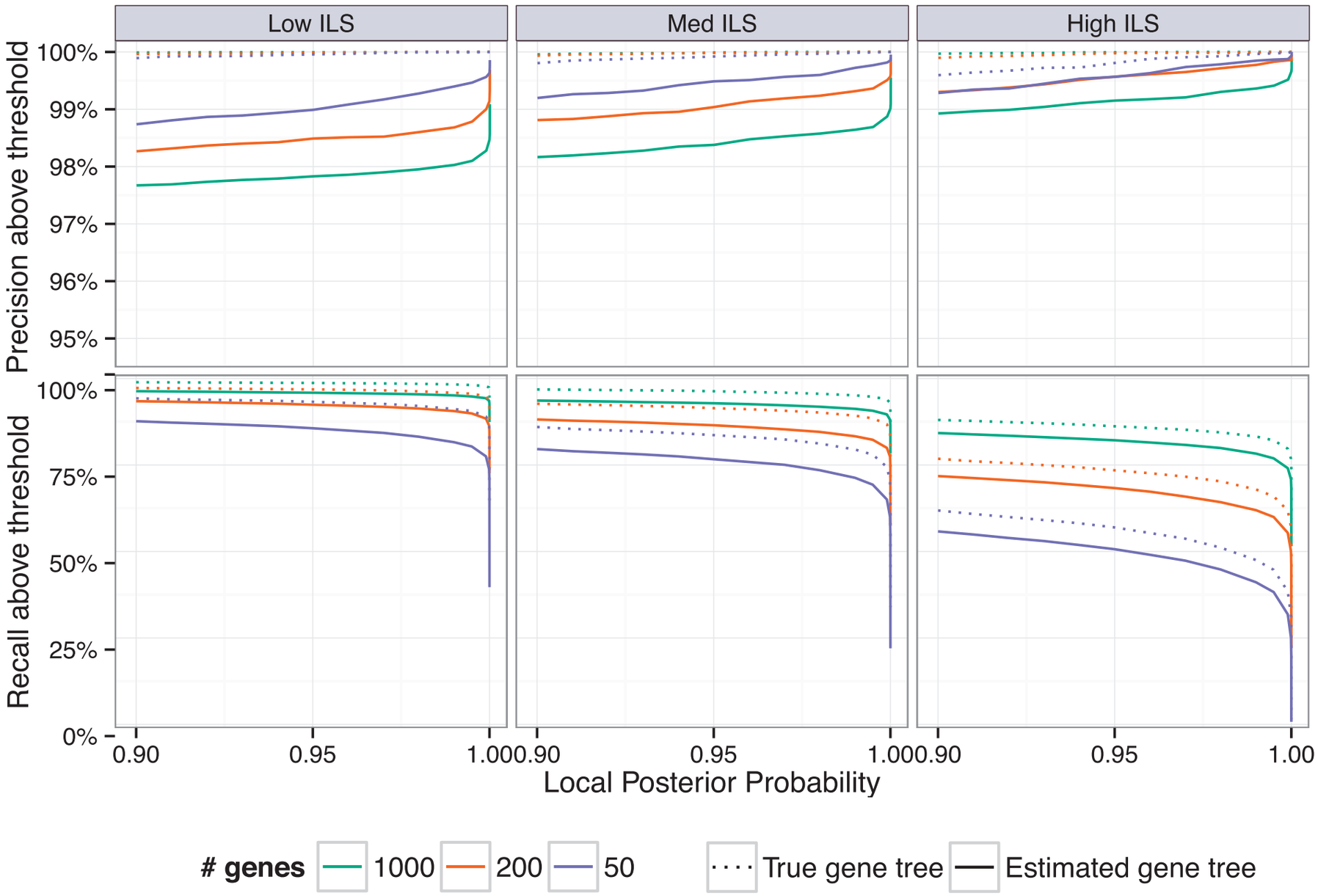
201-taxon datasets (simPhy)

High precision and recall at high support



201-taxon datasets (simPhy)

