Novel computational methods for discordance-aware phylogenomics analysis

Yasamin Tabatabaee

Siebel School of Computing and Data Science

University of Illinois Urbana-Champaign

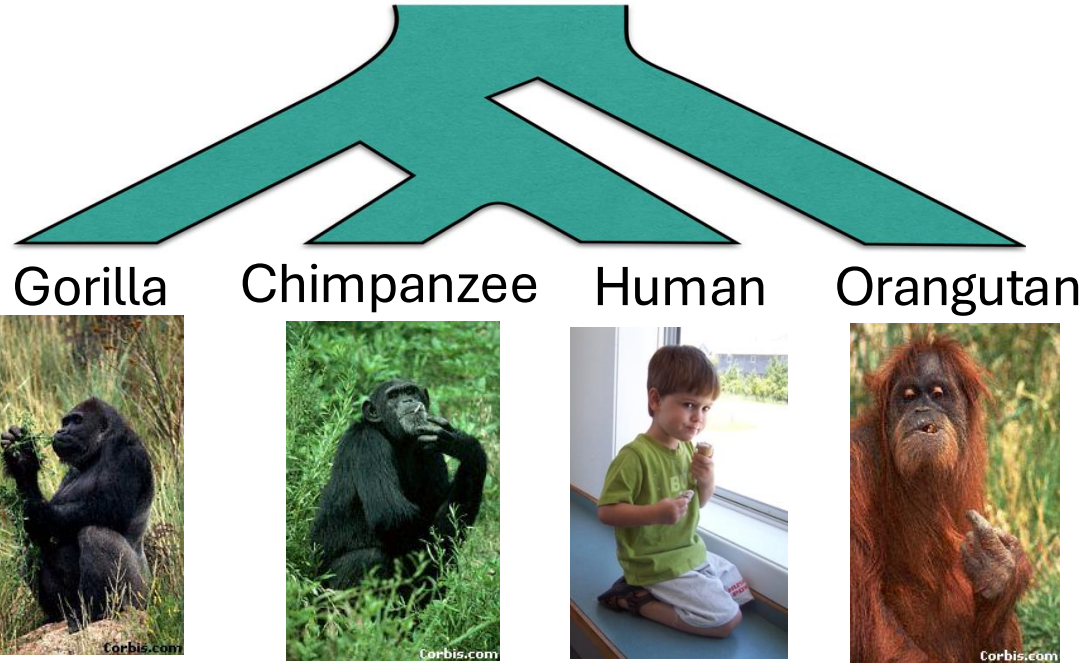
Outline

- Background and Motivation
  - Phylogenomics pipeline
  - Gene tree discordance
  - Species tree estimation
- Overview of Contributions
  - Discordance-aware post-species tree analysis
- Rooting species trees
- Phylogenomic branch length estimation
- Dating species trees and gene trees
- Conclusions

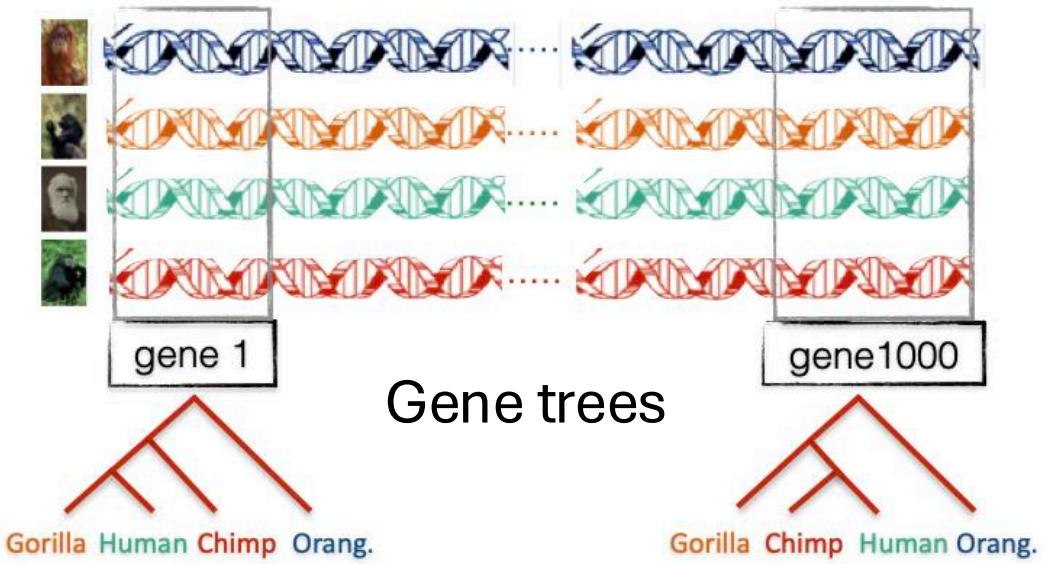
Outline

- **Background and Motivation**
  - **Phylogenomics pipeline**
  - **Gene tree discordance**
  - **Species tree estimation**
- Overview of Contributions
  - Discordance-aware post-species tree analysis
- Rooting species trees
- Phylogenomic branch length estimation
- Dating species trees and gene trees
- Conclusions

Phylogenomics and gene tree discordance

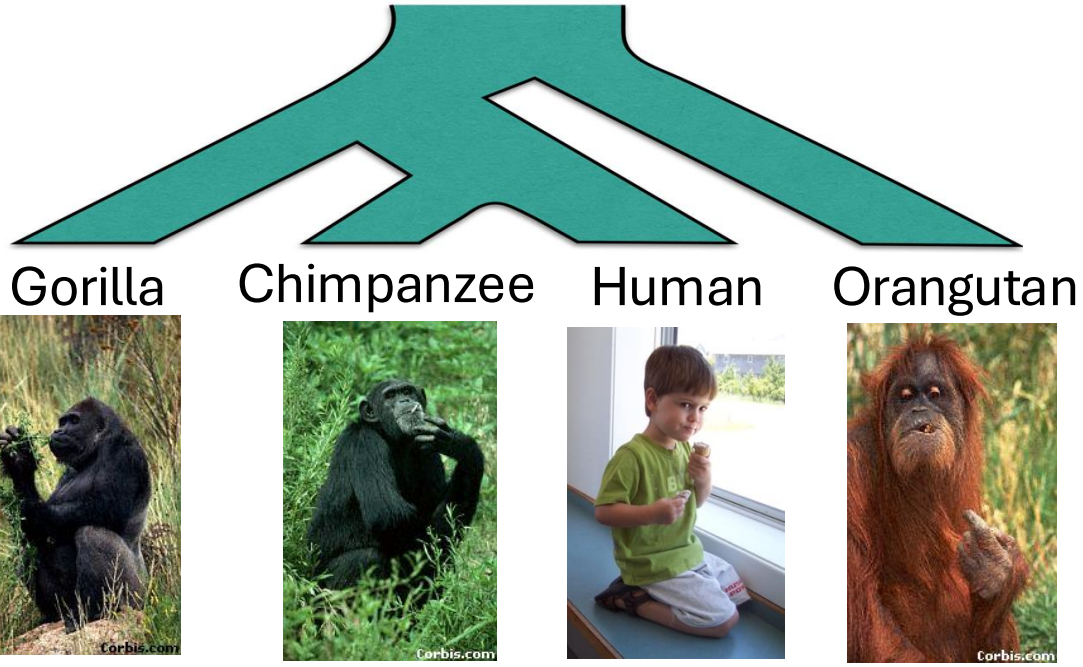


Species tree

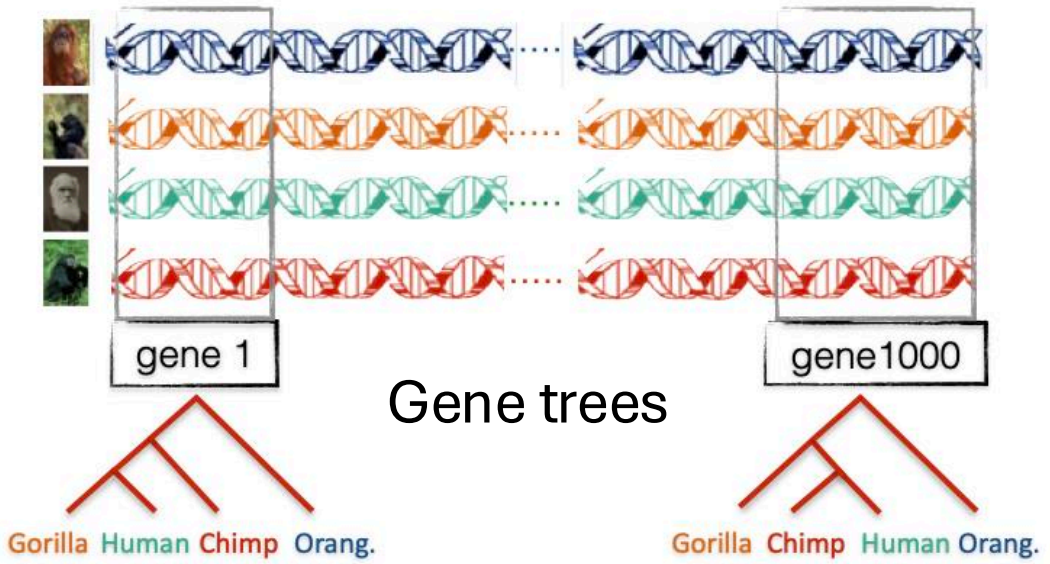


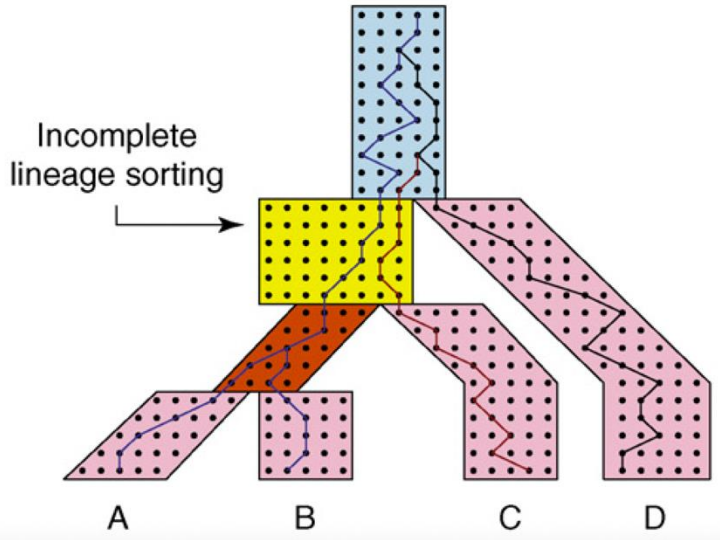
Gene trees

Phylogenomics and gene tree discordance



Species tree





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

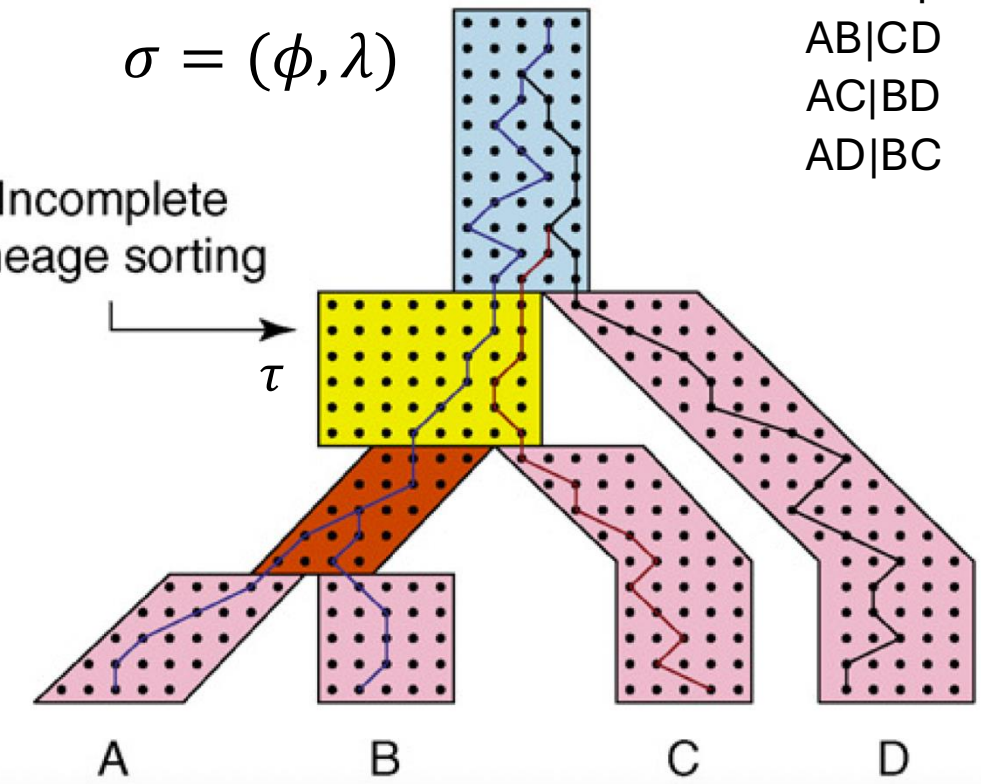
Image Credit: Degnan and Rosenberg, 2009, Trends in Ecology and Evolution

Multi-Species Coalescent (MSC) model

$\sigma = (\phi, \lambda)$

Incomplete lineage sorting

τ



Pamilo and Nei (1988)

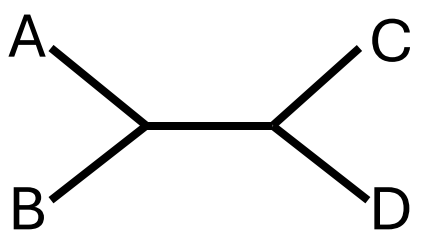
Gene Tree Topologies

AB|CD

AC|BD

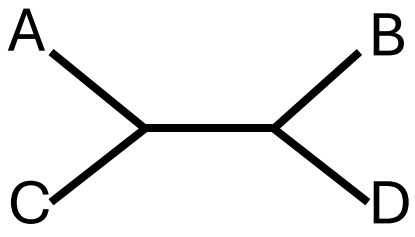
AD|BC

g_1



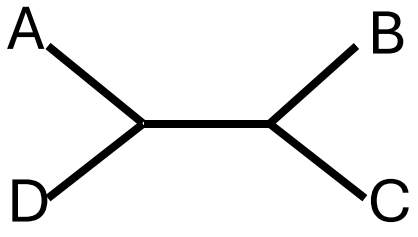
$P(g1|\sigma) = 1 - \frac{2}{3}e^{-\tau}$

g_2



$P(g2|\sigma) = \frac{1}{3}e^{-\tau}$

g_3



$P(g3|\sigma) = \frac{1}{3}e^{-\tau}$

The model species tree defines a probability distribution on the rooted and unrooted gene tree topologies

ILS is **omnipresent** across the tree of life - most likely for short branches (i.e., rapid radiations) or large population sizes.

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

Other sources of gene tree discordance

Biological sources of gene tree discordance

- Incomplete lineage sorting (ILS)
- Gene duplication and loss (GDL)
- Horizontal gene transfer (HGT)
- Hybridization
- Introgression
- Gene flow
- …

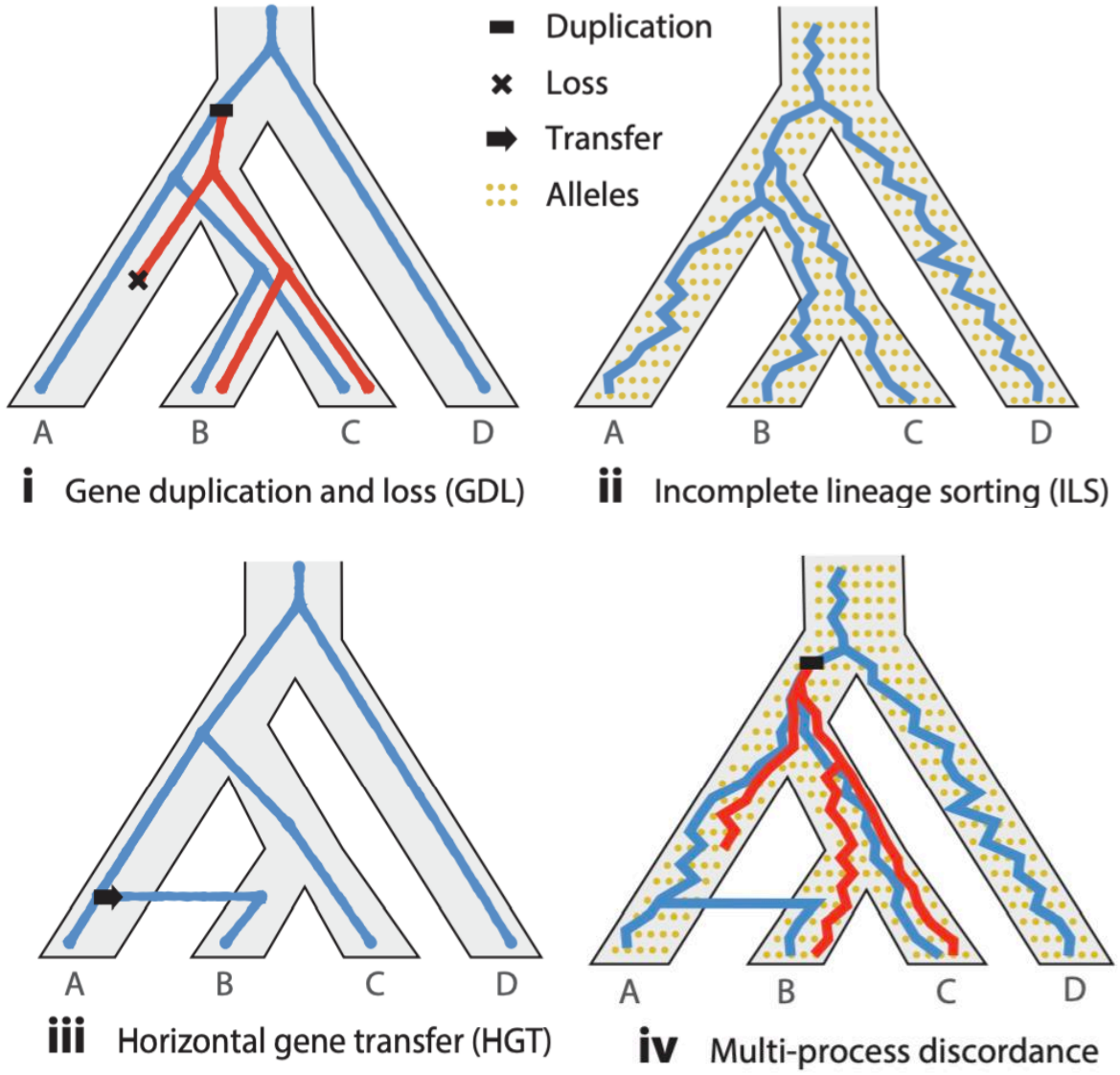
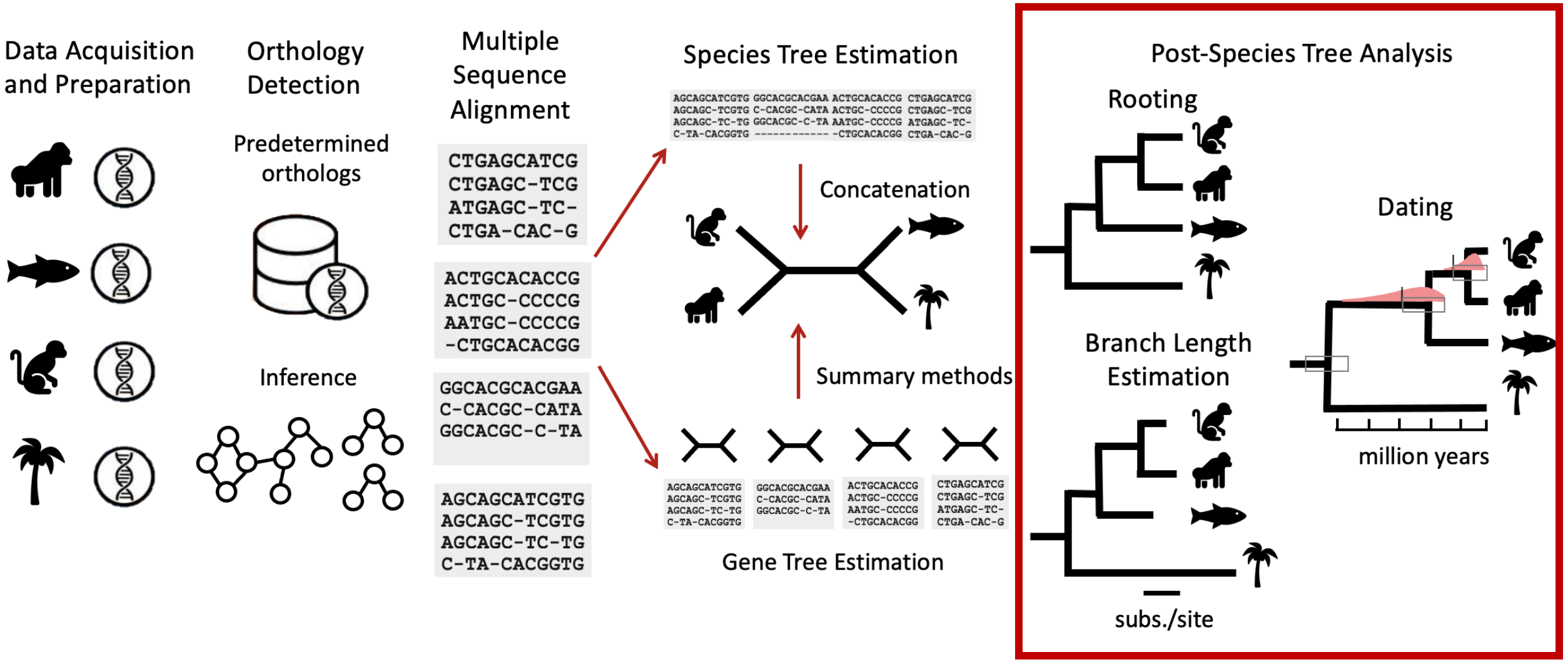
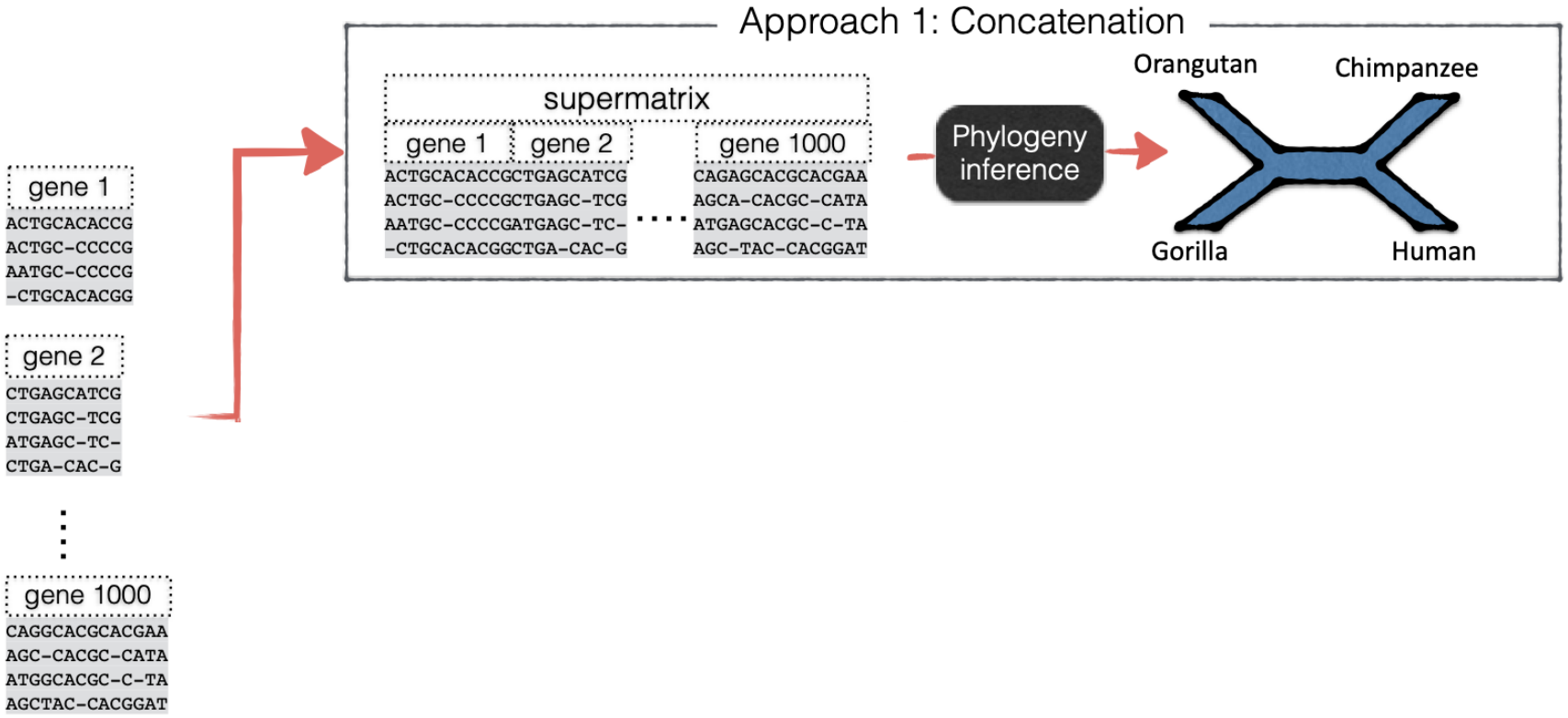


Image Credit: Mirarab et al., 2021, *Annual Review of Ecology, Evolution, and Systematics*

Phylogenomics Pipeline



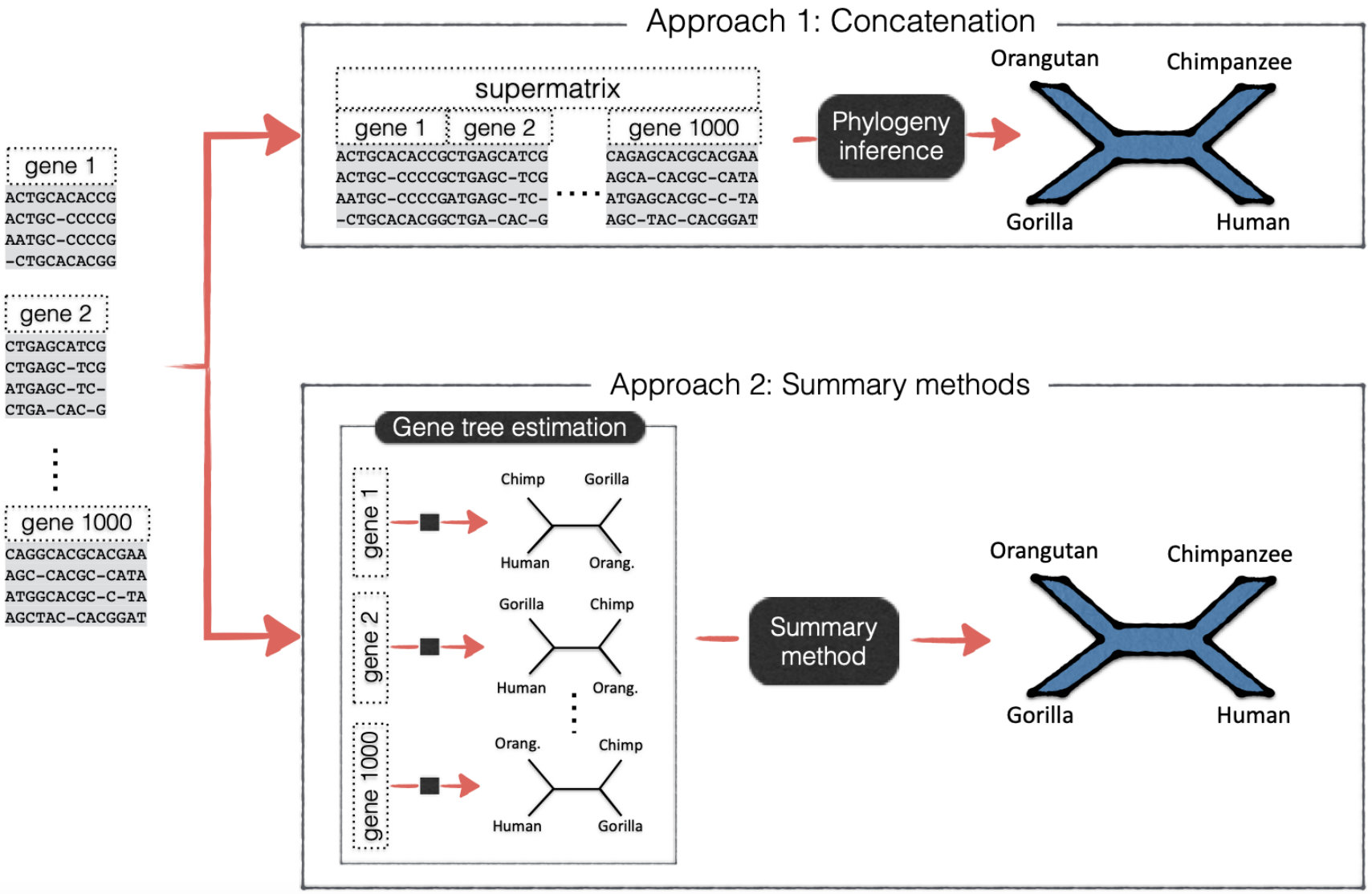
Species tree estimation



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

Image Credit: Siavash Mirarab

Species tree estimation

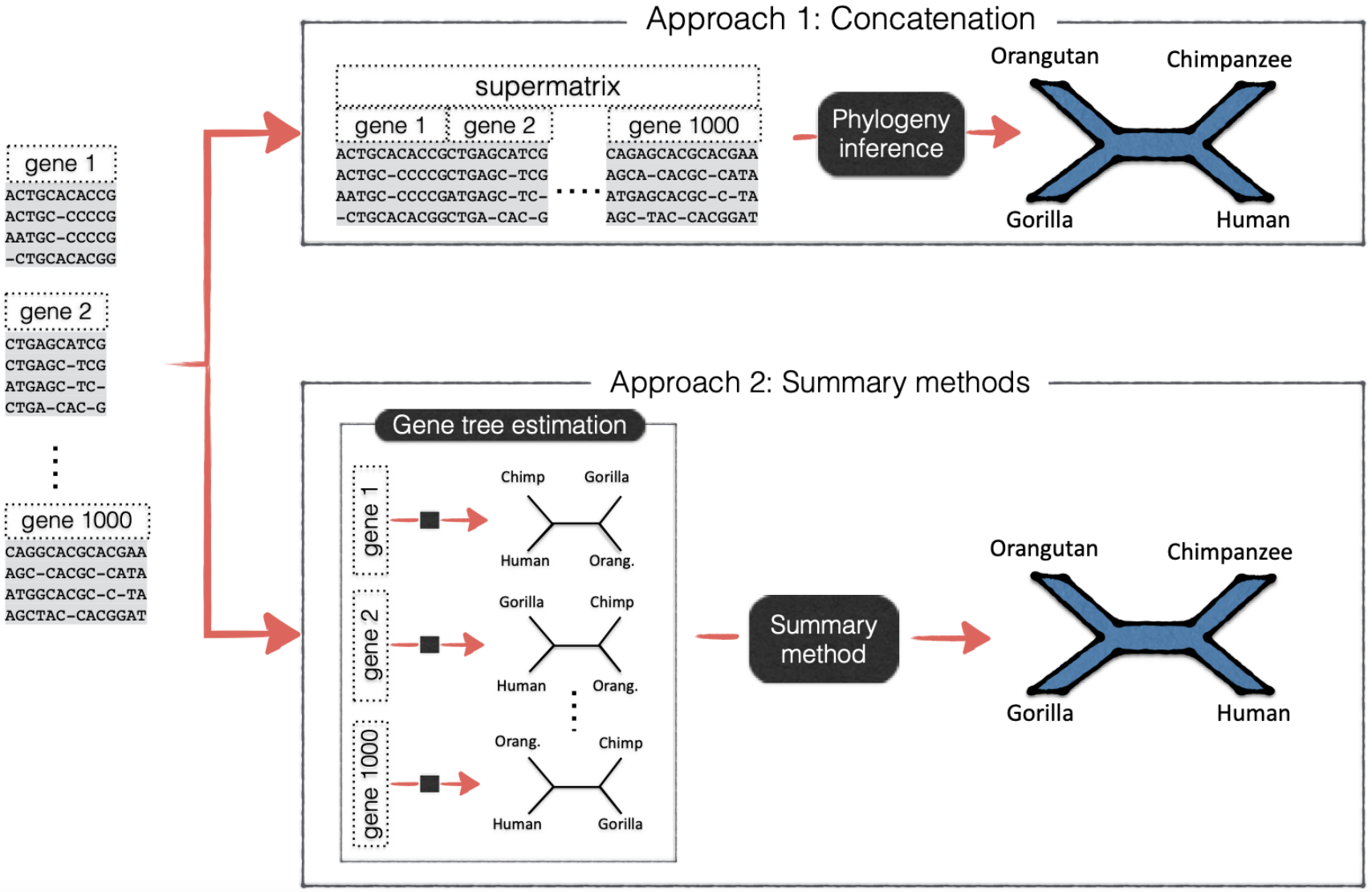


Maximum Likelihood, e.g.
RAxML [Stamatakis, 2014]
FastTree [Price et al, 2010]

