Phylogenomic Branch Length Estimation using Quartets

Yasamin Tabatabaee[1], Chao Zhang[2], Tandy Warnow[1], Siavash Mirarab[3]

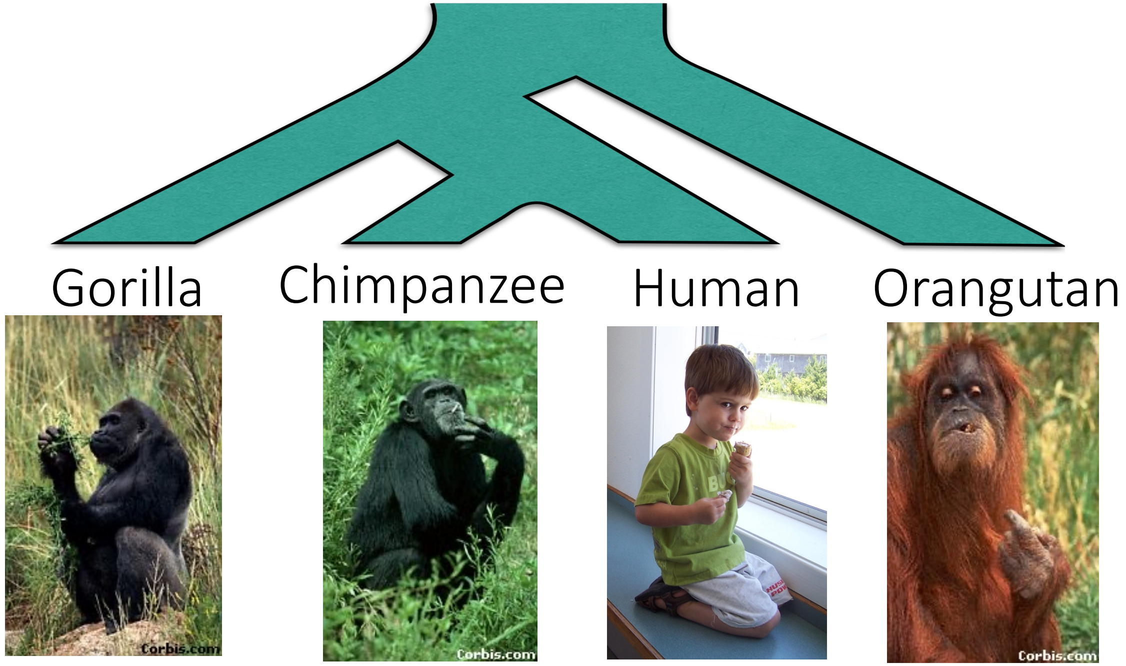
[1] University of Illinois at Urbana-Champaign, [2] University of California, Berkeley,

[3] University of California, San Diego

ISMB/ECCB

July 27, 2023

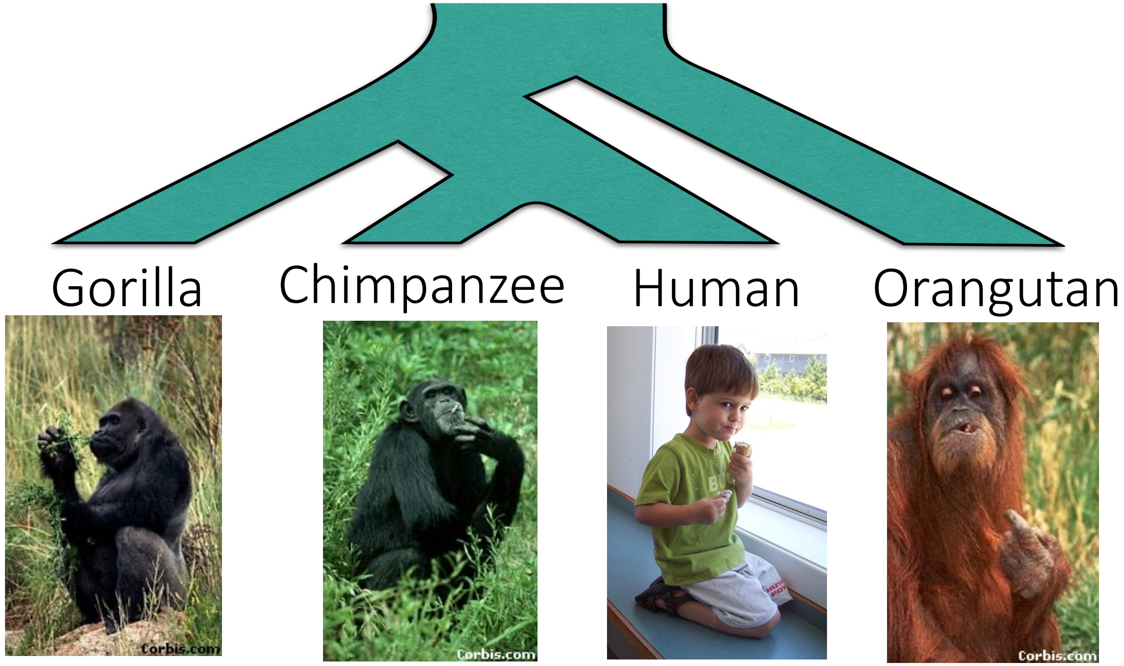
Phylogenomics and gene tree discordance



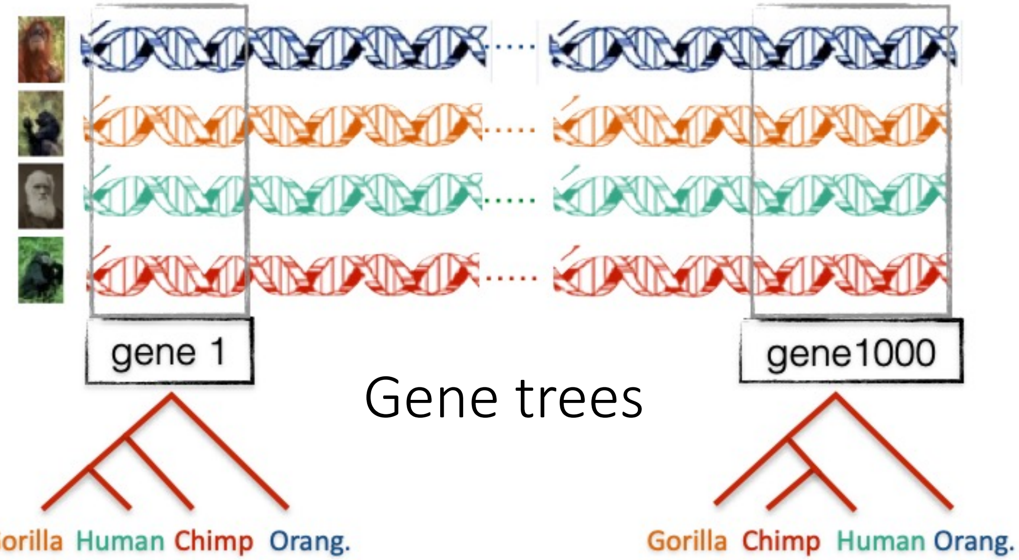
Gorilla Chimpanzee Human Orangutan

Species tree

Phylogenomics and gene tree discordance

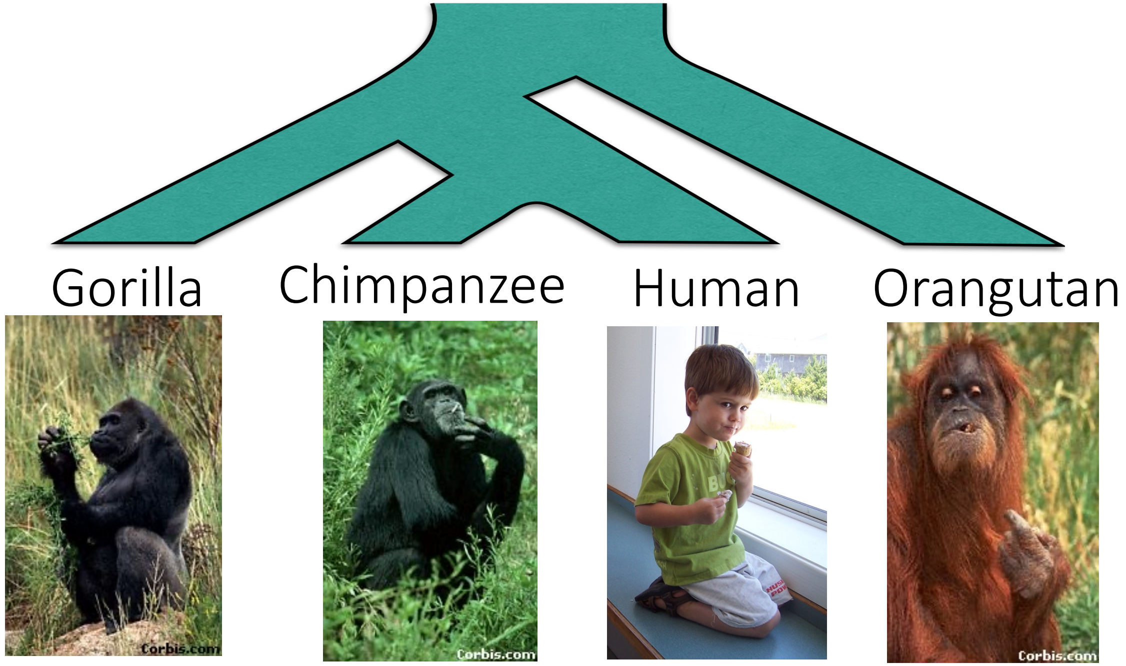


Species tree

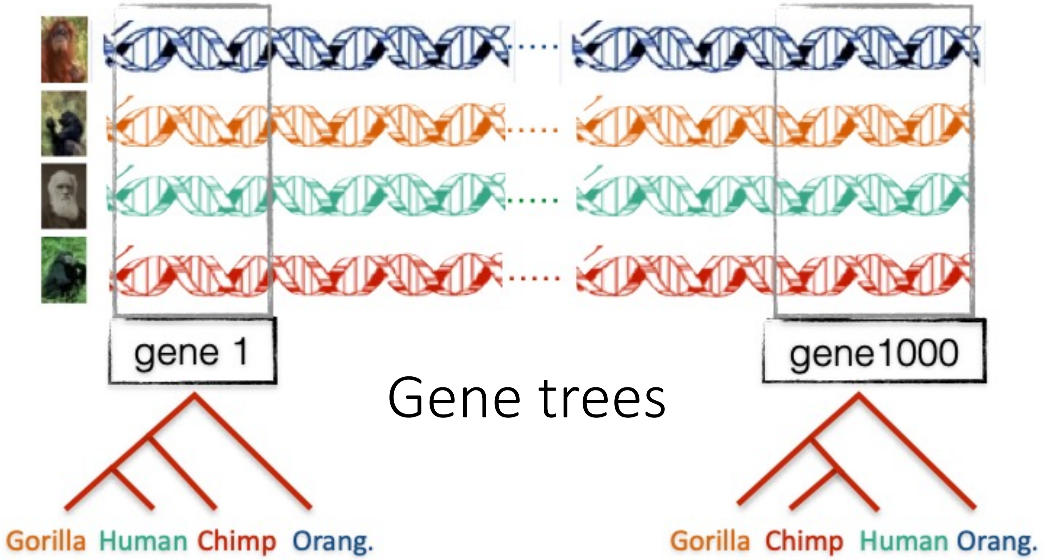


Gene trees

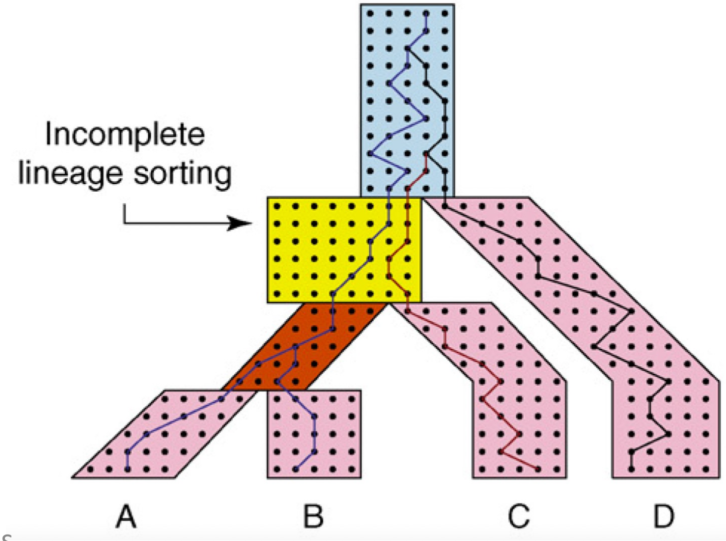
Phylogenomics and gene tree discordance



Species tree



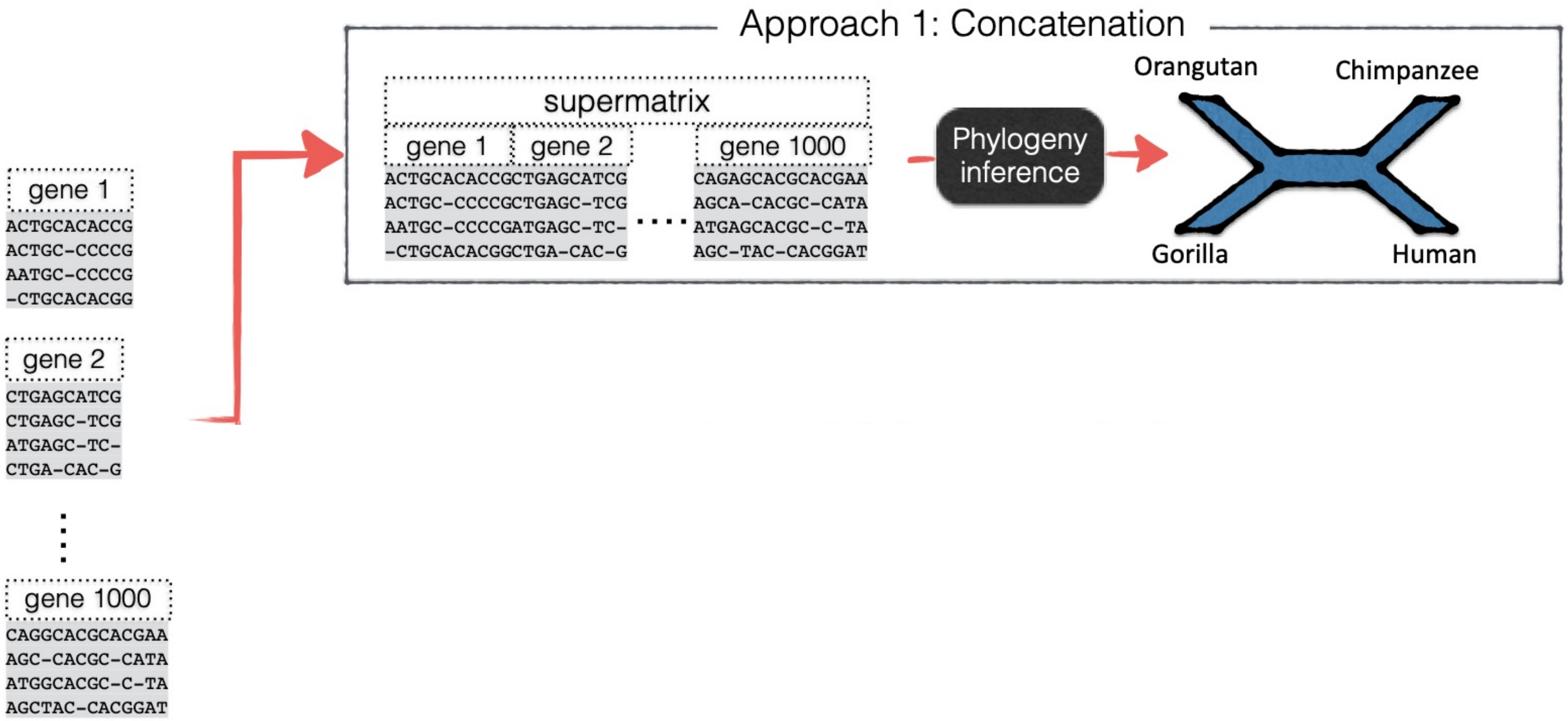
Gene trees



- Incomplete lineage sorting (ILS) is a major cause of gene tree discordance.
- ILS can be modeled by the multi-species coalescent (MSC) model.

Image Credit: Degnan and Rosenberg, 2009, Trends in Ecology and Evolution

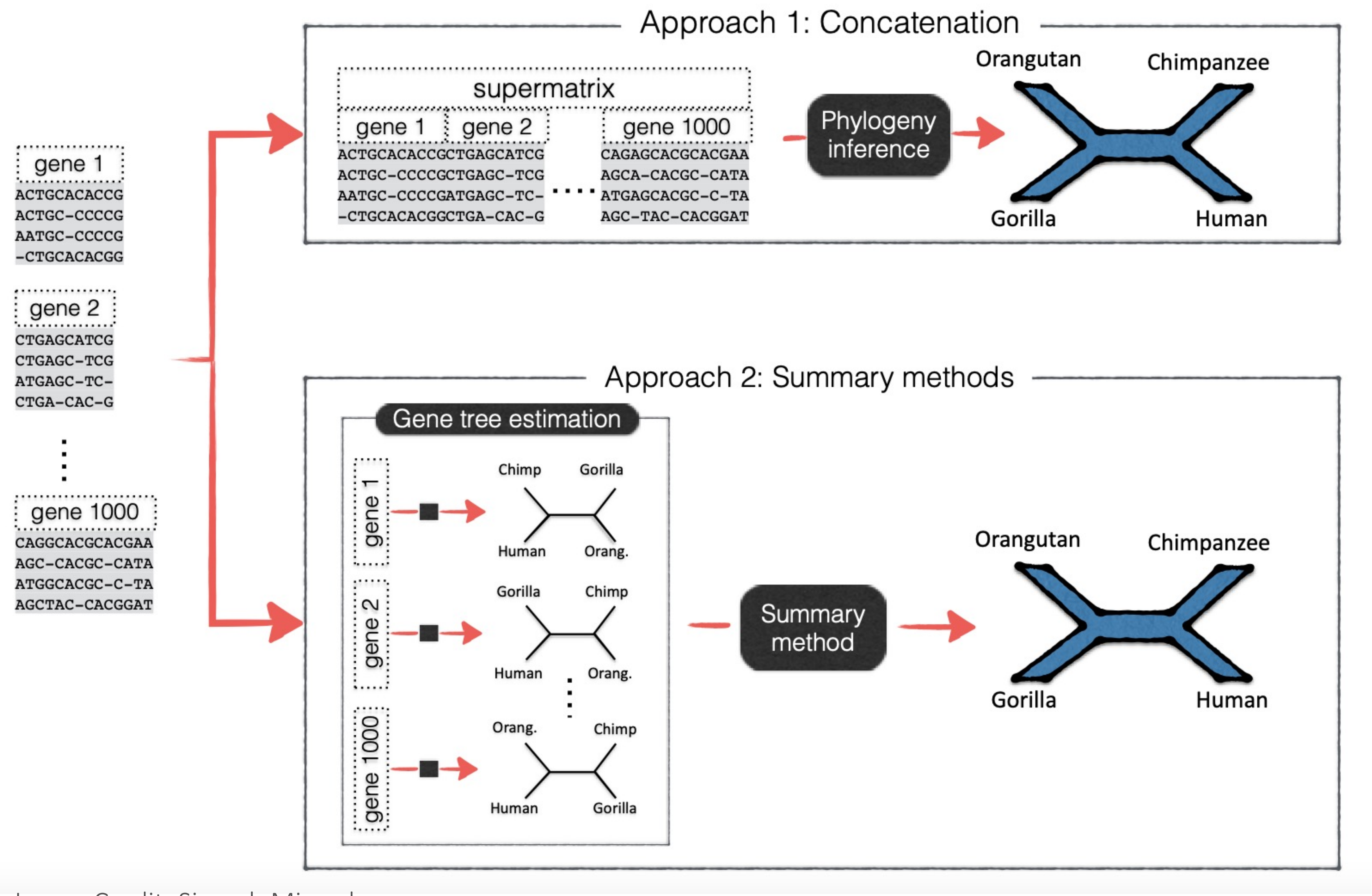
Species tree and branch length estimation



Maximum Likelihood, e.g.
RAxML [Stamatakis, 2014]
FastTree [Price et al, 2010]

Image Credit: Siavash Mirarab

Species tree and branch length estimation



Maximum Likelihood, e.g.
RAxML [Stamatakis, 2014]
FastTree [Price et al, 2010]

ASTRAL [Mirarab et al, 2014]
MP-EST [Liu et al, 2010]
...