High precision and recall at high support



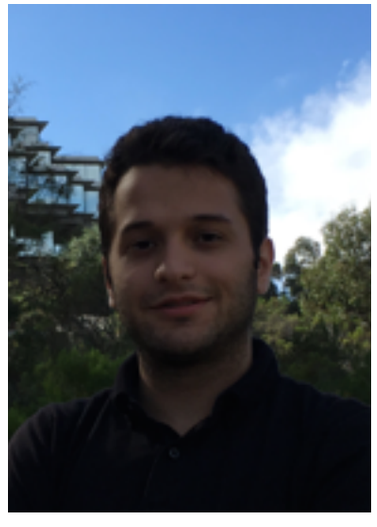
201-taxon datasets (simPhy)

Summary

- Both branch length and support can be computed quickly
  - a function of the observed amount of gene tree discordance
    - support is also a function of the number of genes
- Local posterior probability outperforms bootstrapping
  - Requires strong assumptions (to be relaxed in future)
- Branch length accuracy depends on the gene tree accuracy
- All available at https://github.com/smirarab/ASTRAL

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Tandy Warnow



Erfan Sayyari

NSF

ALFRED P. SLOAN FOUNDATION

HHMI
HOWARD HUGHES MEDICAL INSTITUTE

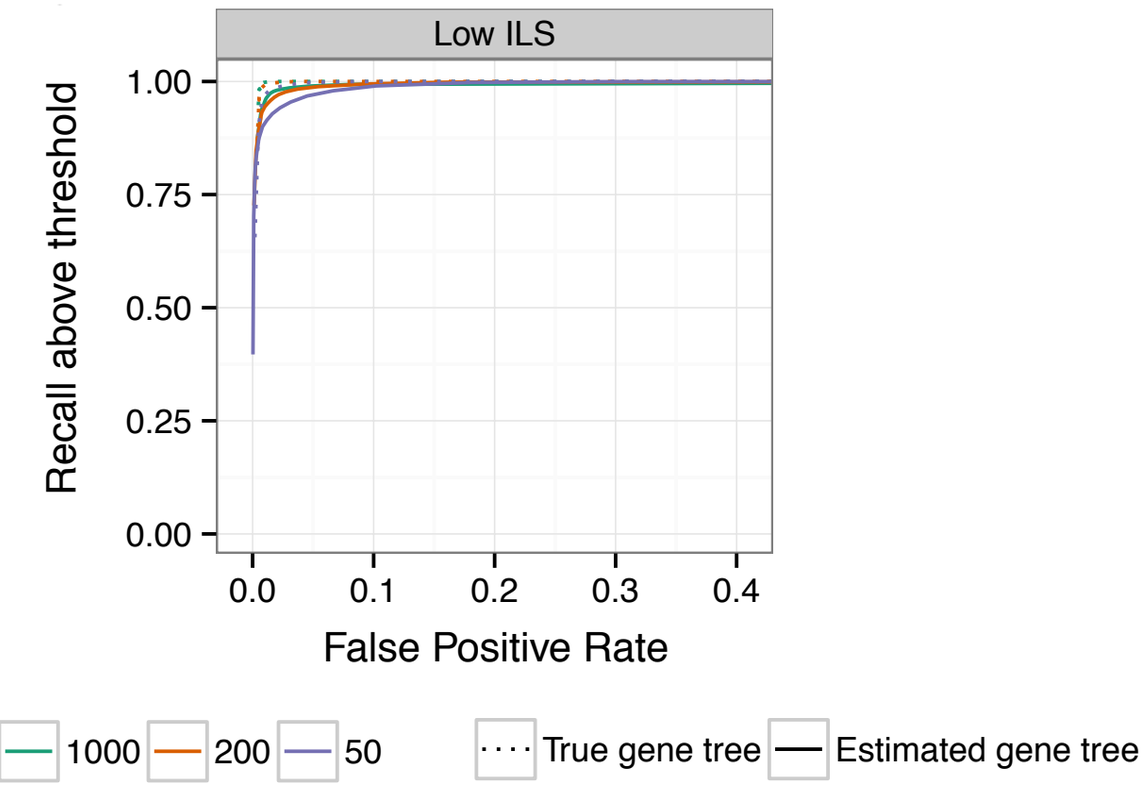
Results (A200)

Results (A200)

- recall (the percentage of all true branches that have support ≥s),
- false positive rate (FPR) (the percentage of all false branches that have support ≥s).