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

Image Credit: Siavash Mirarab

Species tree estimation



Concatenation

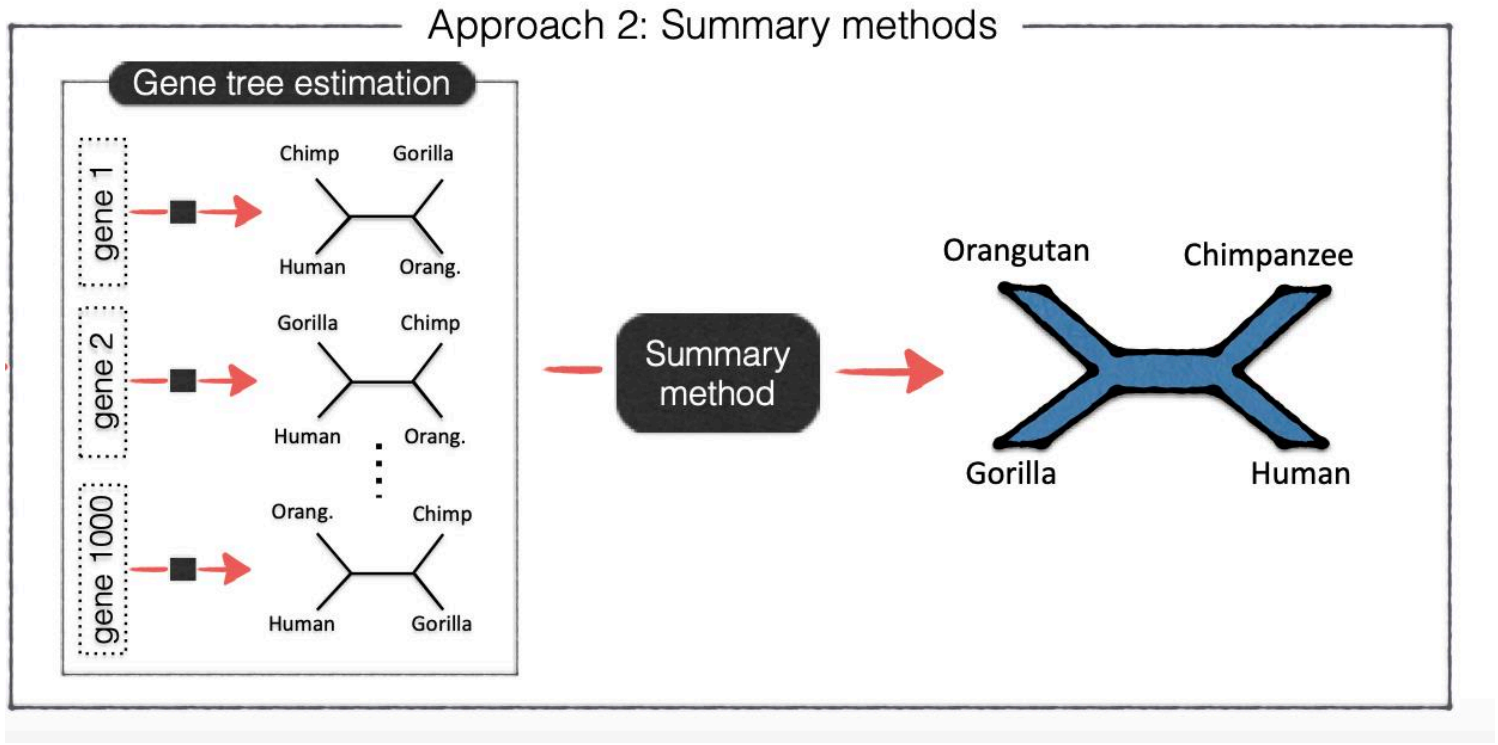
- ignores gene tree heterogeneity
- statistically inconsistent and positively misleading [Roch et al, 2019]
- simulation results has been mixed [e.g., Molloy and Warnow, 2018]

Summary methods

- more scalable
- more accurate when ILS is at least moderate
- can have guarantees of statistical consistency (e.g., [Mirarab et al, 2014])

Image Credit: Siavash Mirarab

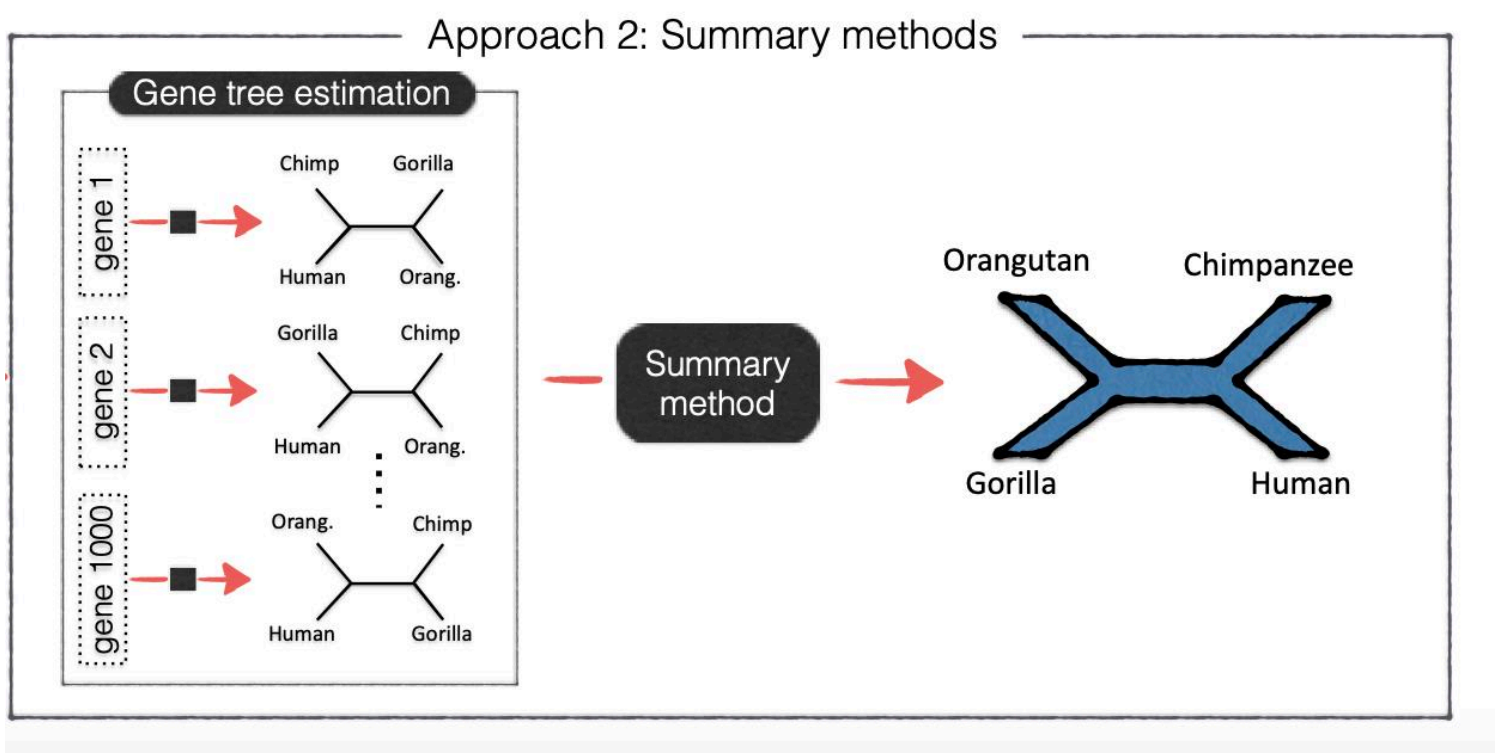
Post-species tree analysis



- Summary methods usually produce an unrooted species tree topology without useful branch lengths

Image Credit: Siavash Mirarab

Post-species tree analysis



- Summary methods usually produce an unrooted species tree topology without useful branch lengths

- Downstream analysis need rooted species trees with branch lengths in mutation-units or time unit

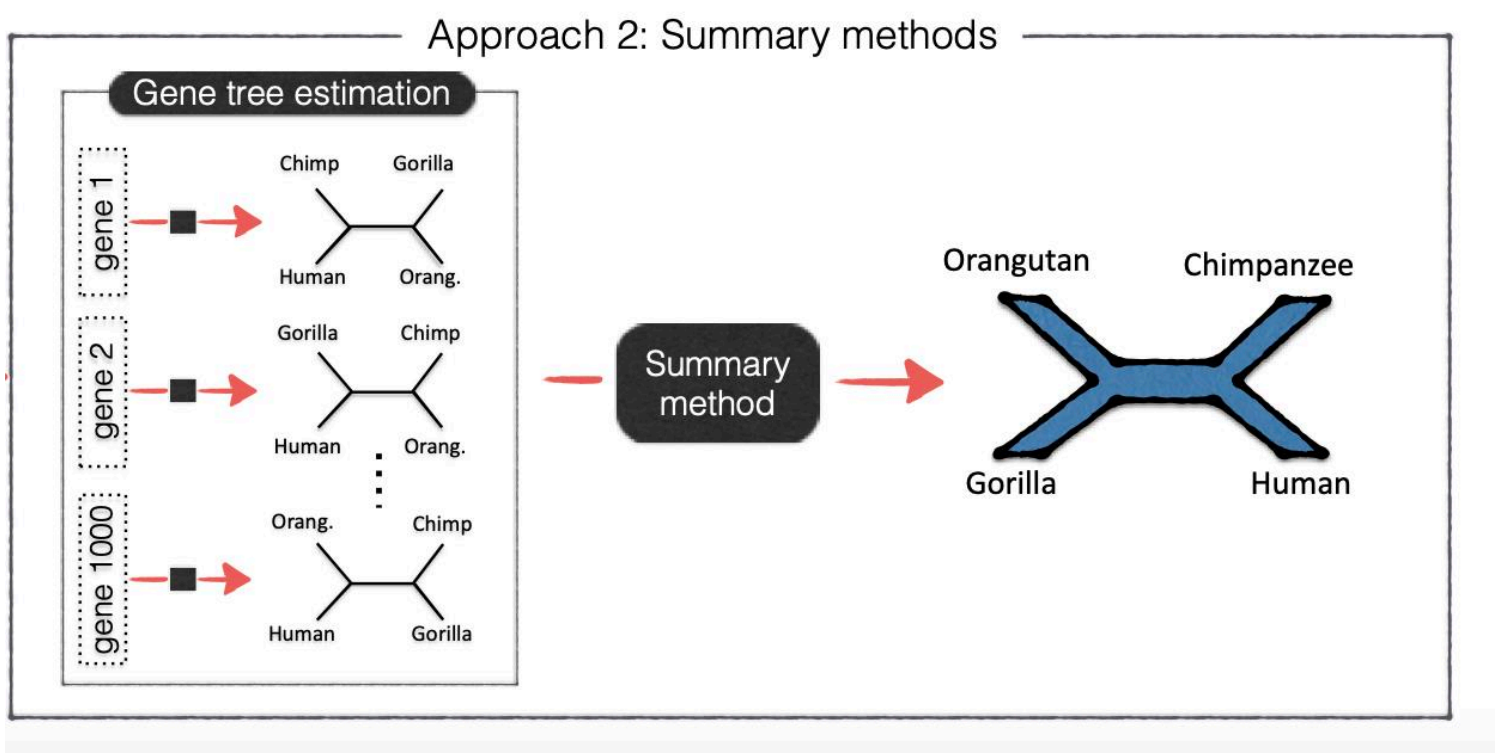
- **Two-step approach**:

1. infer the topology with summary methods (e.g. ASTRAL, MP-EST)

2. infer the root, branch lengths and dates on that fixed topology with additional tools

Image Credit: Siavash Mirarab

Post-species tree analysis



- Summary methods usually produce an unrooted species tree topology without useful branch lengths

- Downstream analysis need rooted species trees with branch lengths in mutation-units or time unit

- **Two-step approach**:

1. infer the topology with summary methods (e.g. ASTRAL, MP-EST)
2. infer the root, branch lengths and dates on that fixed topology with additional tools

  - Based on some form of concatenation analysis
  - Ignores heterogeneity

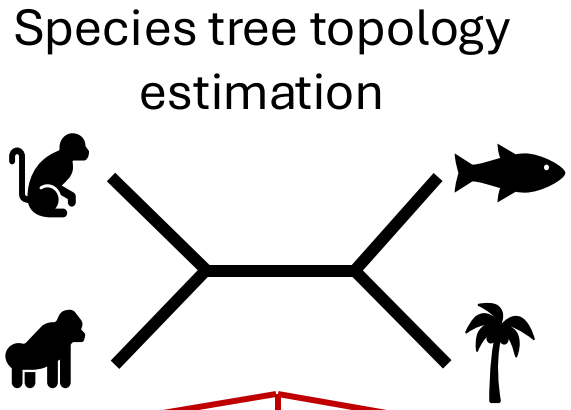
Image Credit: Siavash Mirarab

Outline

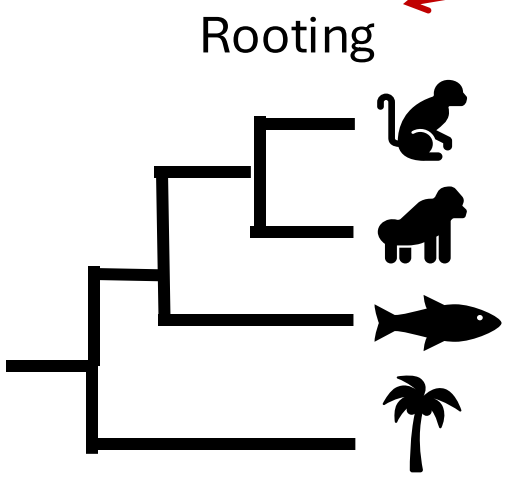
- Background and Motivation
  - Phylogenomics pipeline
  - Gene tree discordance
  - Species tree estimation
- **Overview of Contributions**
  - **Discordance-aware post-species tree analysis**
- Rooting species trees
- Phylogenomic branch length estimation
- Dating species trees and gene trees
- Conclusions

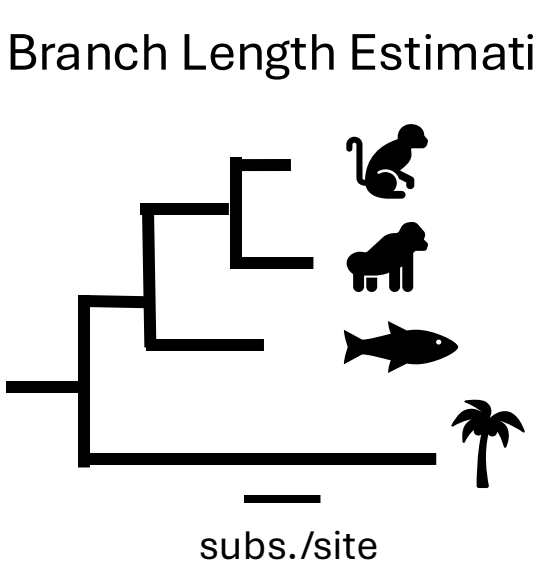
Discordance-aware post-species tree analysis

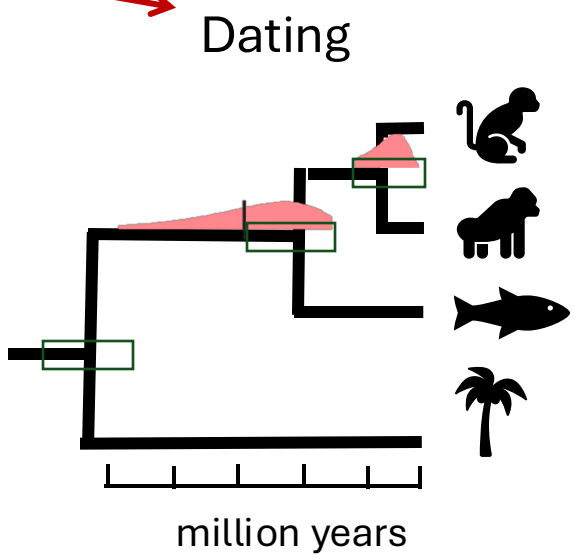
- Many methods for species tree topology estimation exist that account for sources of gene tree discordance [e.g. ASTRAL family, ASTRID, MP-EST, etc]



- Most downstream analysis need a rooted species trees with branch lengths and dates



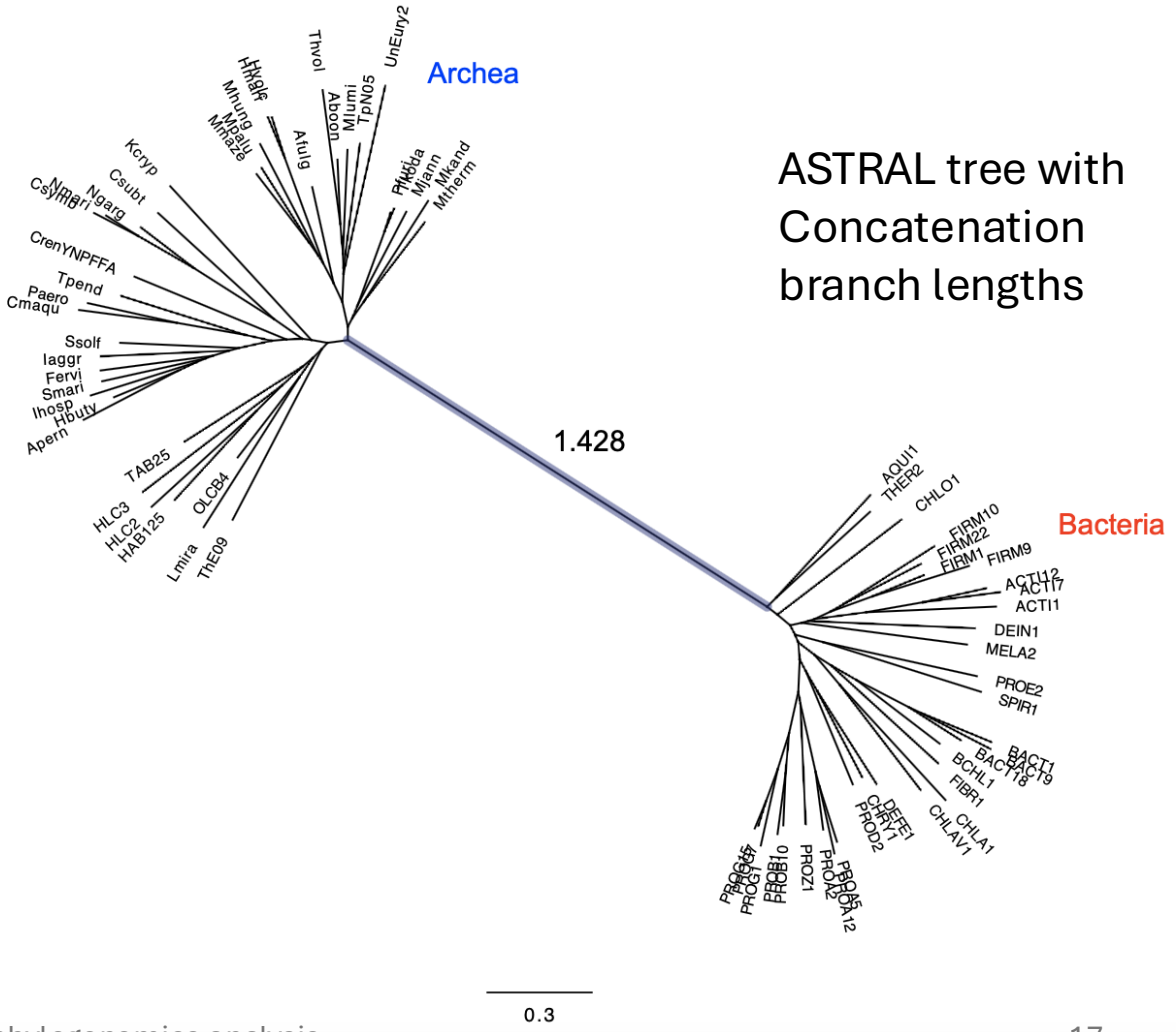




Motivating example: debate about the Archaea-Bacteria (AB) branch length at the root of the tree of life

- Long-standing hypothesis that domains Archaea and Bacteria are separated by a long branch

- Recent studies using large groups of marker genes estimated much lower divergences with concatenation (e.g. Zhu et al, Nature Communications 2019)



Motivating example: debate about the Archaea-Bacteria (AB) branch length at the root of the tree of life

- Moody et al. (2022) suggested that concatenation can substantially underestimate branch lengths in the face of heterogeneity due to HGT

- We need discordance-aware methods that account for heterogeneity…

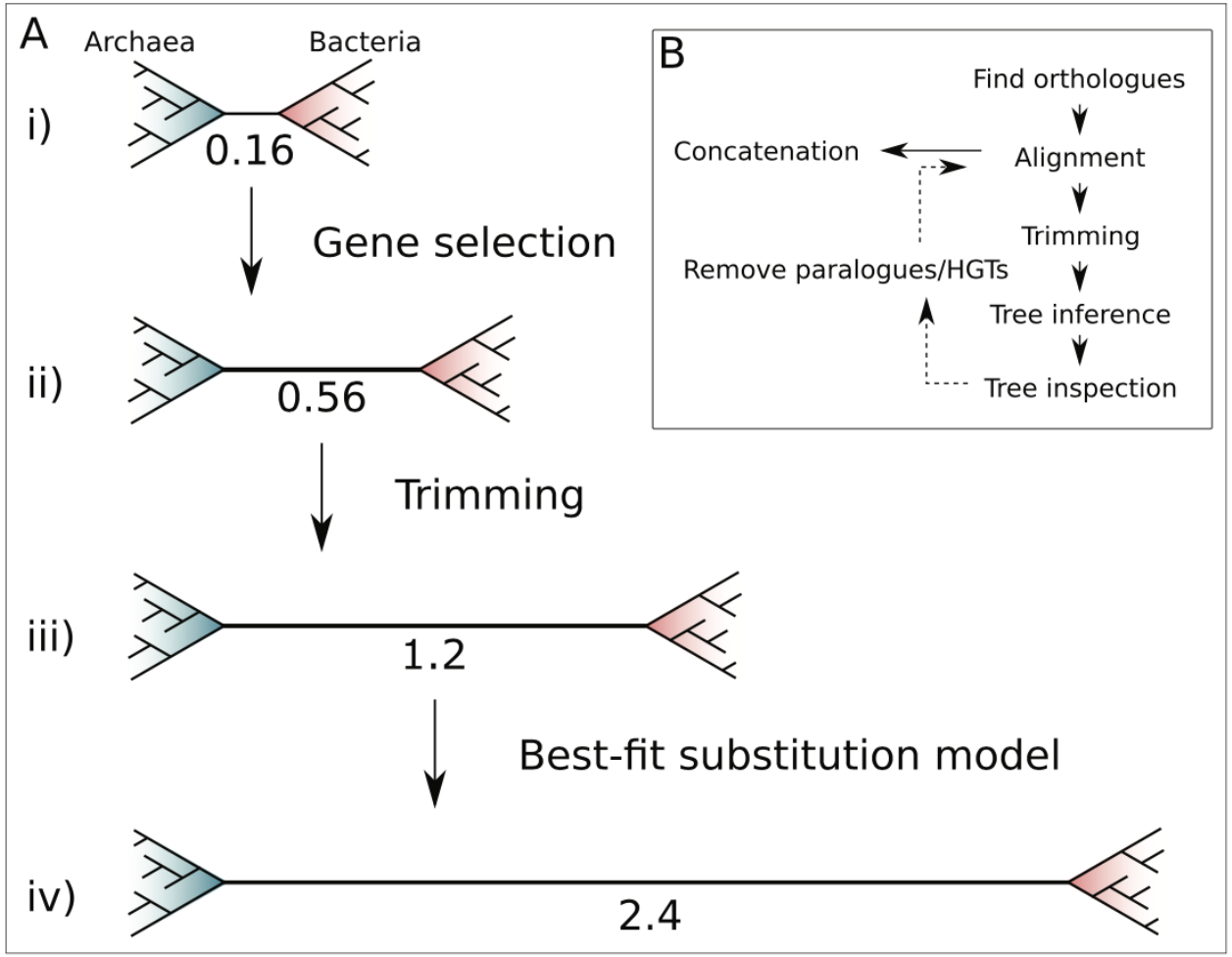
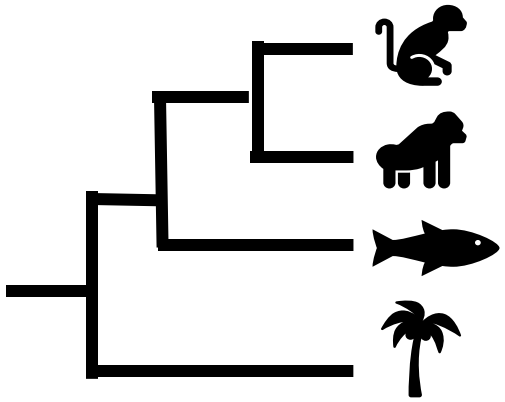


Image Credit: Moody et al., 2022, "An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *Elife*
"

Contributions

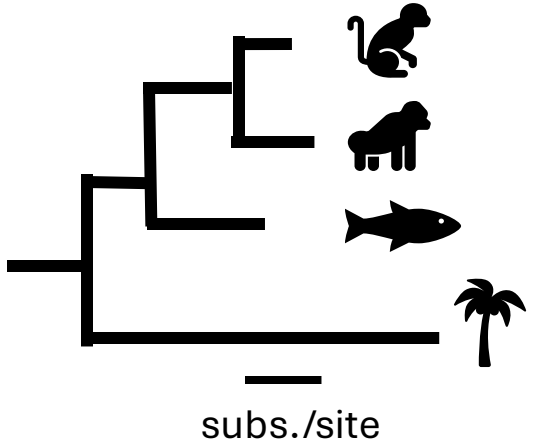
- Discordance-aware methods for post-species tree analysis

Rooting



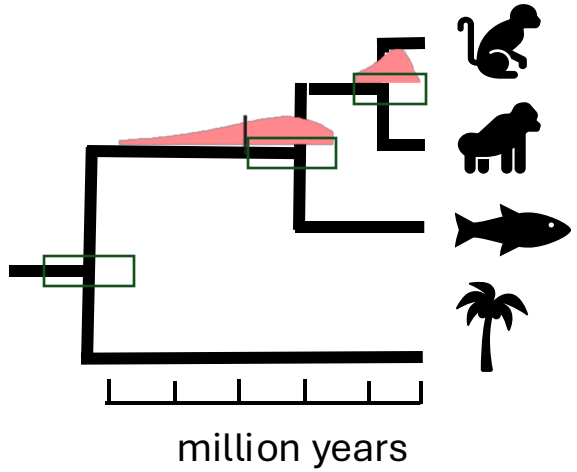
- Quintet Rooting (QR) [Tabatabaee et al., ISMB and Bioinformatics 2022]
- QR-STAR [Tabatabaee et al., RECOMB and JCB 2023]