Image Credit: Siavash Mirarab

Species tree and branch length estimation

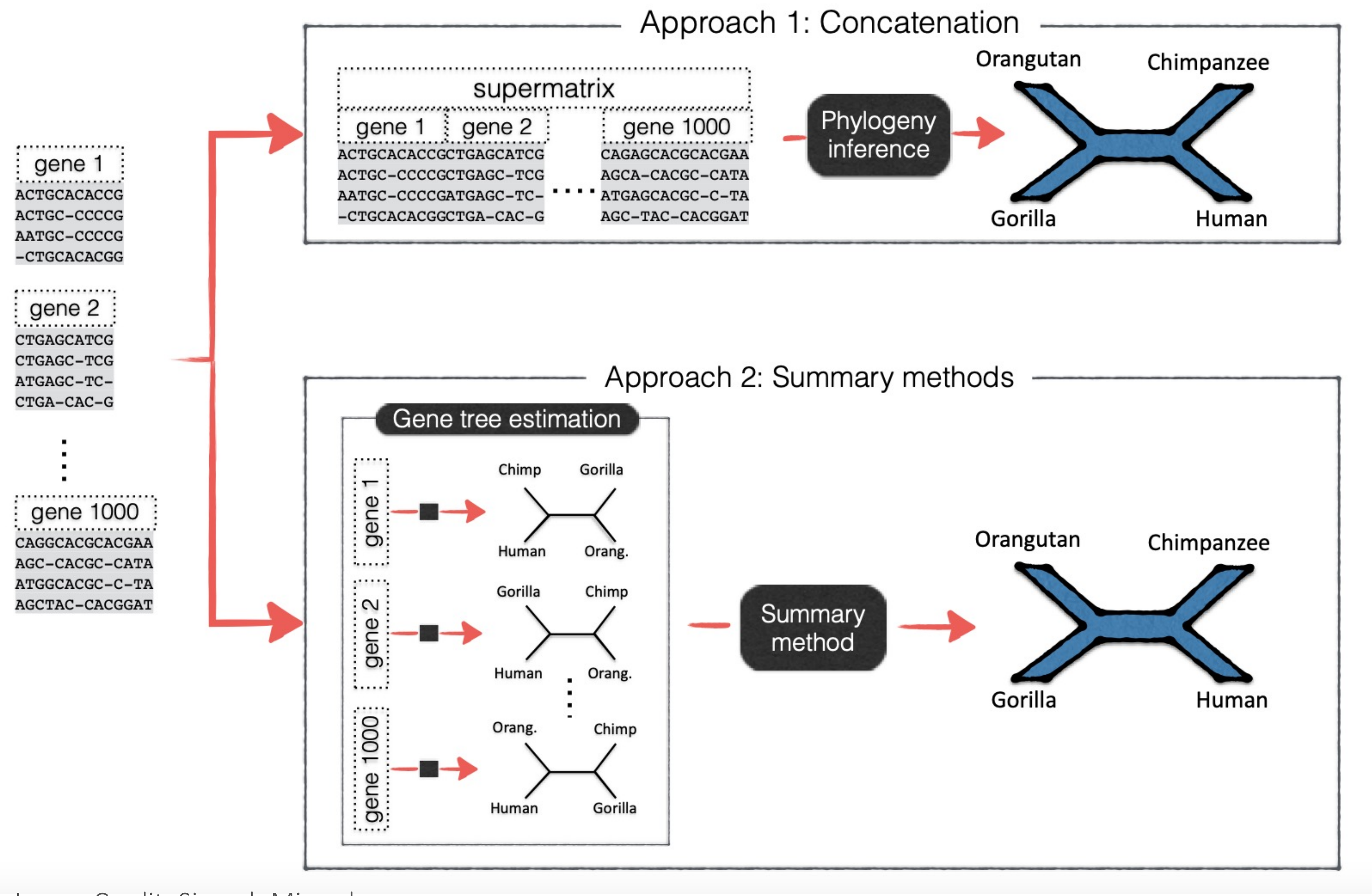


Image Credit: Siavash Mirarab

- Summary methods are more scalable and more accurate when ILS is high, but produce branch lengths in coalescent units (CU)

- CU branch lengths are not useful for most downstream analysis

- **Two-step approach**:

1. infer the topology with summary methods (e.g. ASTRAL, MP-EST)
2. infer branch lengths on that fixed topology with concatenation

Species tree and branch length estimation

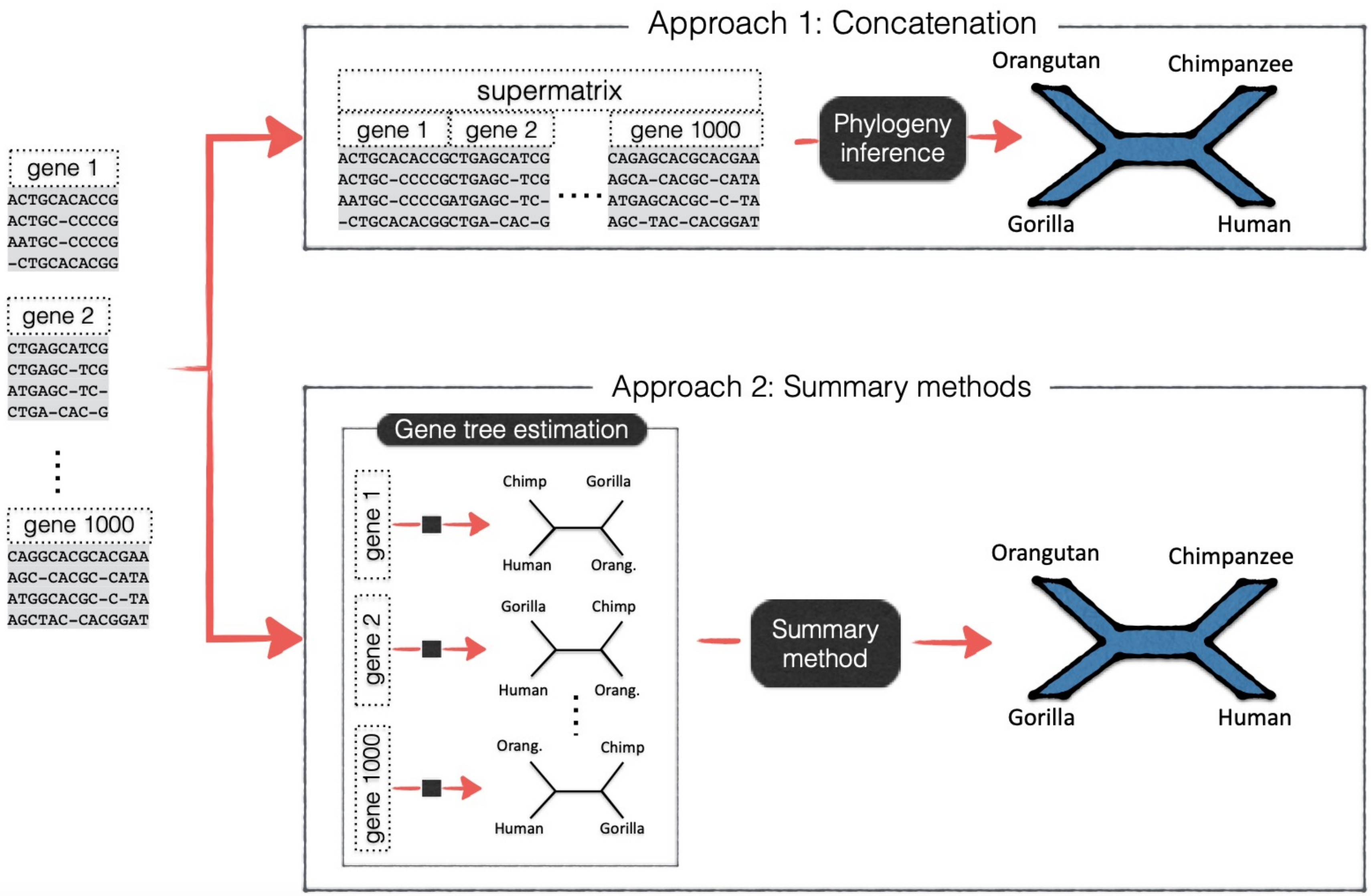


Image Credit: Siavash Mirarab

- Summary methods are more scalable and more accurate when ILS is high, but produce branch lengths in coalescent units (CU)

- CU branch lengths are not useful for most downstream analysis

- **Two-step approach**:

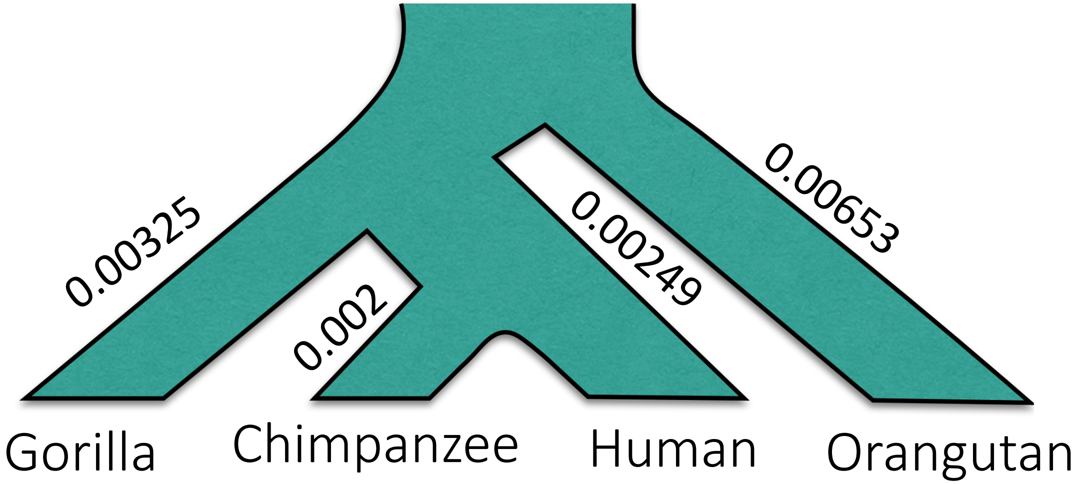
1. infer the topology with summary methods (e.g. ASTRAL, MP-EST)
2. infer branch lengths on that fixed topology with concatenation

ignores heterogeneity across the genome

Branch lengths are necessary for downstream analysis

- Most downstream analysis need branch lengths in the unit of the expected number of substitutions per sites (SU)

- Applications of SU branch lengths
  - Dating
  - Comparative genomics
  - Species delimitation
  - Detecting and characterizing selection
  - …



- CU branch lengths do not directly lead to SU branch lengths and are only inferable for internal branches

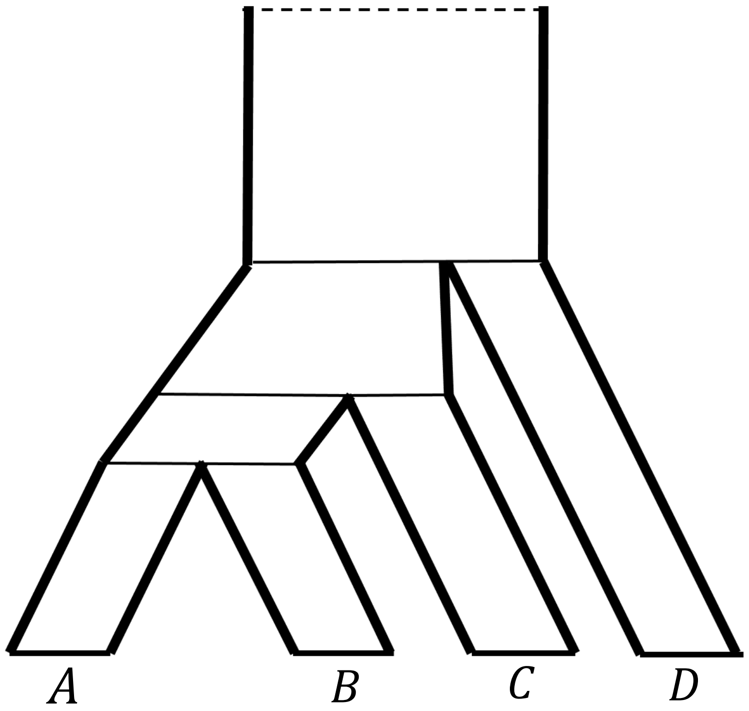
Our motivation

Can we design a branch length estimation method that…

- estimates branch lengths in substitution units (SU)
- addresses gene tree heterogeneity due to ILS and variation in mutation rates
- has strong theoretical foundation based on the MSC
- is scalable to large genome-wide datasets with hundreds to thousands of genes and species?

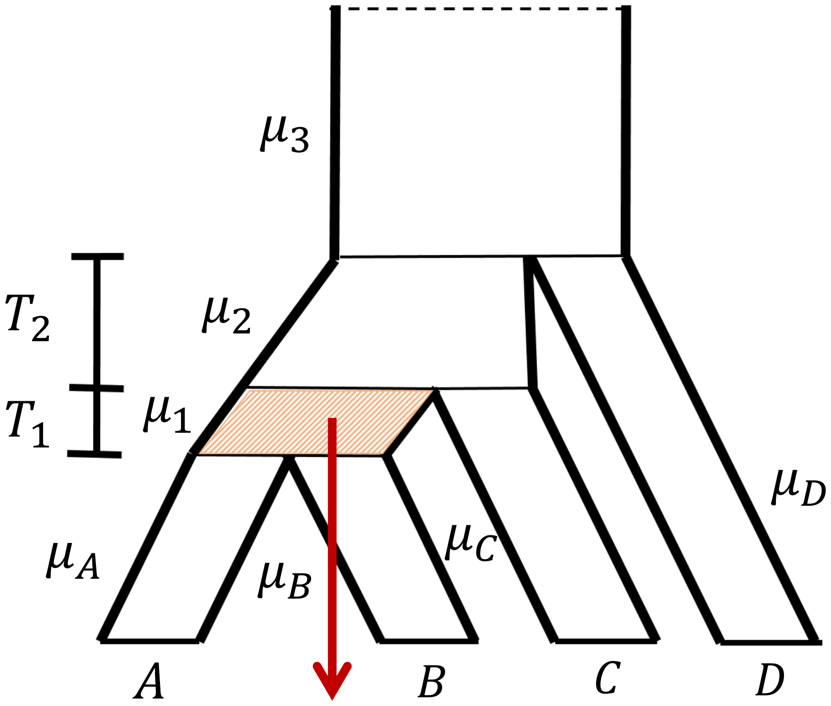
MSC+Substitution model

species tree S



MSC+Substitution model

species tree S

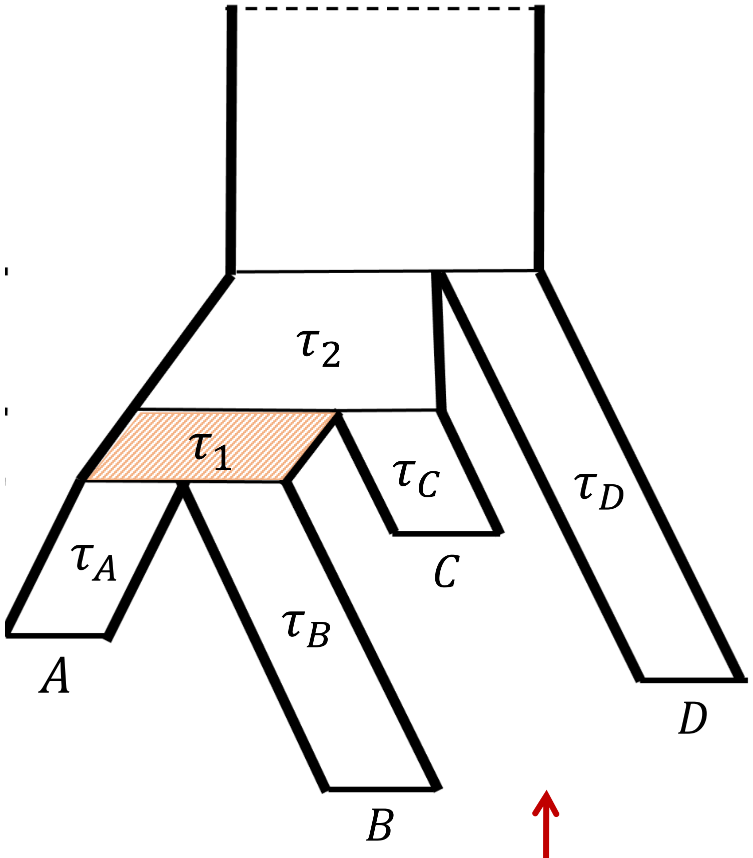


Parameters of branch 1:

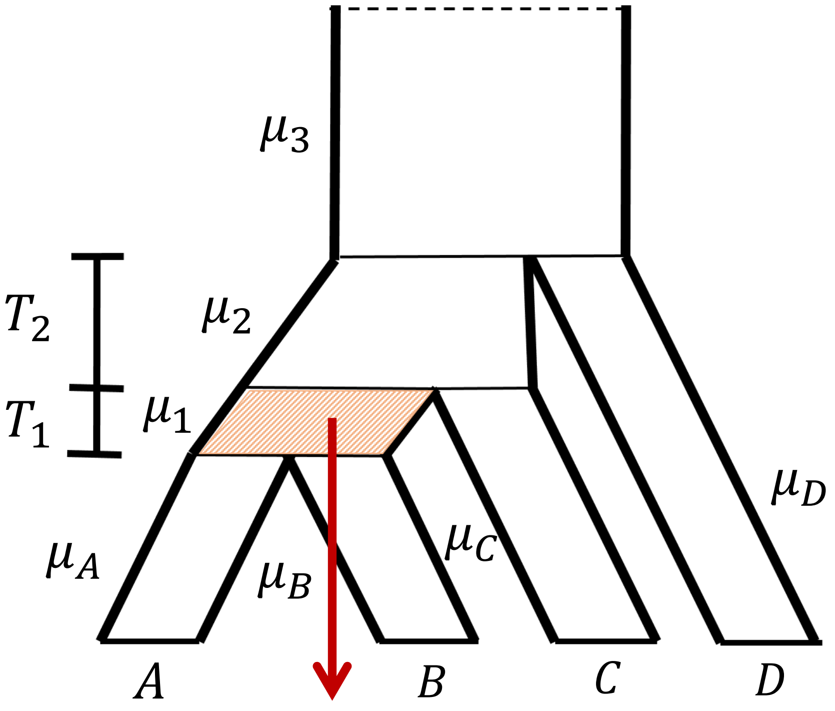
μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1.\mu_1$	Branch length in substitution units (SU)

MSC+Substitution model

S with SU branch lengths



species tree S

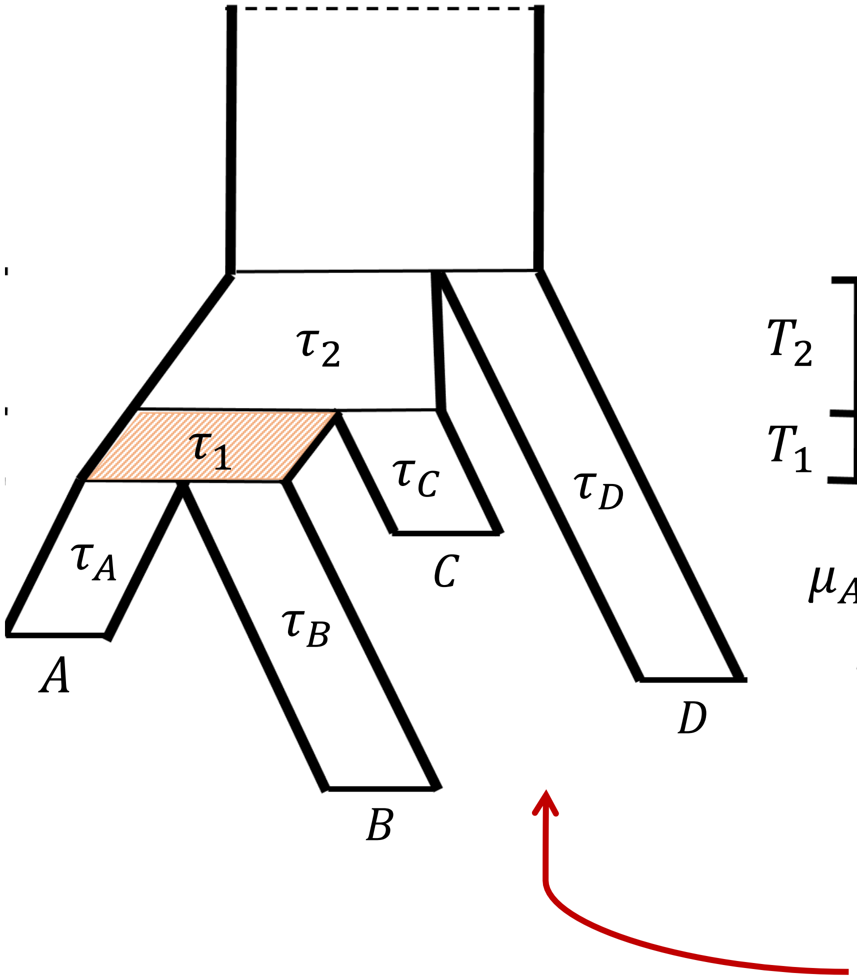


Parameters of branch 1:

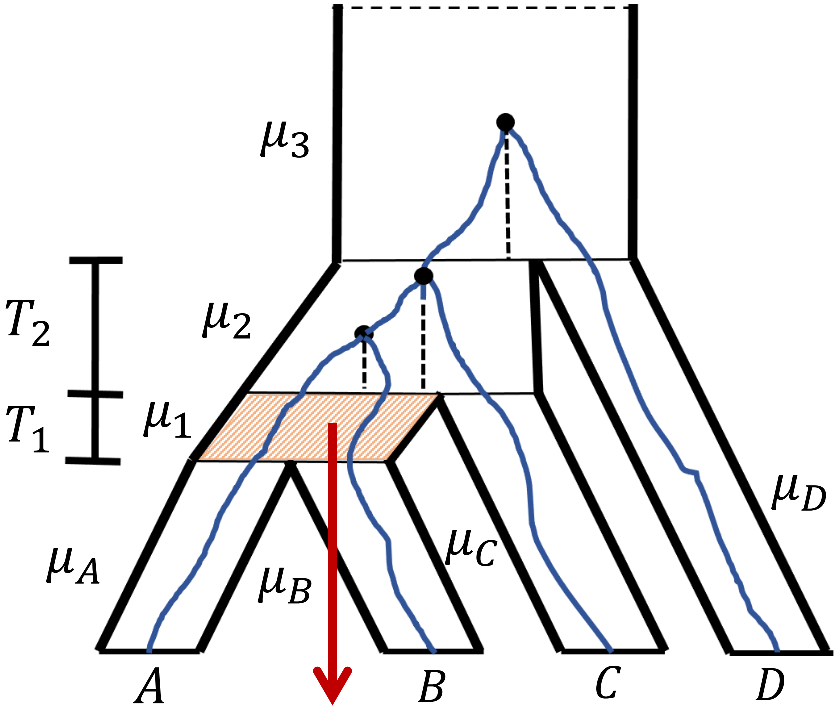
μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1.\mu_1$	Branch length in substitution units (SU)