Results (A200)

- recall (the percentage of all true branches that have support ≥s),
- false positive rate (FPR) (the percentage of all false branches that have support ≥s).

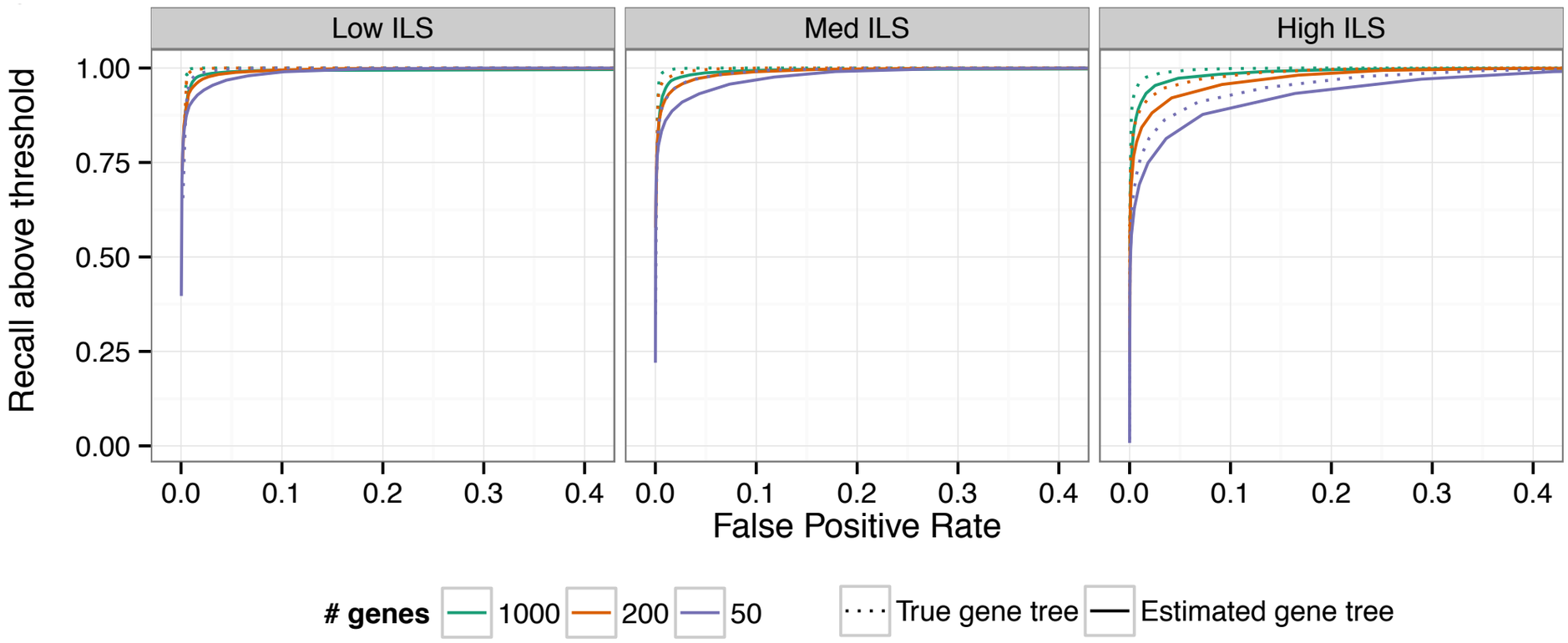


Results (A200)

- recall (the percentage of all true branches that have support ≥s),
- false positive rate (FPR) (the percentage of all false branches that have support ≥s).

Results (A200)

- recall (the percentage of all true branches that have support ≥s),
- false positive rate (FPR) (the percentage of all false branches that have support ≥s).