Branch Length Estimation



- CASTLES [Tabatabaee et al., ISMB and Bioinformatics 2023]
- CASTLES-Pro [Tabatabaee et al., 2025, submitted]

Dating

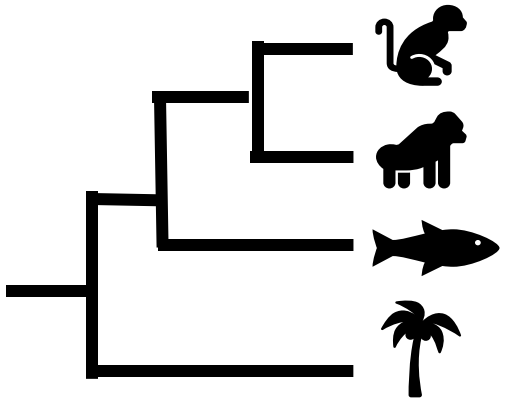


- Coalescent-based dating [Tabatabaee et al., 2025, submitted]

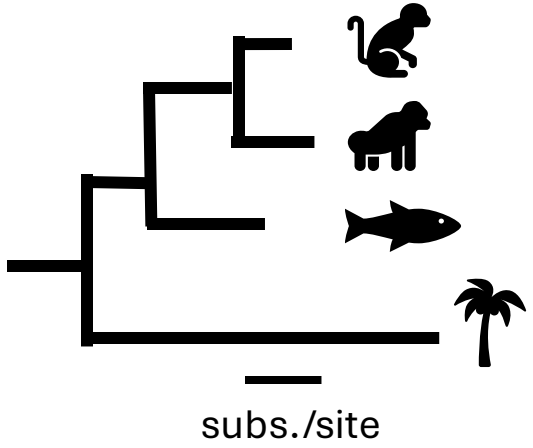
Contributions

- Discordance-aware methods for post-species tree analysis

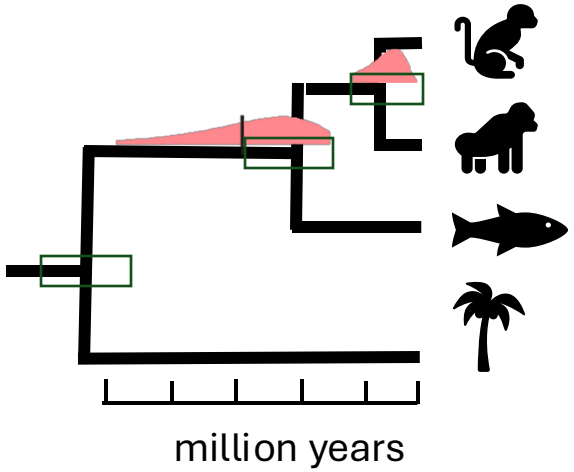
Rooting



Branch Length Estimation



Dating

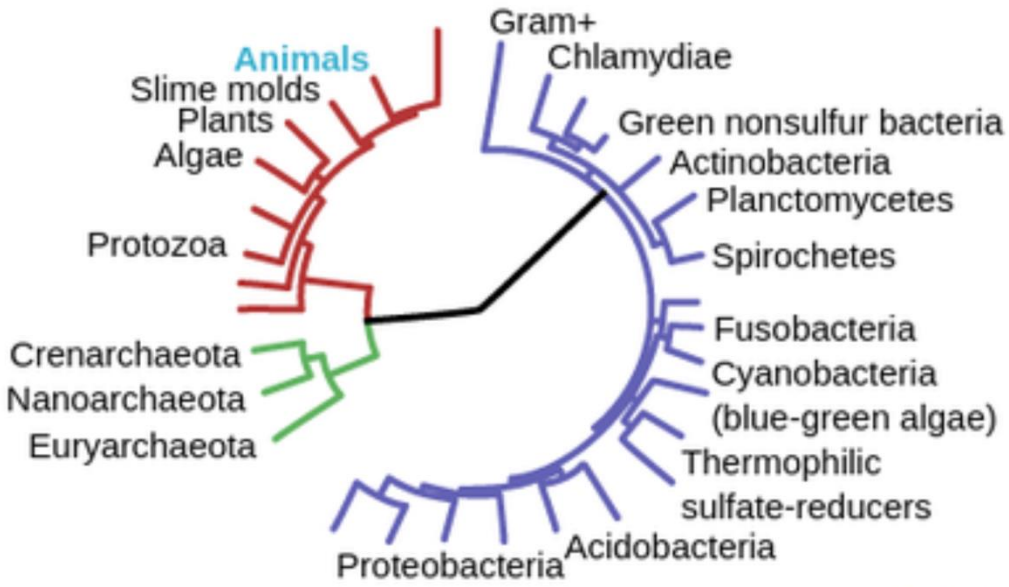


- More scalable
- Better theoretical/statistical properties
- Can be more accurate in the presence of gene tree discordance

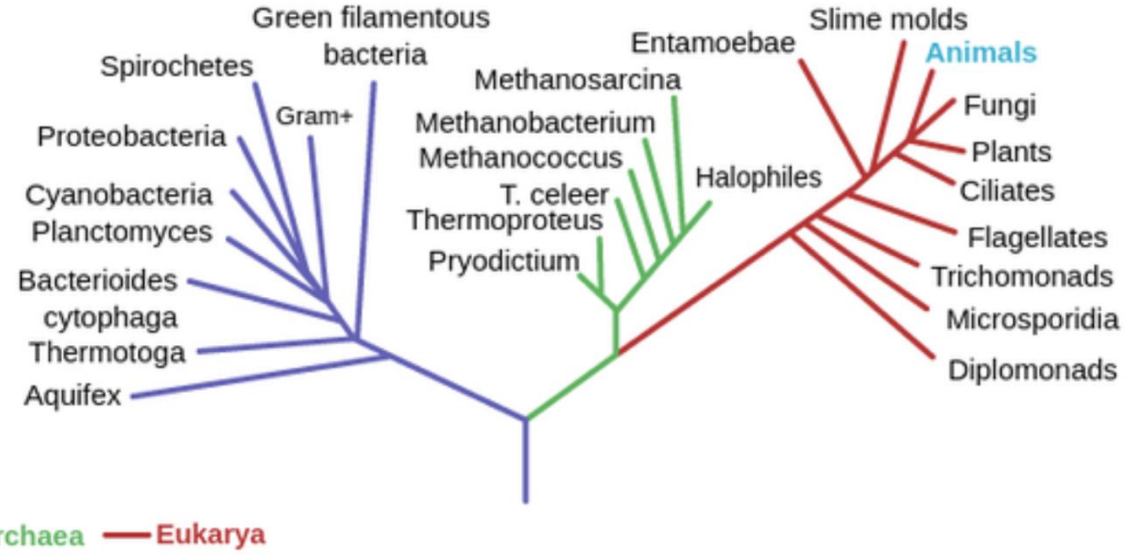
Outline

- Background and Motivation
  - Phylogenomics pipeline
  - Gene tree discordance
  - Species tree estimation
- Overview of Contributions
  - Discordance-aware post-species tree analysis
- **Rooting species trees**
- Phylogenomic branch length estimation
- Dating species trees and gene trees
- Conclusions

Rooting species trees



Unrooted species tree



Rooted species tree

Image Credit: https://theory.labster.com/phylogenetic-tree/, modification of work by Eric Gaba

Why Rooting Species Trees?

- Multiple applications throughout biology
- Understanding
  - Adaptation
  - Biodiversity
  - Phylogeography
  - Co-Evolution
- Most species tree estimation methods don't produce rooted trees


BioEssays

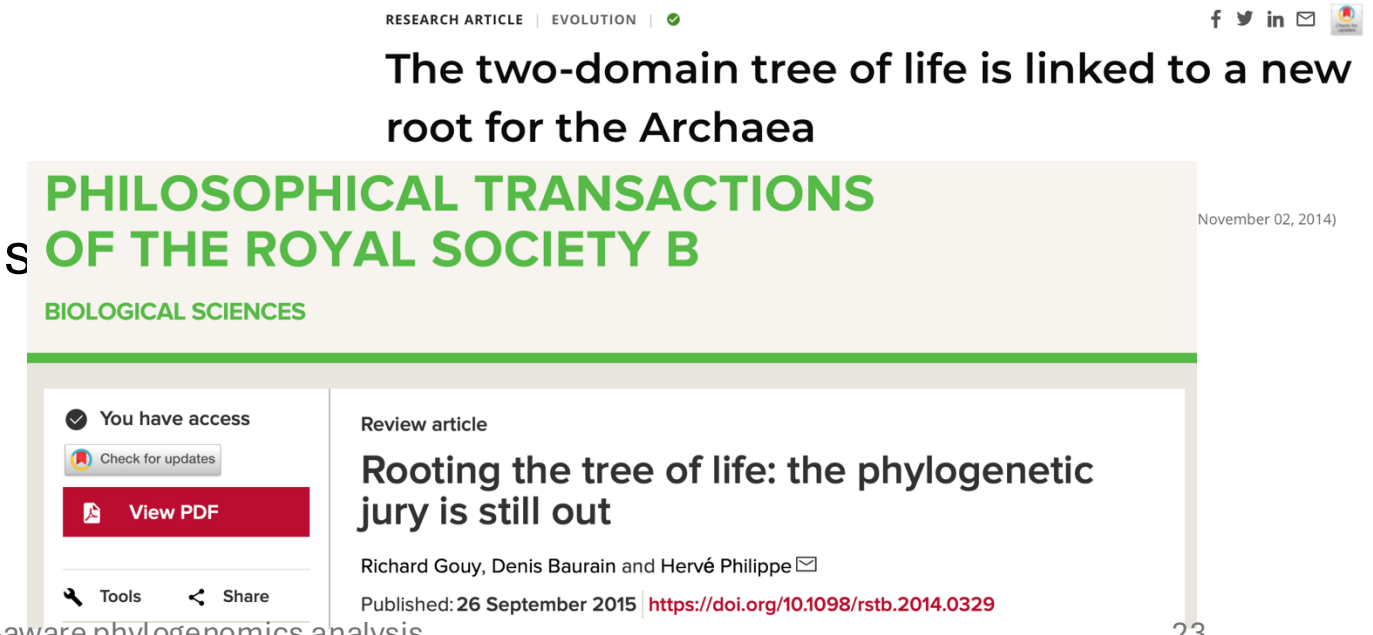
Problems and Paradigms | Full Access

Where is the root of the universal tree of life?

Patrick Forterre, Hervé Philippe

First published: 23 September 1999 |

PNAS ARTICLES FRONT MATTER AUTHORS TOPICS +


RESEARCH ARTICLE | EVOLUTION |

The two-domain tree of life is linked to a new root for the Archaea

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

BIOLOGICAL SCIENCES

November 02, 2014)

You have access

Check for updates

View PDF

Tools Share

Review article

Rooting the tree of life: the phylogenetic jury is still out

Richard Gouy, Denis Baurain and Hervé Philippe

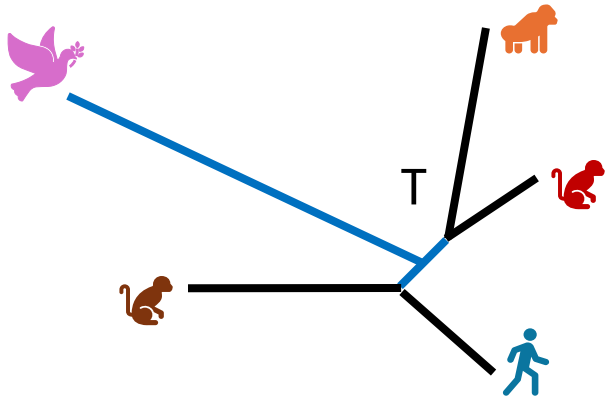
Published: 26 September 2015 https://doi.org/10.1098/rstb.2014.0329

Common Approaches for Rooting Species Trees

Problem: Find the root position in a given unrooted species tree T.

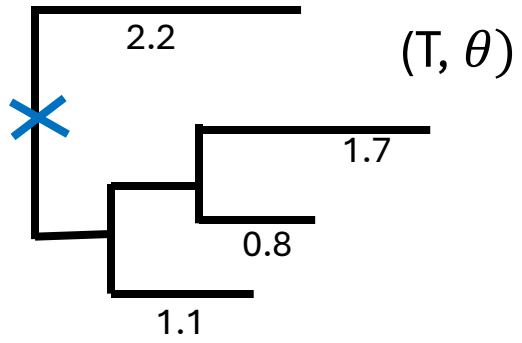
Outgroup Rooting

- Needs prior information about taxa
- Selecting a proper outgroup can be challenging



Distance-Based

- Species tree with branch lengths (e.g. Midpoint, MAD, MinVar, …)
- Most are sensitive to deviations from the molecular clock



These methods do not account for the biological processes that create discordance between gene trees and species trees.

ADR: Identifiability of **Unrooted** Topology under MSC

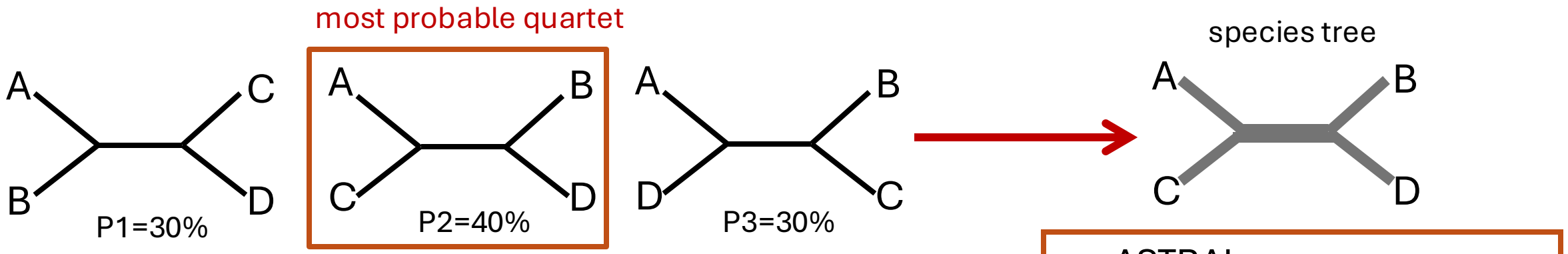
***Theorem*: For 4 or more species, the unrooted topology of the species tree is identifiable from the probability distribution of the unrooted gene trees.** [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]

ADR: Identifiability of **Unrooted** Topology under MSC

***Theorem*: For 4 or more species, the unrooted topology of the species tree is identifiable from the probability distribution of the unrooted gene trees.** [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]

Key property: For 4 species, the most probable unrooted gene tree has the same topology as the unrooted species tree

- Does not hold for more than 4 species



Statistically consistent **quartet-based** species tree estimation methods

ASTRAL [Mirarab et al, 2014]
BUCKy-pop [Larget et al, 2010]
wQFM [Mahbub et al, 2021]
…

ADR: Identifiability of **Rooted** Topology under MSC

***Theorem*: For 5 or more species, the rooted topology of the species tree is identifiable from the probability distribution of the unrooted gene trees.** [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]

ADR: Identifiability of **Rooted** Topology under MSC

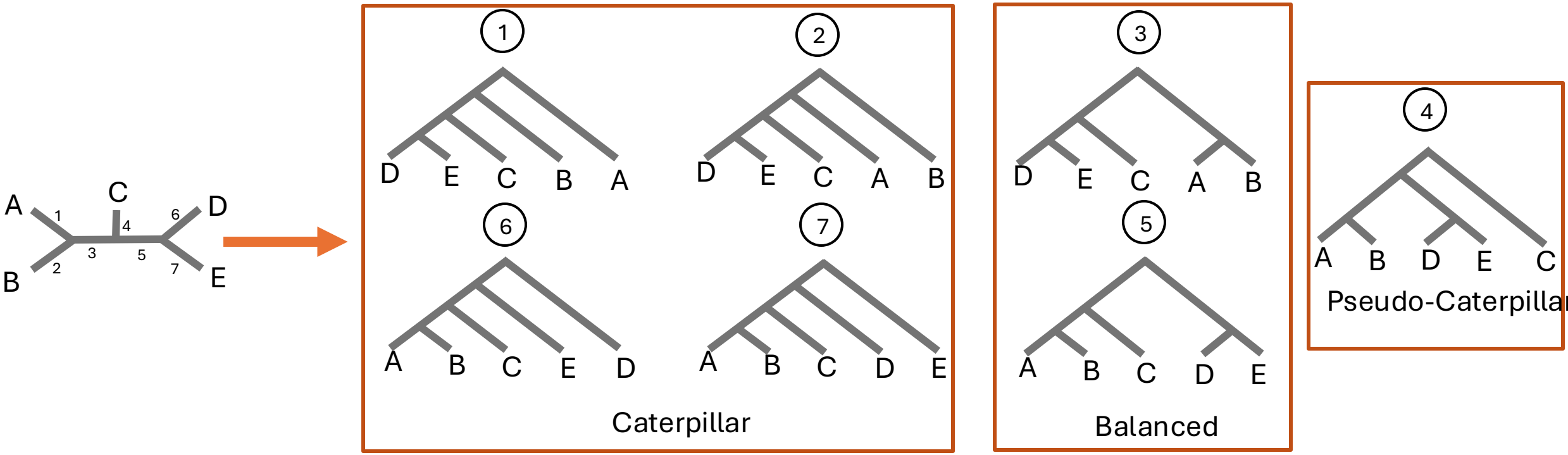
***Theorem*: For 5 or more species, the rooted topology of the species tree is identifiable from the probability distribution of the unrooted gene trees. [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]**

- ADR derive linear invariants and inequalities on the probability distribution of unrooted gene trees.
- They prove that these inequalities and invariants suffice to identify the rooted species tree

If the probability distribution of unrooted gene trees is exactly known, there will be exactly one rooted species tree topology satisfying all invariants and inequalities.

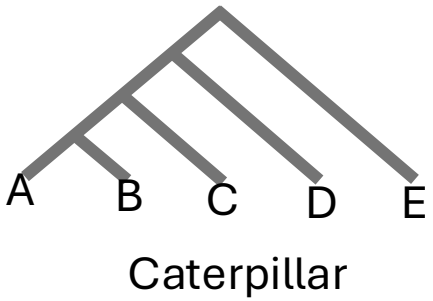
Properties of Quintet Trees

- There are **105** rooted binary trees and **15** unrooted binary trees on 5 taxa
- Each unrooted 5-taxon tree can be rooted on any of its **7** edges
- Rooted 5-taxon trees fall into **three** different shapes: caterpillar, balanced and pseudo-caterpillar [Rosenberg, 2007]

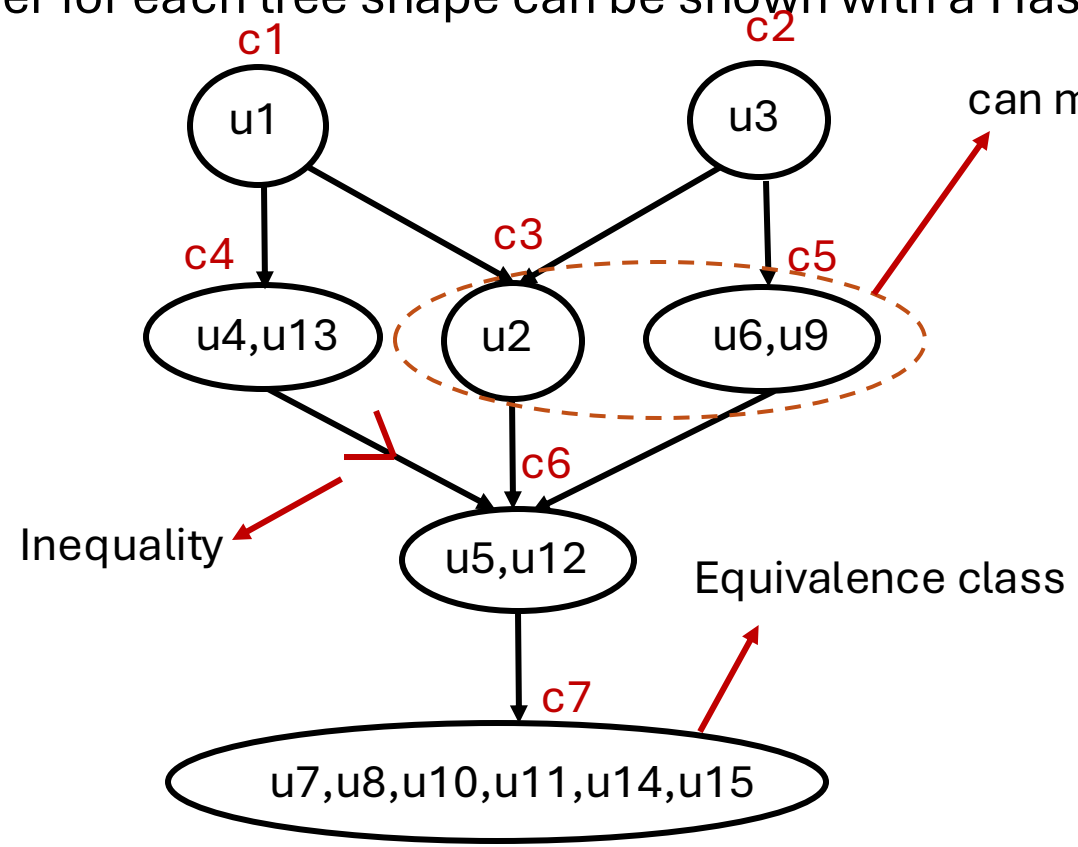


ADR Invariants & Inequalities

- ADR invariants and inequalities define a partial order on the distribution of unrooted gene trees $\vec{u}$
- The partial order for each tree shape can be shown with a Hasse diagram



Caterpillar



- 15 5-taxon unrooted topologies $T_1, \ldots, T_{15}$
- $u_i = \mathbb{P}(T_i)$

Example of invariants:

- $u_4 = u_{13}$

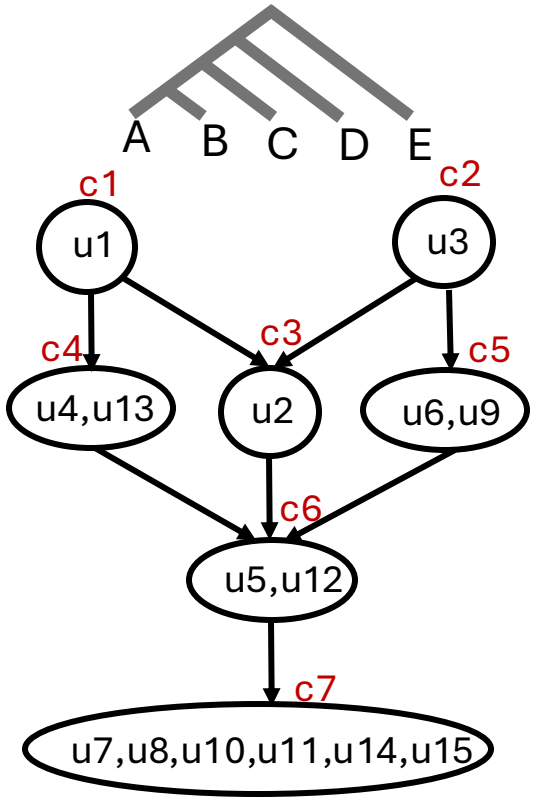
Example of inequalities:

- $u_4 > u_5$

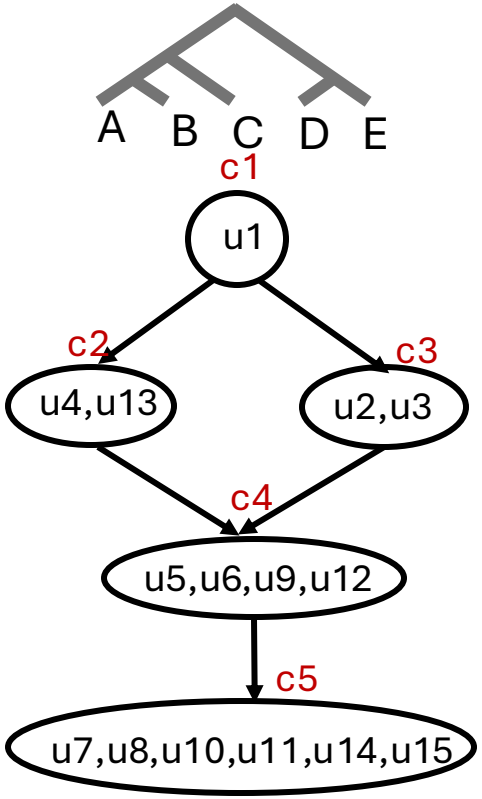
- Equivalence classes that are not related by inequalities can merge for some values of branch lengths

Tabatabaee et al., "Quintet Rooting: rooting species trees under the multi-species coalescent model". Bioinformatics and ISMB 2022

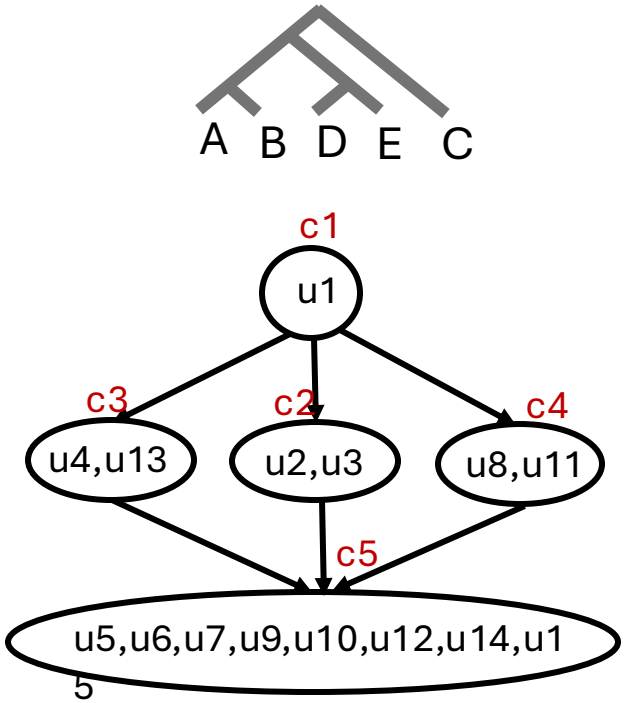
ADR Invariants & Inequalities



Caterpillar



Balanced



Pseudo-Caterpillar

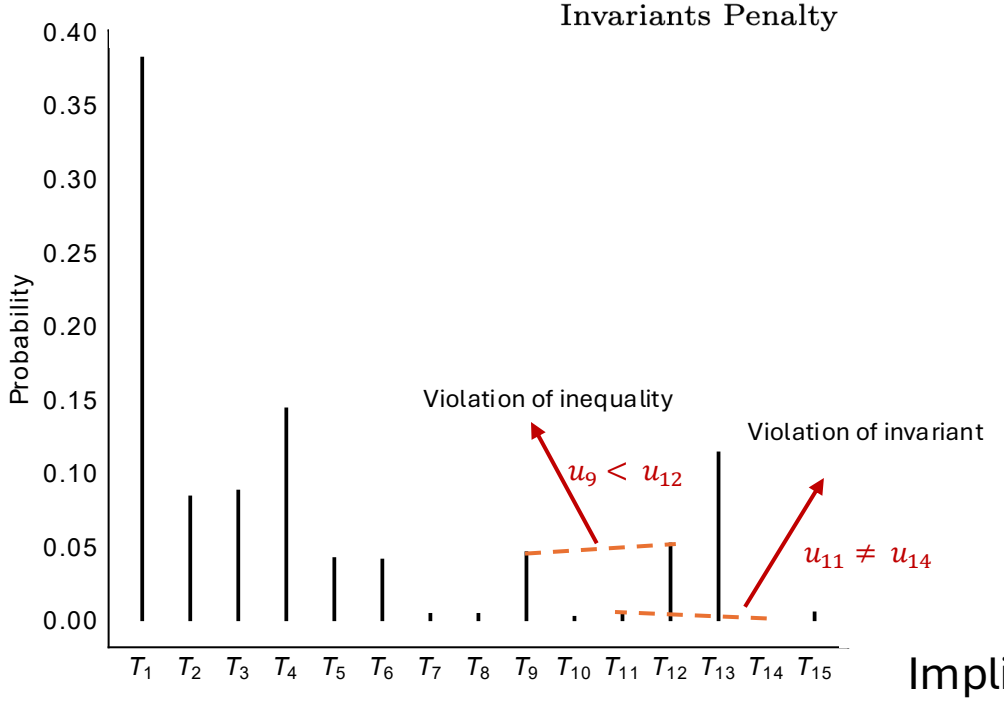
- According to ADR theory, each 105 rooted binary tree corresponds to a unique Hasse diagram
- The shape of this diagram only depends on the topological shape of the tree
- The rooted tree is identifiable from the probability distribution $\vec{u} = (u_1, u_2, \ldots, u_{15})$

Tabatabaee et al., "Quintet Rooting: rooting species trees under the multi-species coalescent model". Bioinformatics and ISMB 2022

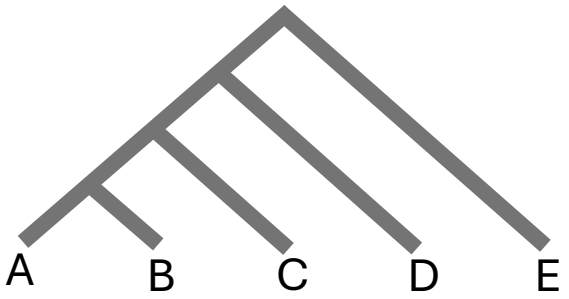
Cost : Fitness between a Tree and a Distribution

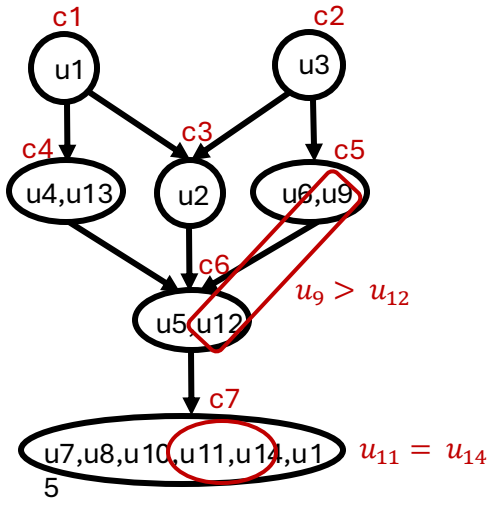
- Measures the fitness between a distribution and a tree (i.e. its partial order)
- Linear combination of invariant and inequality penalty terms

$$\mathrm{Cost}(r, \vec{\hat{u}}) = \underbrace{\sum_{c \in C_r} \frac{1}{|c|} \sum_{u_a, u_b \in c} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c > c' \in C_r} \frac{1}{|c'|} \sum_{u_a \in c, u_b \in c'} \max(0, \hat{u}_b - \hat{u}_a)}_{\text{Inequalities Penalty}} .$$



Estimated gene tree distribution $\vec{u}$





A model tree R and its partial order

Implied by the distribution:

- $u_{11} \neq u_{14}$
- $u_9 < u_{12}$

violations

Implied by the partial order:

- $u_{11} = u_{14}$
- $u_9 > u_{12}$

Tabatabaee et al., "Quintet Rooting: rooting species trees under the multi-species coalescent model". Bioinformatics and ISMB 2022

QR: Rooting Species Trees under MSC

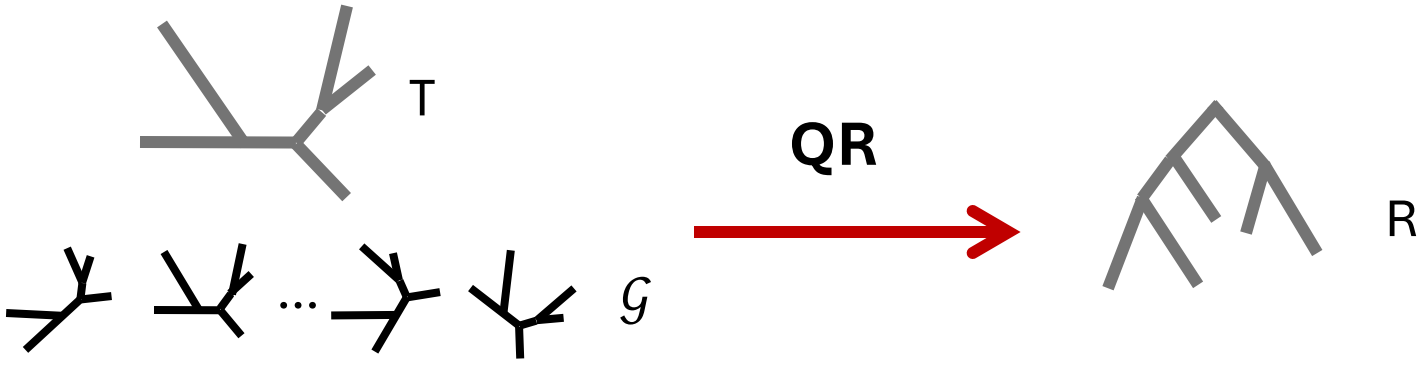
Input

- An unrooted species tree T.
- A set of k unrooted single-copy gene trees $\mathcal{G}$ on $\mathcal{L}(T)$.
- A cost function $Cost(R, \vec{u})$.