MSC+Substitution model

S with SU branch lengths



species tree S

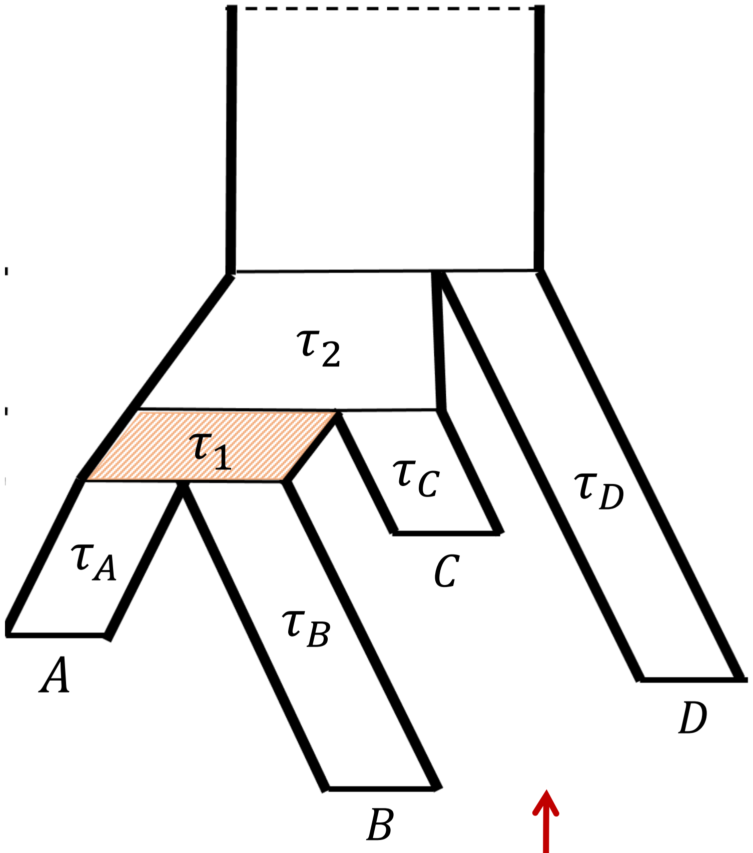


Parameters of branch 1:

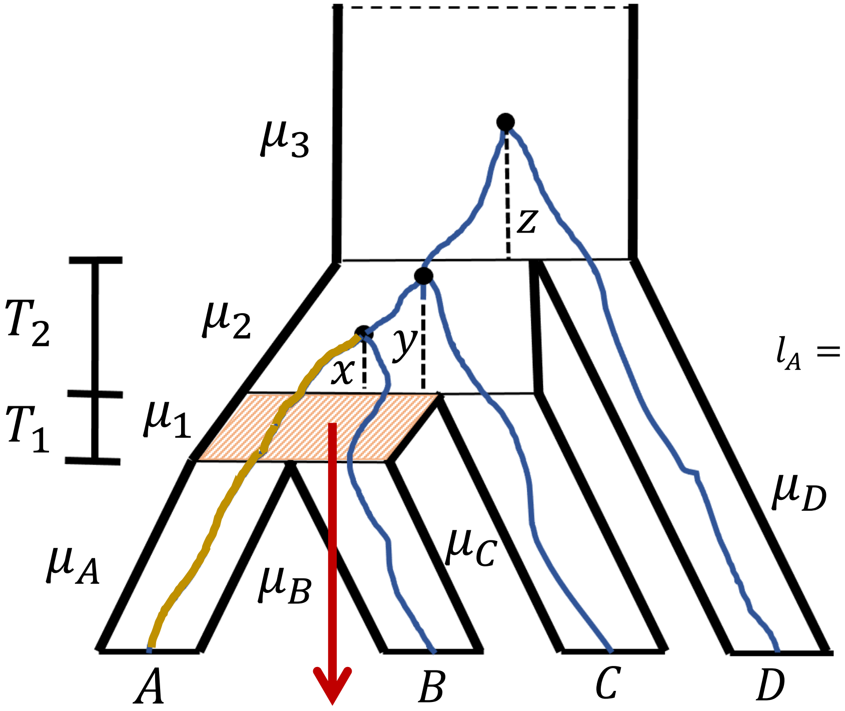
μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1 . \mu_1$	Branch length in substitution units (SU)

MSC+Substitution model

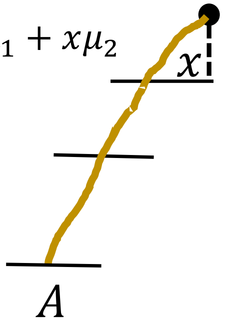
S with SU branch lengths



species tree S



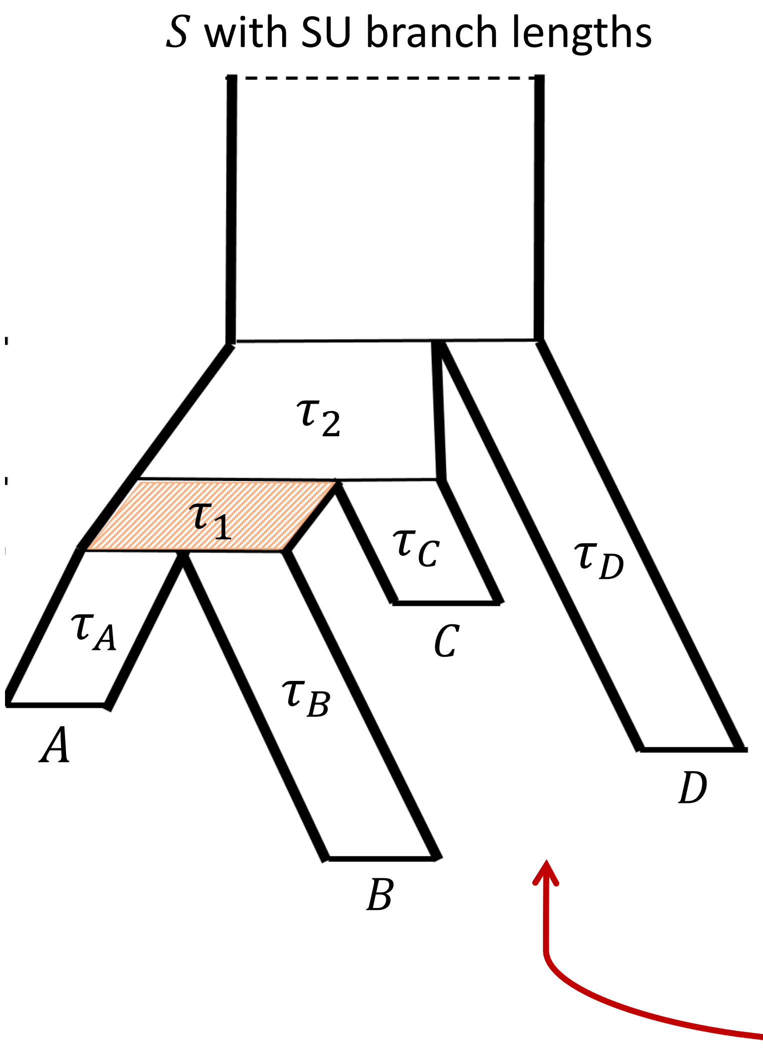
$l_A = T_A\mu_A + T_1\mu_1 + x\mu_2$

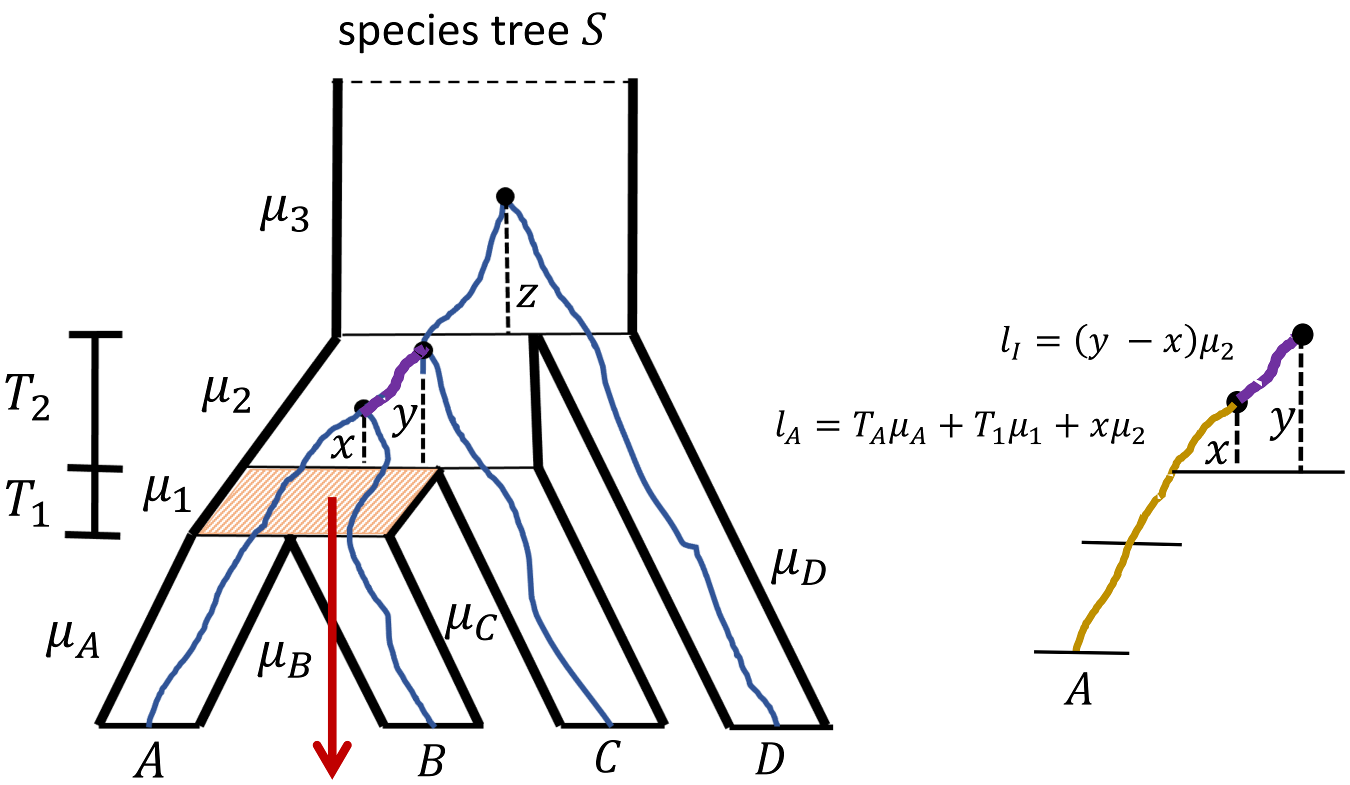


Parameters of branch 1:

μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1 . \mu_1$	Branch length in substitution units (SU)

MSC+Substitution model

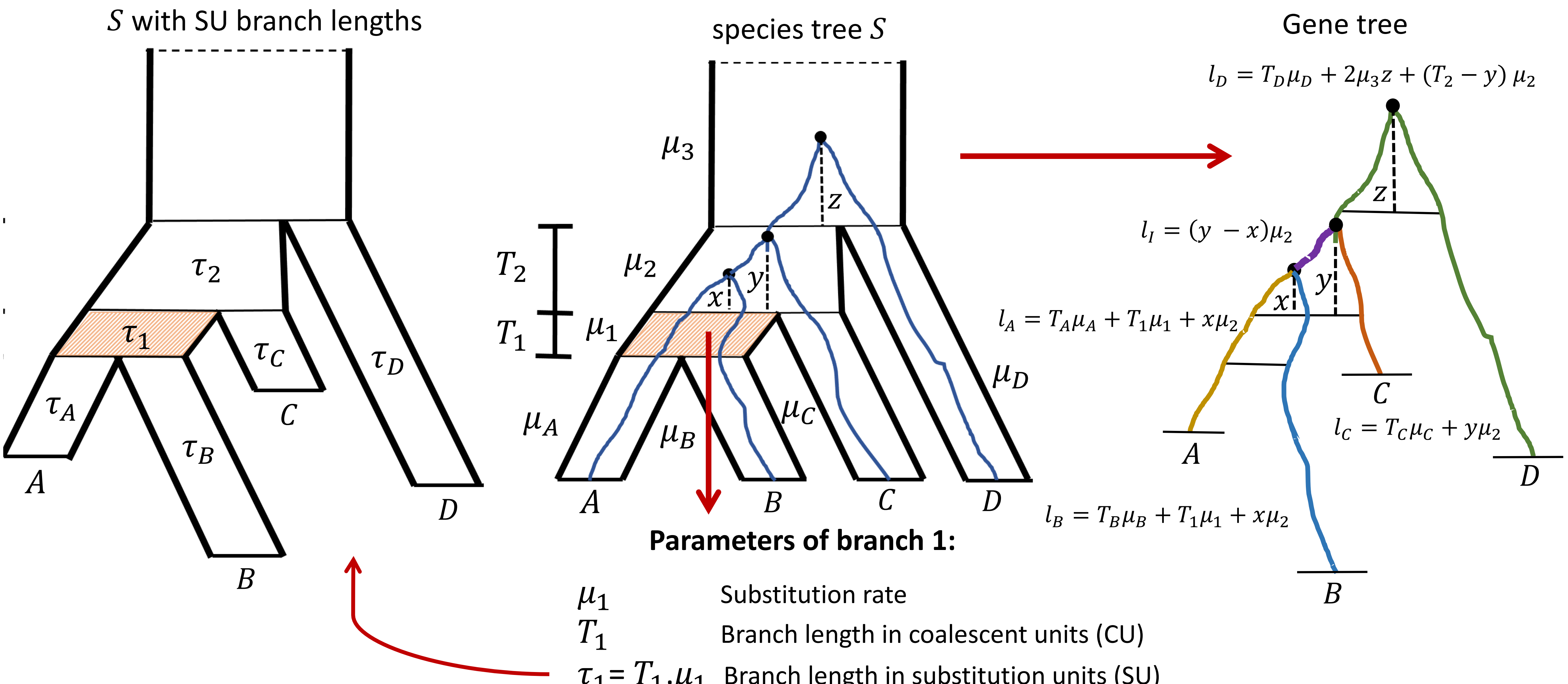




Parameters of branch 1:

μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1.\mu_1$	Branch length in substitution units (SU)

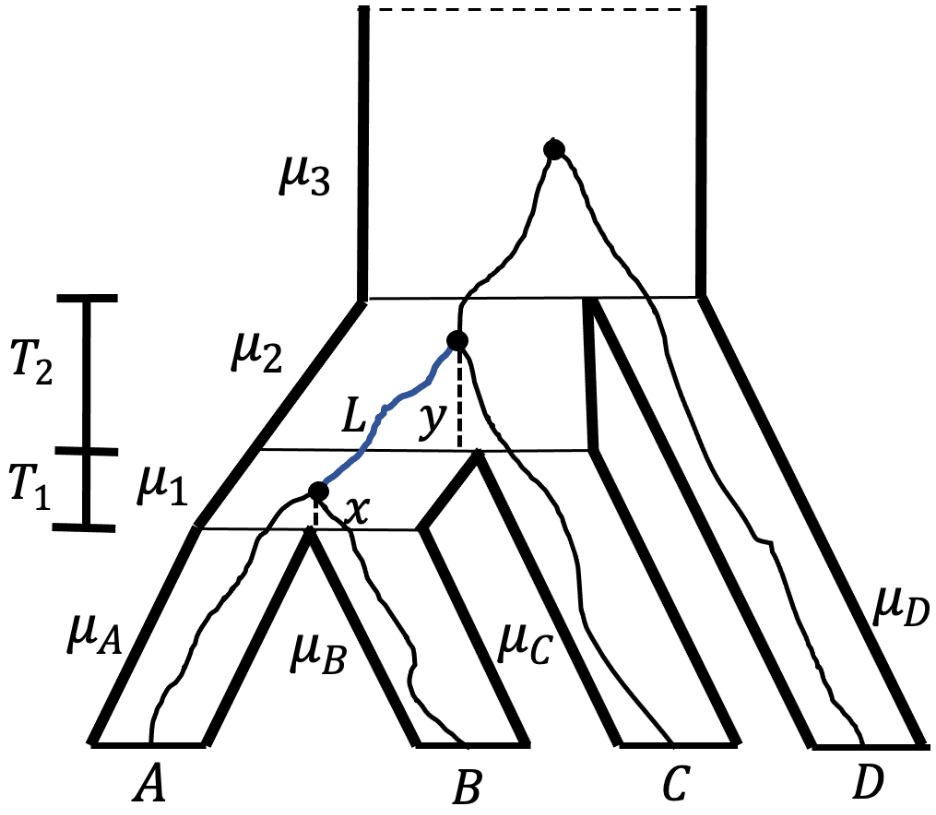
MSC+Substitution model



Parameters of branch 1:

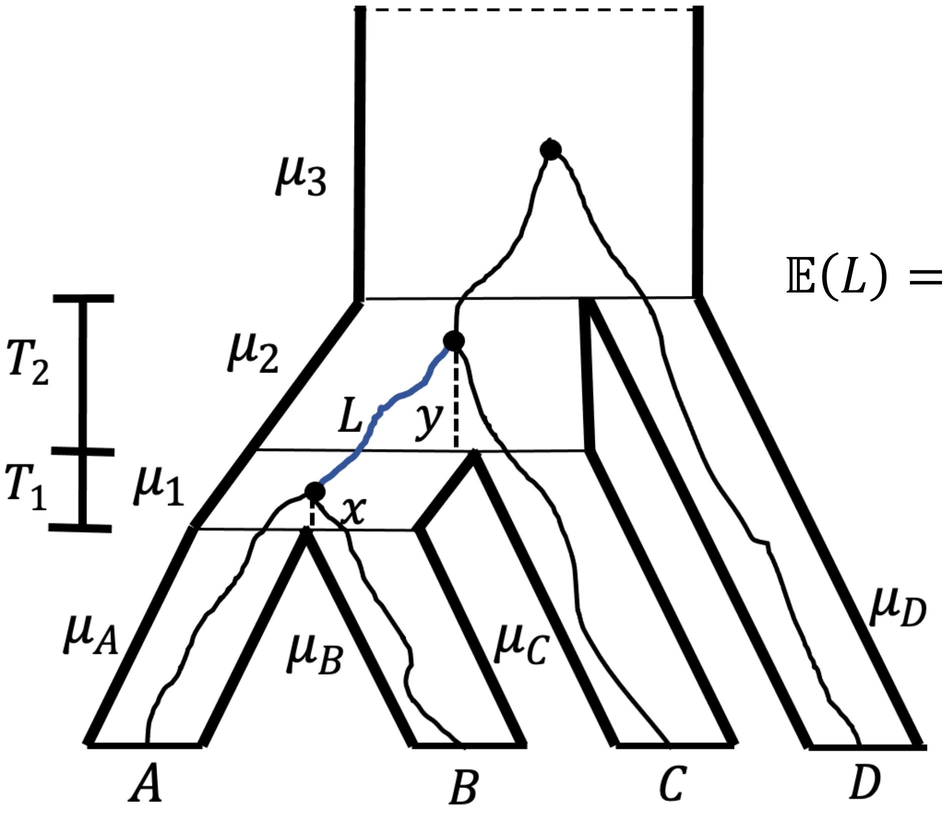
μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1.\mu_1$	Branch length in substitution units (SU)

Expected quartet branch lengths under MSC



$$L = (T_1 - x)\mu_1 + y\mu_2$$

Expected quartet branch lengths under MSC



$$L = (T_1 - x)\mu_1 + y\mu_2$$

$$\mathbb{E}(L) = \int_0^{T_1}\int_0^{T_2} e^{-x}\, e^{-y}\big((T_1 - x)\mu_1 + y\mu_2\big)dydx$$

↑

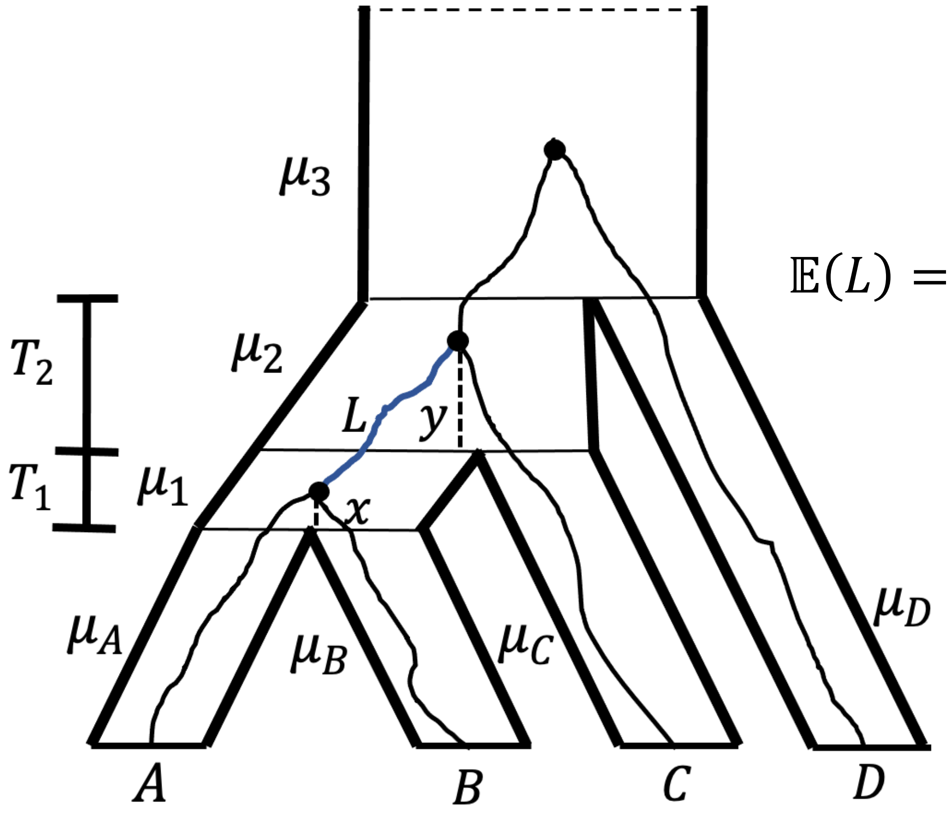
$k = 2$ lineages not coalescing in an interval with length x