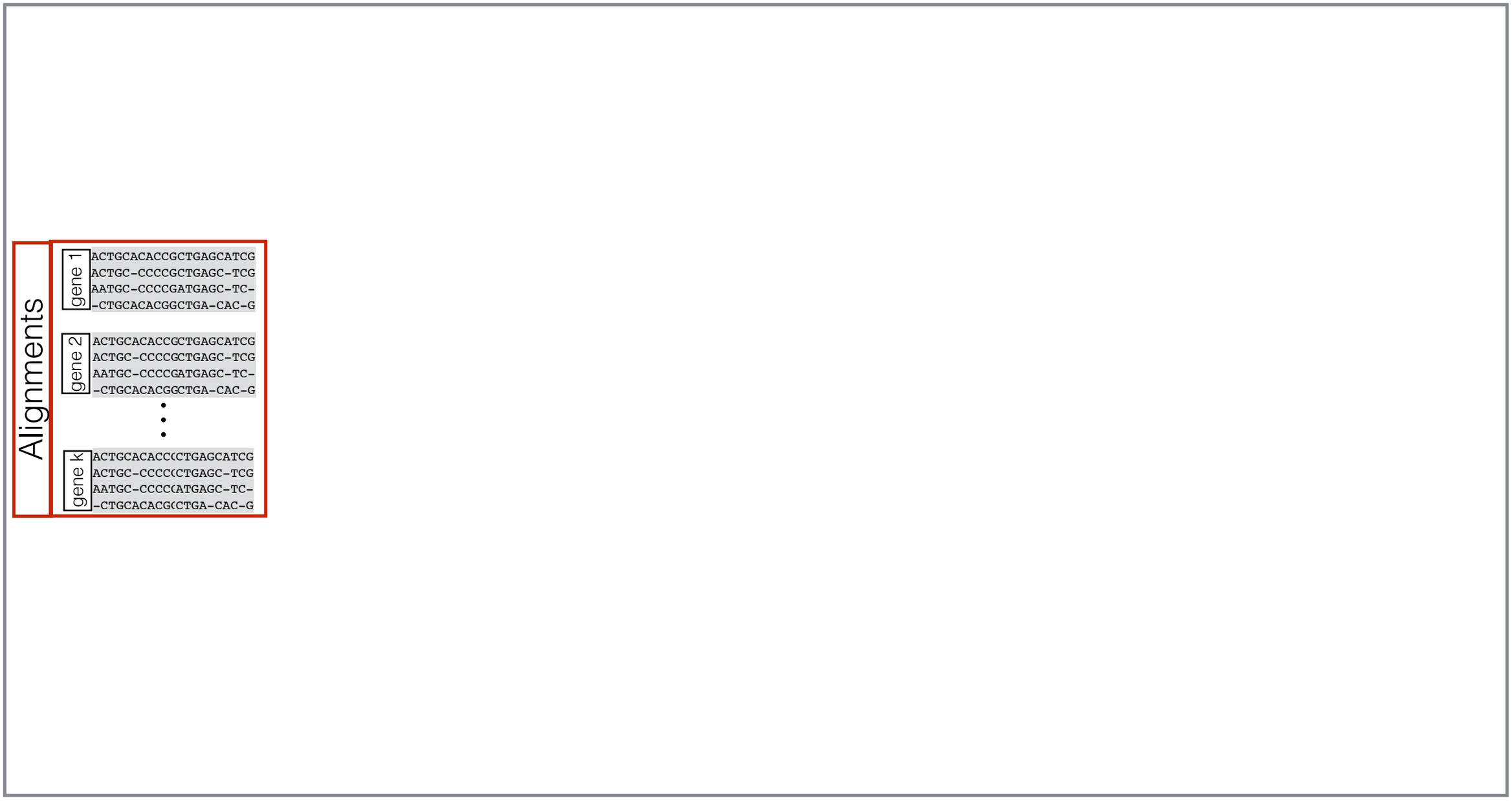
MLBS Procedure

MLBS Procedure

- First bootstrap each gene