Output

- A rooted version of T that minimizes

$$Score(R,T) = \sum_{q \in Q^*} Cost(q, \overrightarrow{u_q})$$

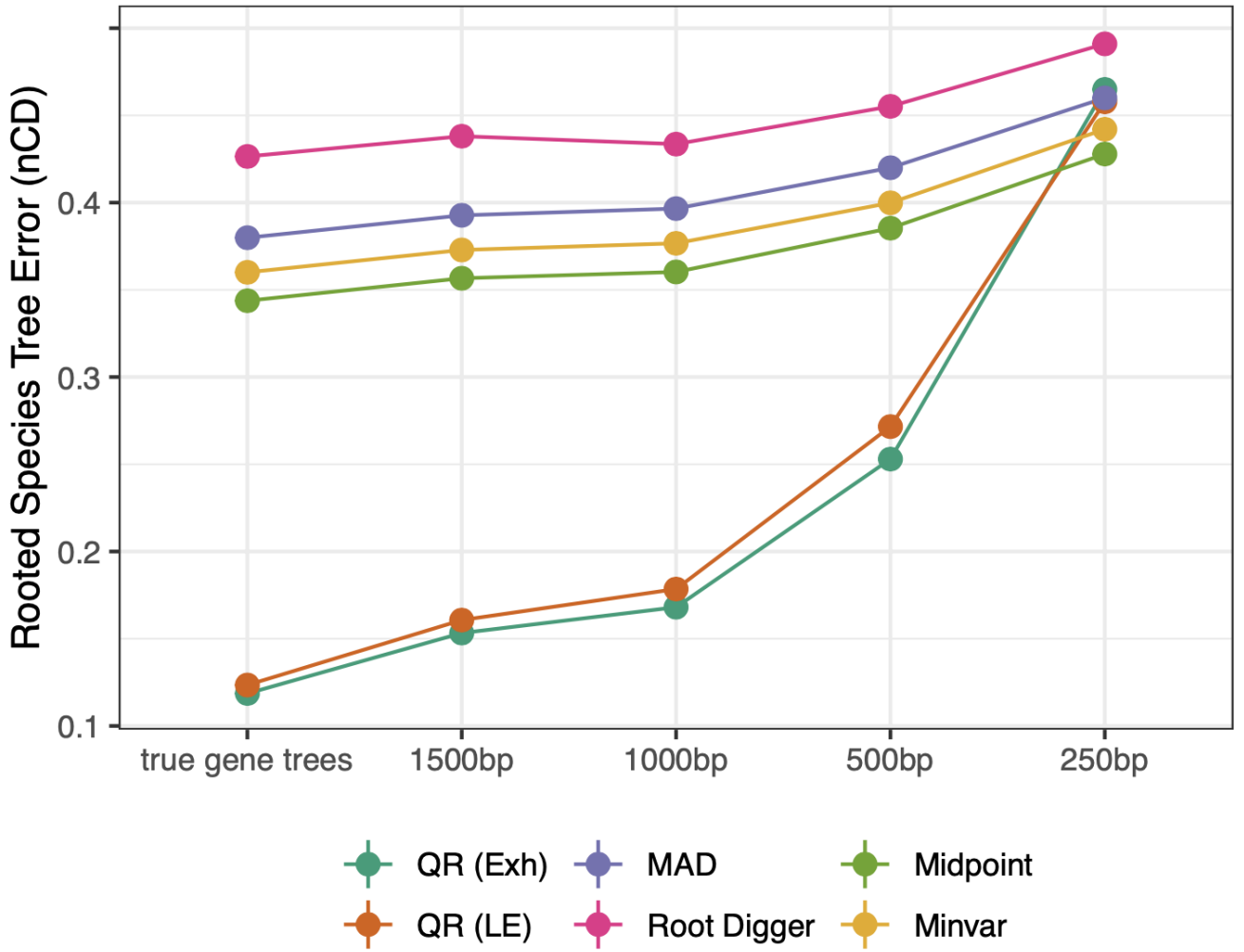


Tabatabaee et al., "Quintet Rooting: rooting species trees under the multi-species coalescent model". Bioinformatics and ISMB 2022

QR provides accurate rooting of species trees in the presence of ILS

- Simulated dataset with moderate ILS, 1000 gene, 10-taxon trees [Mirarab et al (2014)]

- nCD: normalized clade distance, extension of Robinson-Foulds (RF) distance for rooted trees.



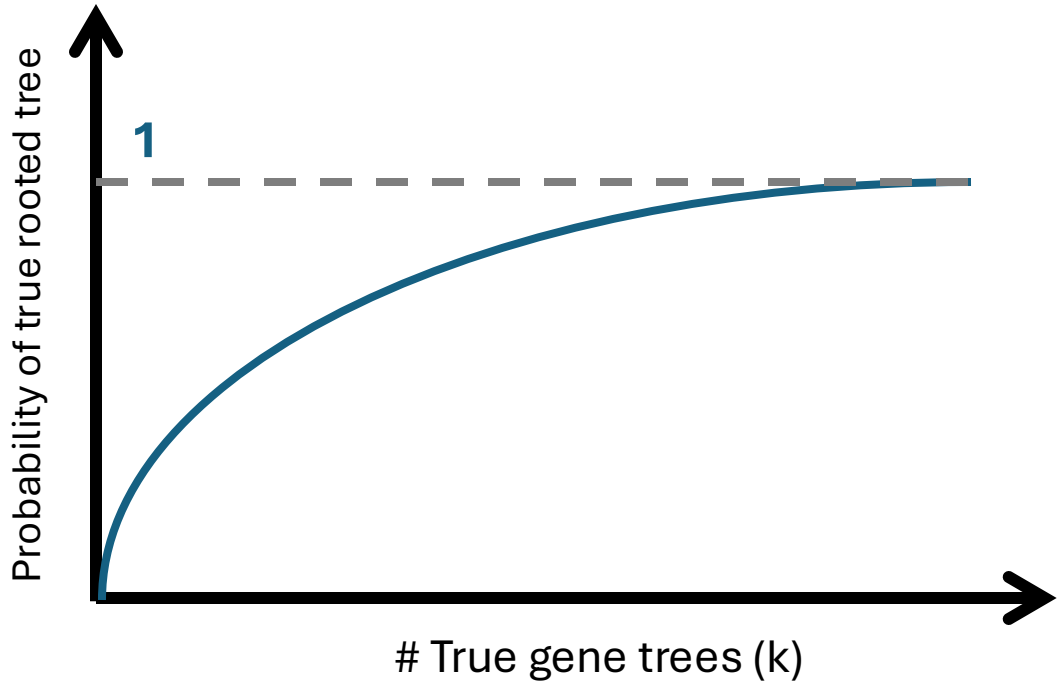
- QR uses gene trees as input, other methods use a species tree with concatenation branch lengths

Tabatabaee et al., "Quintet Rooting: rooting species trees under the multi-species coalescent model". Bioinformatics and ISMB 2022

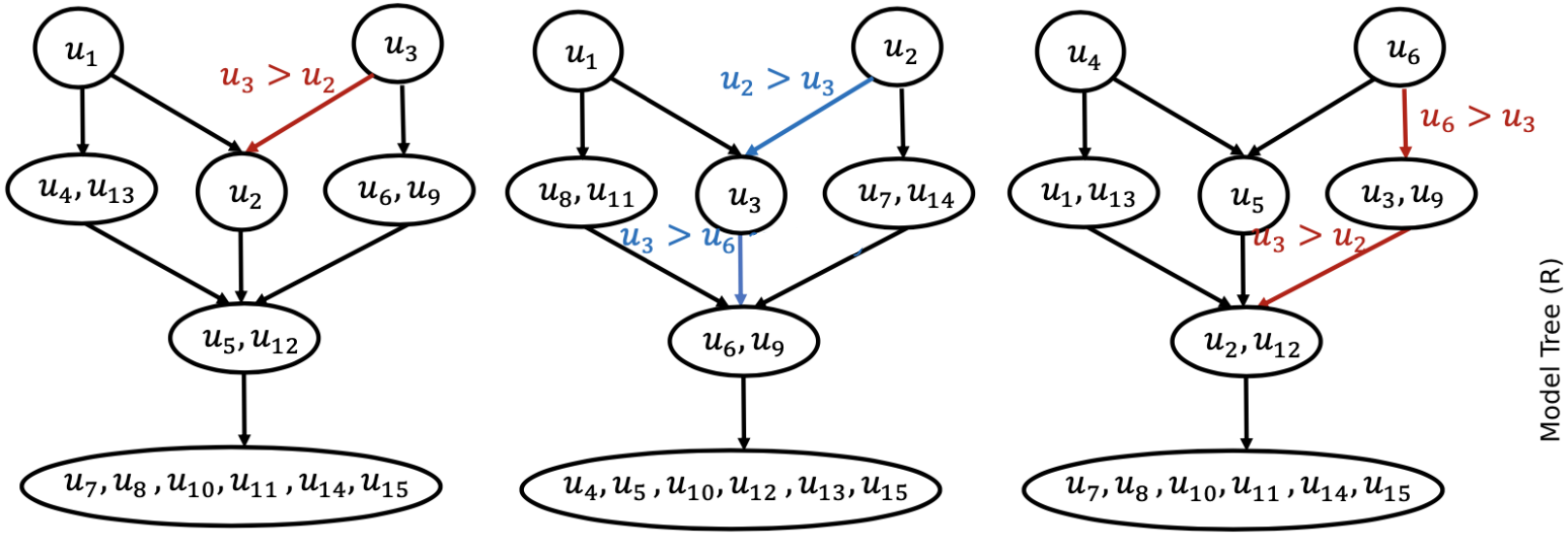
Statistical Consistency

An estimation method is statistically consistent under a model, if its output converges to the true parameter as the number of input samples increase. **(based on proof)**

- Several methods proven statistically consistent estimators of ***unrooted*** species tree topology under MSC (ASTRAL, ASTRID, BUCKy-pop)
- No consistency result for ***rooting*** methods



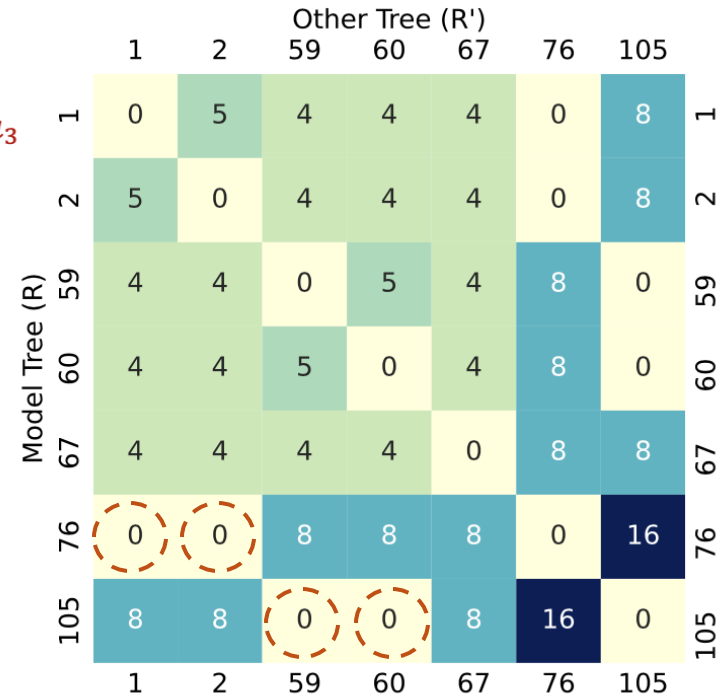
Is QR statistically consistent?

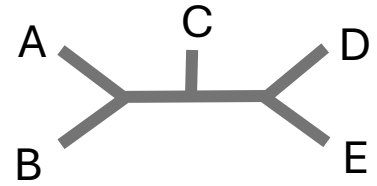


$R_1 = (((a, b), c), d), e)$ $R_4 = (((a, b), d), e), c)$ $R_7 = (((a, c), b), d), e)$

$$V(R_1, R_4) = \{\{2, 3\}\} \rightarrow |V(R_1, R_4)| = 1$$
$$V(R_7, R_4) = \{\{2, 3\}, \{3, 6\}\} \rightarrow |V(R_7, R_4)| = 2$$

Heatmap showing |V(R,R')|





- There are pairs of trees whose partial orders have no conflicts → QR is not consistent.

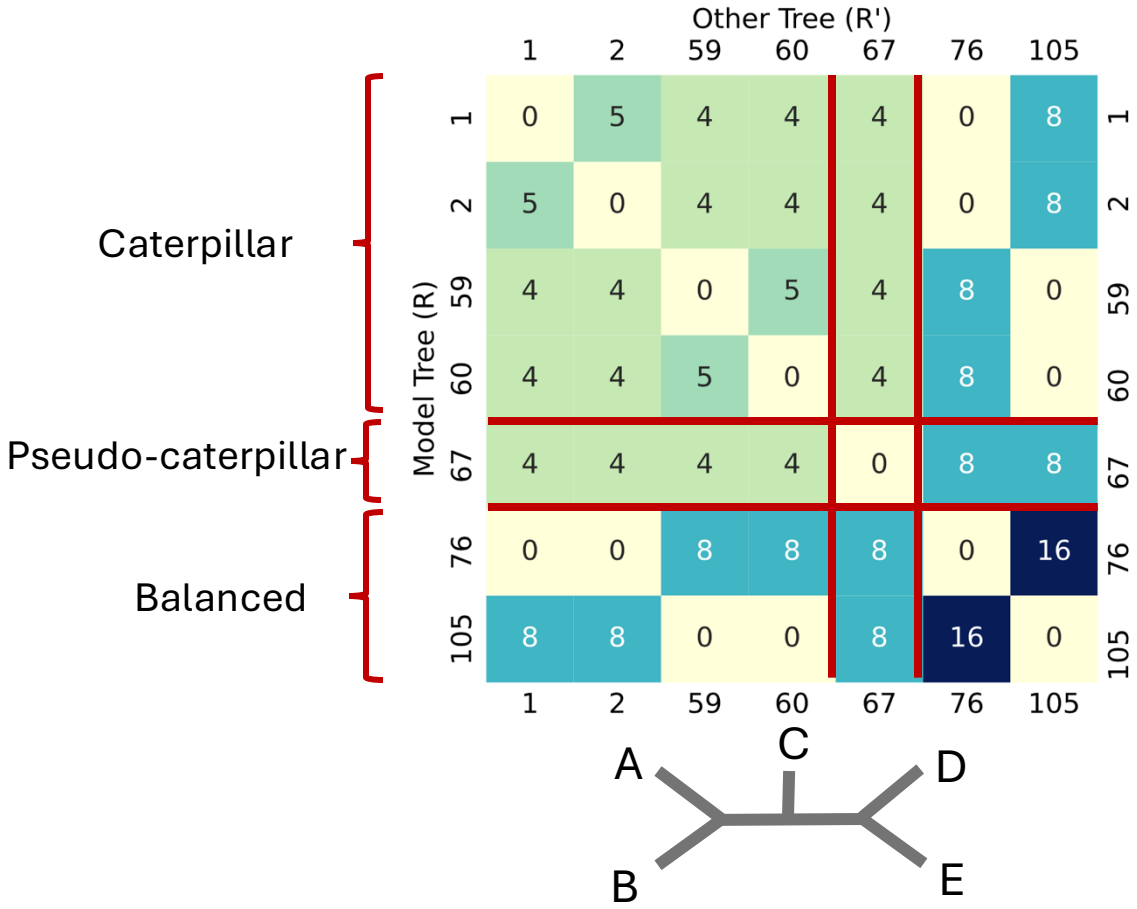
Tabatabaee et al., "Statistically consistent rooting of species trees under the multispecies coalescent model". RECOMB and JCB 2023

Key Idea behind QR-STAR

- Pairs of trees with the same rooted topological shape (caterpillar, balanced, pseudo-caterpillar) always have conflicting distributions

- Idea:
  - Determine the topological shape of each quintet
  - Incorporate the topological shape in the cost function

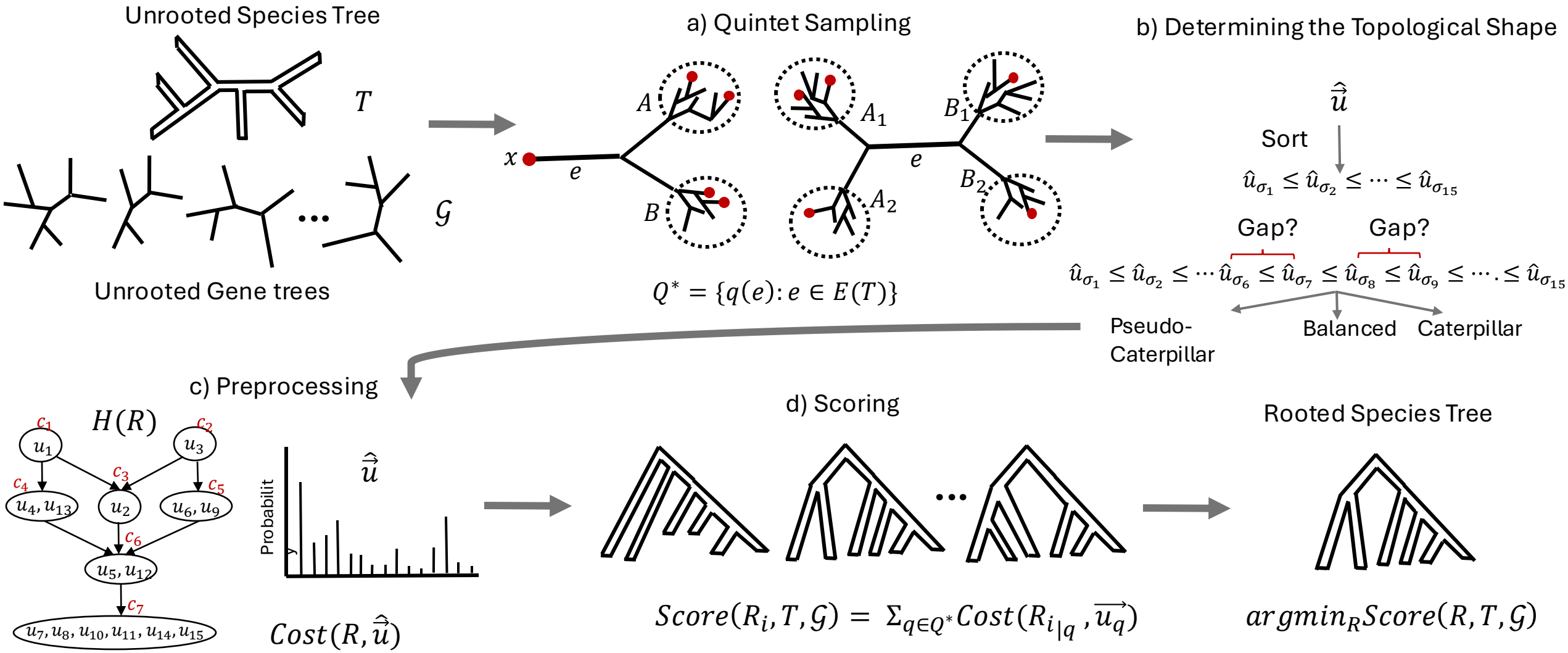
Heatmap of the number of conflicts between pairs of trees with the same unrooted topology



Tabatabaee et al., "Statistically consistent rooting of species trees under the multispecies coalescent model". RECOMB and JCB 2023

QR-STAR Algorithm

Runtime: O(nk)



Tabatabaee et al., "Statistically consistent rooting of species trees under the multispecies coalescent model". RECOMB and JCB 2023

Theoretical guarantees of QR-STAR

Definition 4. Let T be an unrooted tree and Q be a set of quintets of taxa from $\mathcal{L}(T)$. We say Q is "root-identifying" if every rooted tree R with topology T is identifiable from T and the set of rooted quintet trees in $\{R|_q : q \in Q\}$, that is, no two rooted trees with topology T induce the same set of rooted quintet trees on Q.

Theorem 2 (Statistical Consistency of QR-STAR). Let R be an MSC model species tree with $n \geq 5$ leaves and let T denote its unrooted topology. Given T and a set $\mathcal{G}$ of unrooted true gene trees on the leafset $\mathcal{L}(T)$, QR-STAR is a statistically consistent estimator of the rooted version of T under the MSC, if the set of sampled quintets Q is root-identifying.

Corollary. QR-STAR with linear encoding is polynomial time and statistically consistent under the MSC.

Tabatabaee et al., "Statistically consistent rooting of species trees under the multispecies coalescent model". RECOMB and JCB 2023

Statistical Consistency of QR-STAR (Proof Sketch)

- As the number of input gene trees increase
  - Probability that the first step of QR-STAR correctly determines the rooted shape of each quintet converges to 1
  - The cost of true rooted quintet becomes arbitrarily close to zero
  - The cost of any other rooted quintet is bounded away from zero, where the bound depends on the *path length parameter* of the model tree $h(R)$
  - The set of quintets sampled in QR-STAR is selected so that each two different rooted trees define different set of quintets
- Therefore, the probability that QR-STAR correctly roots the given unrooted tree converges to 1

Tabatabaee et al., "Statistically consistent rooting of species trees under the multispecies coalescent model". RECOMB and JCB 2023

Sample Complexity of QR-STAR

Theorem 3 (Sample Complexity of QR-STAR). Let R be an MSC model species tree with $n \geq 5$ leaves and let T denote its unrooted topology. Given T, Q (a root-identifying set of sampled quintet trees), $\delta > 0$, and k true unrooted gene trees on the leafset of T, QR-STAR returns the true tree R with probability at least $1-\delta$, when the number of gene trees satisfies:

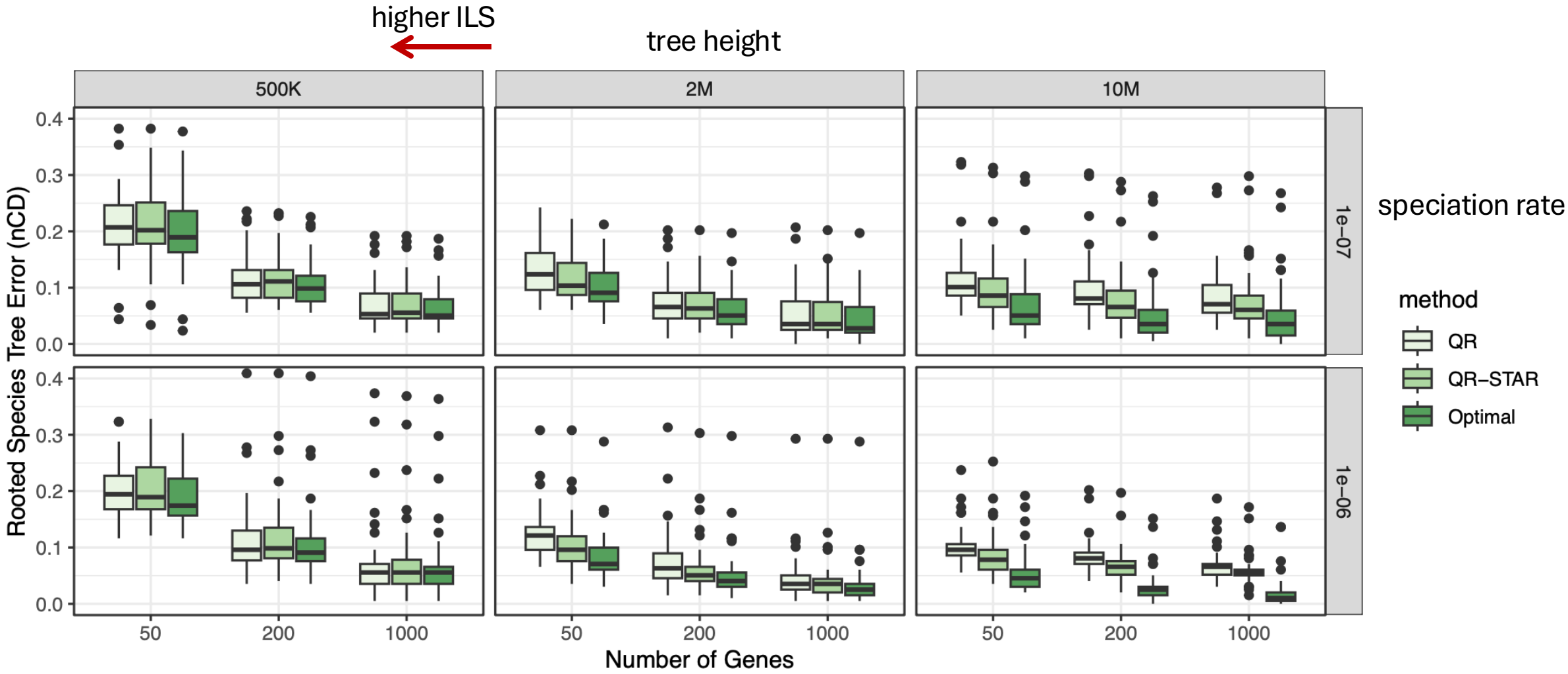
$$k > 2 \times 36^2 \ln\left(\frac{30|Q|}{\delta}\right) \frac{e^{6g}}{(1-e^{-f})^4} \tag{23}$$

where f and g are the lengths of the shortest internal branch and the longest internal path in R, respectively. When the linear encoding is used so that $|Q| = 2n-3$ and in the limit of small f, QR-STAR returns the true rooted tree with probability at least $1-\delta$ when the number k of gene trees satisfies:

$$k = \Omega(f^{-4} e^{6g} (\ln(n) - \ln(\delta))). \tag{24}$$

Tabatabaee et al., "Statistically consistent rooting of species trees under the multispecies coalescent model". RECOMB and JCB 2023

QR-STAR improves over QR in most conditions



- 200-taxon simulated ILS datasets [Mirarab and Warnow (2015)], rooting the ASTRAL species tree

Tabatabaee et al., "Statistically consistent rooting of species trees under the multispecies coalescent model". RECOMB and JCB 2023

Outline

- Background and Motivation
  - Phylogenomics pipeline
  - Gene tree discordance
  - Species tree estimation
- Overview of Contributions
  - Discordance-aware post-species tree analysis
- Rooting species trees
- **Phylogenomic branch length estimation**
- Dating species trees and gene trees
- Conclusions