- Under MSC, waiting times before coalescent events are exponential random variables with rate $\lambda = \binom{k}{2}$ where k is the number of lineages entering an interval

↓

$$f_X(x) = \binom{k}{2} e^{-\binom{k}{2}x}$$

Expected quartet branch lengths under MSC



$$L = (T_1 - x)\mu_1 + y\mu_2$$

$$\mathbb{E}(L) = \int_0^{T_1}\int_0^{T_2} e^{-x}\, e^{-y}\big((T_1 - x)\mu_1 + y\mu_2\big)dydx$$

↑

$k = 2$ lineages not coalescing in an interval with length x

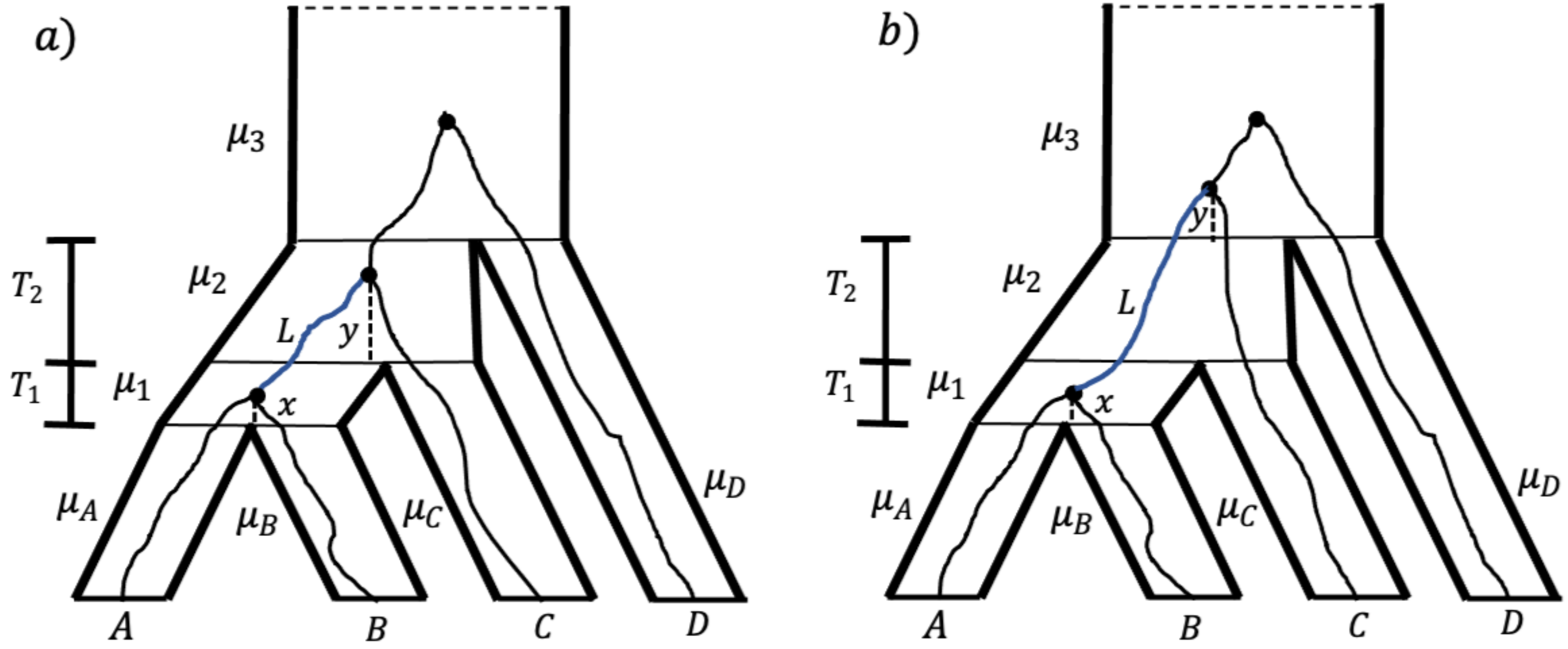
- Under MSC, waiting times before coalescent events are exponential random variables with rate $\lambda = \binom{k}{2}$ where k is the number of lineages entering an interval

↓

$$f_X(x) = \binom{k}{2} e^{-\binom{k}{2}x}$$

What about other patterns of coalescence?

What about other patterns of coalescence?

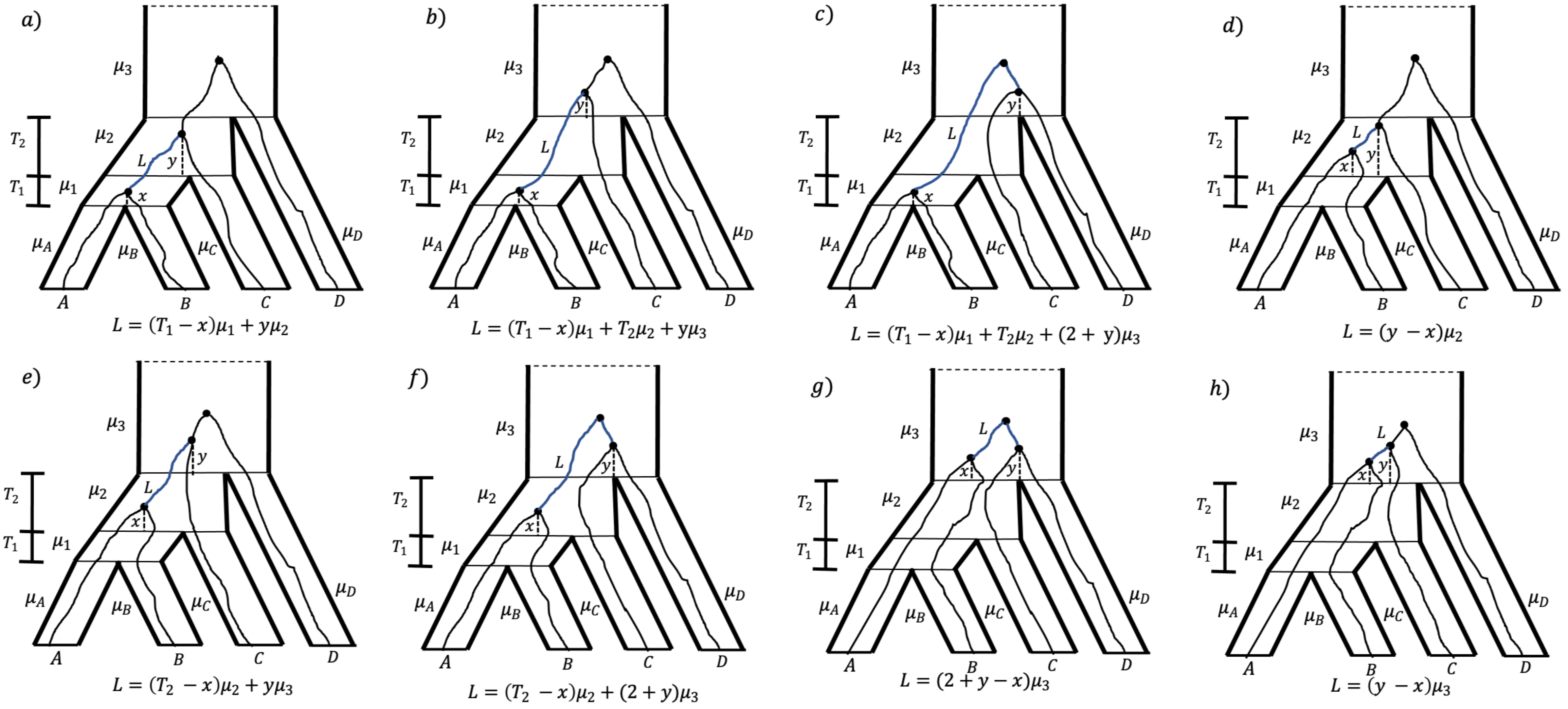


$$L = (T_1 - x)\mu_1 + y\mu_2$$

$$L = (T_1 - x)\mu_1 + T_2\mu_2 + y\mu_3$$

different patterns → different expected lengths

Scenarios for gene tree matching the unbalanced species tree



$L = (T_1 - x)\mu_1 + y\mu_2$

$L = (T_1 - x)\mu_1 + T_2\mu_2 + y\mu_3$

$L = (T_1 - x)\mu_1 + T_2\mu_2 + (2 + y)\mu_3$

$L = (y - x)\mu_2$

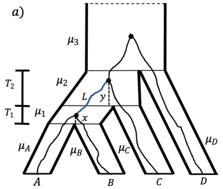
$L = (T_2 - x)\mu_2 + y\mu_3$

$L = (T_2 - x)\mu_2 + (2 + y)\mu_3$

$L = (2 + y - x)\mu_3$

$L = (y - x)\mu_3$

Expected quartet branch lengths under MSC

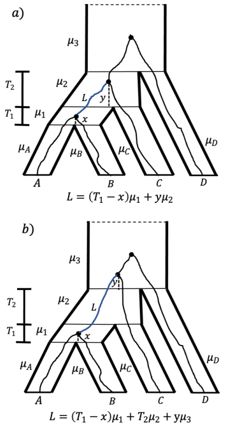


$L = (T_1 - x)\mu_1 + y\mu_2$

$$L_I = (\int_0^{T_1}\int_0^{T_2} e^{-x}\, e^{-y}\big((T_1 - x)\mu_1 + y\mu_2\big)dydx$$

scenario (a)

Expected quartet branch lengths under MSC



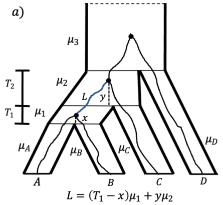
$$L_I = (\int_0^{T_1} \int_0^{T_2} e^{-x}\, e^{-y}\big((T_1 - x)\mu_1 + y\mu_2\big)dydx$$

scenario (a)

$$+2e^{-T_2} \int_0^{T_1} \int_0^{\infty} e^{-x}\, e^{-3y}\big((T_1 - x)\mu_1 + T_2\mu_2 + y\mu_3\big)dydx$$

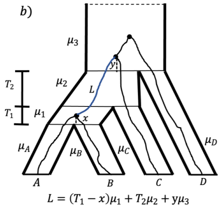
scenario (b)

Expected quartet branch lengths under MSC



$$L_I = (\int_0^{T_1} \int_0^{T_2} e^{-x}\, e^{-y} \big((T_1 - x)\mu_1 + y\mu_2\big) dydx$$

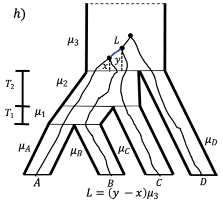
scenario (a)



$$+2e^{-T_2} \int_0^{T_1} \int_0^{\infty} e^{-x}\, e^{-3y} \big((T_1 - x)\mu_1 + T_2\mu_2 + y\mu_3\big) dydx$$

scenario (b)

$$\vdots$$



$$+4e^{-T_1} e^{-3T_2} \int_0^{\infty} \int_x^{\infty} e^{-6x}\, e^{-3(y-x)} (y - x)\mu_3 dydx) \,/ (1 - \frac{2}{3} e^{-T_1})$$

scenario (h)

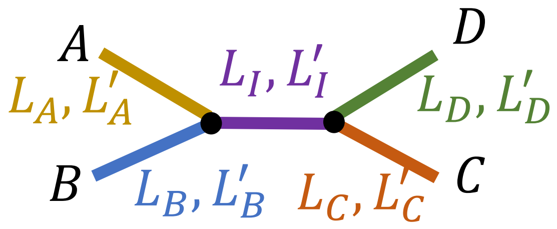
Expected value of internal branch length conditioned on gene tree matching the species tree

$$= \boxed{\frac{(e^{-3T_2} + 3e^{-T_2} - 6e^{T_1 - T_2})(\mu_2 - \mu_3) + 6(1 - e^{T_1} + T_1 e^{T_1})\mu_1}{2(3e^{T_1} - 2)} + \mu_2}$$

How do we infer the species tree parameters?

- We derive expected values for all branches (internal and terminal), for both matching and non-matching gene trees.

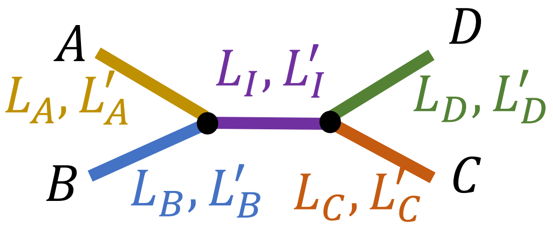
- The parameters of the species tree can be estimated from these 10 equations.



How do we infer the species tree parameters?

- We derive expected values for all branches (internal and terminal), for both matching and non-matching gene trees.

- The parameters of the species tree can be estimated from these 10 equations.



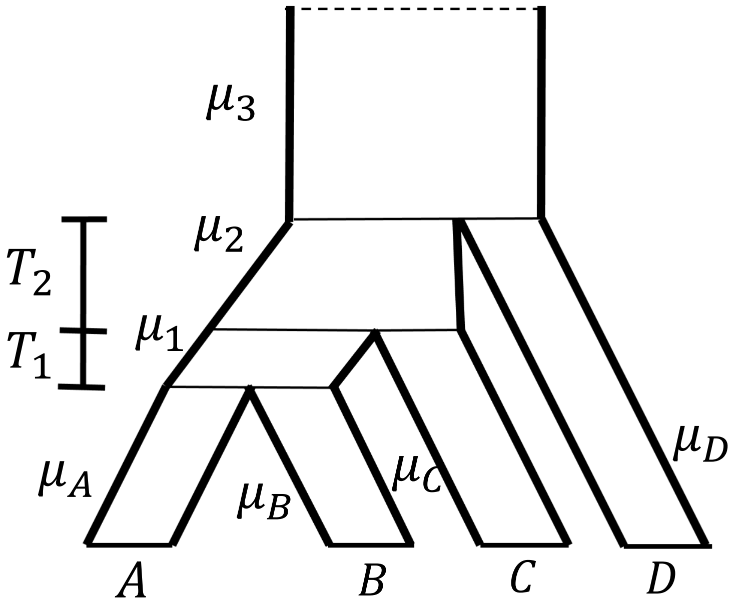
- **Challenge**: Solving these systems of equations directly can cause numerical instabilities and may not produce optimal solutions.

- We use simplifications that give analytical formulas for every branch of a quartet tree.

Simplifications (example)

Expected lengths for internal branch

$$\begin{cases} L_I = \dfrac{(e^{-3T_2} + 3e^{-T_2} - 6e^{T_1-T_2})(\mu_2 - \mu_3) + 6(1\ - e^{T_1} + T_1 e^{T_1})\mu_1}{2(3e^{T_1} - 2)} + \mu_2 \\ L_I' = \mu_2 + \frac{1}{2}(\mu_2 - \mu_3)\ (e^{-3T_2} - e^{-T_2}) \end{cases}$$

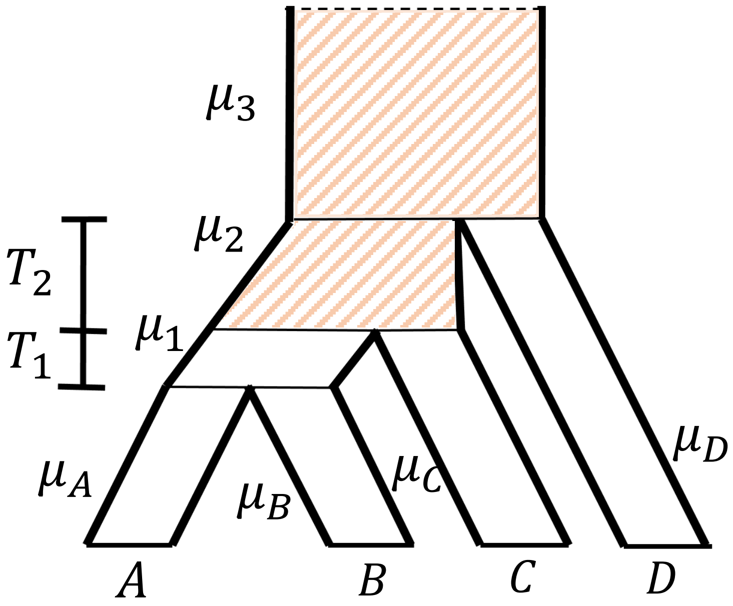


Simplifications (example)

Expected lengths for internal branch

$$\begin{cases} L_I = \dfrac{(e^{-3T_2} + 3e^{-T_2} - 6e^{T_1 - T_2})(\mu_2 - \mu_3) + 6(1 - e^{T_1} + T_1 e^{T_1})\mu_1}{2(3e^{T_1} - 2)} + \mu_2 \\ L_I' = \mu_2 + \frac{1}{2}(\mu_2 - \mu_3)\,(e^{-3T_2} - e^{-T_2}) \end{cases}$$

Local clock assumption: $\mu_2 = \mu_3$

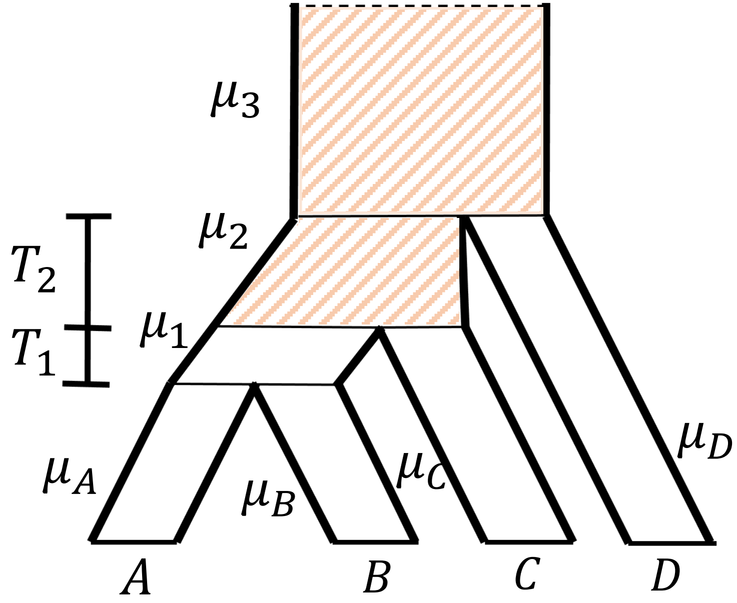


Simplifications (example)