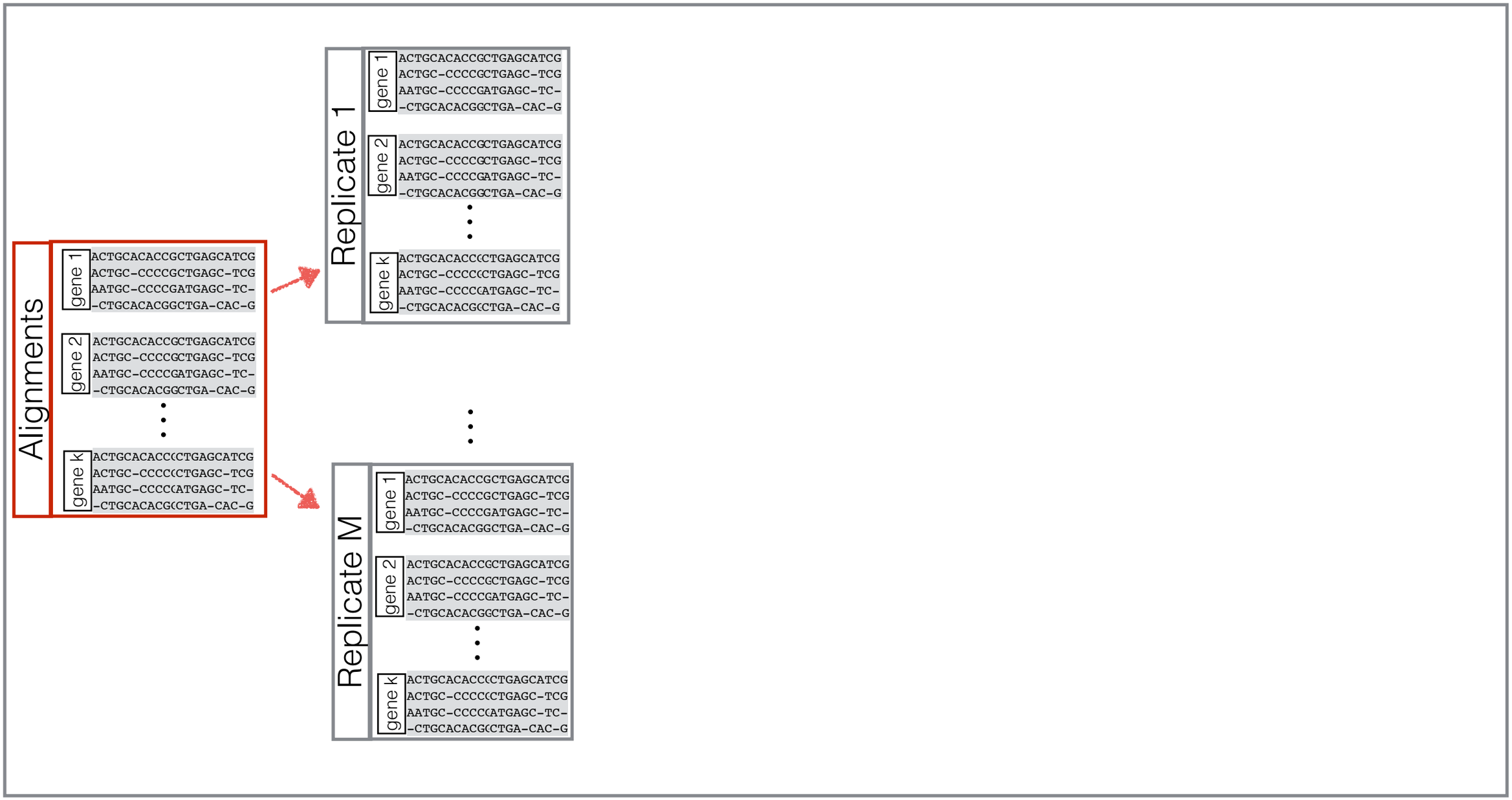
MLBS Procedure

- First bootstrap each gene



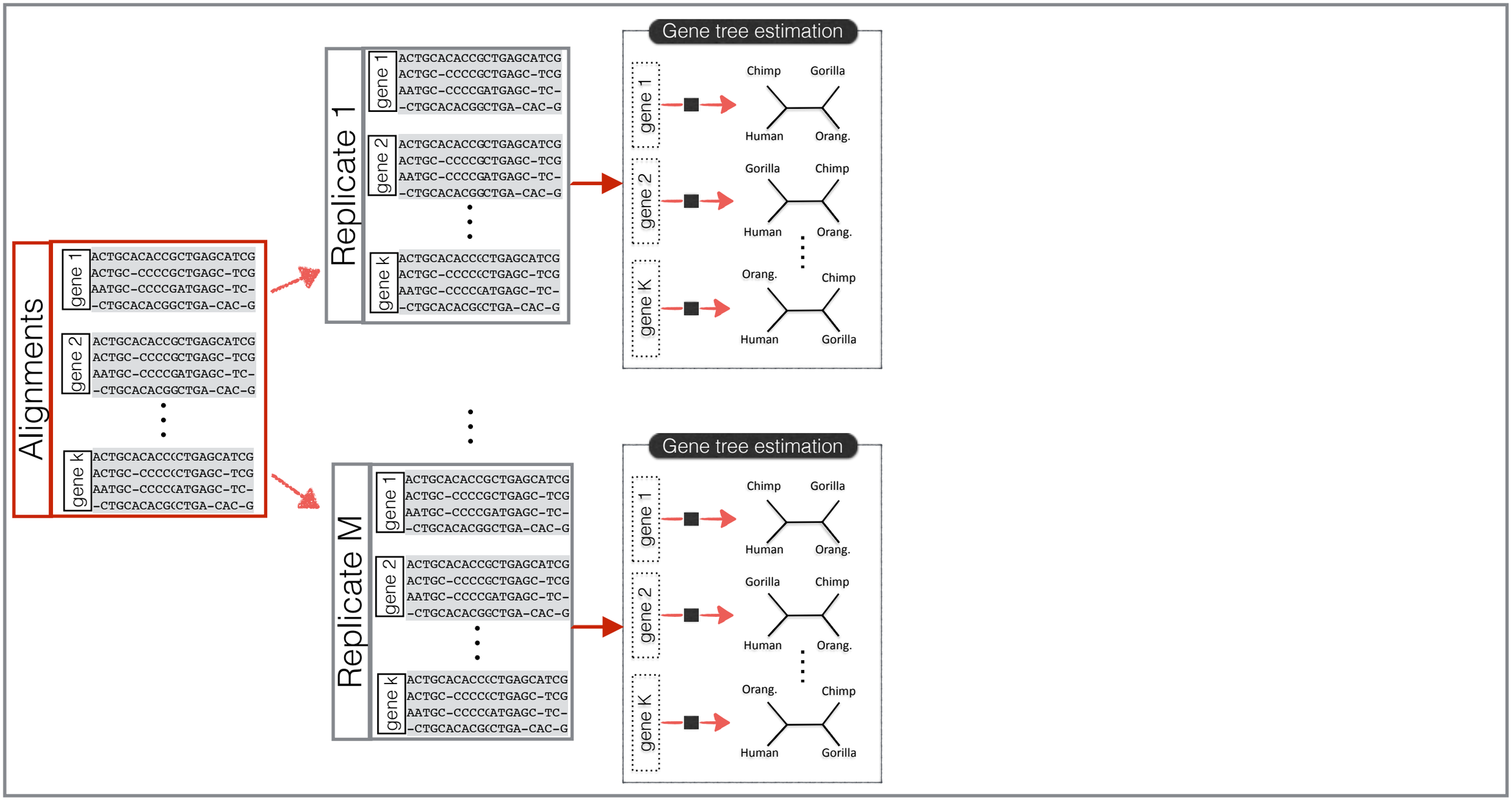