Species tree branch length estimation

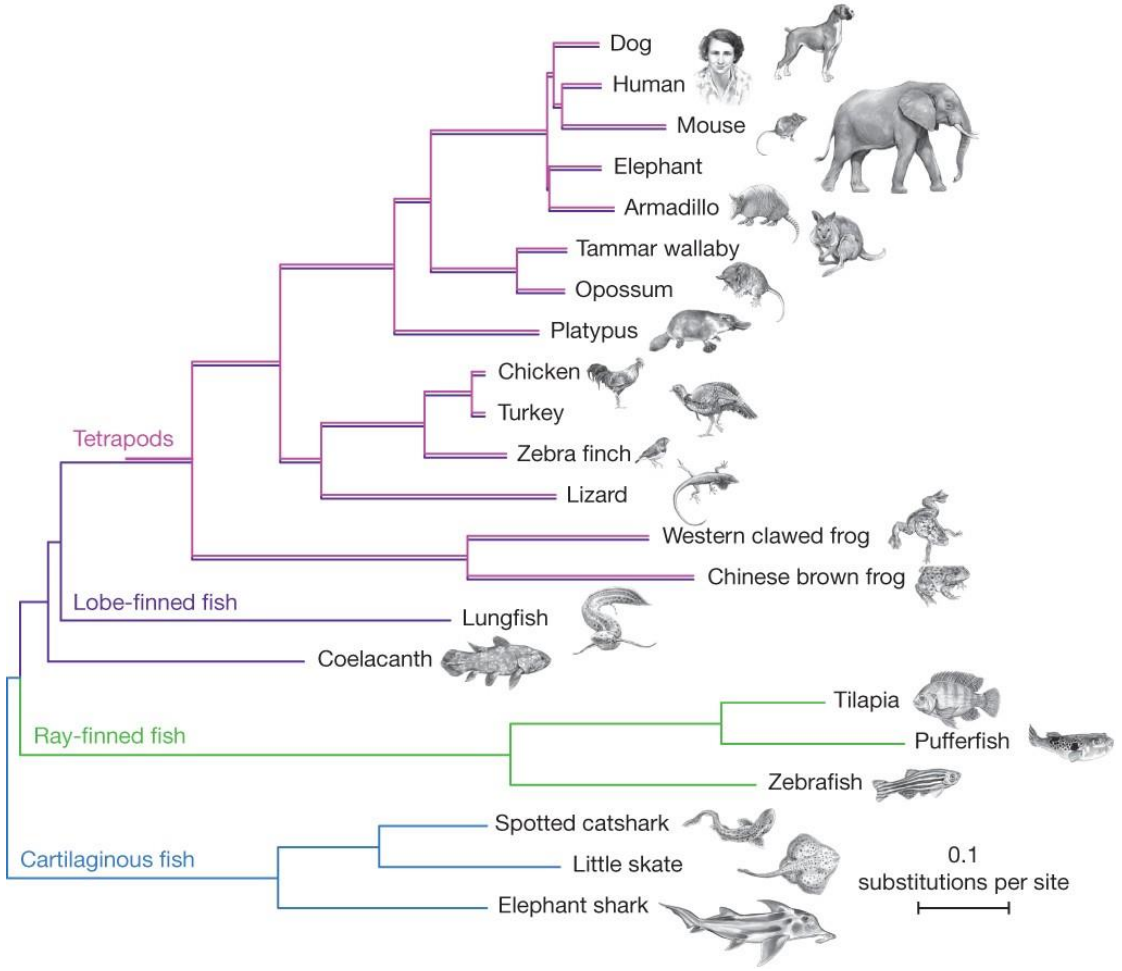


Image Credit: https://plato.stanford.edu/entries/phylogenetic-inference/

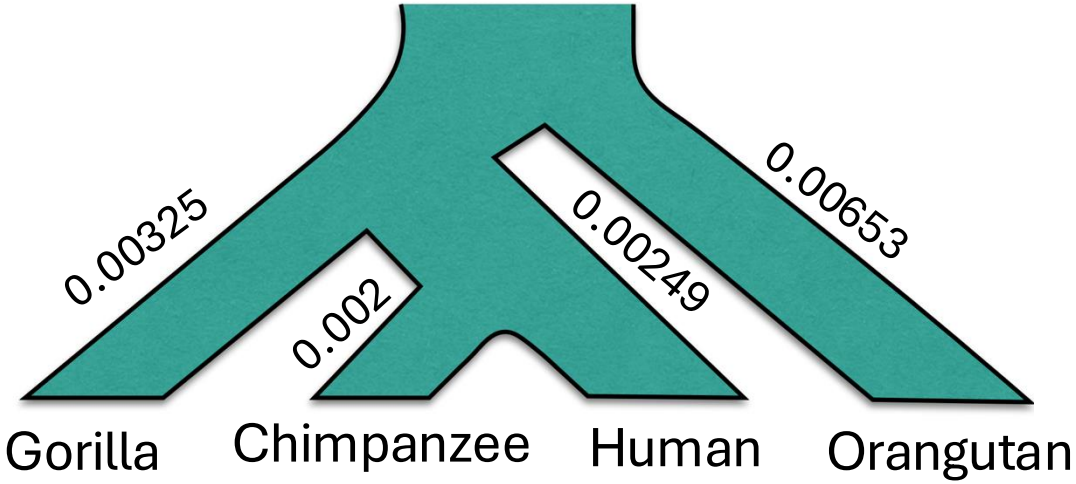
Branch lengths are necessary for downstream analysis

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

$\mu = 1/12$

AGCAGCATCGTG
AGCAGCGTCGTG

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



- Branch lengths inferred by summary methods do not directly lead to SU branch lengths and are only inferable for internal branches

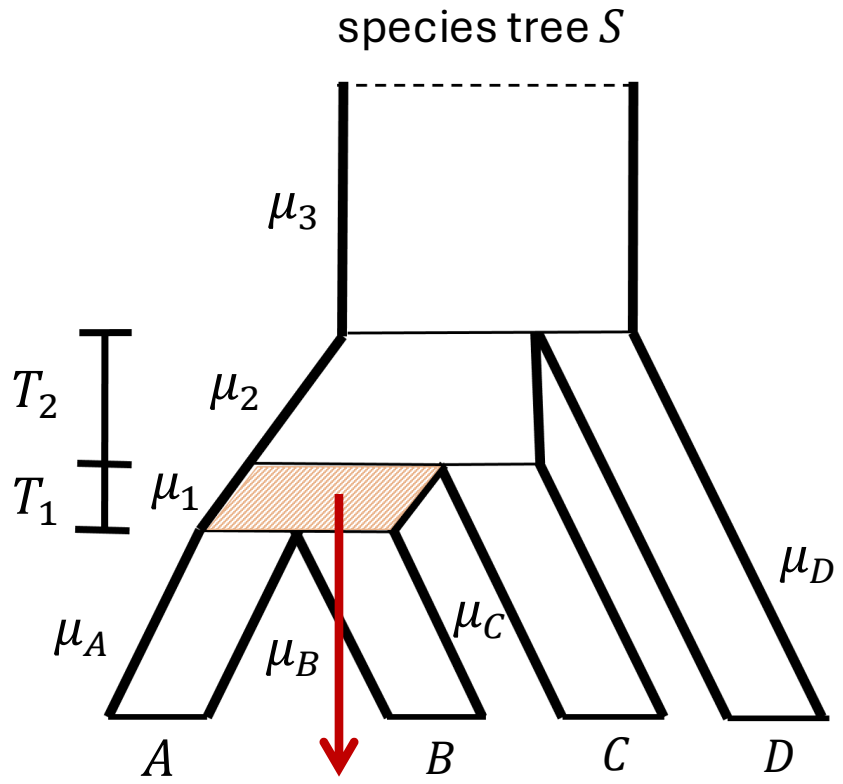
Our motivation for CASTLES

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?

Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

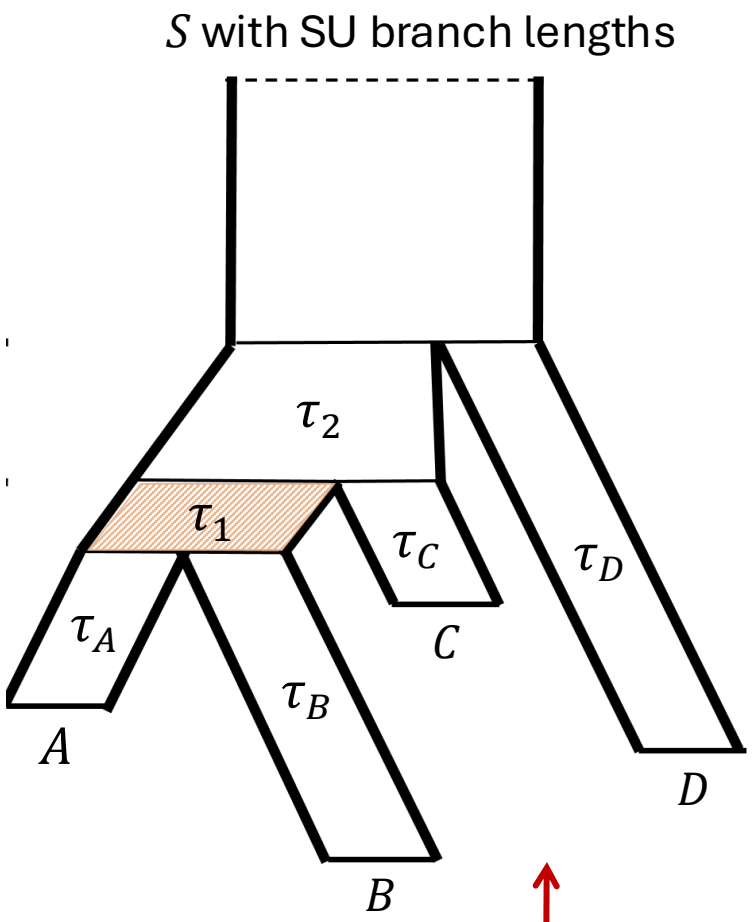
MSC+Substitution model

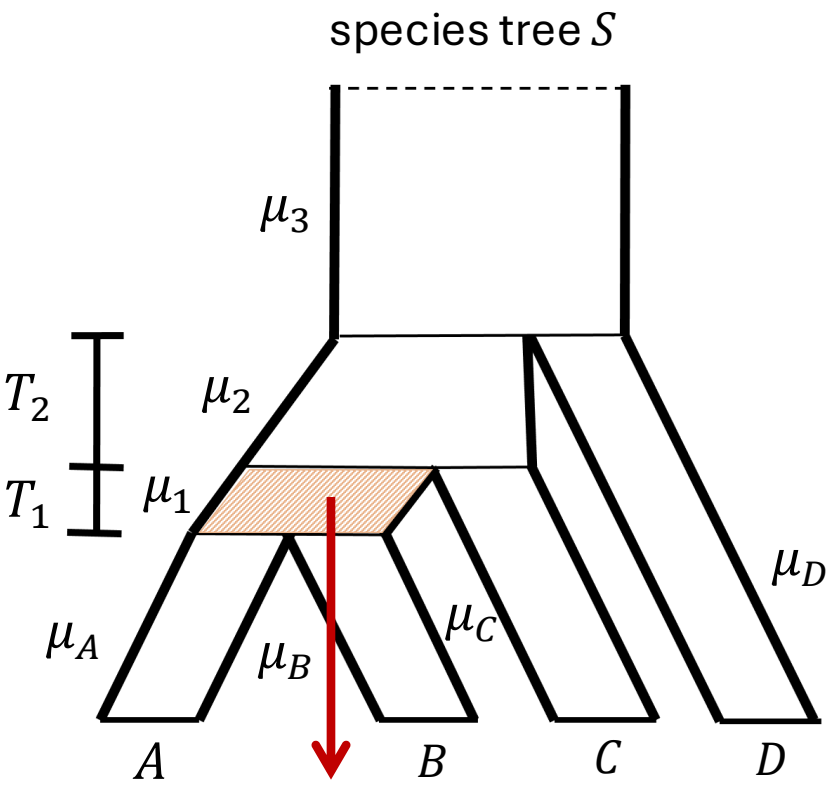


Parameters of branch 1:

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

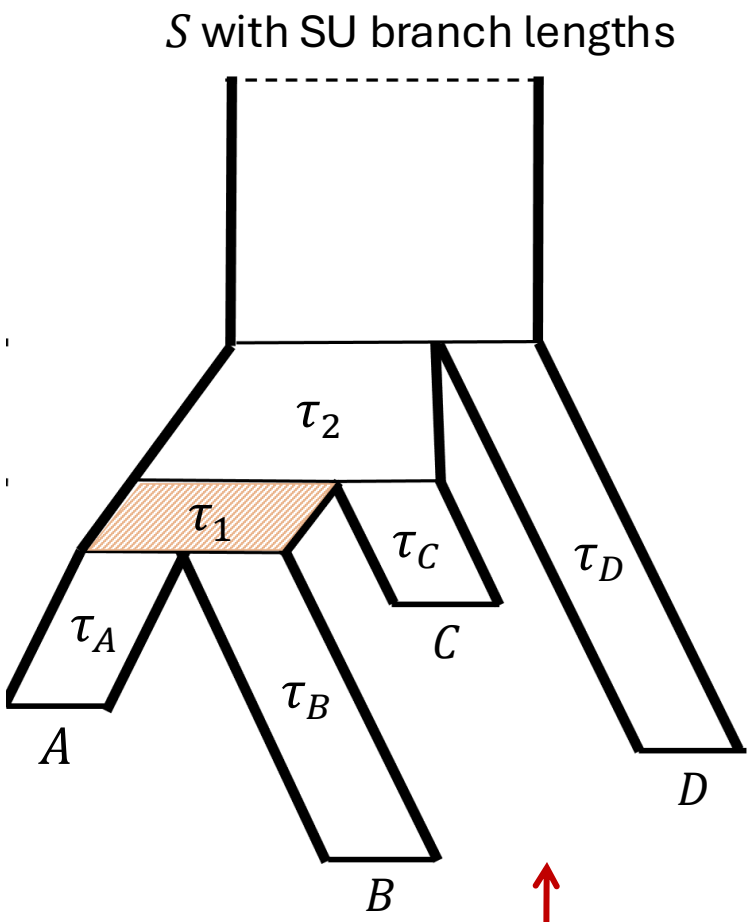


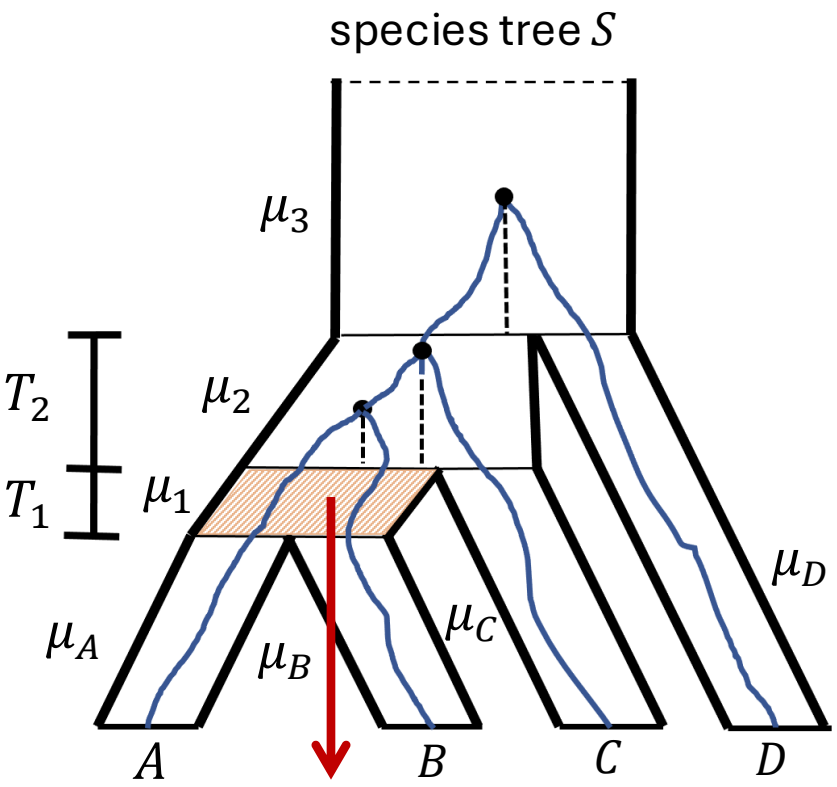


Parameters of branch 1:

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

MSC+Substitution model

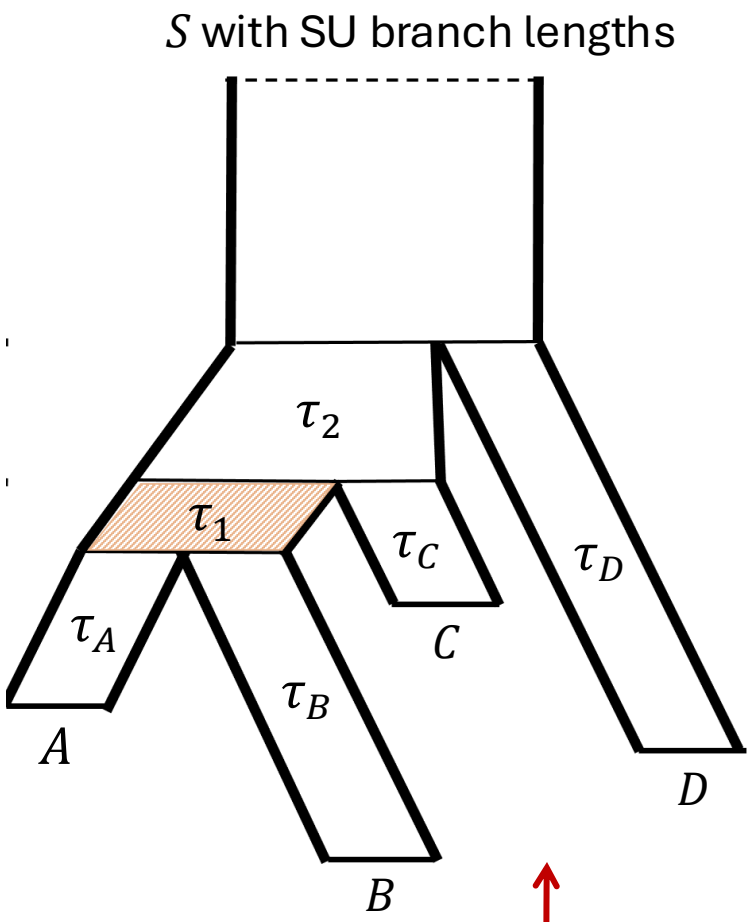


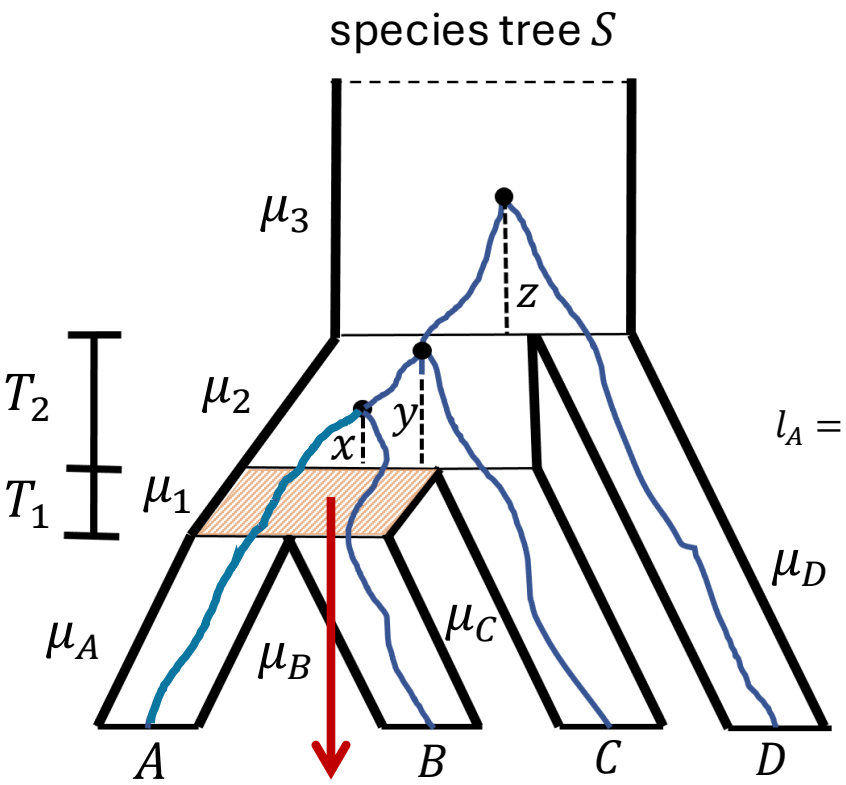


Parameters of branch 1:

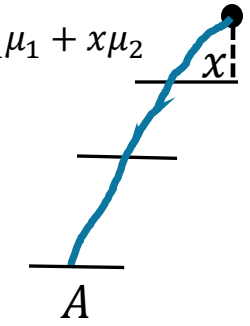
μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1 . \mu_1$	Branch length in substitution units (SU)

MSC+Substitution model





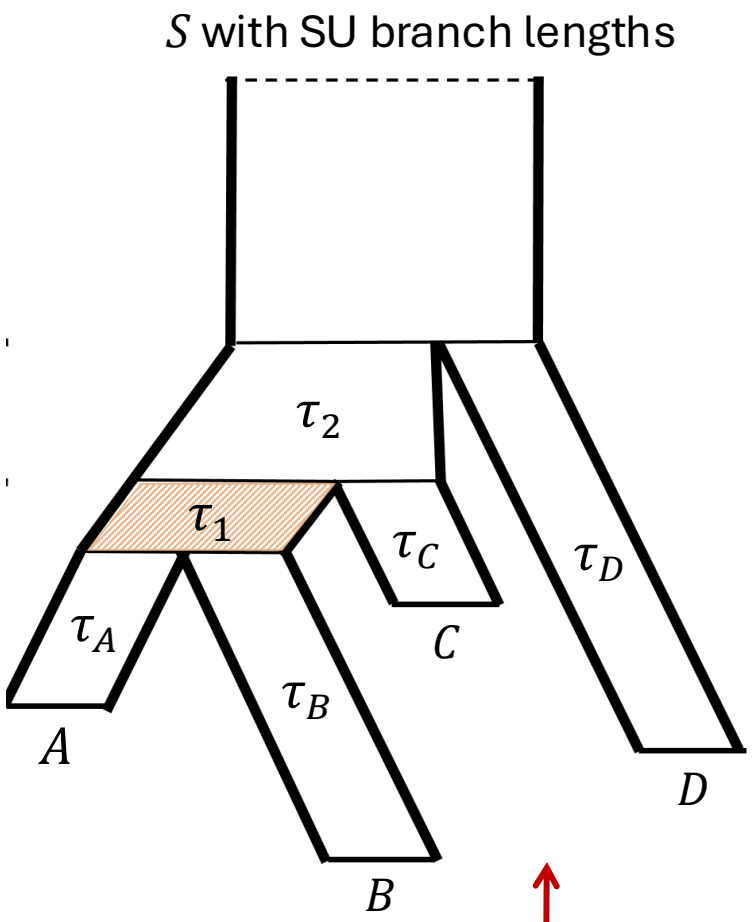
$l_A = T_A\mu_A + T_1\mu_1 + x\mu_2$

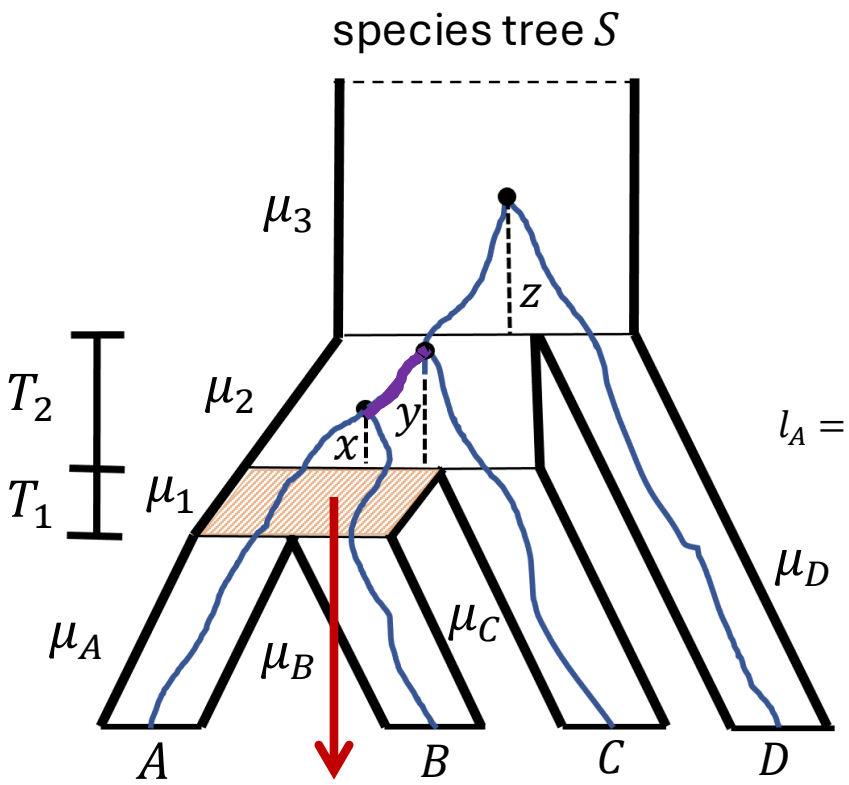


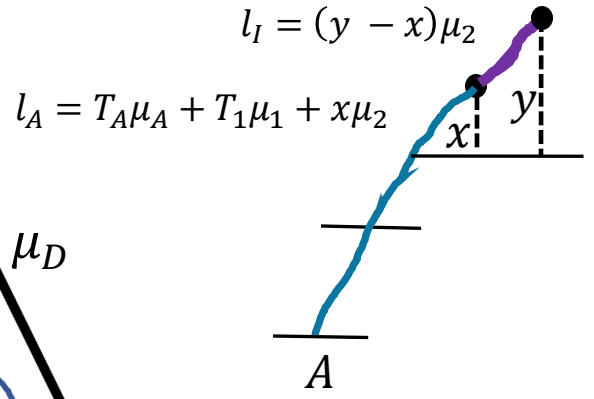
Parameters of branch 1:

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







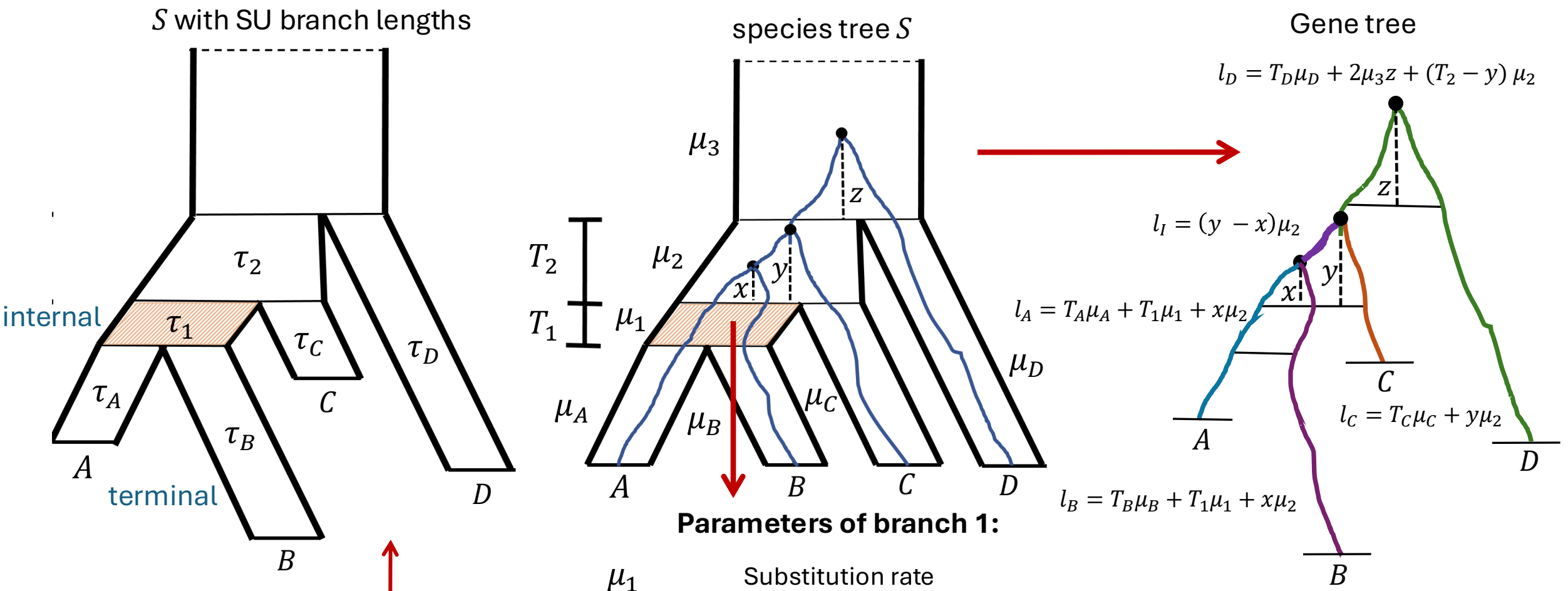
Parameters of branch 1:

μ_1 Substitution rate

T_1 Branch length in coalescent units (CU)

$\tau_1 = T_1 . \mu_1$ Branch length in substitution units (SU)