Expected lengths for internal branch

$$L_I = \frac{(e^{-3T_2} + \cancel{3e^{-T_2} - 6e^{T_1-T_2})(\mu_2 - \mu_3}) + 6(1 - e^{T_1} + T_1 e^{T_1})\mu_1}{2(3e^{T_1} - 2)} + \mu_2$$

$$L_I' = \mu_2 + \frac{1}{2}\cancel{(\mu_2 - \mu_3)\,(e^{-3T_2} - e^{-T_2})}$$



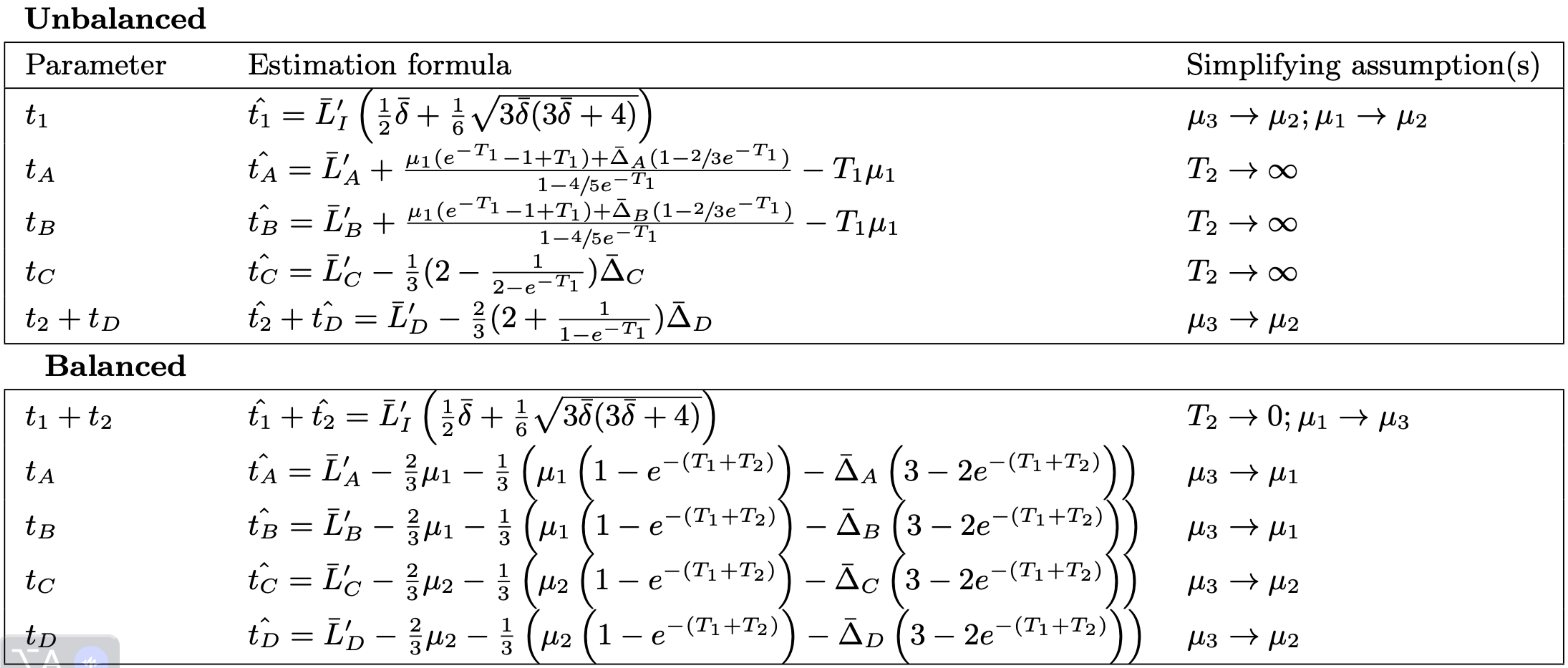
Local clock assumption: $\mu_2 = \mu_3$

Simplified formulas

$$\lim_{\mu_2 \to \mu_3} L_I = \frac{3\mu_1(e^{-T_1} - 1 + T_1)}{3 - 2e^{-T_1}} + \mu_2$$

$$\lim_{\mu_2 \to \mu_3} L_I' = \mu_2$$

Summary of SU branch length formulas



Unbalanced

Parameter	Estimation formula	Simplifying assumption(s)
t_1	$\hat{t_1} = \bar{L}'_I \left(\frac{1}{2}\bar{\delta} + \frac{1}{6}\sqrt{3\bar{\delta}(3\bar{\delta}+4)}\right)$	$\mu_3 \to \mu_2; \mu_1 \to \mu_2$
t_A	$\hat{t_A} = \bar{L}'_A + \frac{\mu_1(e^{-T_1}-1+T_1)+\bar{\Delta}_A(1-2/3e^{-T_1})}{1-4/5e^{-T_1}} - T_1\mu_1$	$T_2 \to \infty$
t_B	$\hat{t_B} = \bar{L}'_B + \frac{\mu_1(e^{-T_1}-1+T_1)+\bar{\Delta}_B(1-2/3e^{-T_1})}{1-4/5e^{-T_1}} - T_1\mu_1$	$T_2 \to \infty$
t_C	$\hat{t_C} = \bar{L}'_C - \frac{1}{3}(2 - \frac{1}{2-e^{-T_1}})\bar{\Delta}_C$	$T_2 \to \infty$
$t_2 + t_D$	$\hat{t_2} + \hat{t_D} = \bar{L}'_D - \frac{2}{3}(2 + \frac{1}{1-e^{-T_1}})\bar{\Delta}_D$	$\mu_3 \to \mu_2$

Balanced

$t_1 + t_2$	$\hat{t_1} + \hat{t_2} = \bar{L}'_I \left(\frac{1}{2}\bar{\delta} + \frac{1}{6}\sqrt{3\bar{\delta}(3\bar{\delta}+4)}\right)$	$T_2 \to 0; \mu_1 \to \mu_3$
t_A	$\hat{t_A} = \bar{L}'_A - \frac{2}{3}\mu_1 - \frac{1}{3}\left(\mu_1\left(1-e^{-(T_1+T_2)}\right) - \bar{\Delta}_A\left(3-2e^{-(T_1+T_2)}\right)\right)$	$\mu_3 \to \mu_1$
t_B	$\hat{t_B} = \bar{L}'_B - \frac{2}{3}\mu_1 - \frac{1}{3}\left(\mu_1\left(1-e^{-(T_1+T_2)}\right) - \bar{\Delta}_B\left(3-2e^{-(T_1+T_2)}\right)\right)$	$\mu_3 \to \mu_1$
t_C	$\hat{t_C} = \bar{L}'_C - \frac{2}{3}\mu_2 - \frac{1}{3}\left(\mu_2\left(1-e^{-(T_1+T_2)}\right) - \bar{\Delta}_C\left(3-2e^{-(T_1+T_2)}\right)\right)$	$\mu_3 \to \mu_2$
t_D	$\hat{t_D} = \bar{L}'_D - \frac{2}{3}\mu_2 - \frac{1}{3}\left(\mu_2\left(1-e^{-(T_1+T_2)}\right) - \bar{\Delta}_D\left(3-2e^{-(T_1+T_2)}\right)\right)$	$\mu_3 \to \mu_2$

CASTLES

Coalescent-**A**ware **S**pecies **T**ree **L**ength **E**stimation in **S**ubstitution-units

Input:

- Rooted species tree *topology* S
- A set of gene trees $\mathcal{G}$ with SU branch lengths

Output:

- Species tree S with SU branch lengths

CASTLES

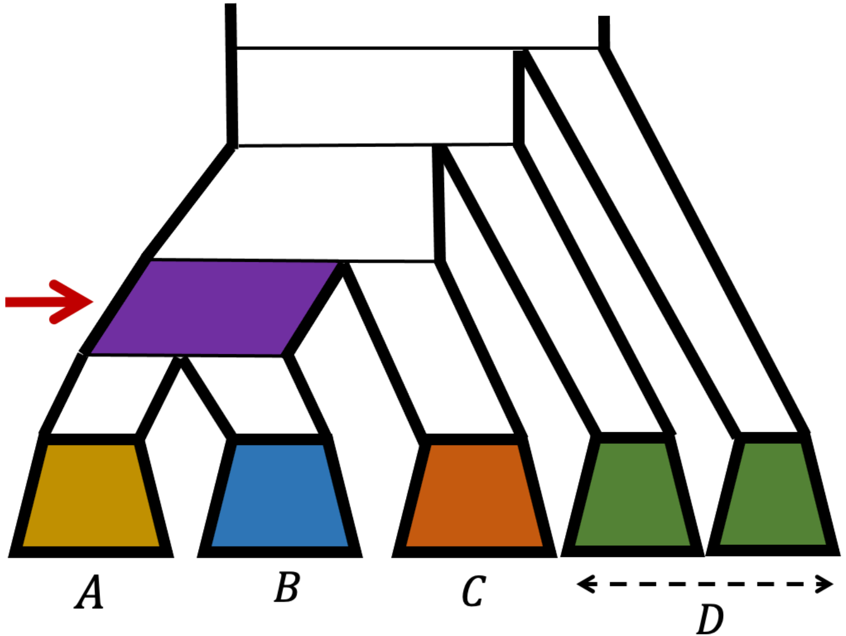
Coalescent-**A**ware **S**pecies **T**ree **L**ength **E**stimation in **S**ubstitution-units

Input:

- Rooted species tree *topology* S
- A set of gene trees $\mathcal{G}$ with SU branch lengths

Output:

- Species tree S with SU branch lengths



Quadripartition $AB|CD$

CASTLES

Coalescent-**A**ware **S**pecies **T**ree **L**ength **E**stimation in **S**ubstitution-units

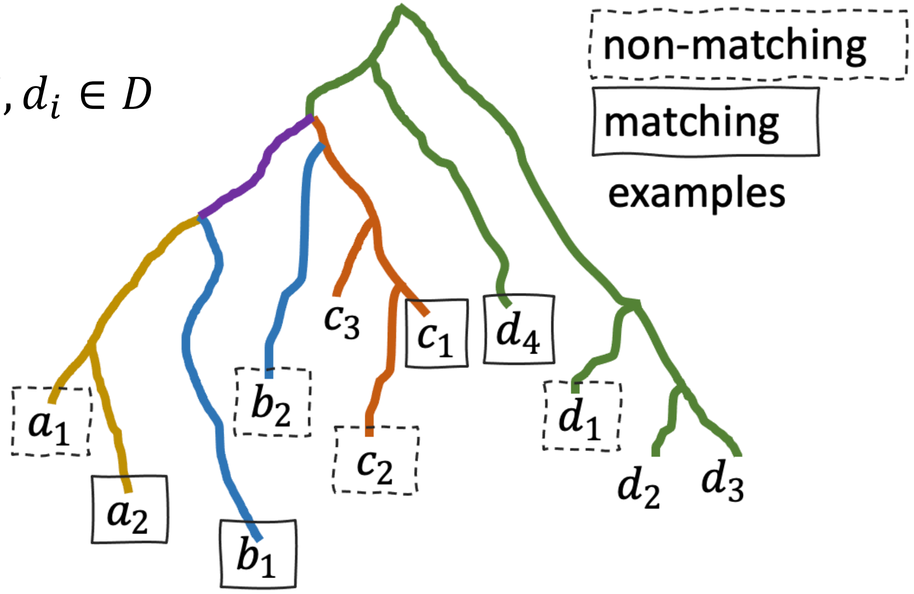
Input:

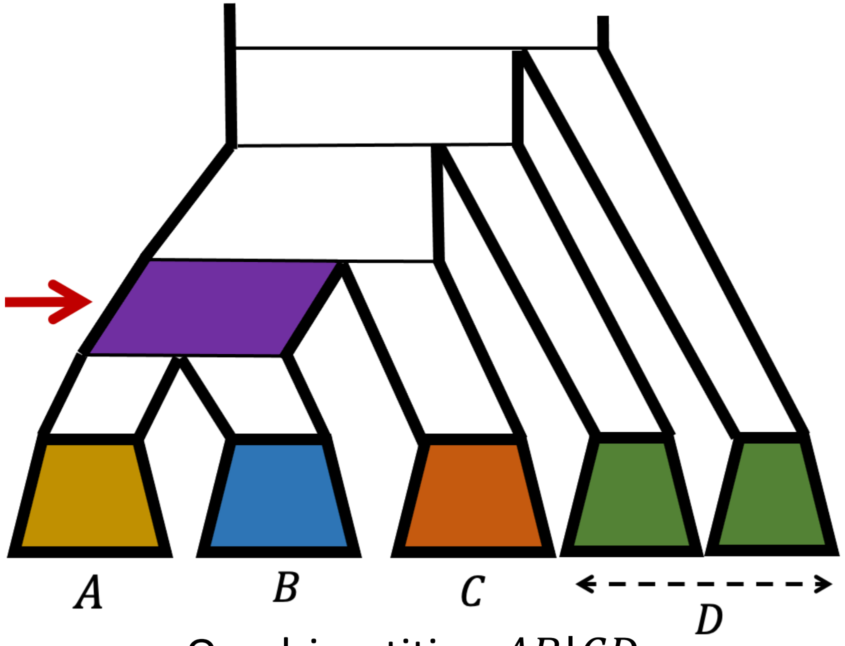
- Rooted species tree *topology* S
- A set of gene trees $\mathcal{G}$ with SU branch lengths

Output:

- Species tree S with SU branch lengths

Quartets: $a_i \in A, b_i \in B, c_i \in C, d_i \in D$





Quadripartition $AB|CD$

CASTLES

Coalescent-**A**ware **S**pecies **T**ree **L**ength **E**stimation in **S**ubstitution-units

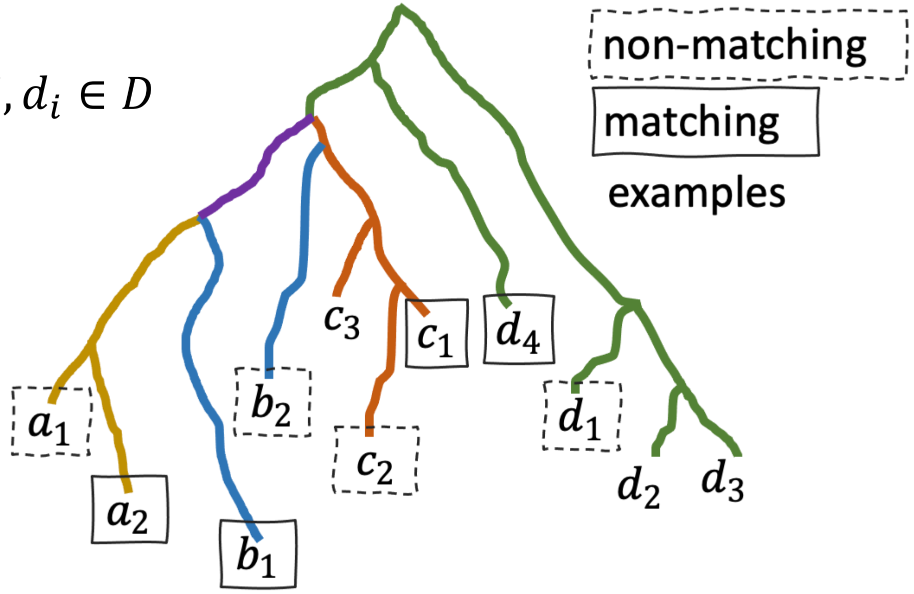
Input:
- Rooted species tree *topology* S
- A set of gene trees $\mathcal{G}$ with SU branch lengths

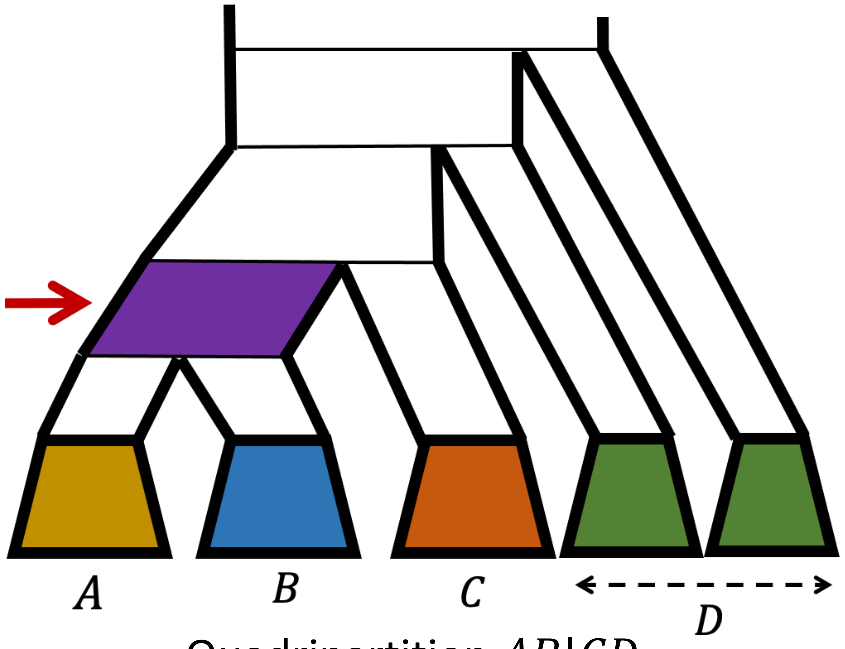
Output:
- Species tree S with SU branch lengths

Quartets: $a_i \in A, b_i \in B, c_i \in C, d_i \in D$

- We average branch lengths over all quartets with an $O(n^2k)$ dynamic programming

n species, k genes

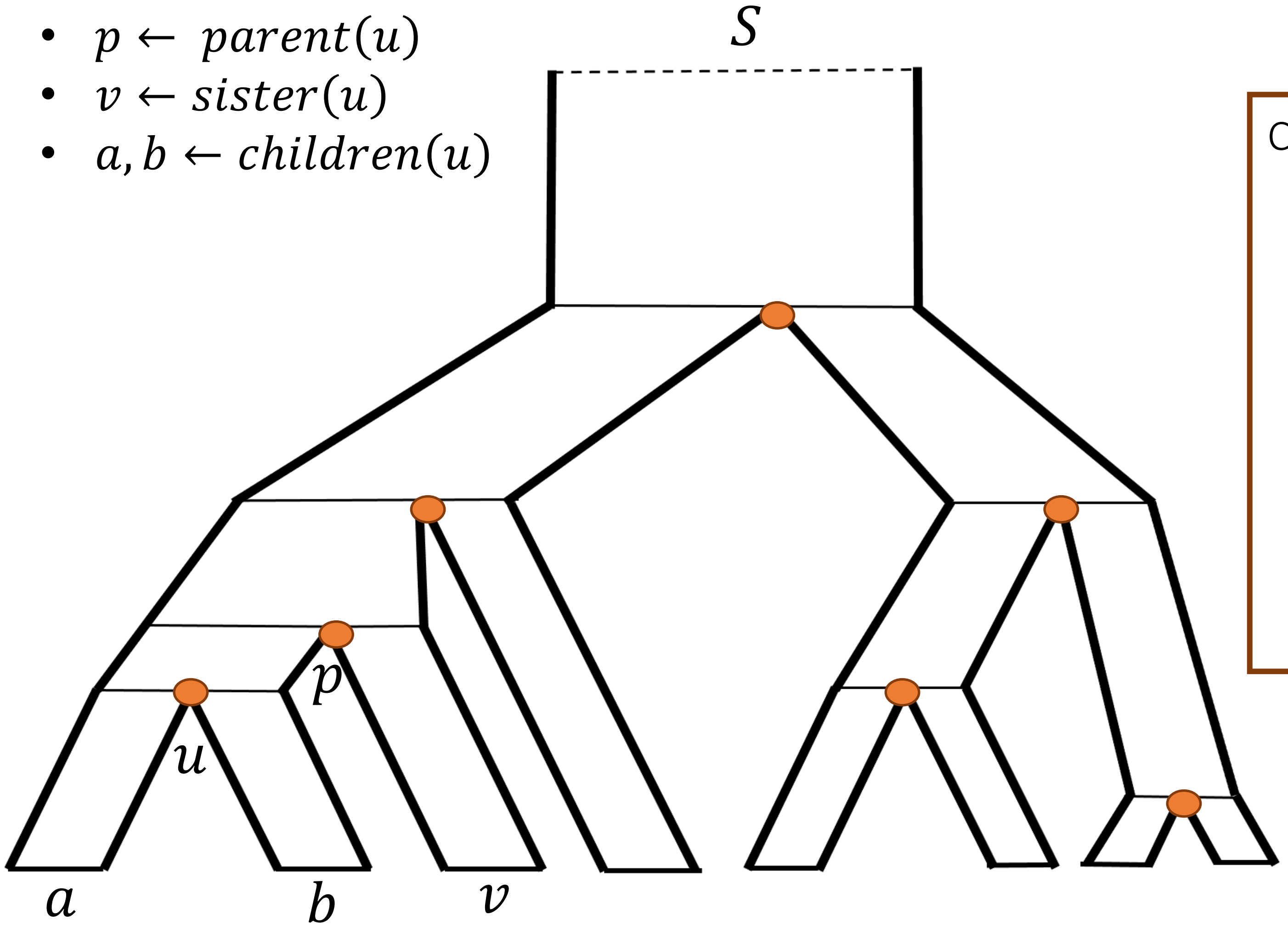




Quadripartition $AB|CD$

Large tree algorithm

- $p \leftarrow parent(u)$
- $v \leftarrow sister(u)$
- $a, b \leftarrow children(u)$

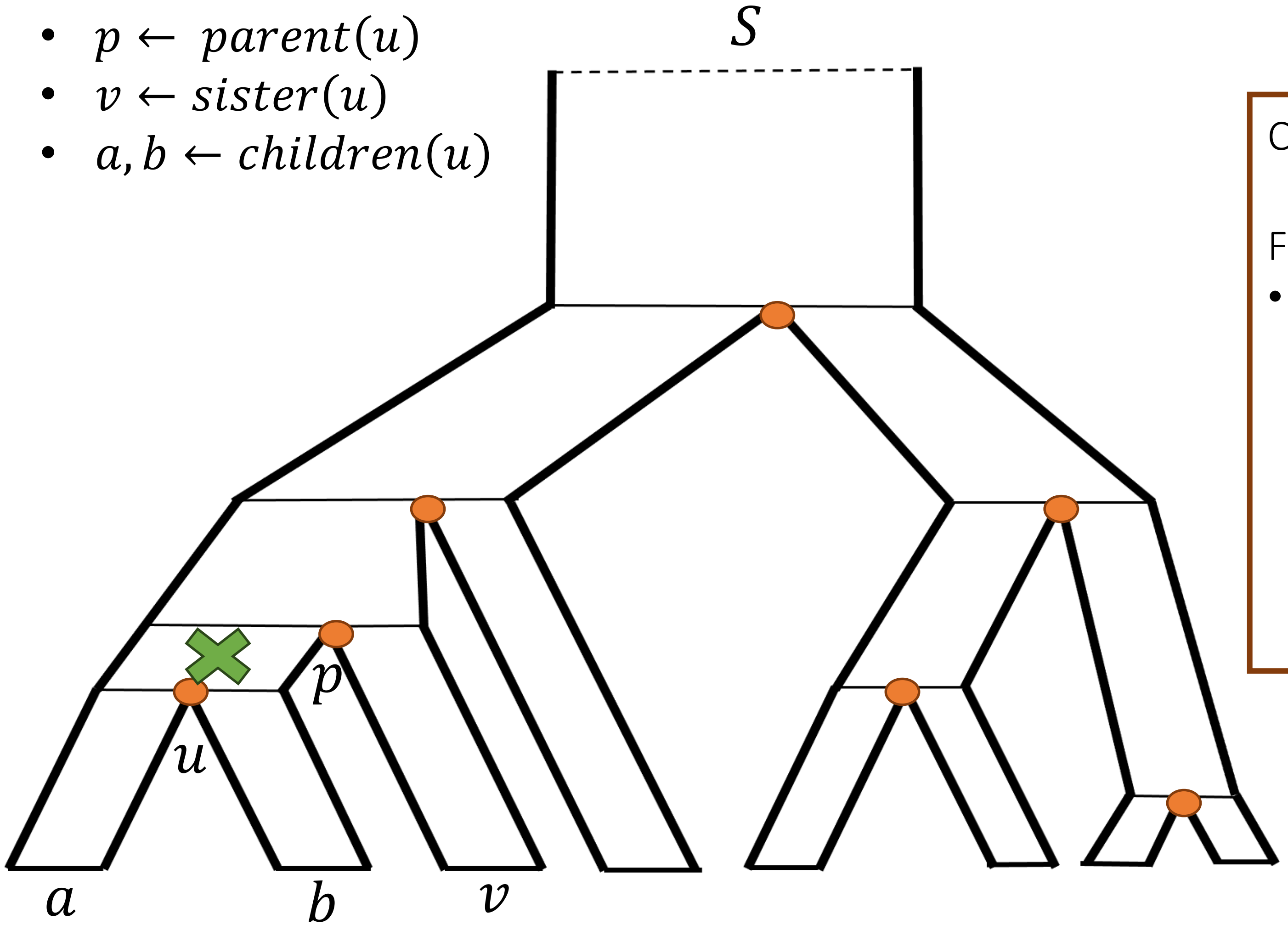


$O(n^2k)$ dynamic programming

Calculate $\overline{L_a}, \overline{L_b}, \overline{L_v}, \overline{L_p}, \overline{L'_a}, \overline{L'_b}, \overline{L'_v}, \overline{L'_p}$ for each branch