MLBS Procedure

- First bootstrap each gene



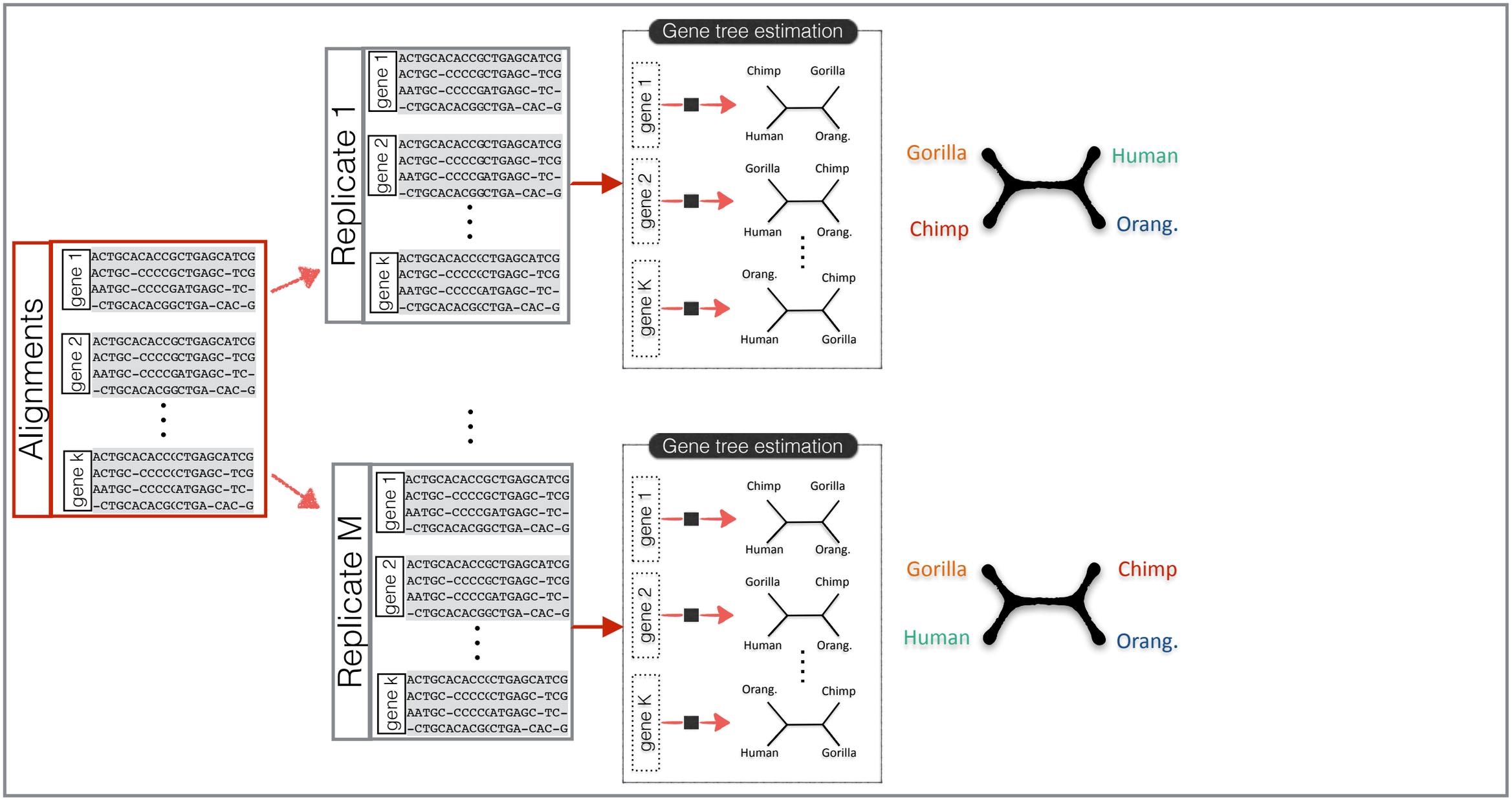