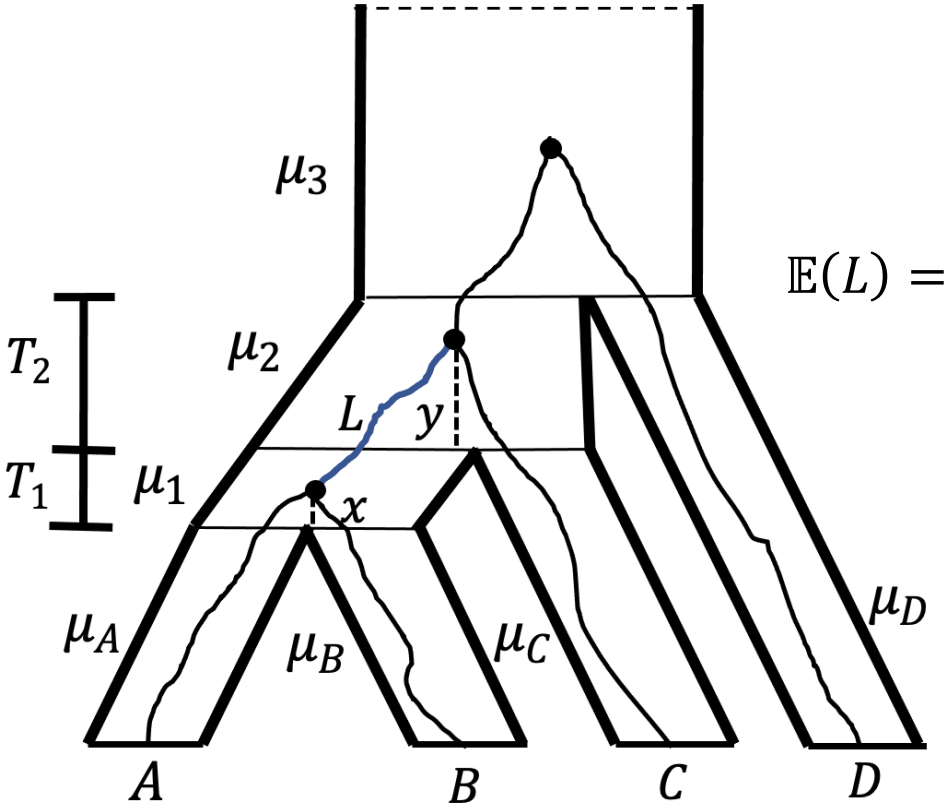
MSC+Substitution model



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



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

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

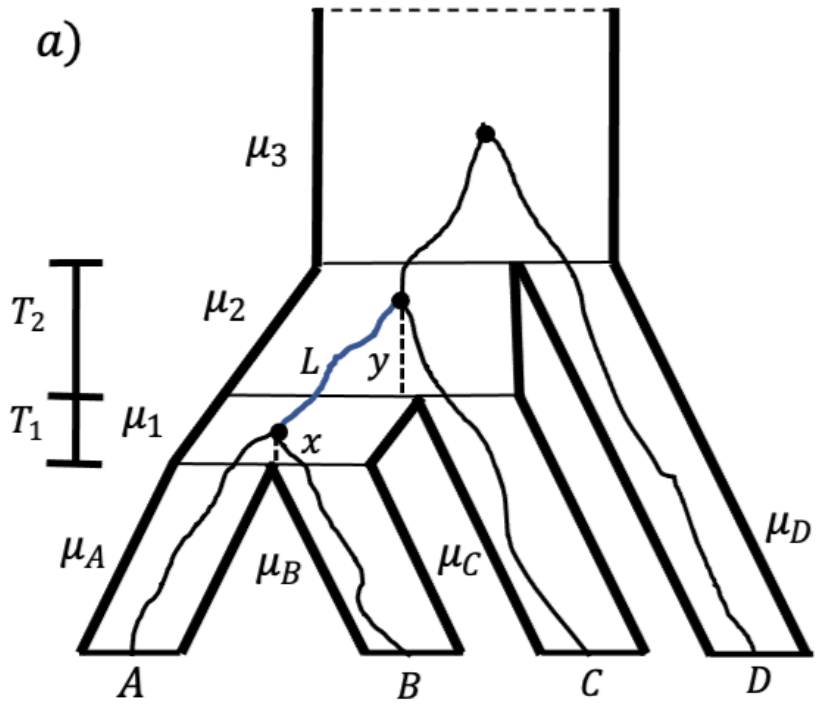
$k = 2$ lineages not coalescing in an interval with length x

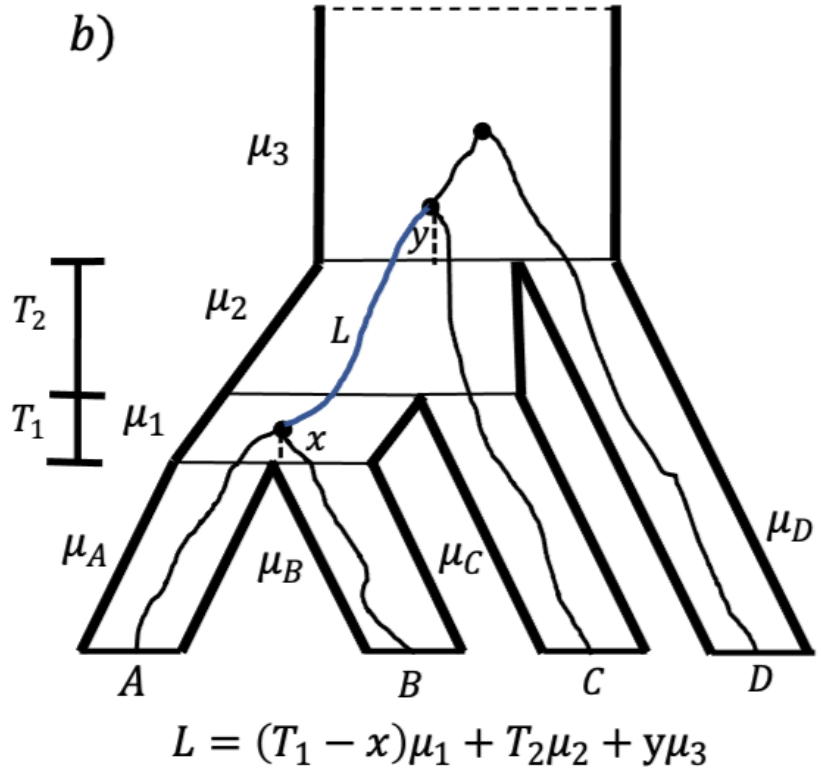
- Under MSC, waiting times before coalescent events are exponential random variables with rate $\lambda = \binom{k}{2}$ where k is the number of lineages entering an interval

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

Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

What about other patterns of coalescence?

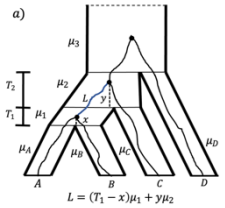




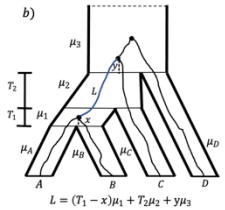
different patterns ⟶ different expected lengths

Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

Expected quartet branch lengths under MSC

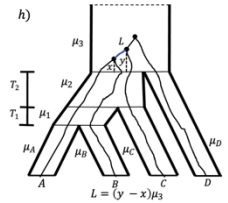


$$L_I = (\int_0^{T_1}\int_0^{T_2} e^{-x}\, e^{-y}\big((T_1 - x)\mu_1 + y\mu_2\big)dydx$$ scenario (a)



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

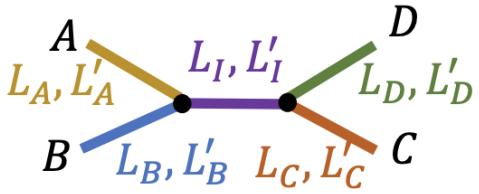
$$= \frac{(e^{-3T_2} + 3e^{-T_2} - 6e^{T_1 - T_2})(\mu_2 - \mu_3) + 6(1 - e^{T_1} + T_1 e^{T_1})\mu_1}{2(3e^{T_1} - 2)} + \mu_2$$

Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

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.



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

Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

CASTLES

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

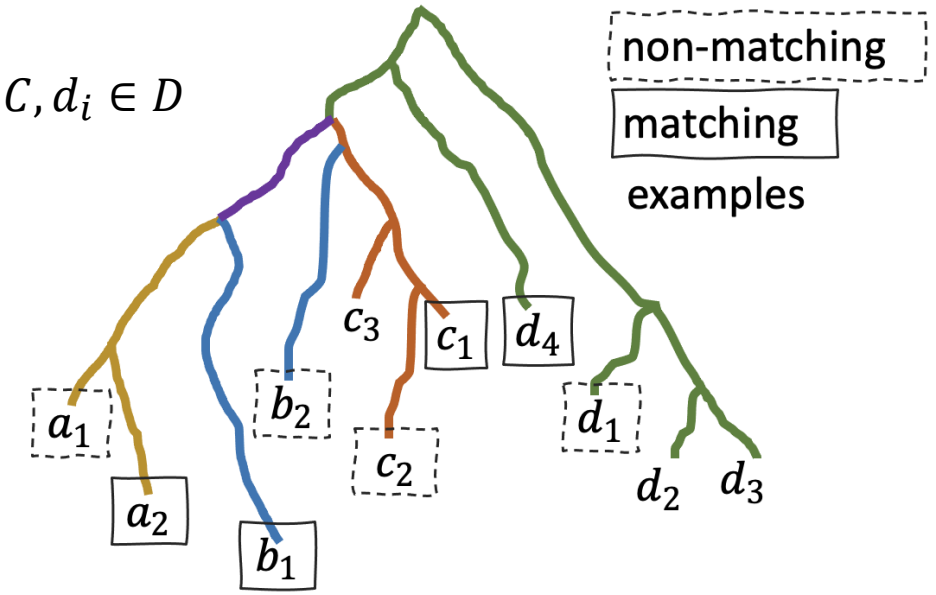
Input:
- Rooted species tree *topology* S
- A set of gene trees $\mathcal{G}$ with SU branch lengths

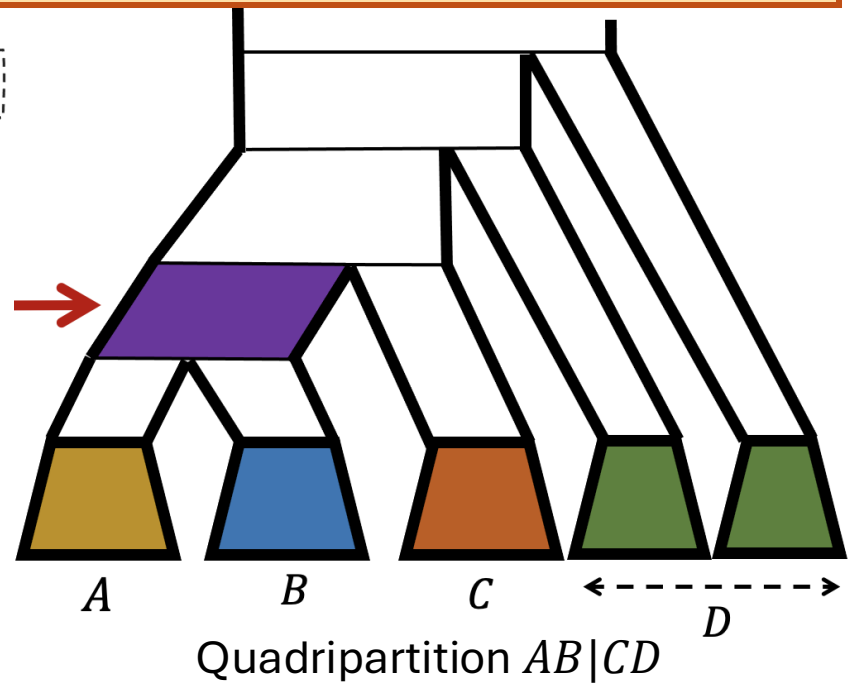
Output:
- Species tree S with SU branch lengths

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





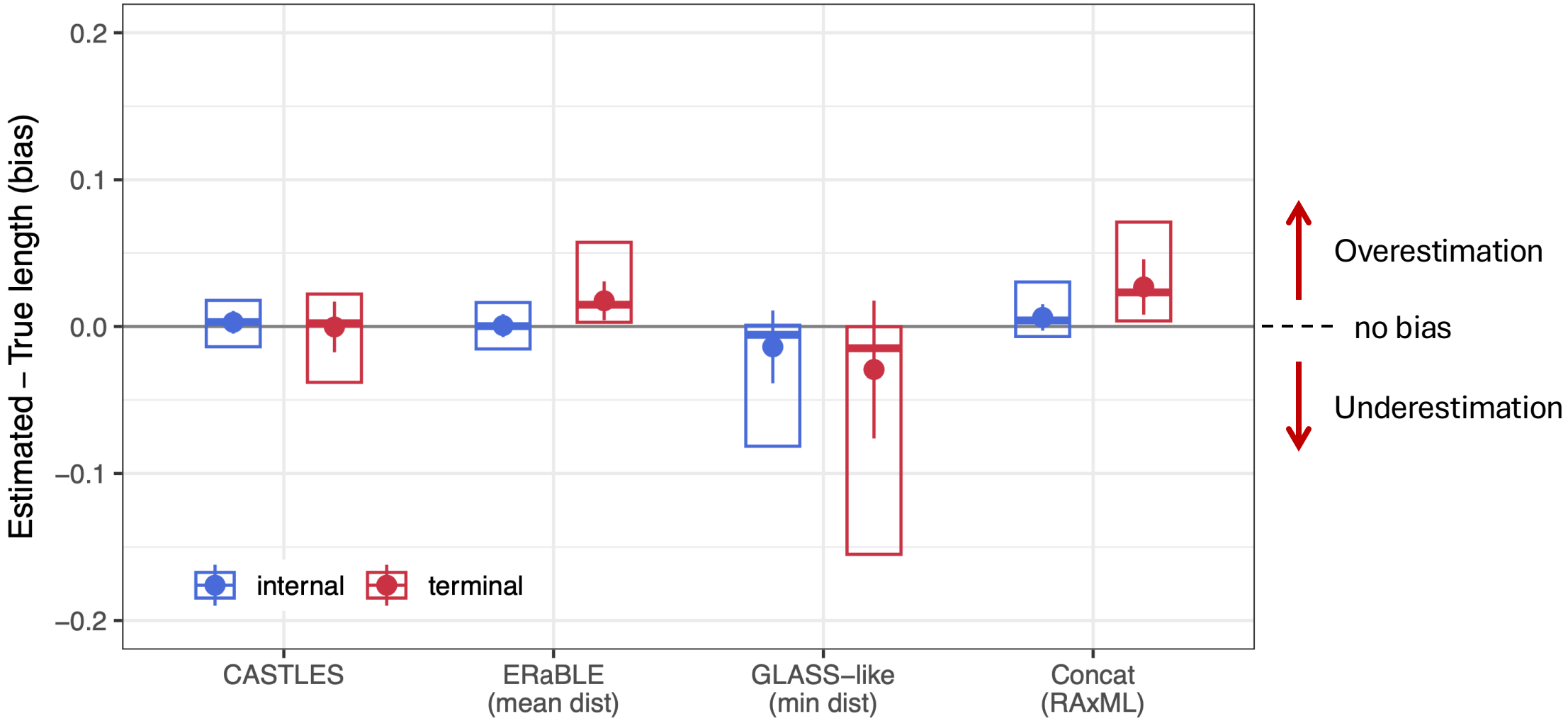
Quadripartition $AB|CD$

Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

CASTLES is less biased than other methods

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

Bias is higher for **terminal** branches.

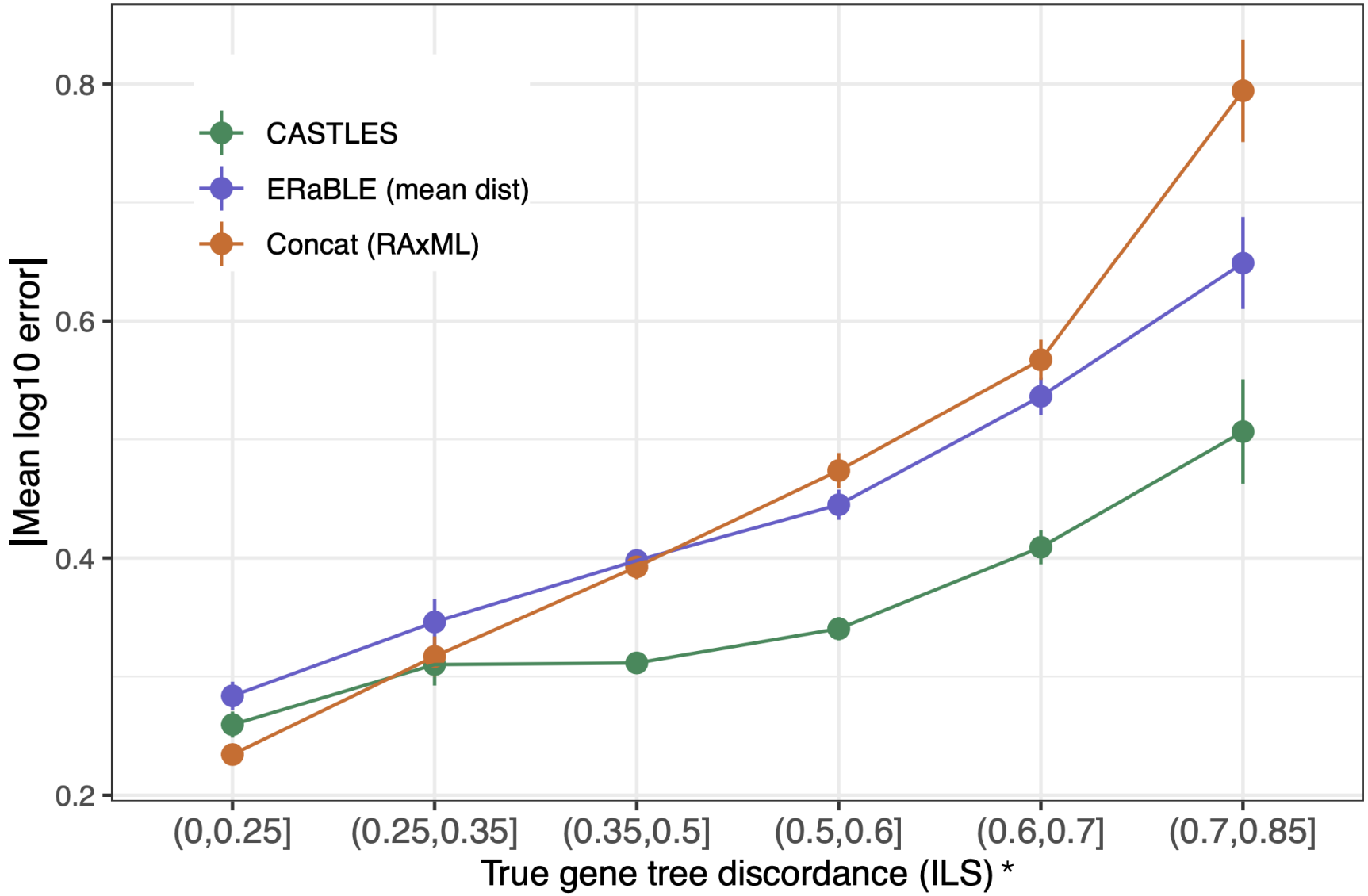


CASTLES's advantage increases with ILS

- 30-taxon ILS simulated dataset with 500 estimated gene trees [Mai et al (2017)]

|log10 (est. length / true length)|

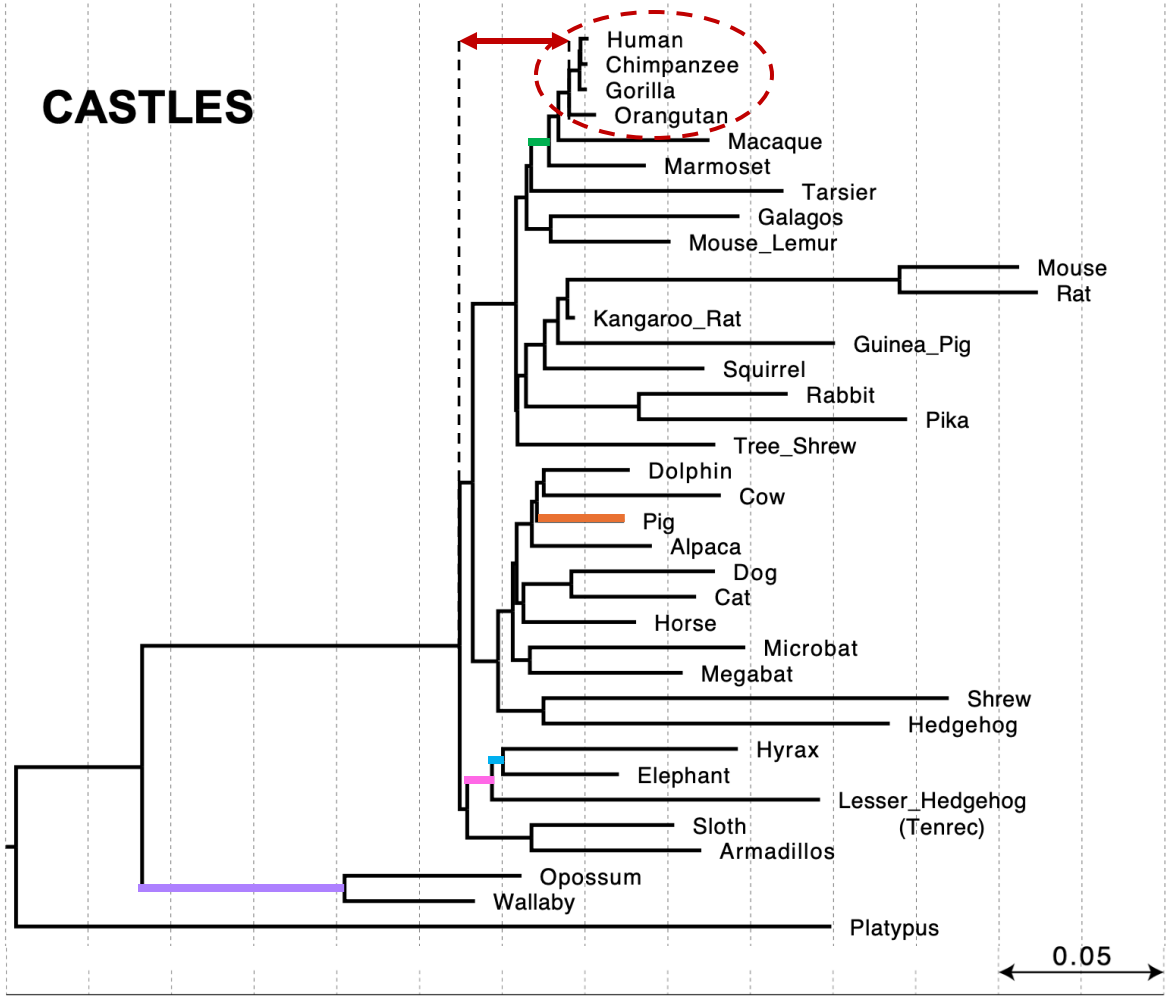
* Average RF distance between model species tree and true gene trees

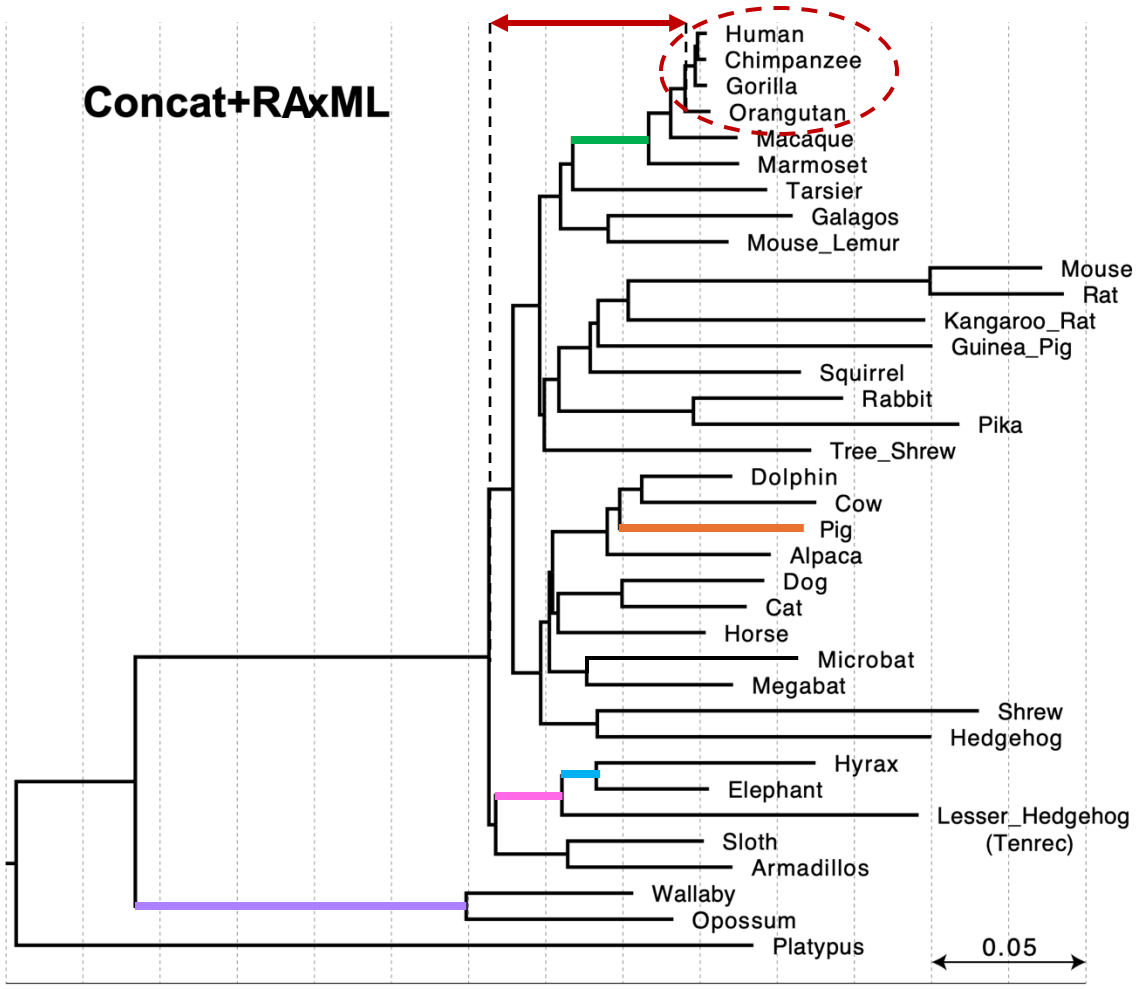


Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

CASTLES produces shorter branches than concatenation on mammalian dataset with ILS

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





Tabatabaee et al., "Phylogenomic branch length estimation using quartets". ISMB and Bioinformatics 2023

What about gene duplication and loss?

- CASTLES and other branch length estimation methods cannot directly work with multi-copy gene trees
- While specialized tools for species tree estimation under GDL exist [i.e., ASTRAL-Pro, FastMulRFS, DISCO+X,…], no pipeline/method was previously proposed for branch length estimation under GDL
- Restriction to single-copy genes leads to loss of signal [e.g., Wickett et al., 2014]

9,683 multi-copy → 424 single-copy

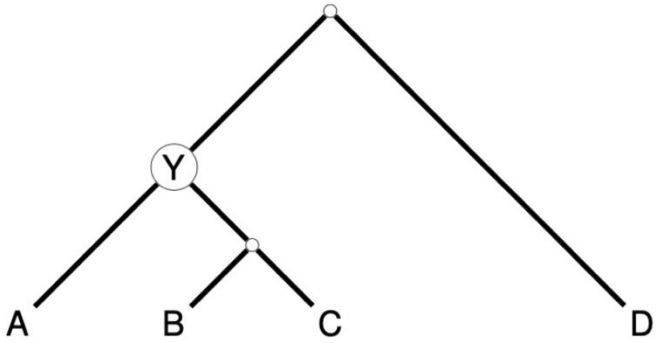
- How can we extend CASTLES to handle multi-copy genes?

Tabatabaee et al., "Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss". 2025, Submitted

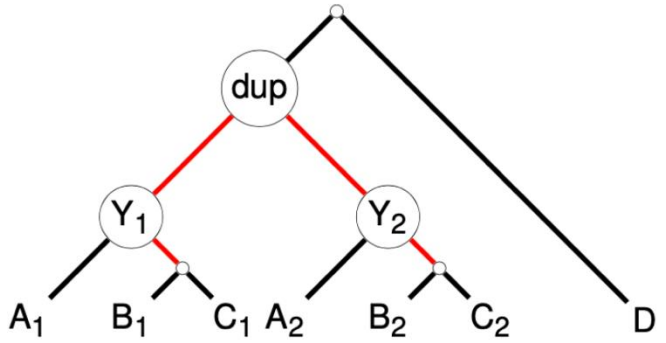
CASTLES-Pro

"Pro" stands for PaRalogs and Orthologs…

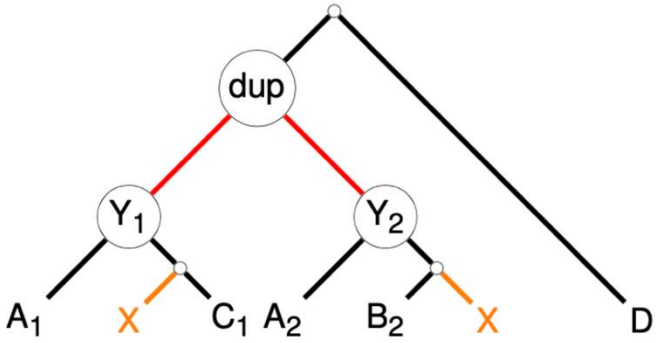
- Extends CASTLES to work with multi-copy genes
- Only considers quartets devoid of paralogous genes
- Improves upon CASTLES by modifying some of its equations
- Relaxes some of the approximations used in CASTLES and is more theoretically appealing
- Implements a weighting scheme to account for uneven rates of duplication



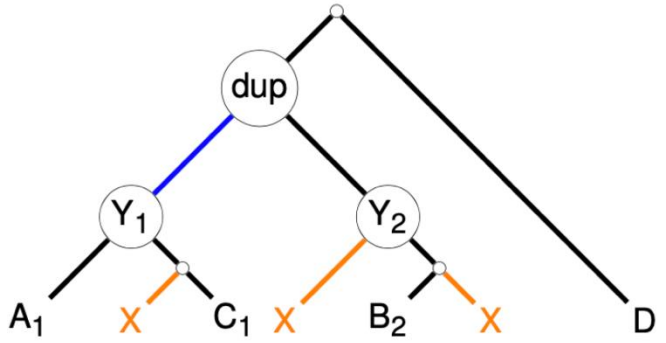
(a) Species tree T^*



(b) Gene tree M_1 with one duplication.



(c) Gene tree M_2 with one duplication and two losses.