Large tree algorithm

- $p \leftarrow parent(u)$
- $v \leftarrow sister(u)$
- $a, b \leftarrow children(u)$



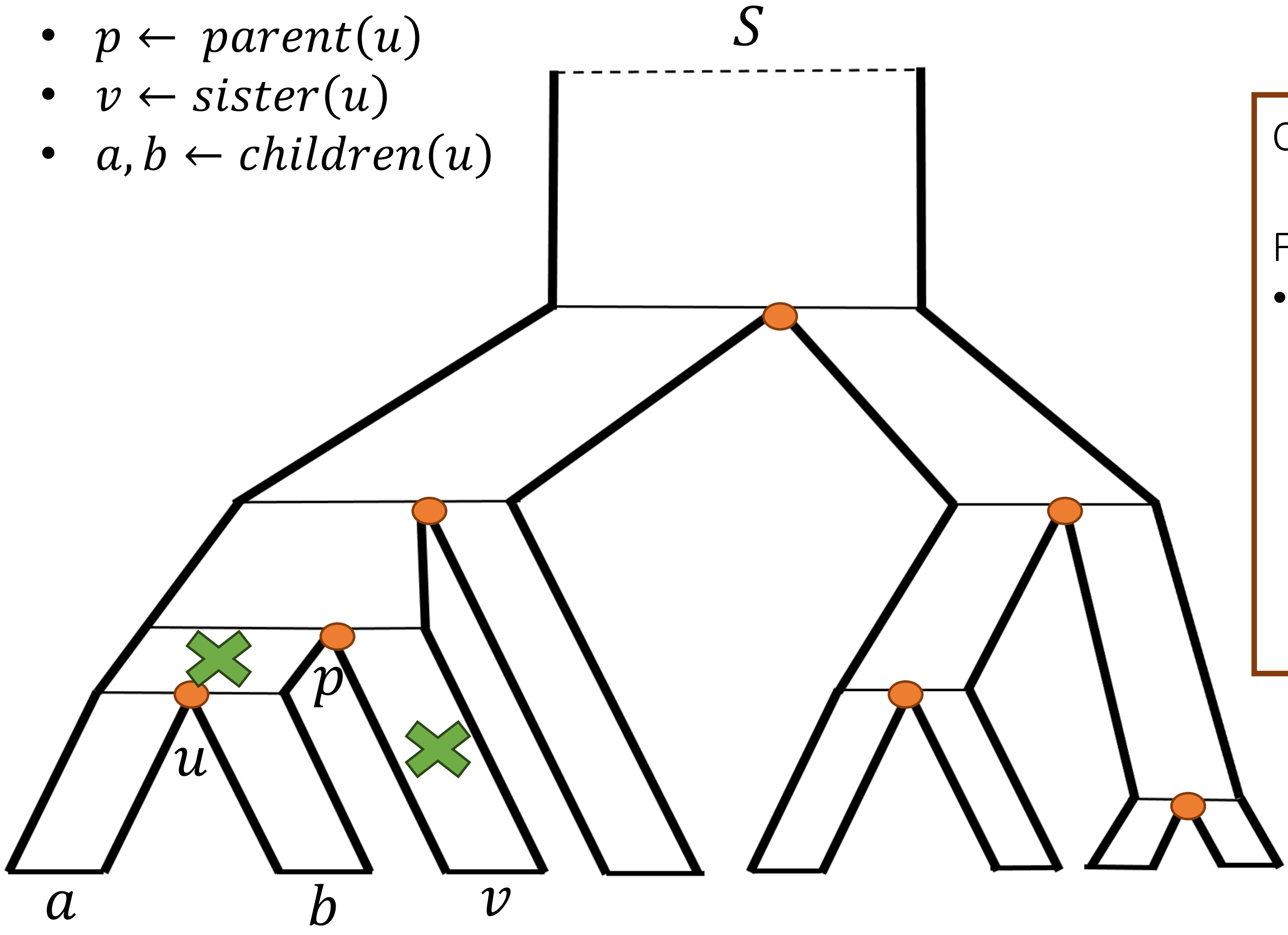
$O(n^2k)$ dynamic programming

```
Calculate $\overline{L_a}, \overline{L_b}, \overline{L_v}, \overline{L_p}, \overline{L'_a}, \overline{L'_b}, \overline{L'_v}, \overline{L'_p}$ for each branch

For $u$ ∈ post order traversal of internal nodes of $S$:
    • if $p$ is not root:
        • $t_{p\to u}$ ← internal branch equation
```

Large tree algorithm

- $p \leftarrow\ parent(u)$
- $v \leftarrow sister(u)$
- $a, b \leftarrow children(u)$



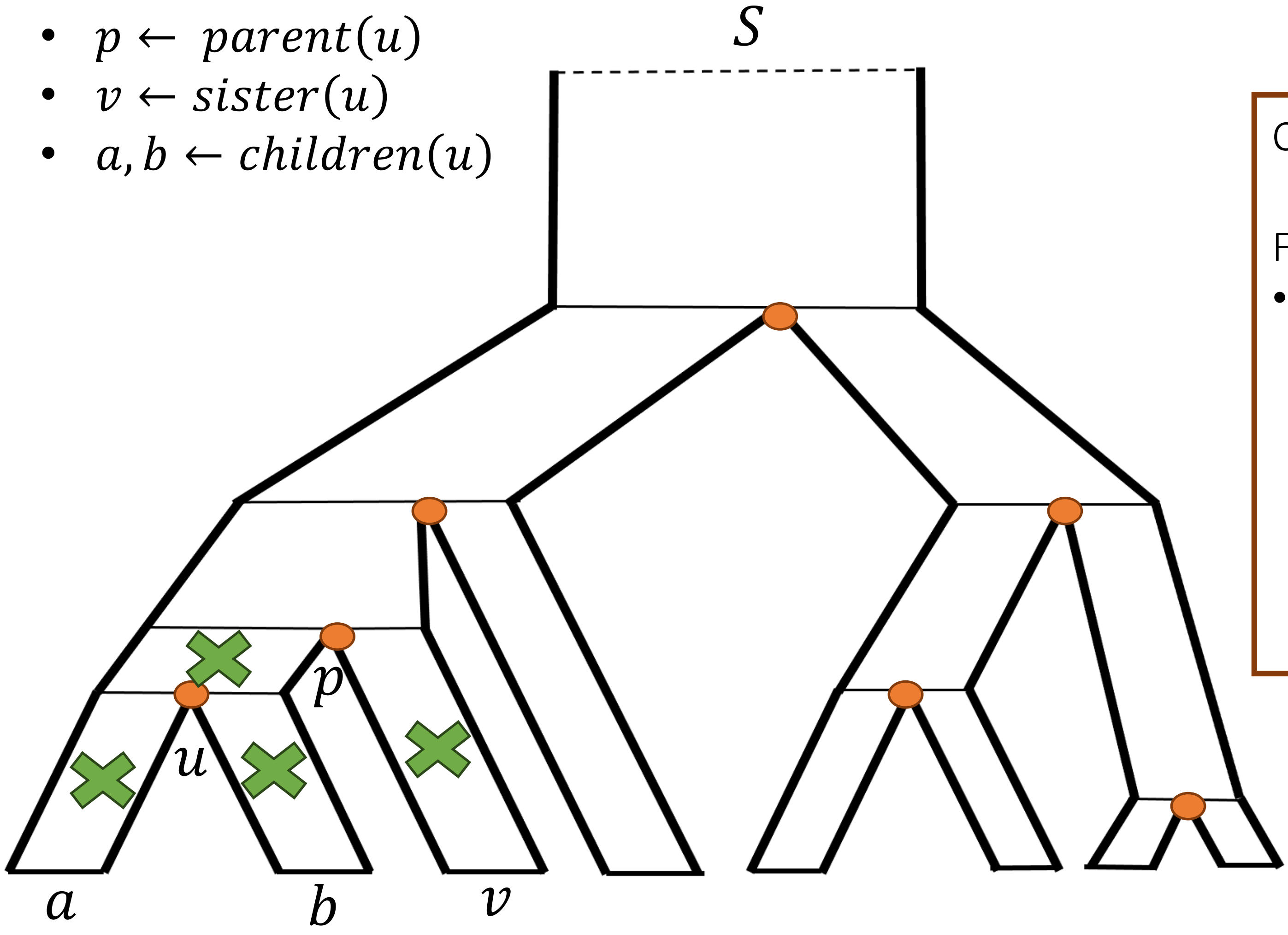
$O(n^2k)$ dynamic programming

Calculate $\overline{L_a}, \overline{L_b}, \overline{L_v}, \overline{L_p}, \overline{L'_a}, \overline{L'_b}, \overline{L'_v}, \overline{L'_p}$ for each branch

```
For $u$ ∈ post order traversal of internal nodes of $S$:
• if $p$ is not root:
    • $t_{p\to u}$ ← internal branch equation
    • If $v$ is leaf:
        $t_{p\to v}$ ← terminal middle branch equation
```

Large tree algorithm

- $p \leftarrow\ parent(u)$
- $v \leftarrow sister(u)$
- $a, b \leftarrow children(u)$



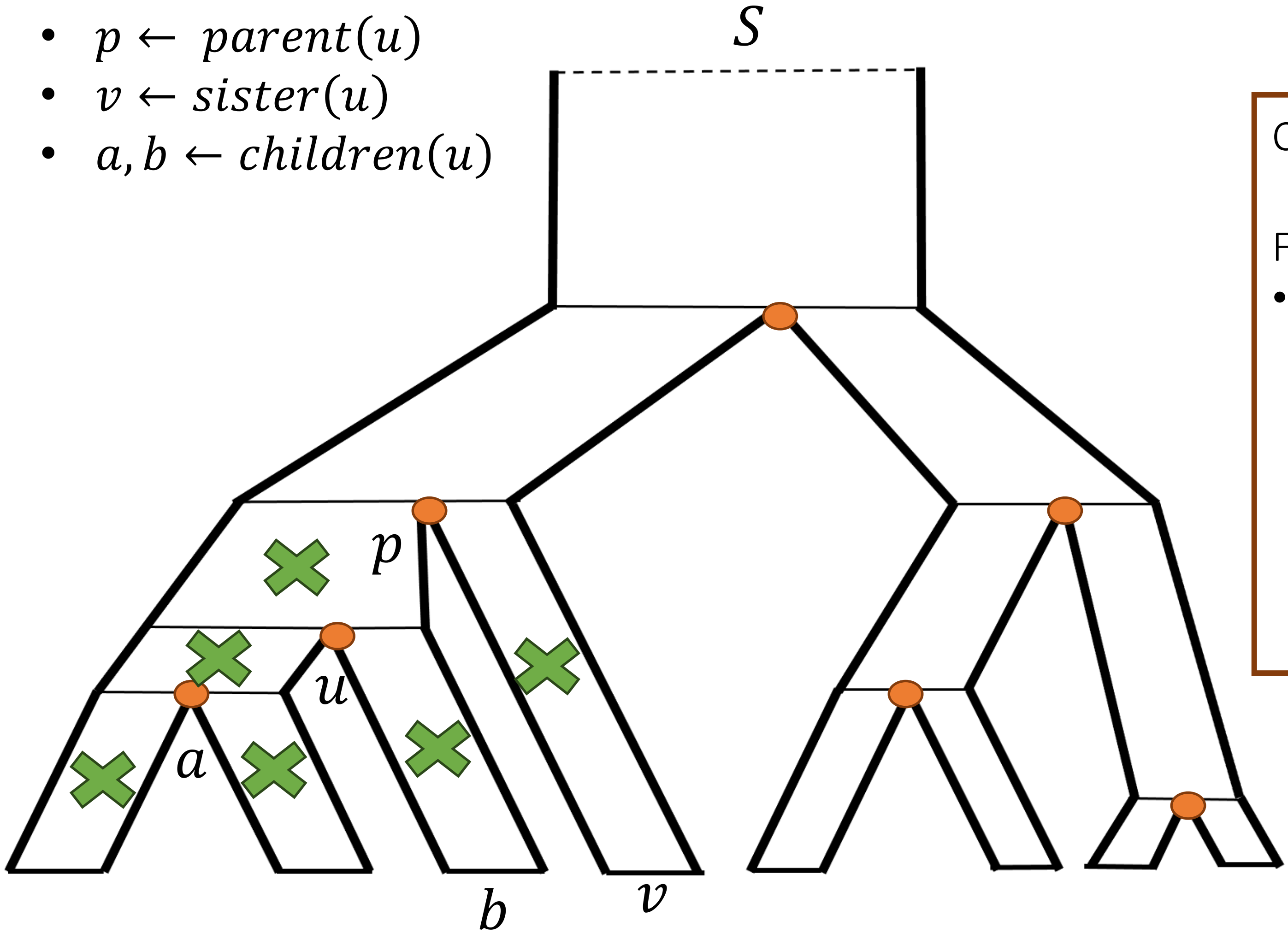
$O(n^2k)$ dynamic programming

```
Calculate $\overline{L_a}, \overline{L_b}, \overline{L_v}, \overline{L_p}, \overline{L'_a}, \overline{L'_b}, \overline{L'_v}, \overline{L'_p}$ for each branch

For $u \in$ post order traversal of internal nodes of $S$:
• if $p$ is not root:
    • $t_{p\to u} \leftarrow$ internal branch equation
    • If $v$ is leaf:
        $t_{p\to v} \leftarrow$ terminal middle branch equation
    • For $w \in children(u)$:
        • If $w$ is leaf:
            $t_{u\to w} \leftarrow$ terminal cherry branch equation
```

Large tree algorithm

- $p \leftarrow parent(u)$
- $v \leftarrow sister(u)$
- $a, b \leftarrow children(u)$



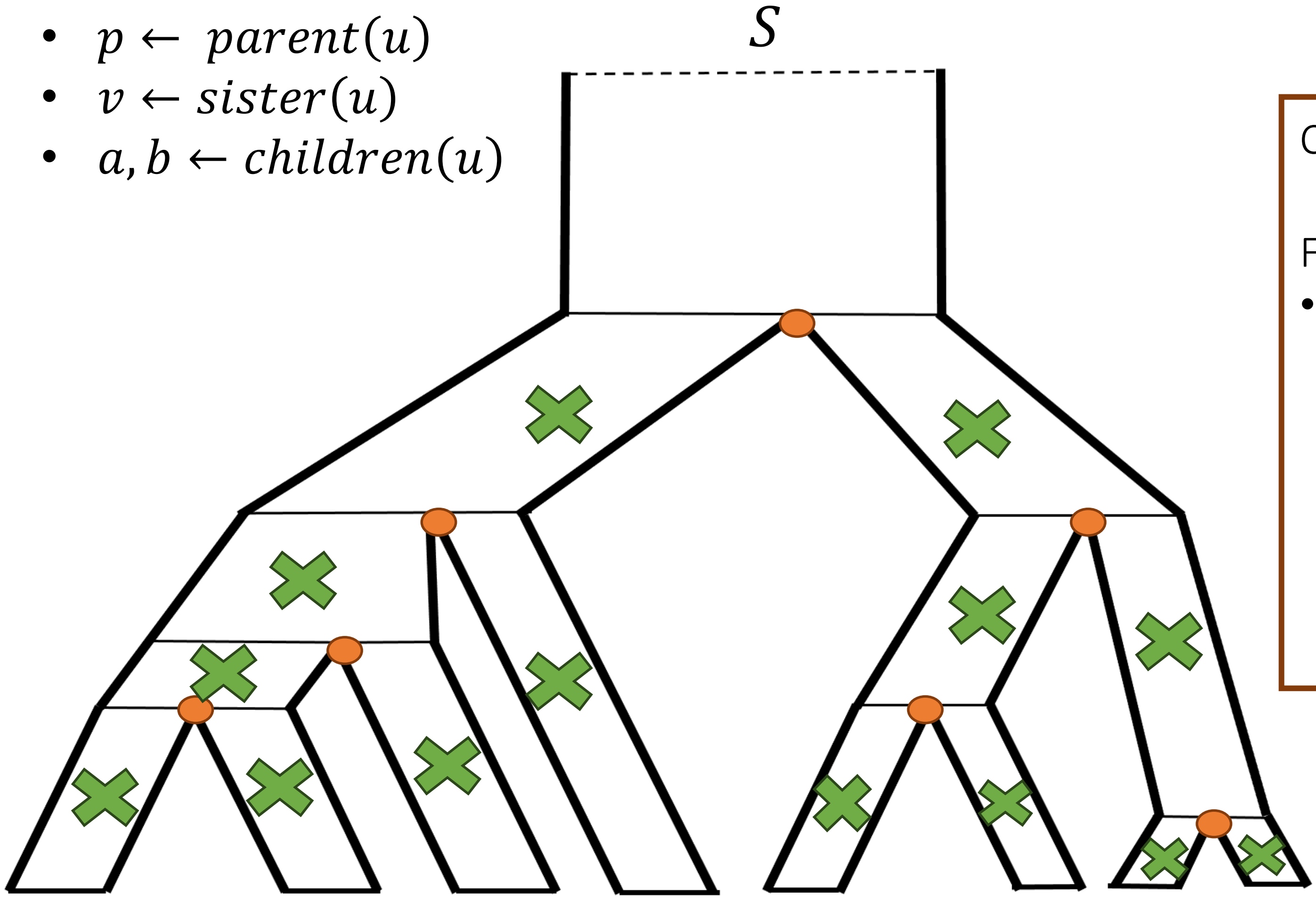
$O(n^2k)$ dynamic programming

```
Calculate $\overline{L_a}, \overline{L_b}, \overline{L_v}, \overline{L_p}, \overline{L'_a}, \overline{L'_b}, \overline{L'_v}, \overline{L'_p}$ for each branch

For $u \in$ post order traversal of internal nodes of $S$:
• if $p$ is not root:
    • $t_{p\to u} \leftarrow$ internal branch equation
    • If $v$ is leaf:
        $t_{p\to v} \leftarrow$ terminal middle branch equation
    • For $w \in children(u)$:
        • If $w$ is leaf:
            $t_{u\to w} \leftarrow$ terminal cherry branch equation
```

Large tree algorithm

- $p \leftarrow parent(u)$
- $v \leftarrow sister(u)$
- $a, b \leftarrow children(u)$



$O(n^2k)$ dynamic programming

```
Calculate $\overline{L_a}, \overline{L_b}, \overline{L_v}, \overline{L_p}, \overline{L'_a}, \overline{L'_b}, \overline{L'_v}, \overline{L'_p}$ for each branch

For $u \in$ post order traversal of internal nodes of $S$:
• if $p$ is not root:
    • $t_{p \to u} \leftarrow$ internal branch equation
    • If $v$ is leaf:
        $t_{p \to v} \leftarrow$ terminal middle branch equation
    • For $w \in children(u)$:
        • If $w$ is leaf:
            $t_{u \to w} \leftarrow$ terminal cherry branch equation
```

Total runtime: $O(n^2k)$

in practice: ~**50s** on 100-taxon tree with 1000 genes

Summary of results so far

- We derived expected branch lengths for matching/non-matching gene trees for an unbalanced/balanced quartet species tree under MSC+Substitution model.

- We presented simplifications that lead to analytical formulas for each branch in the species tree.

- We introduced CASTLES that uses these formulas to estimate branch lengths on a species tree in $O(n^2k)$.

Summary of results so far

- We derived expected branch lengths for matching/non-matching gene trees for an unbalanced/balanced quartet species tree under MSC+Substitution model.
- We presented simplifications that lead to analytical formulas for each branch in the species tree.
- We introduced CASTLES that uses these formulas to estimate branch lengths on a species tree in $O(n^2k)$.

How well does it work in practice?