MLBS Procedure

- First bootstrap each gene



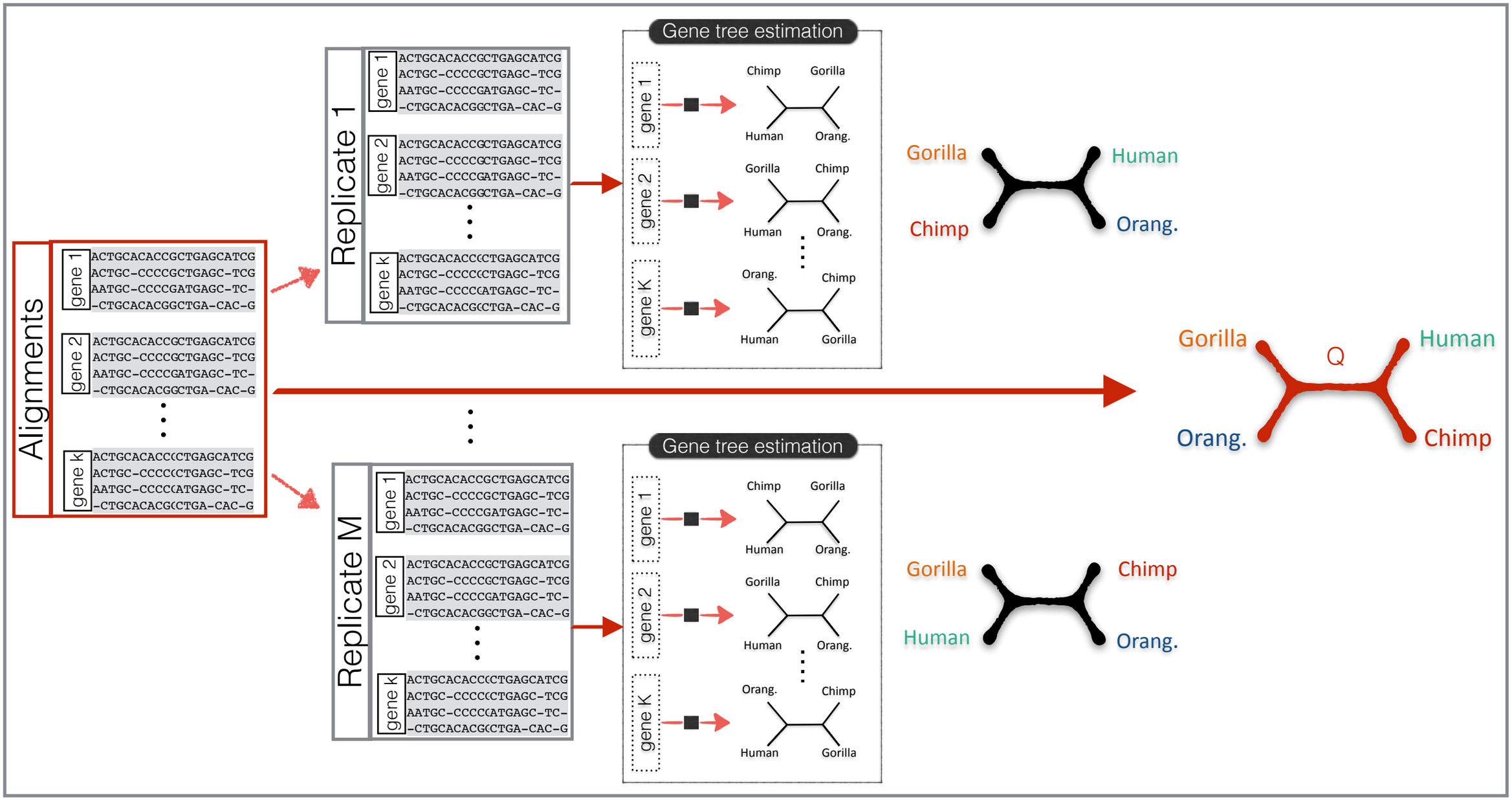
MLBS Procedure

- First bootstrap each gene



MLBS Procedure

- First bootstrap each gene



MLBS Procedure

- First bootstrap each gene