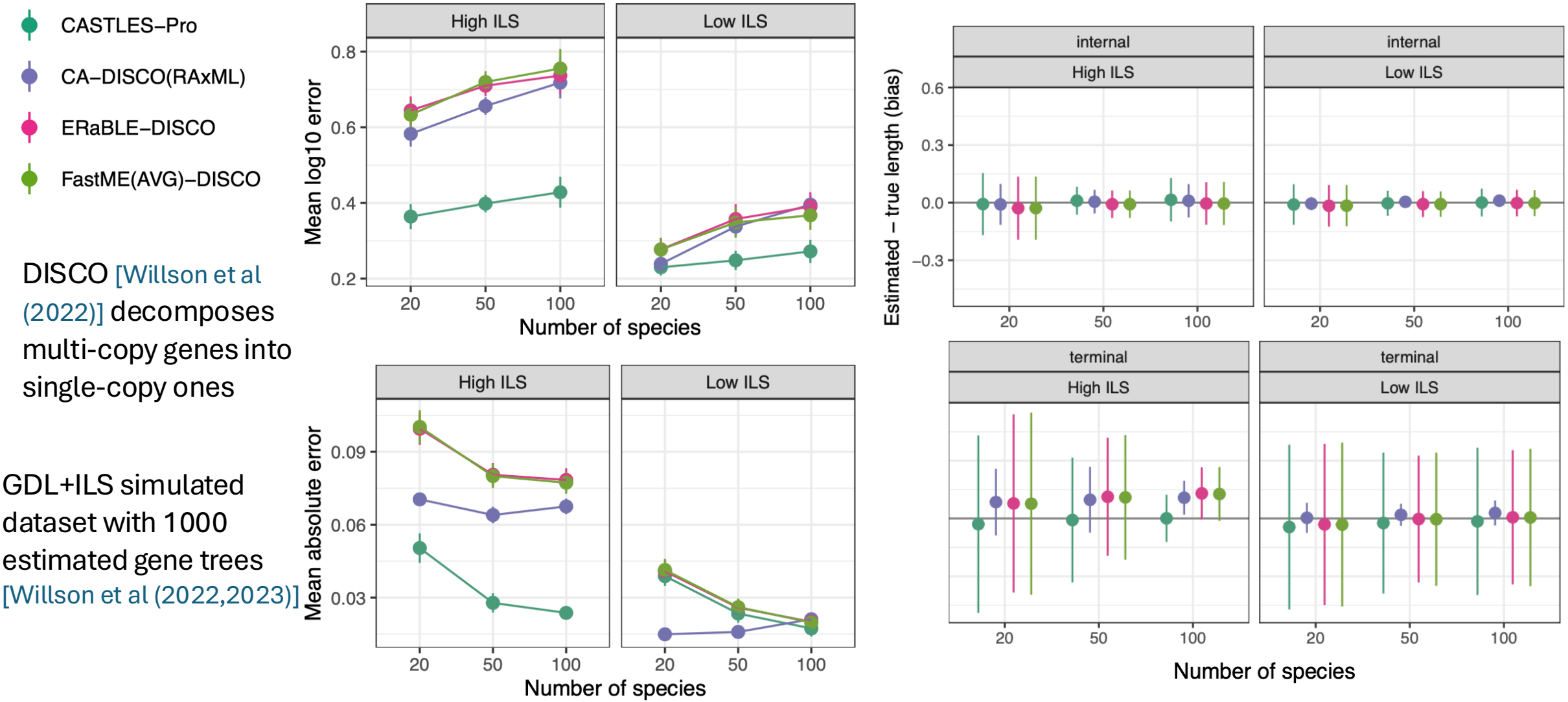


(d) Gene tree with one duplication and three losses.

- A1 and C1 are orthologs
- A1 and B2 are paralogs

Tabatabaee et al., "Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss". 2025, Submitted

CASTLES-Pro enables branch length estimation despite gene duplication and loss



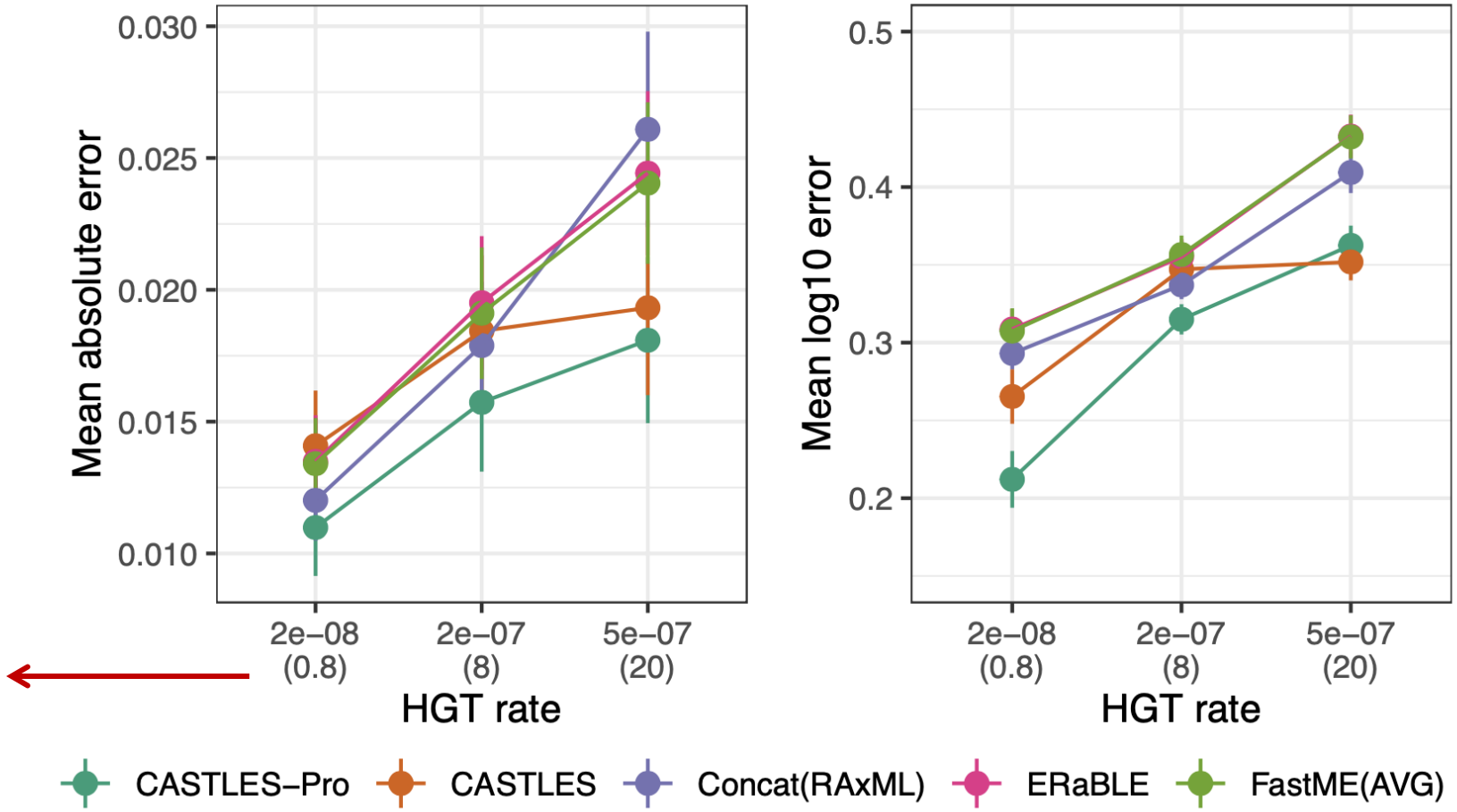
- DISCO [Willson et al (2022)] decomposes multi-copy genes into single-copy ones

GDL+ILS simulated dataset with 1000 estimated gene trees [Willson et al (2022,2023)]

Tabatabaee et al., "Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss". 2025, Submitted

CASTLES-Pro is more robust to HGT in simulations

- 51-taxon HGT+ILS simulated dataset with 1000 estimated gene trees [Davidson et al (2015)]

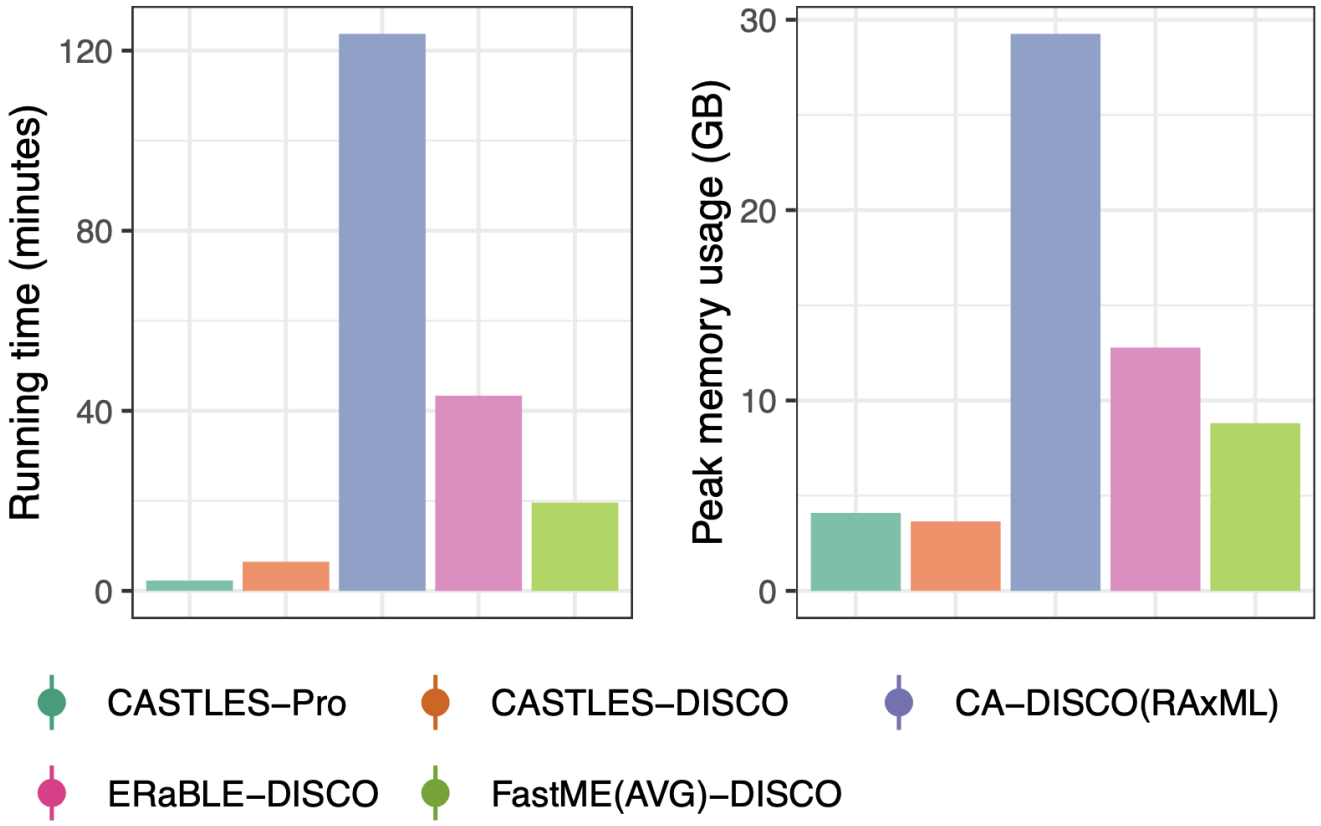


Tabatabaee et al., "Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss". 2025, Submitted

CASTLES-Pro is more scalable than other methods

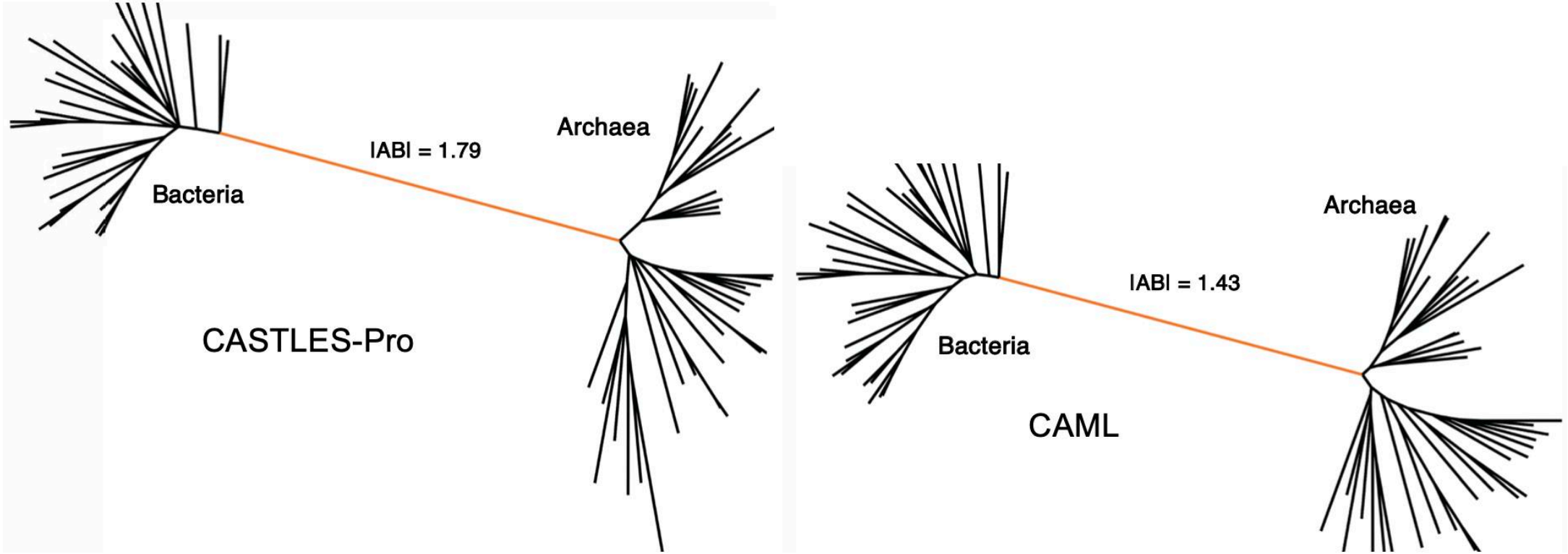
- 101-taxon GDL+ILS simulated dataset with 10,000 estimated gene trees [Willlson et al (2022)]

- CASTLES-Pro takes ~5 minutes on a dataset with 1000 species and 1000 genes and ~1 hour on 10,000 species and 400 genes



Tabatabaee et al., "Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss". 2025, Submitted

CASTLES-Pro produces longer AB branch than concatenation on bacterial datasets

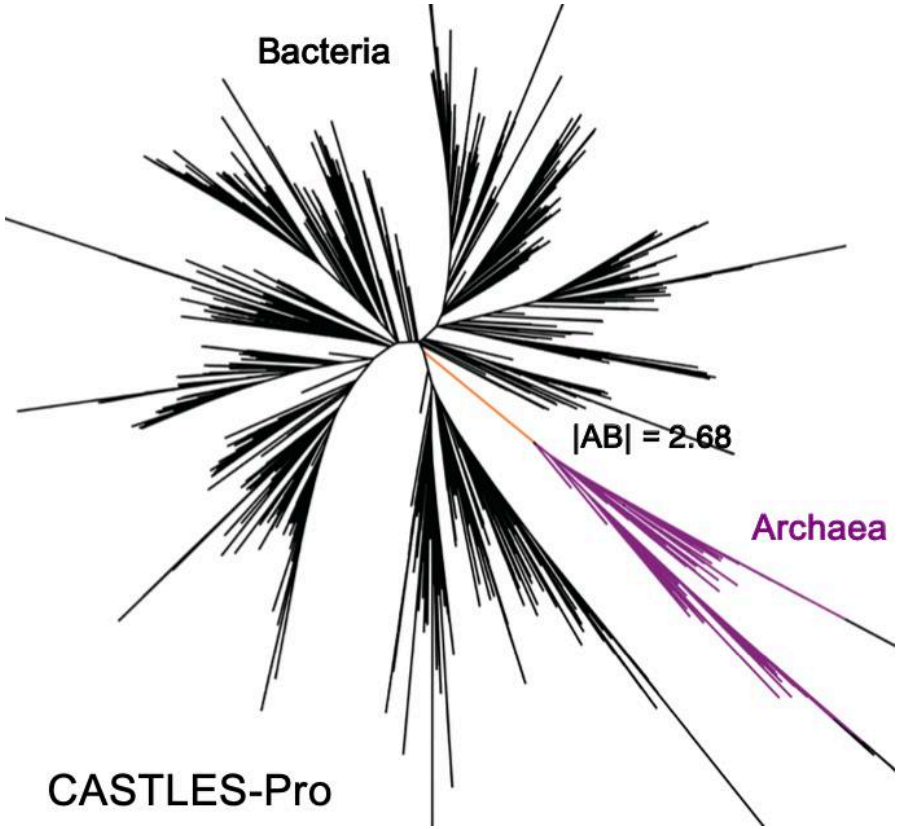


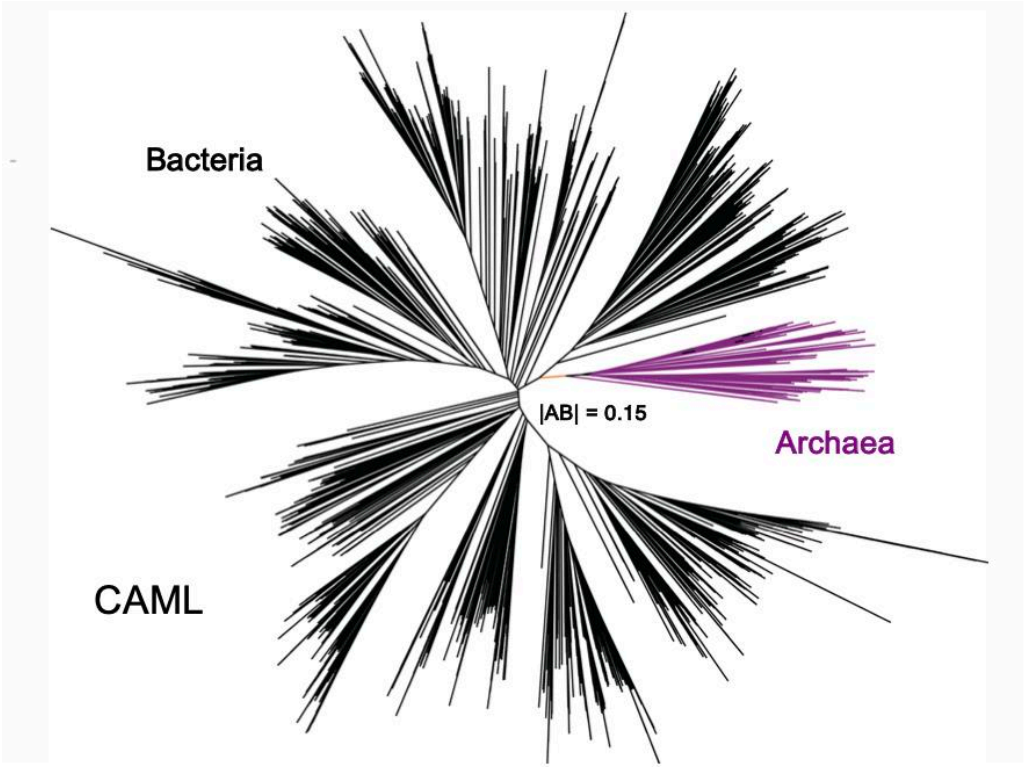
- Bacterial dataset with 72 species and 49 core genes [Williams et al. (2020)].

Tabatabaee et al., "Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss". 2025, Submitted

Arasti et al., "Optimal tree metric matching enables phylogenomic branch length reconciliation". RECOMB 2024

CASTLES-Pro produces longer AB branch than concatenation on bacterial datasets





- Bacterial dataset with 10,575 species and 381 marker genes [Zhu et al. (2019)].

Tabatabaee et al., "Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss". 2025, Submitted
Arasti et al., "Optimal tree metric matching enables phylogenomic branch length reconciliation". RECOMB 2024

Outline

- Background and Motivation
  - Phylogenomics pipeline
  - Gene tree discordance
  - Species tree estimation
- Overview of Contributions
  - Discordance-aware post-species tree analysis
- Rooting species trees
- Phylogenomic branch length estimation
- **Dating species trees and gene trees**
- Conclusions

Dating species trees

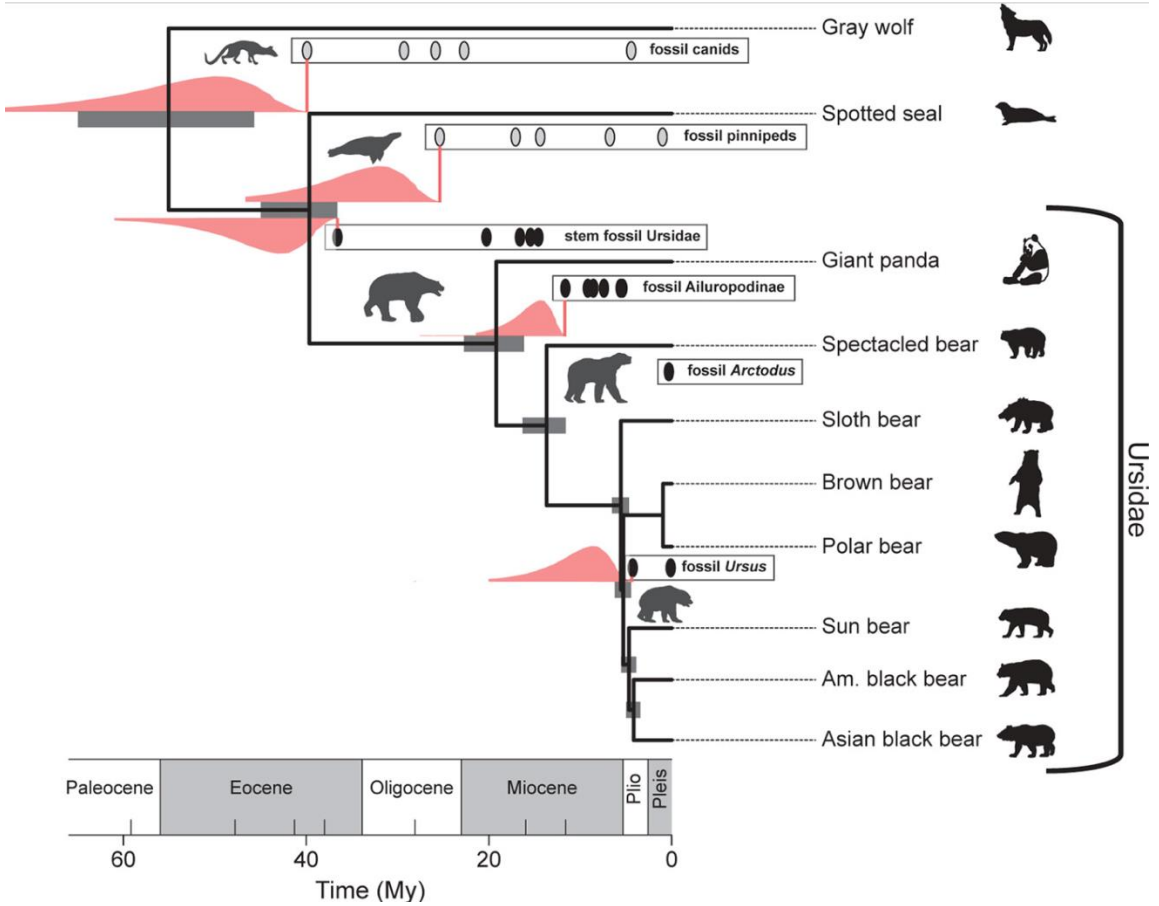


Image Credit: Marshall, 2019, “Using the Fossil Record to Evaluate Timetree Timescales”, *Frontiers in Genetics*

Estimating time trees and molecular dating

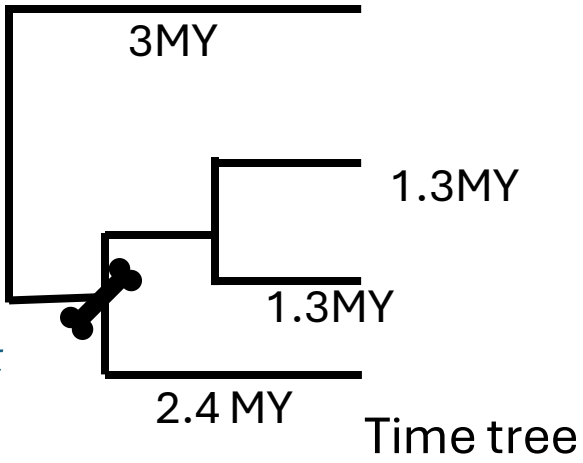
- Molecular dating is the estimation of time-trees from sequence data