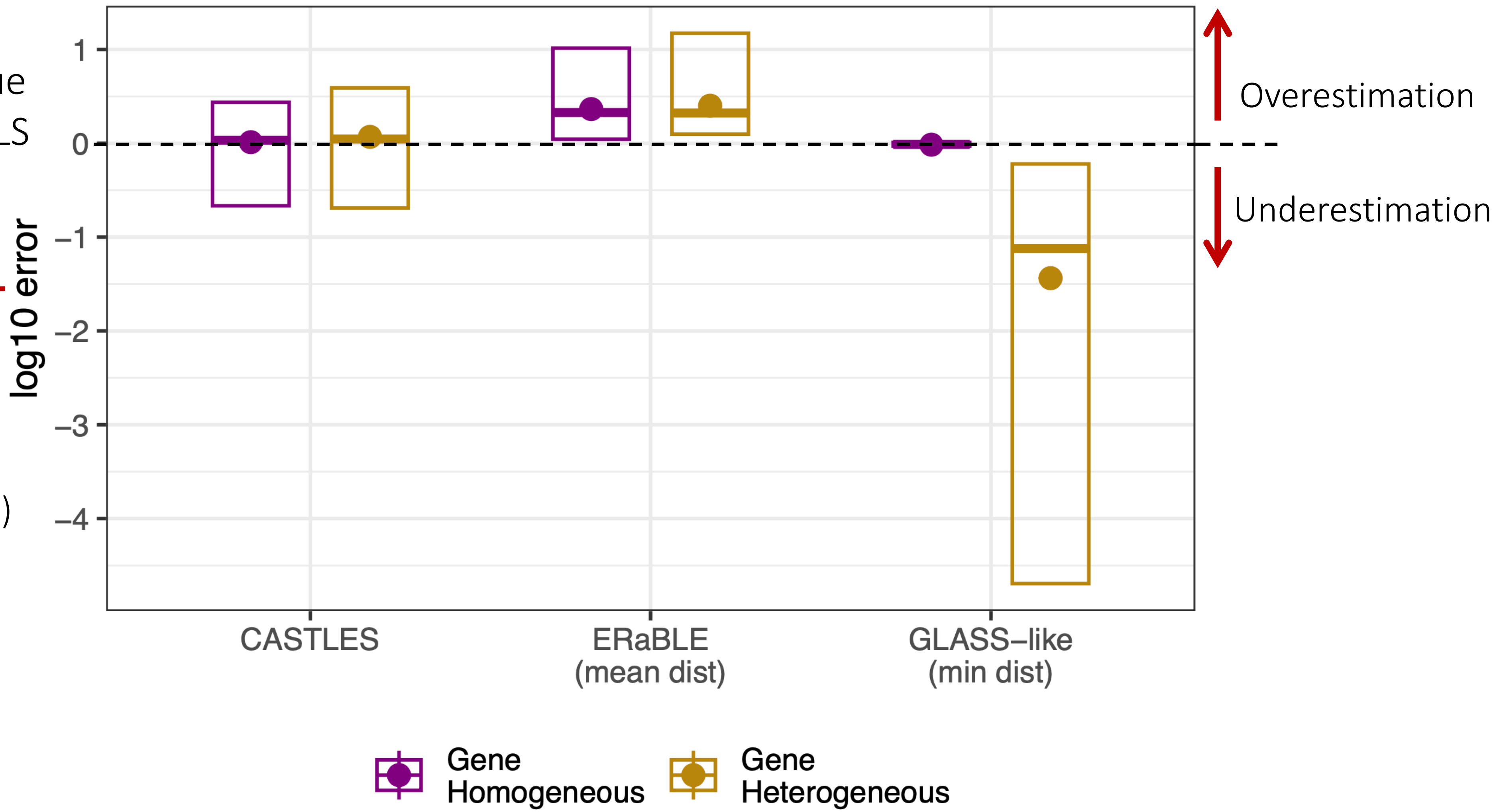
Experimental study

- Modified SimPhy [Mallo et al., 2016] to generate species trees with SU lengths
- Estimating branch lengths on the *true* species tree topology
- Three ILS simulated datasets and a mammalian biological dataset
- Evaluating using bias, absolute error, RMSE, and log error
- **Methods**: Concatenation with RAxML [Stamatakis, 2014], FastME [Lefort et al, 2015] with minimum and average distance matrices and ERaBLE [Binet et al., 2016]

CASTLES is robust to rate heterogeneity across genes

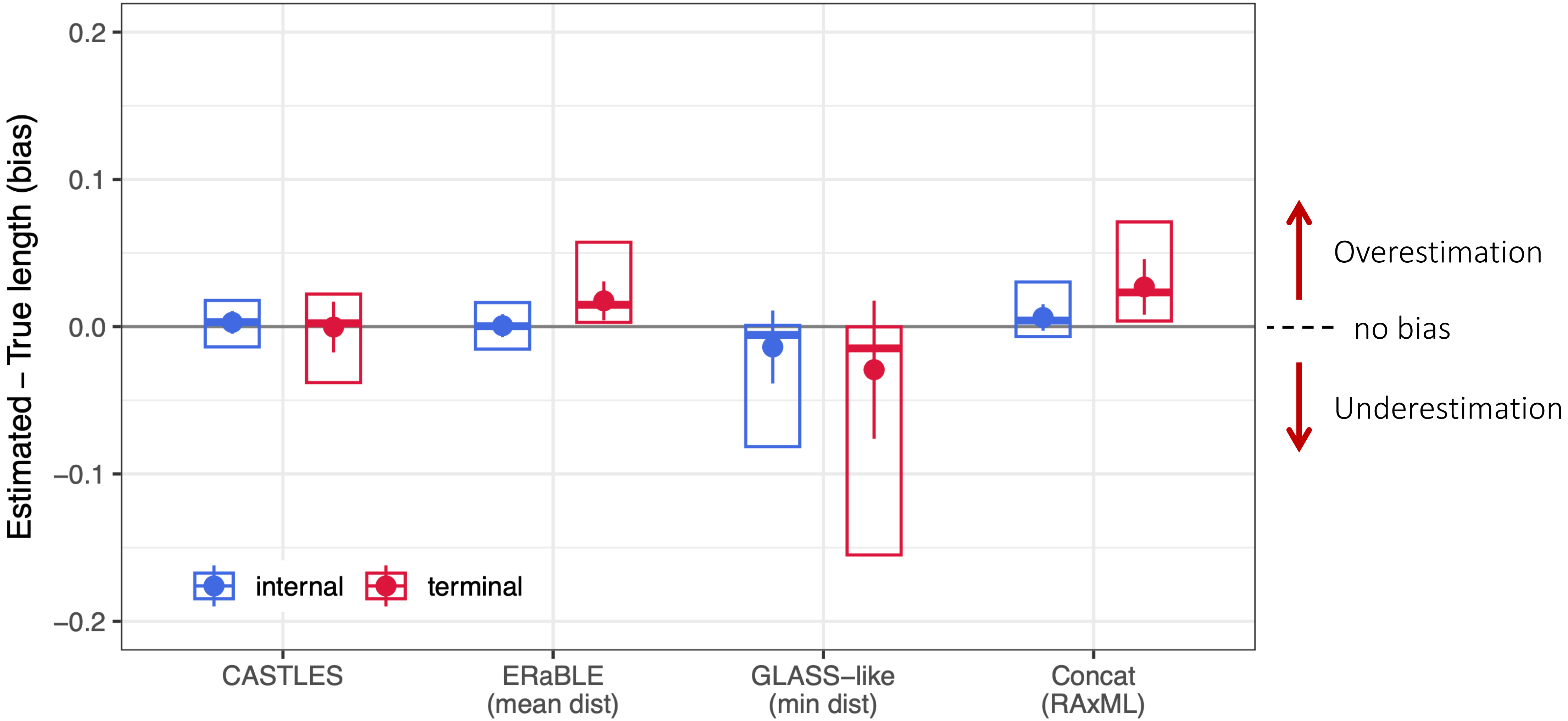
- Quartet ILS simulated dataset with 10,000 true gene trees, moderate ILS

log10 (est. length / true length)



CASTLES is less biased than other methods

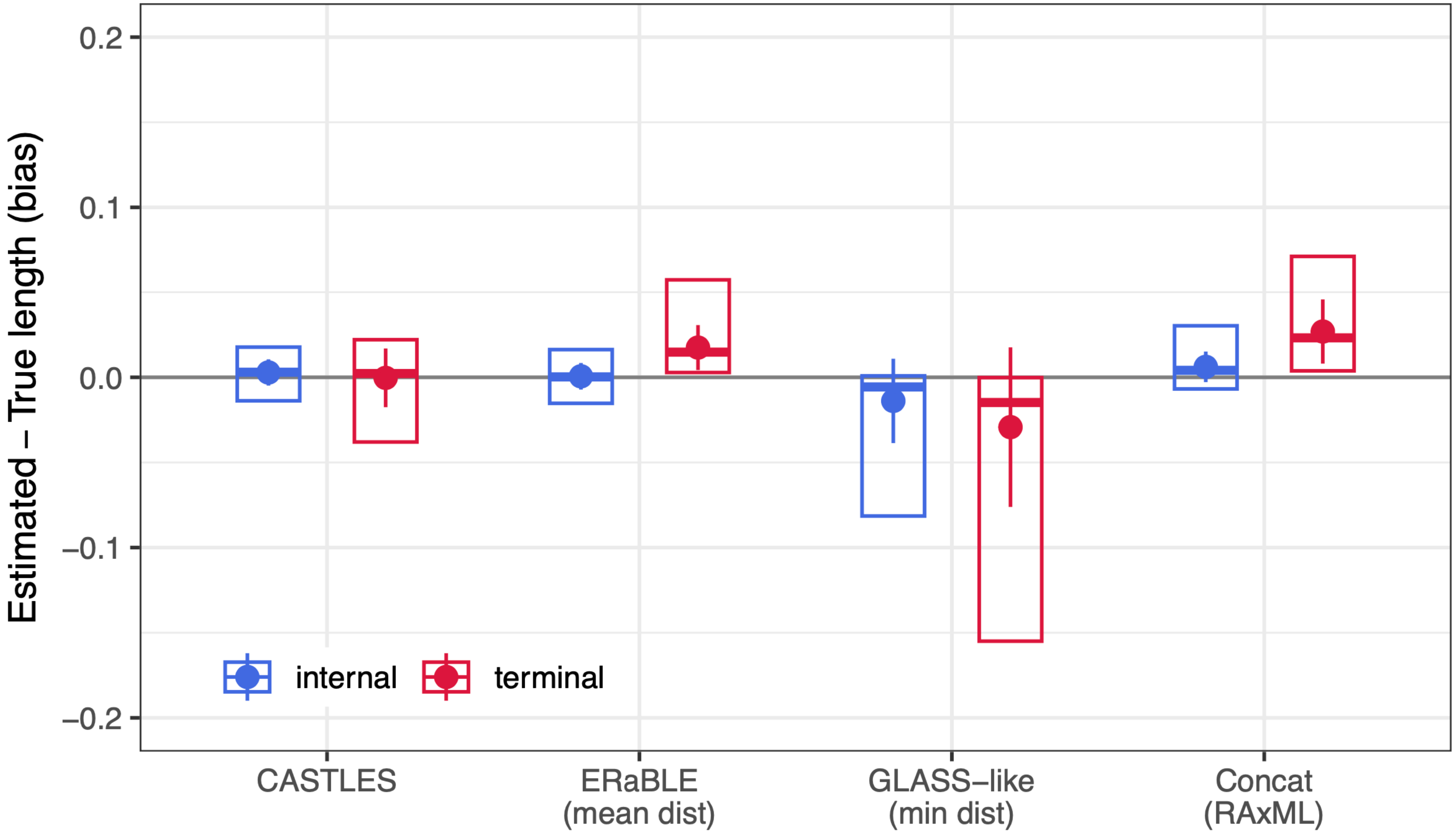
- 100-taxon ILS simulated dataset with 1000 genes, moderate ILS, 200bp sequence length [Zhang et al (2018)]



CASTLES is less biased than other methods

- 100-taxon ILS simulated dataset with 1000 genes, moderate ILS, 200bp sequence length [Zhang et al (2018)]

Bias is higher for **terminal** branches.



Overestimation

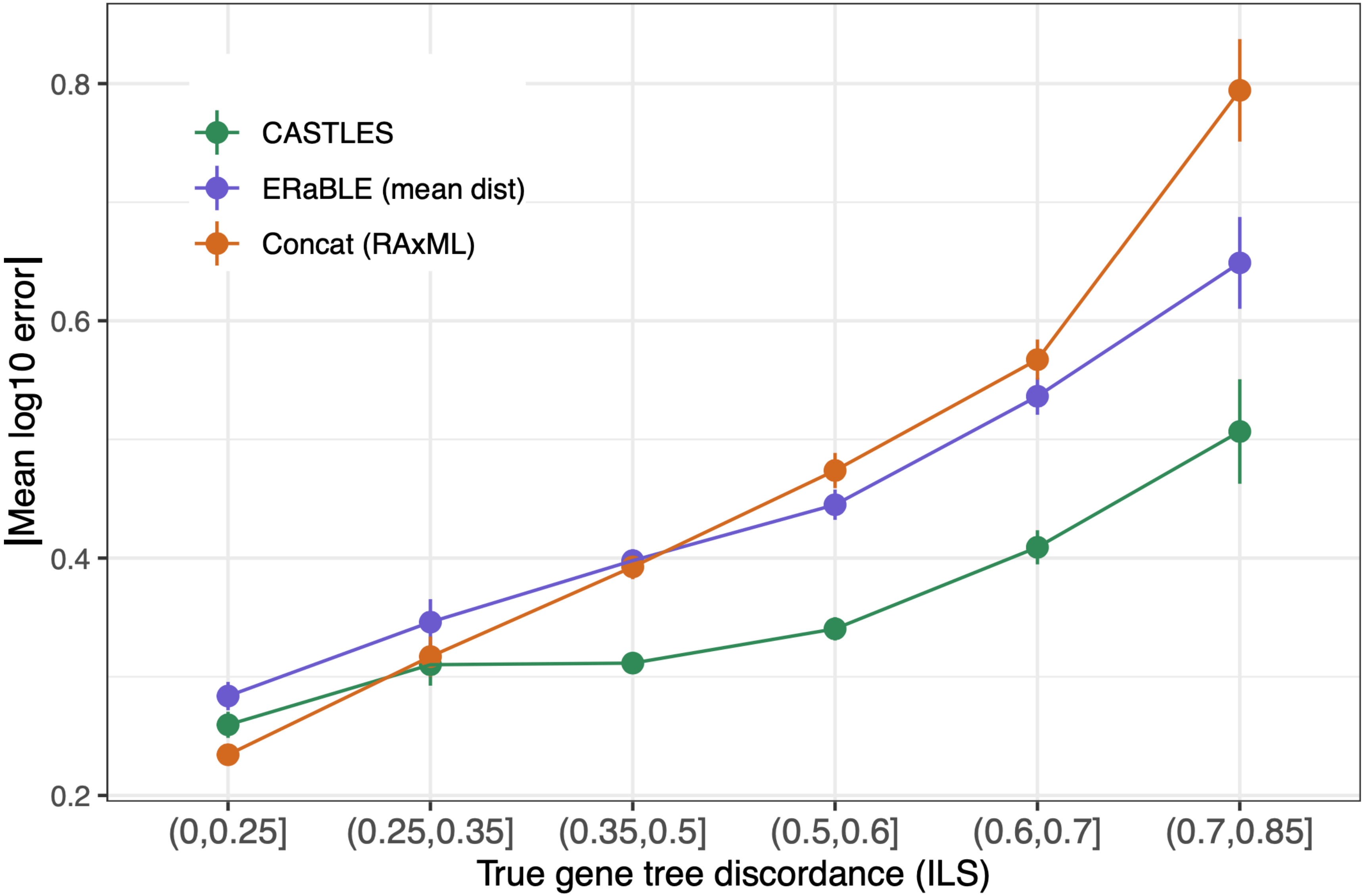
no bias

Underestimation

CASTLES's advantage increases with ILS

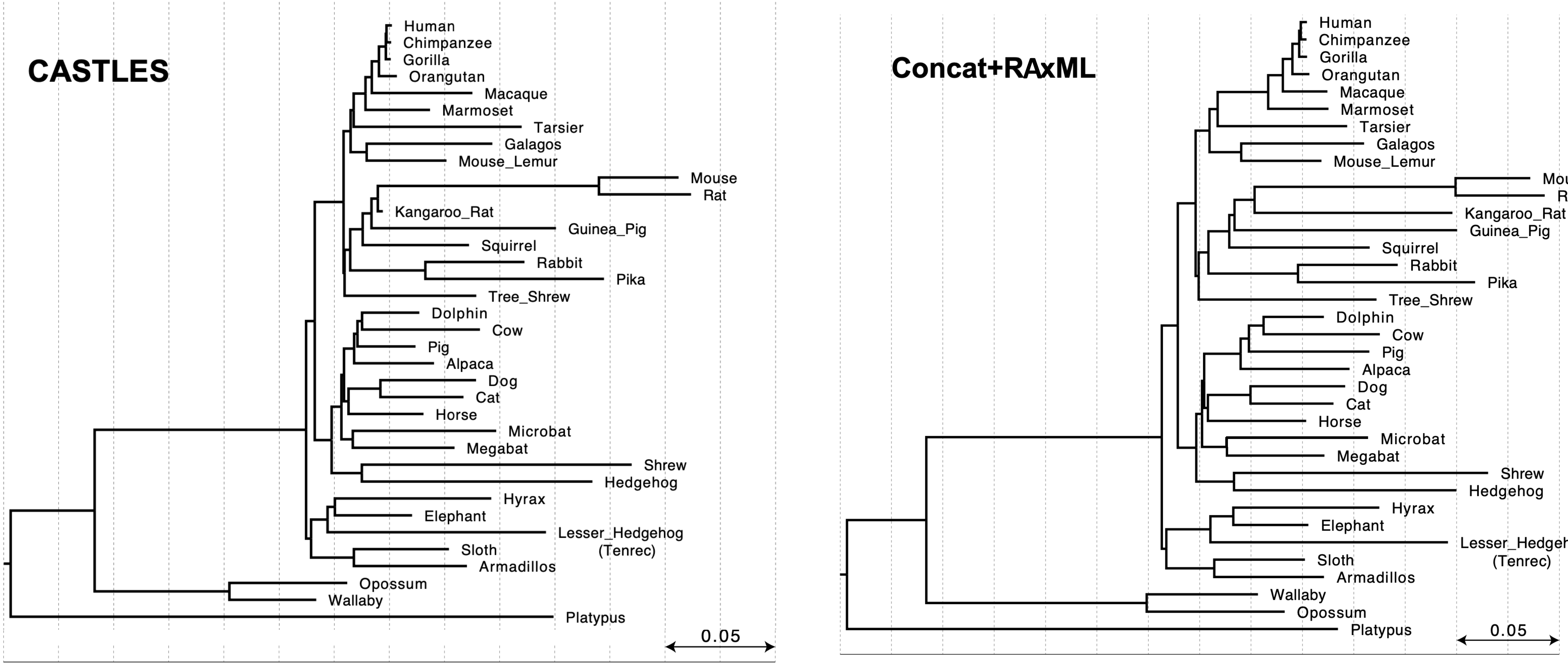
- 30-taxon ILS simulated dataset with 500 genes [Mai et al (2017)]
- Average 0.38 GTEE

|log10 (est. length / true length)|



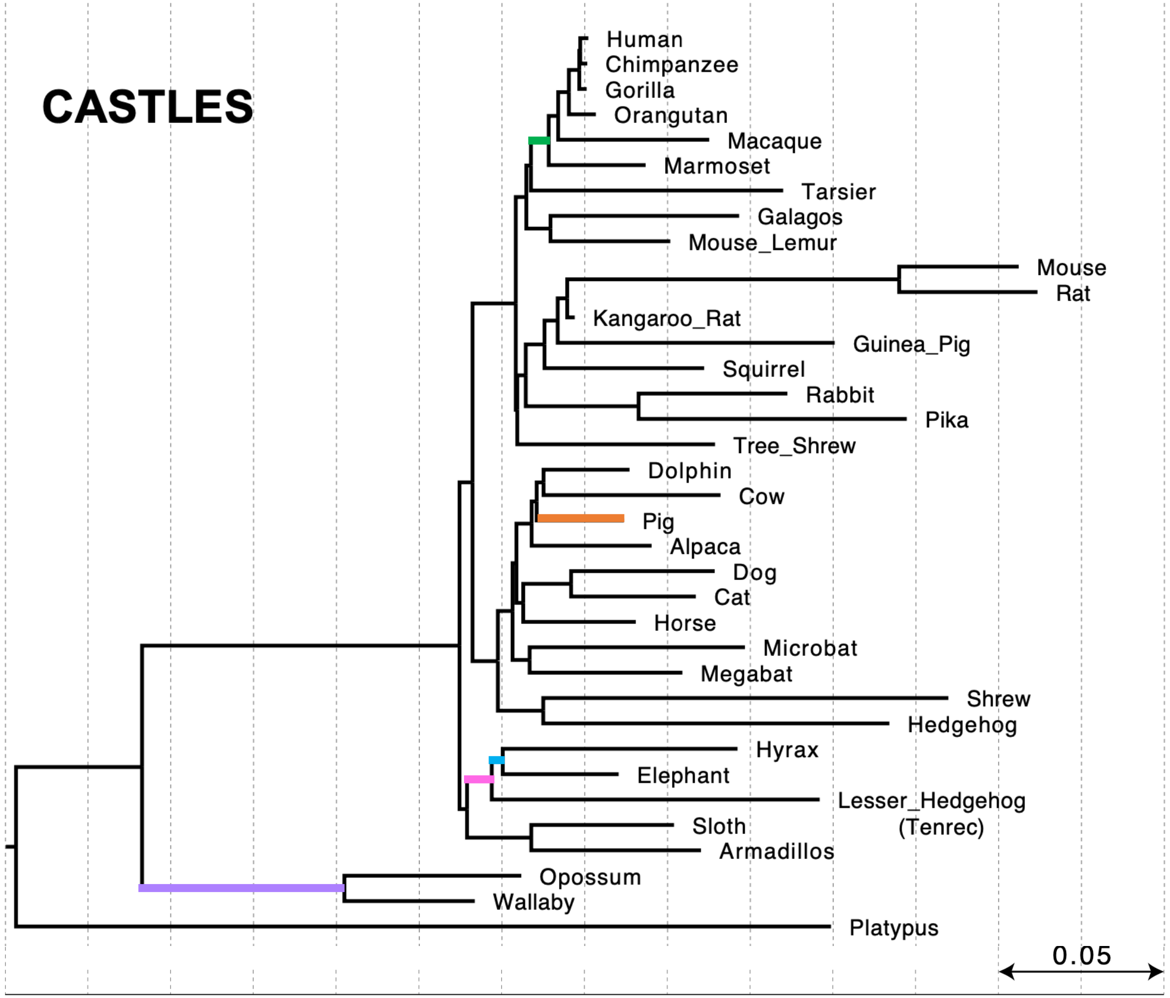
CASTLES produces shorter branches than concatenation on mammalian dataset

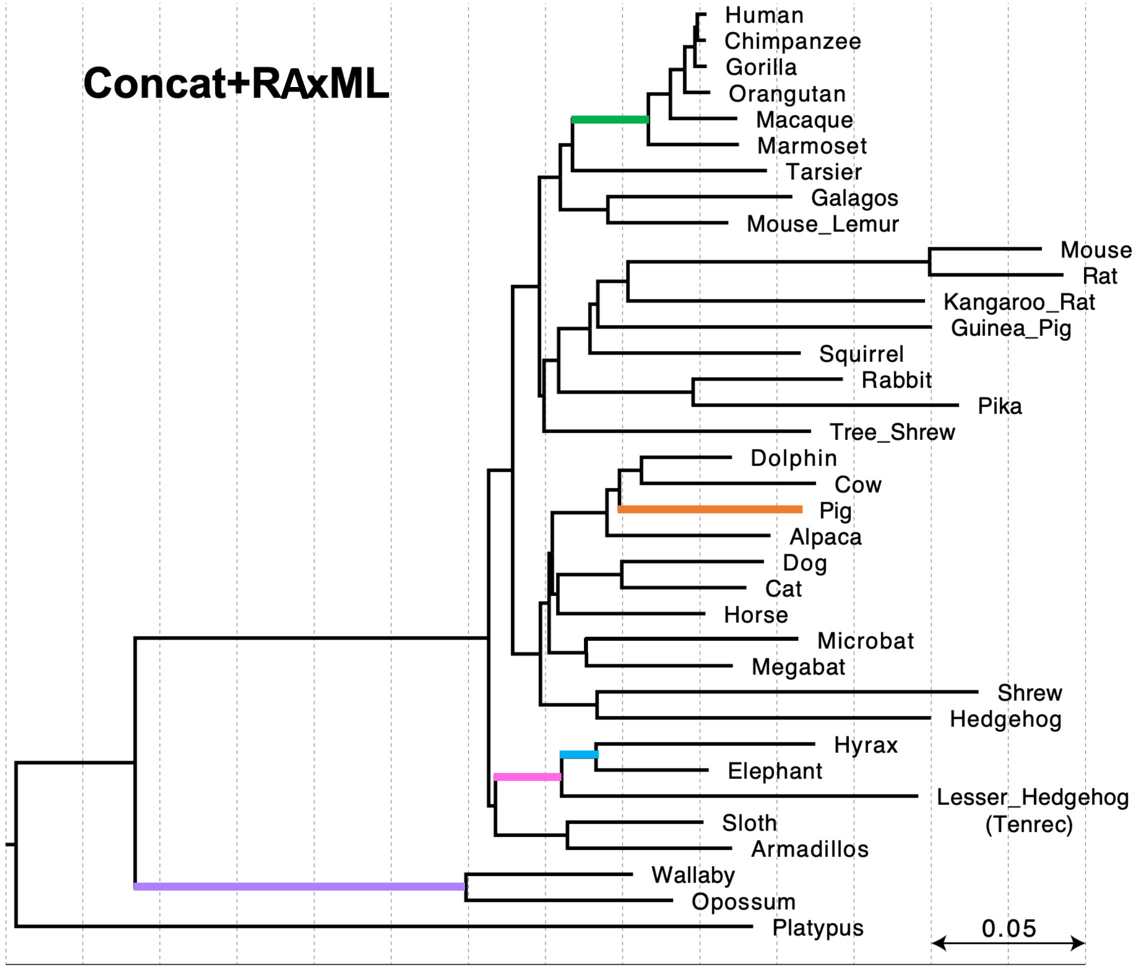
- 37-taxon mammalian biological dataset with 424 genes [Song et al (2012)], ASTRAL species tree



CASTLES produces shorter branches than concatenation on mammalian dataset

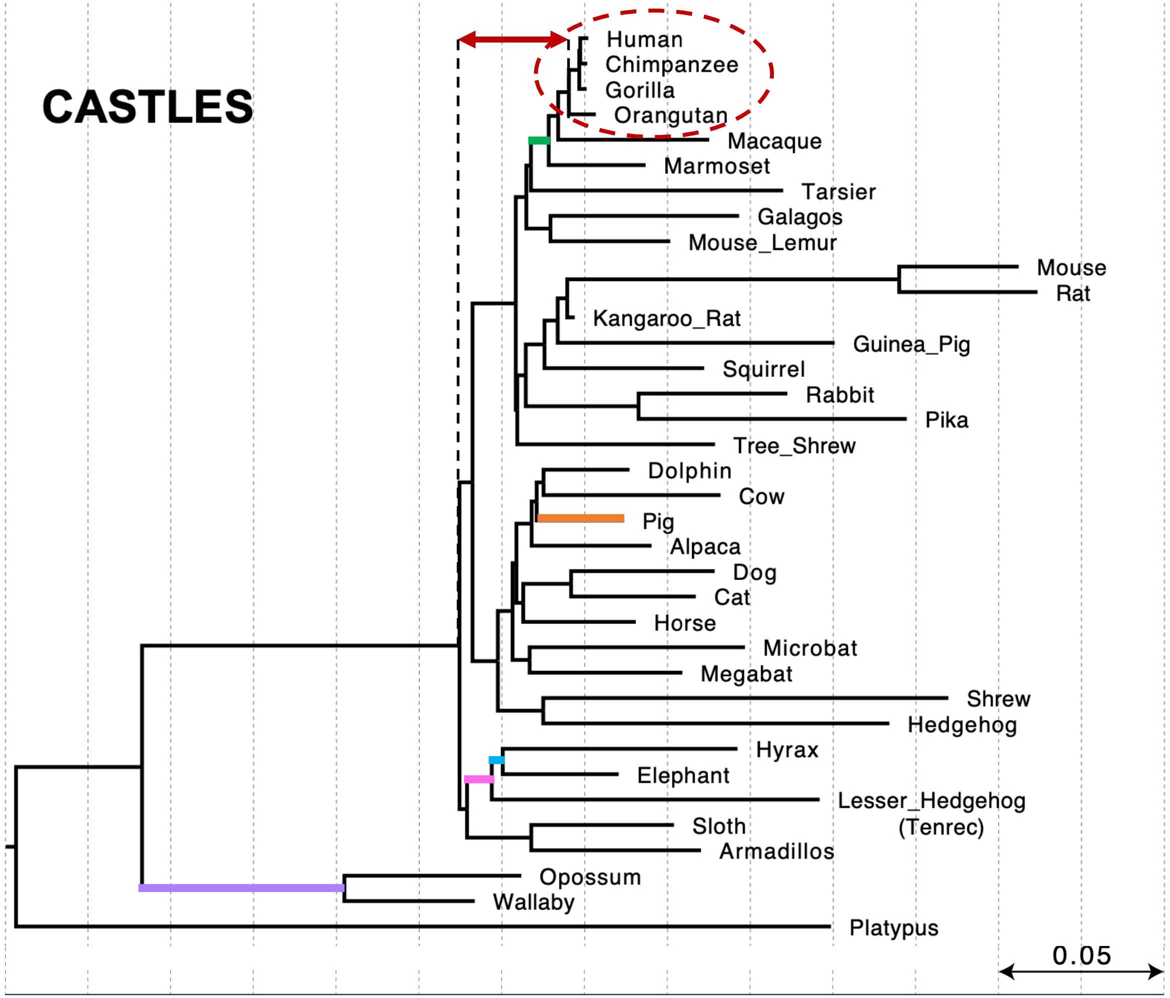
- 37-taxon mammalian biological dataset with 424 genes [Song et al (2012)], ASTRAL species tree

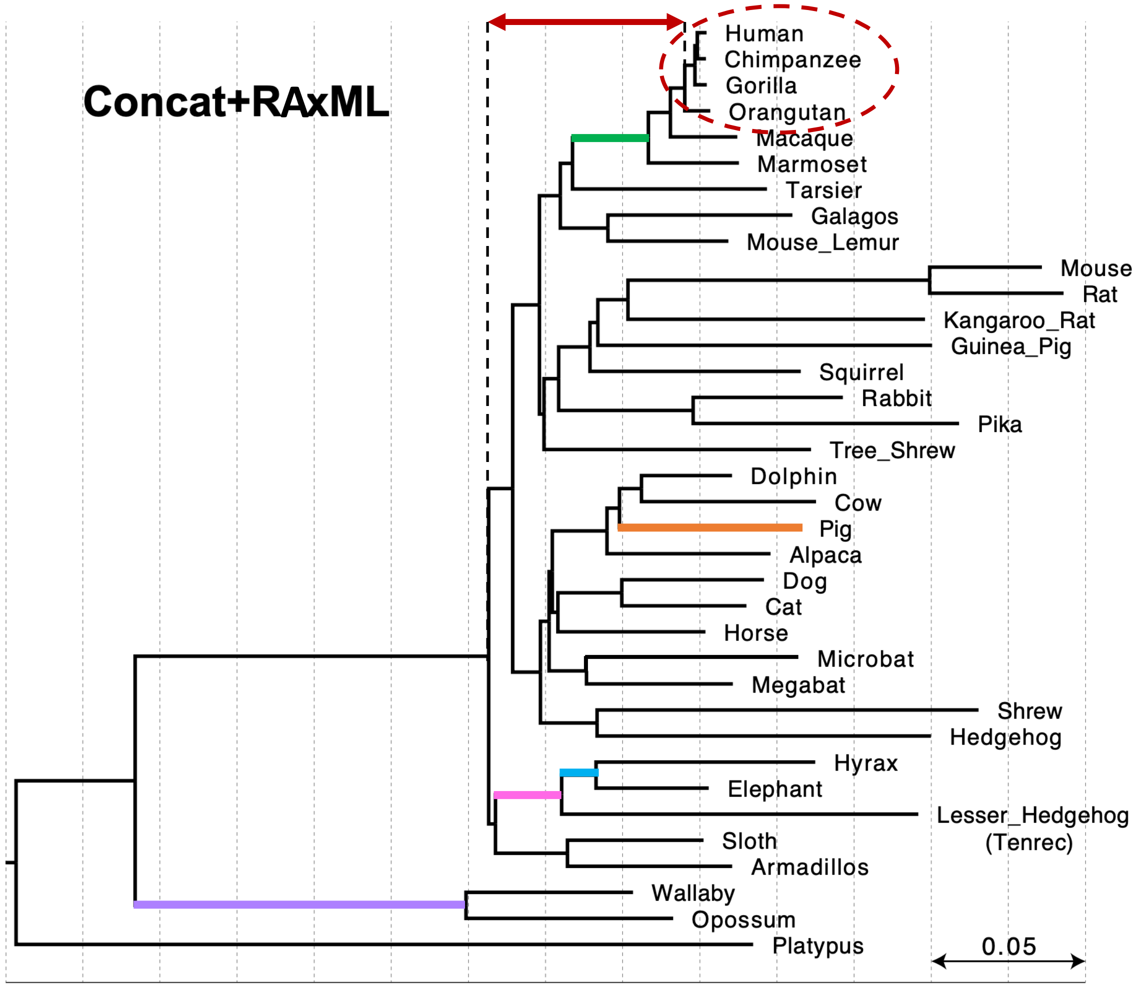




CASTLES produces shorter branches than concatenation on mammalian dataset

- 37-taxon mammalian biological dataset with 424 genes [Song et al (2012)], ASTRAL species tree





Summary & Future Directions

Summary

- CASTLES is a scalable method for estimating branch lengths of a species tree given gene tree branch lengths
- CASTLES addresses gene tree heterogeneity due to ILS, and naturally occurring variations in mutation rates
- CASTLES produces more accurate and less biased branch lengths than prior methods in many model conditions

Future Directions

- Addressing other sources of gene tree discordance, such as gene duplication and loss and horizontal gene transfer
- Evaluating CASTLES on datasets with model misspecification, missing data, etc
- Are SU lengths identifiable under MSC+Substitution model, and is CASTLES statistically consistent?

Acknowledgements

Thank you!

Chao Zhang

Tandy Warnow

Siavash Mirarab

Paper is available at:
https://doi.org/10.1093/bioinformatics/btad221

CASTLES is available on Github:
https://github.com/ytabatabaee/CASTLES

UC San Diego

Berkeley
UNIVERSITY OF CALIFORNIA

COMPUTER SCIENCE

Funding:

NSF

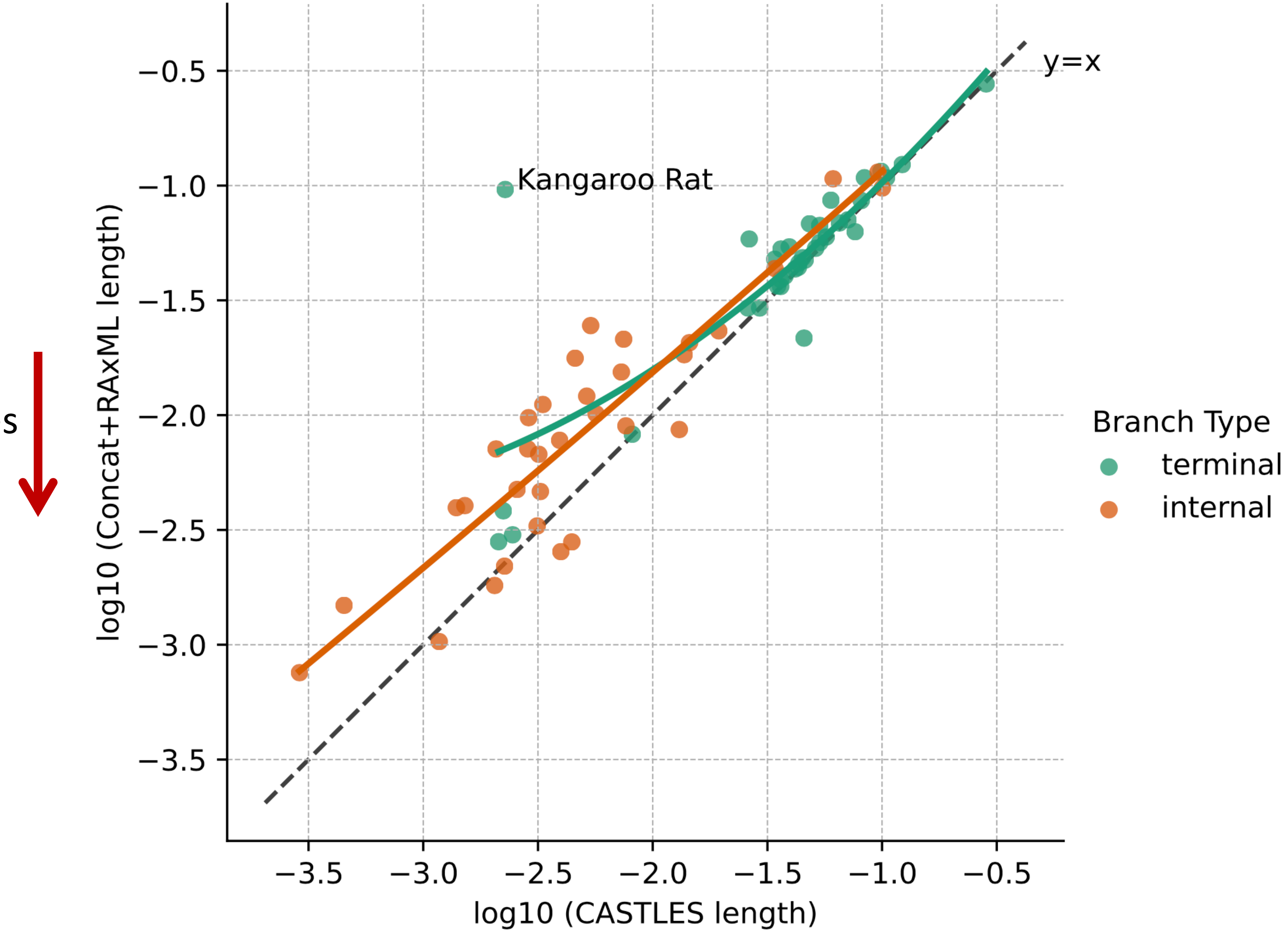
NIH
National Institutes of Health

Computing Resources:
UIUC Campus Cluster

CASTLES produces shorter branches than concatenation on mammalian dataset

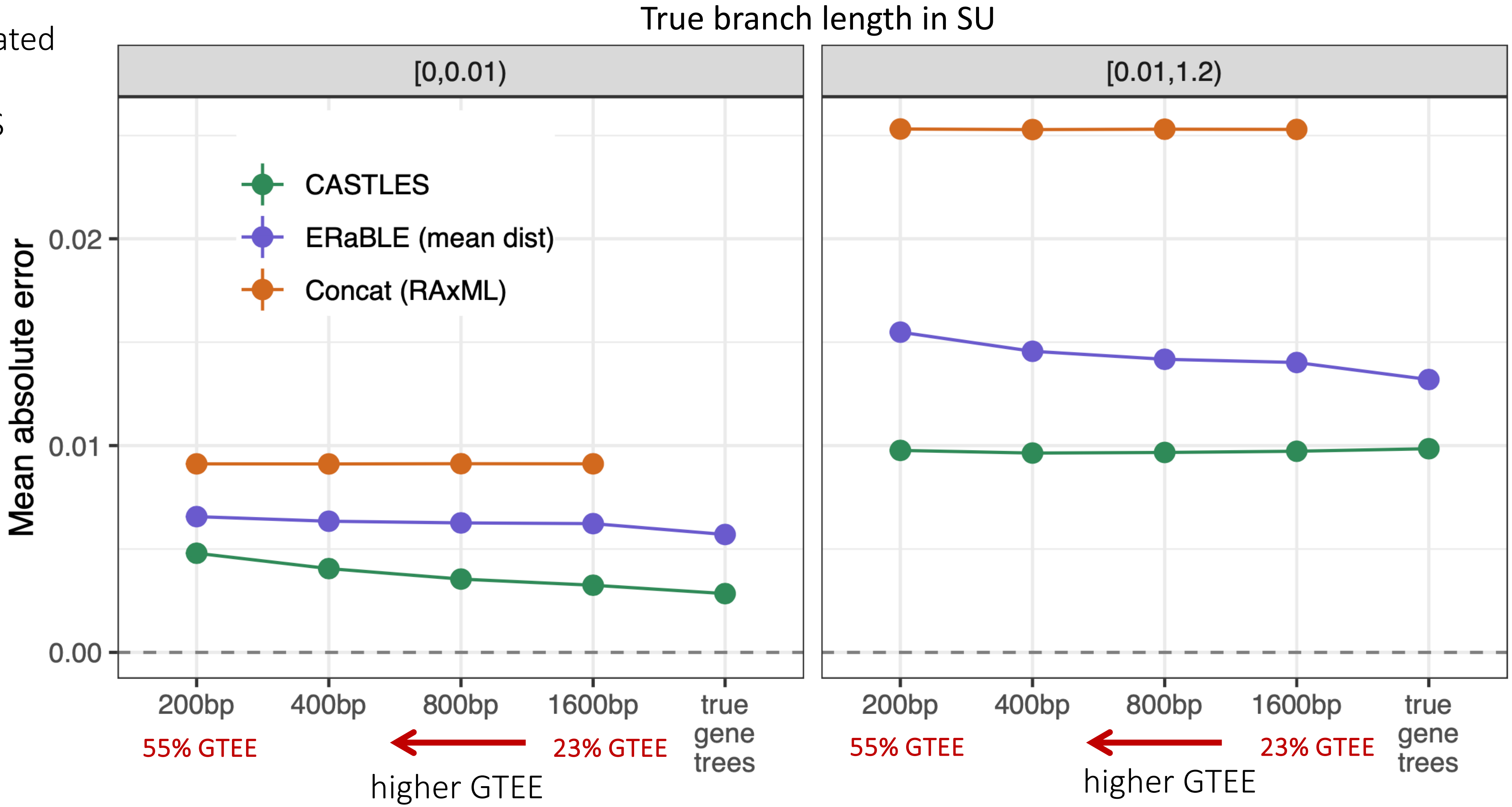
- 37-taxon mammalian biological dataset with 424 genes [Song et al (2012)], ASTRAL species tree

Shorter branches



Phylogenetic signal has relatively small impact on branch length accuracy

- 100-taxon ILS simulated dataset with 1000 genes, moderate ILS [Zhang et al (2018)]


True branch length in SU
[0,0.01)
[0.01,1.2)
CASTLES
ERaBLE (mean dist)
Concat (RAxML)
Mean absolute error
200bp
400bp
800bp
1600bp
true gene trees
55% GTEE
23% GTEE
higher GTEE
200bp
400bp
800bp
1600bp
true gene trees
55% GTEE
23% GTEE
higher GTEE