- Time and mutation rate are inseparable from sequence data and hence we need calibration points

```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

Sequence data

$m = \mu\Delta t$



- Rate variation across branches is modeled with a “clock” model
  - **Strict**: rate variation is constant across branches
  - **Relaxed**: allows rates variation across branches

Calibrations points as time constraints

- Dating methods rely on external information, such as fossil calibrations or sampling times, that specify the timing of some speciation events
- the dates of other nodes are extrapolated from these pre-specified times

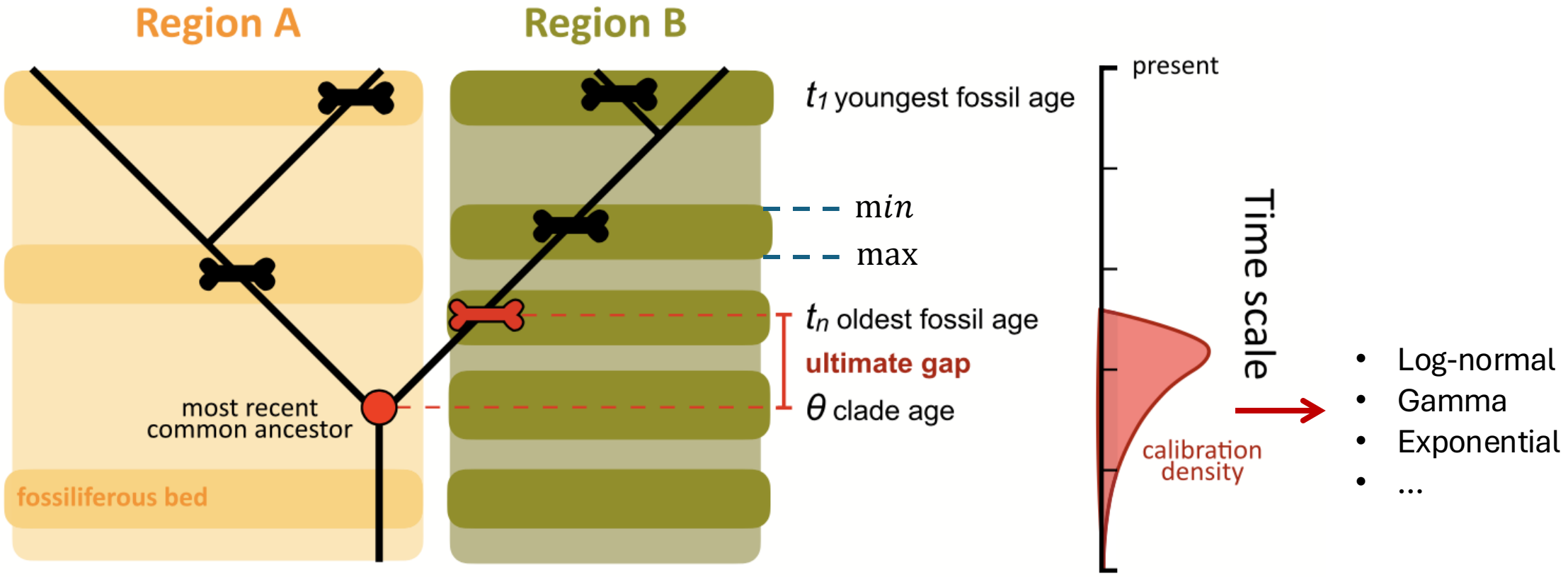
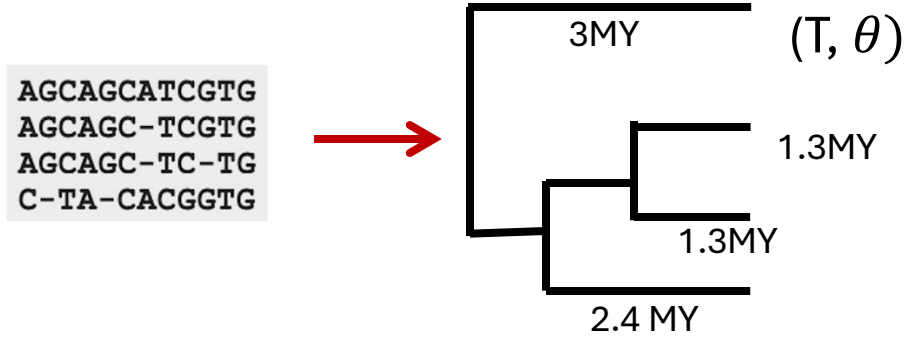


Image Credit: Claramunt, 2022, "CladeDate: Calibration information generator for divergence time estimation", *Methods in Ecology and Evolution*

Existing approaches for dating species trees

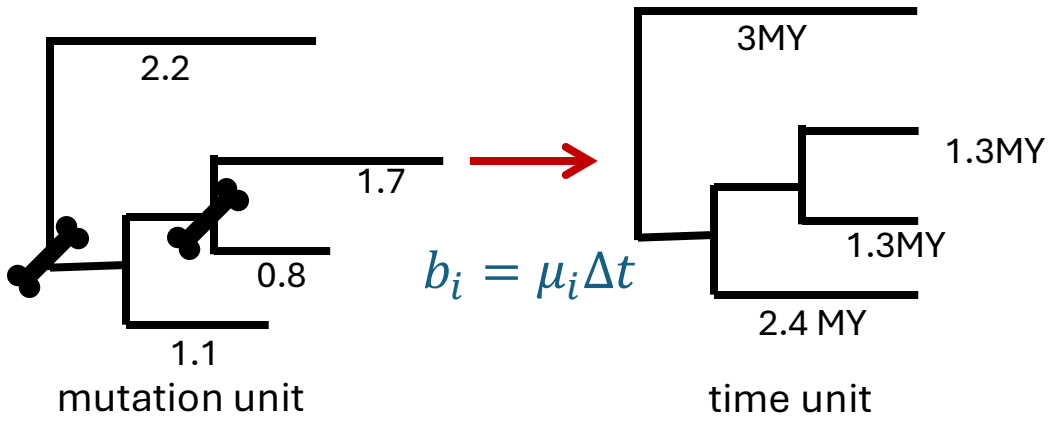
Bayesian

- Estimate the topology, branch lengths and rates jointly from sequence data
- Allow for complex rate and calibration models
- Limited scalability
- BEAST [Drummond and Rambaut, 2007], MCMCTree [Yang and Rannala, 2006], …



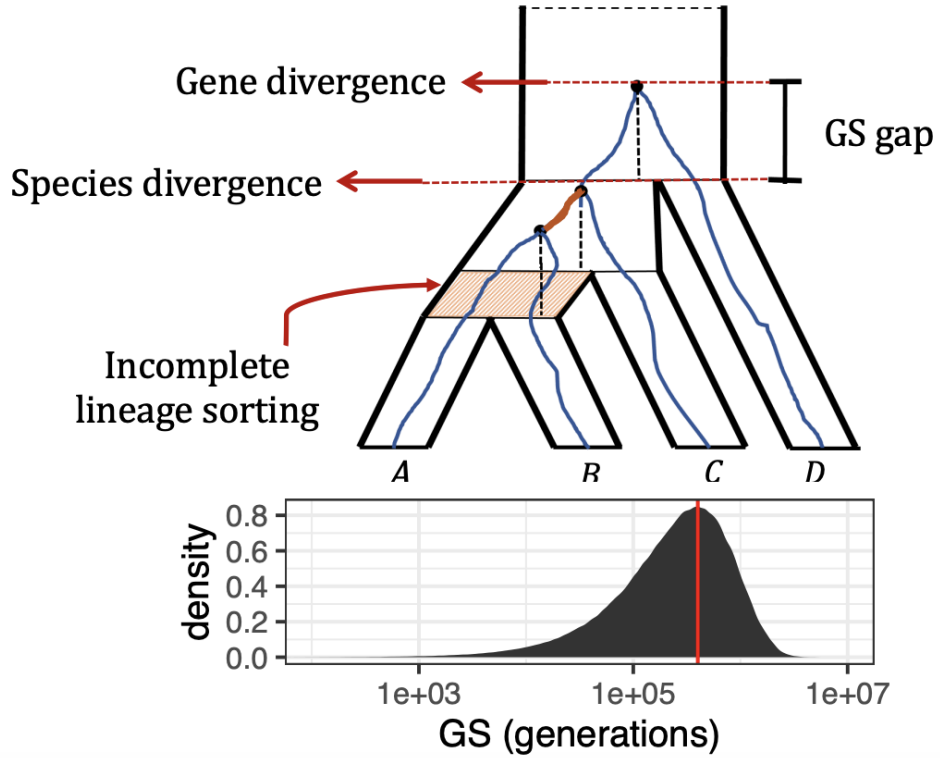
Maximum-likelihood based

- Convert SU branch lengths to time-units
- Find parameters that maximize a constrained likelihood function
- More scalable
- LSD [To et al, 2016], LF [Langley and Fitch, 1974], MD-Cat [Mai et al, 2024], …



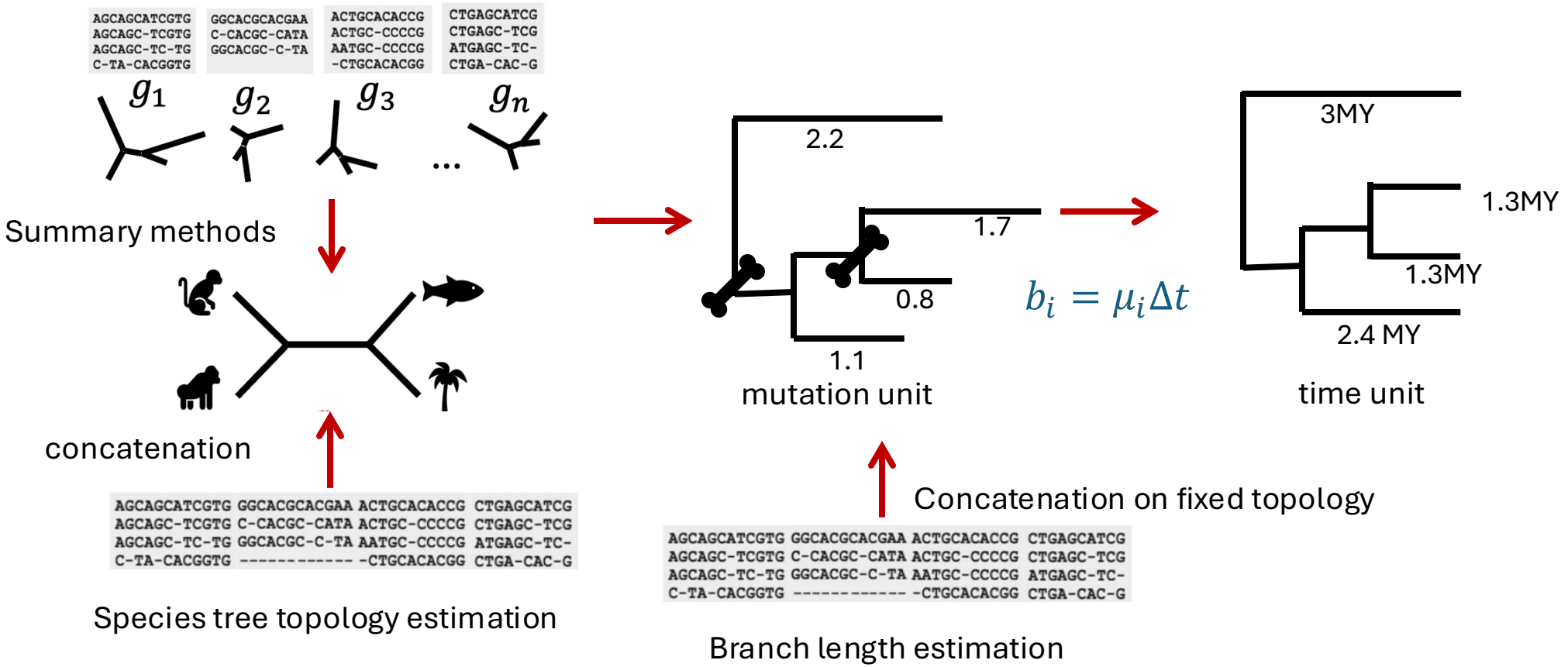
How does gene tree discordance affect dating?

- Previous studies show that gene tree discordance not only affects the patterns of evolution (i.e., tree topology), but also the timing of speciation events
- Species divergence times are systematically overestimated in the presence of ILS [Edwards and Beerli (2000)]
- Davin et al. (2018) show that HGT events can be used to date species tree (especially useful for bacterial organisms)
- How can ILS be used to more accurately date species trees?



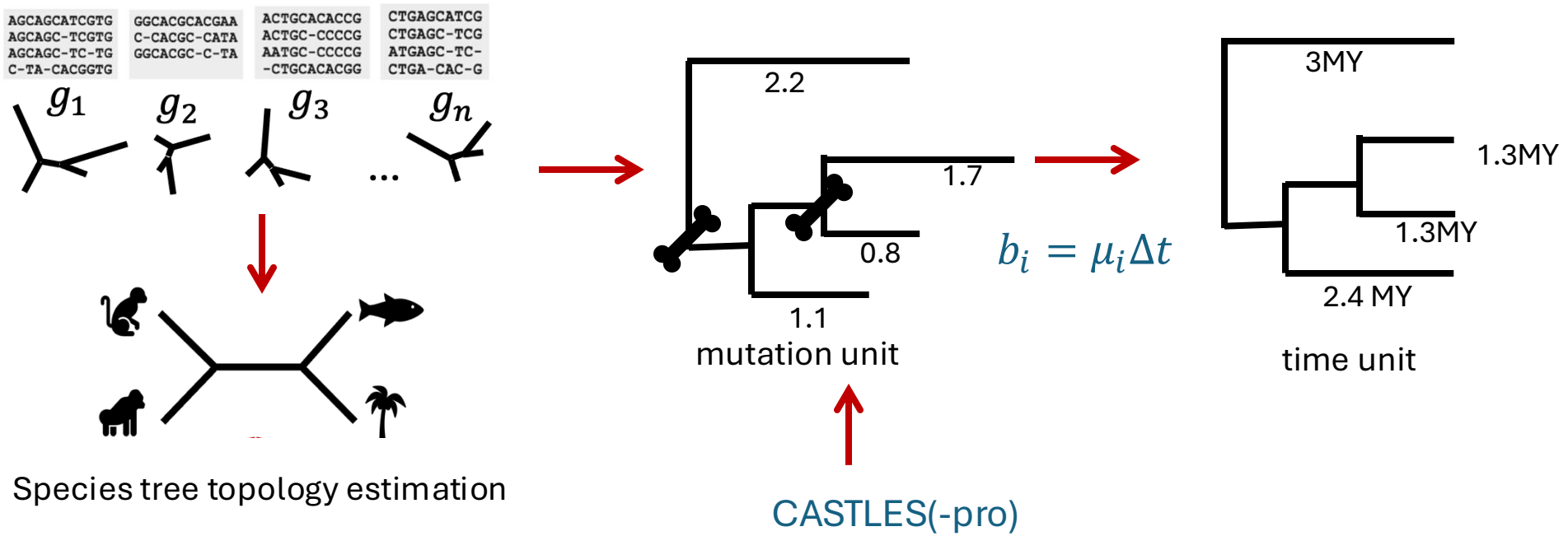
We need dating methods that account for gene tree discordance.

Typical likelihood-based dating pipeline



- Can accounting for gene tree discordance in the branch length estimation step improve the dating pipeline?

Discordance-aware likelihood-based dating pipeline



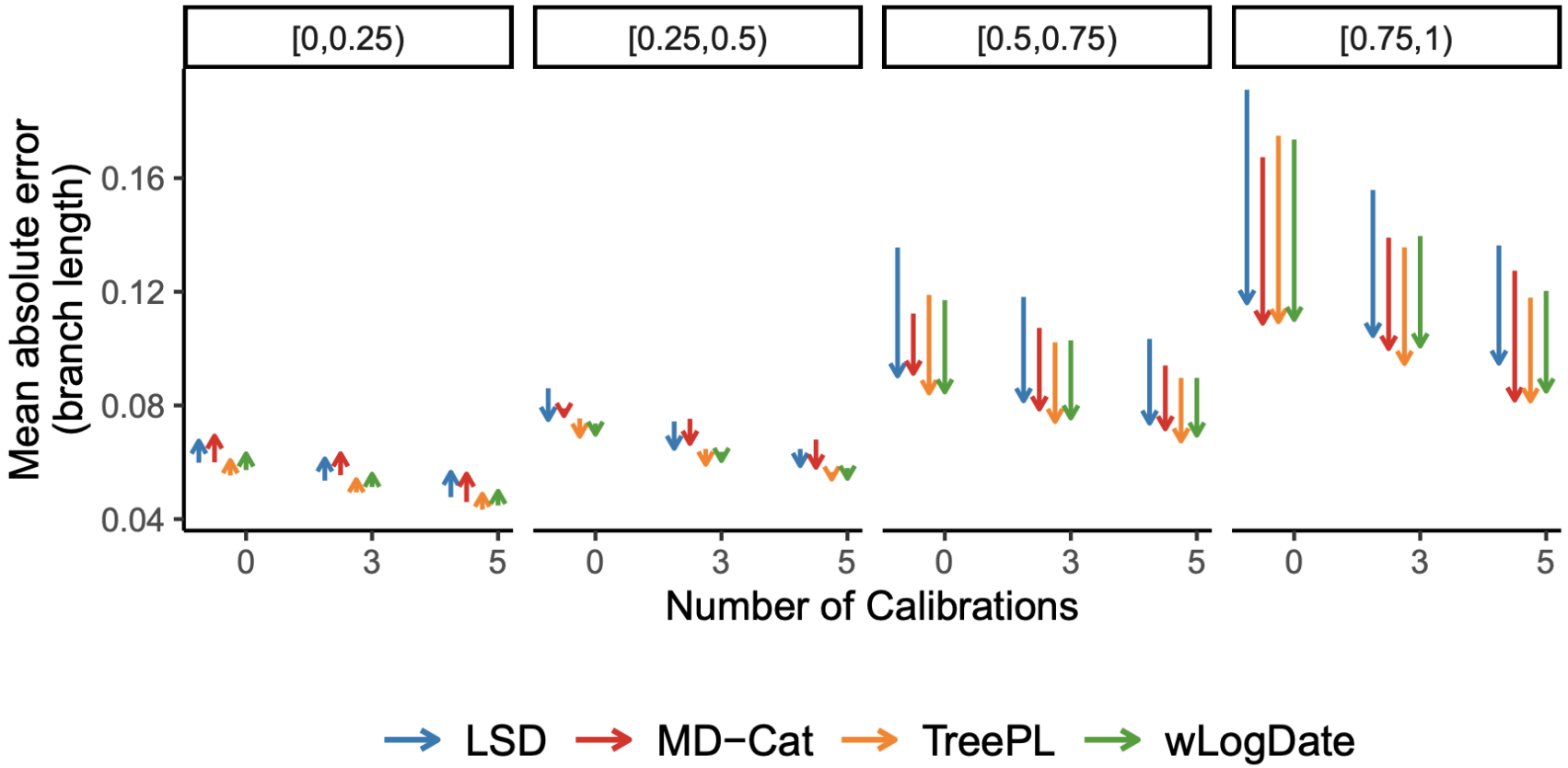
- Can accounting for gene tree discordance in the branch length estimation step improve the dating pipeline?

Coalescent-based branch length estimation improves dating of species trees

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

ML-based dating methods:

- Least-square dating (LSD)
- WLogDate
- treePL
- MD-Cat



- CASTLES improves the accuracy of dating when ILS is at least moderate

Tabatabaee et al., "Coalescent-based branch length estimation improves dating of species trees". 2025, submitted

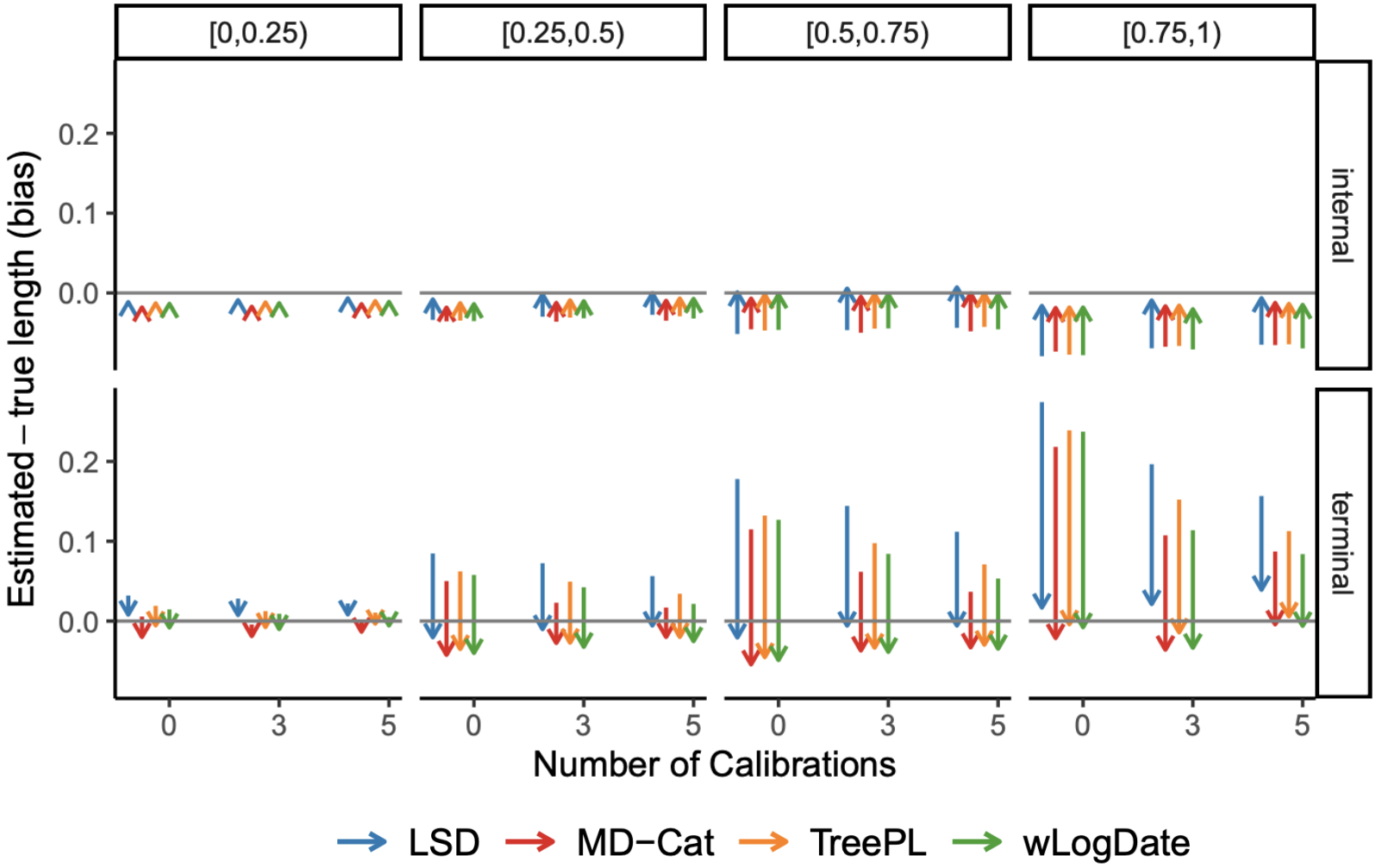
Coalescent-based branch length estimation improves dating of species trees

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

ML-based dating methods:

- Least-square dating (LSD)
- WLogDate
- treePL
- MD-Cat

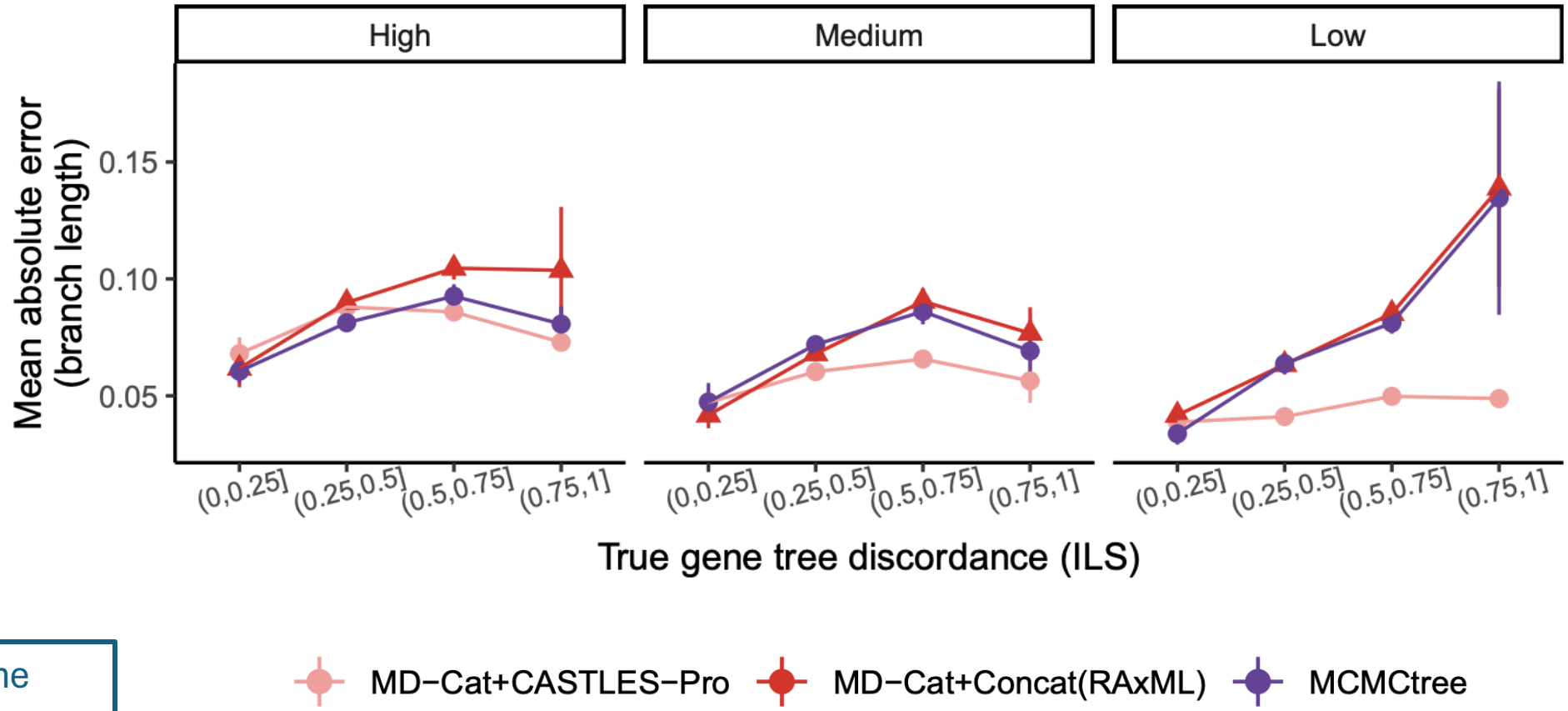
- CASTLES branch lengths reduce the bias of dating



Tabatabaee et al., "Coalescent-based branch length estimation improves dating of species trees". 2025, submitted

Coalescent-based branch length estimation improves dating of species trees

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

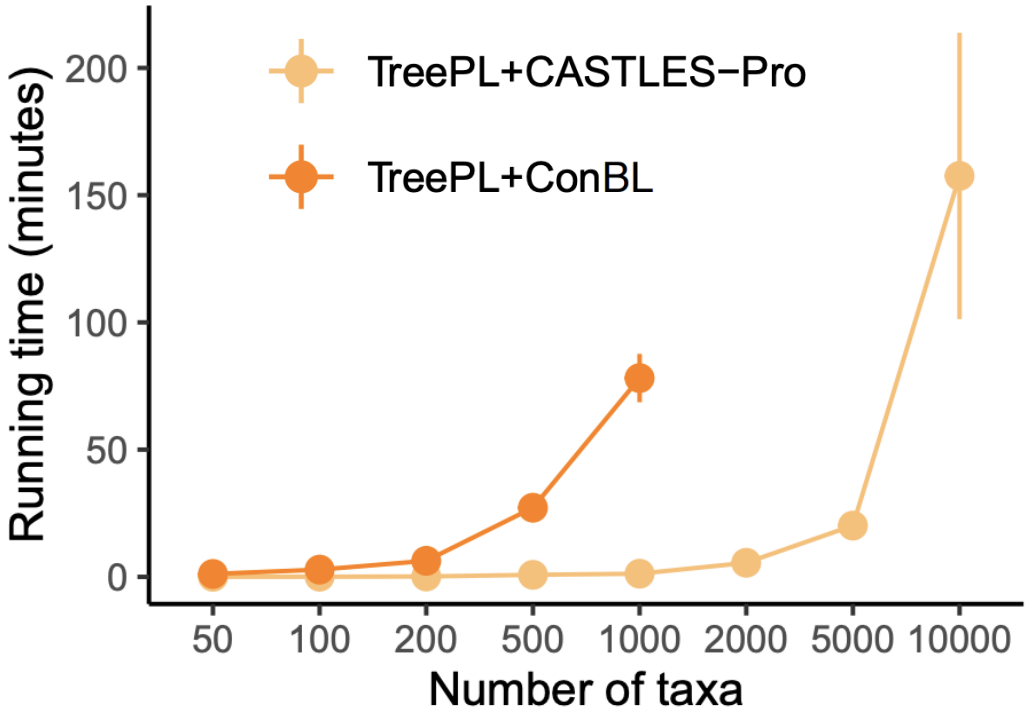


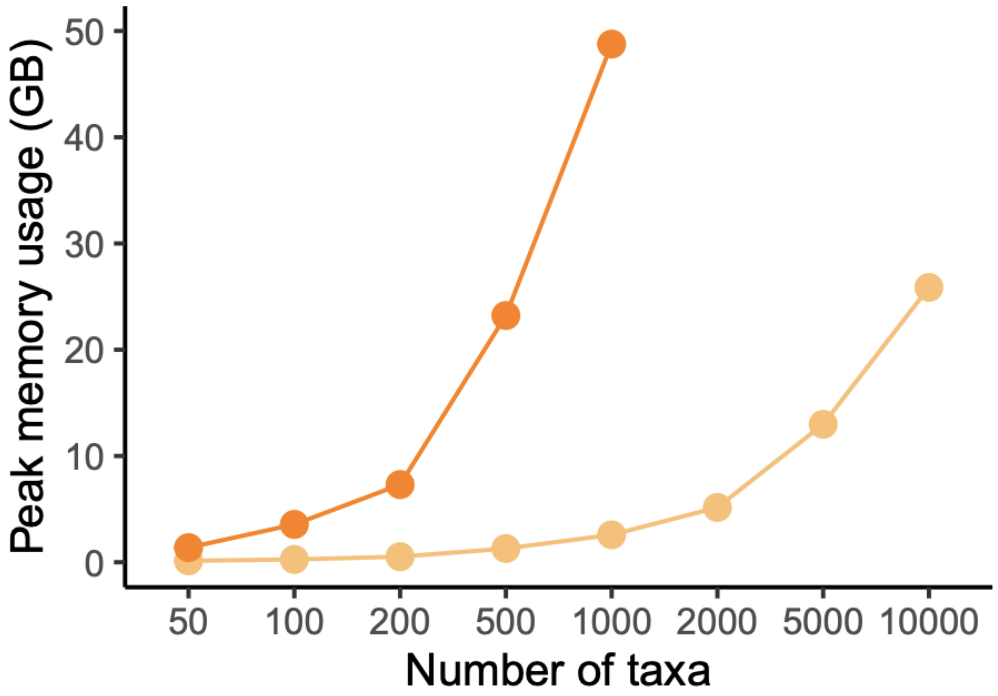
- CASTLES improves the accuracy of dating when ILS is at least moderate

Tabatabaee et al., "Coalescent-based branch length estimation improves dating of species trees". 2025, submitted

Discordance-aware dating scales to tens of thousands of species and genes

- Simulated dataset with 1000 genes, moderate ILS

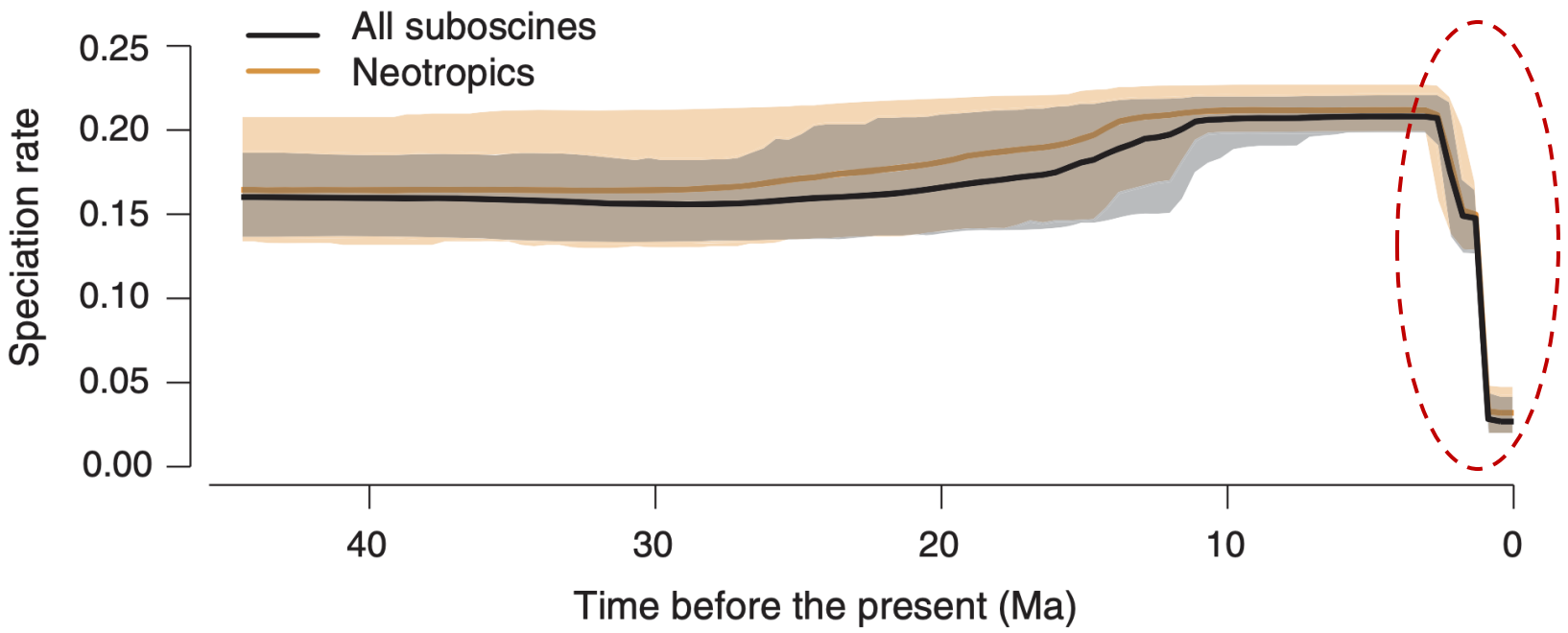




- Concatenation-based dating does not scale beyond 1000 taxa due to memory limit

Tabatabaee et al., “Coalescent-based branch length estimation improves dating of species trees”. 2025, submitted

Overestimation bias of concatenation can impact diversification analysis



- Harvey et al., (2020) studied the diversification of suboscines, the largest tropical bird radiation.
- Diversification rates are stable over most of the history of the group aside from a drop within the past 2 Ma
- 1683 species and 2389 genes

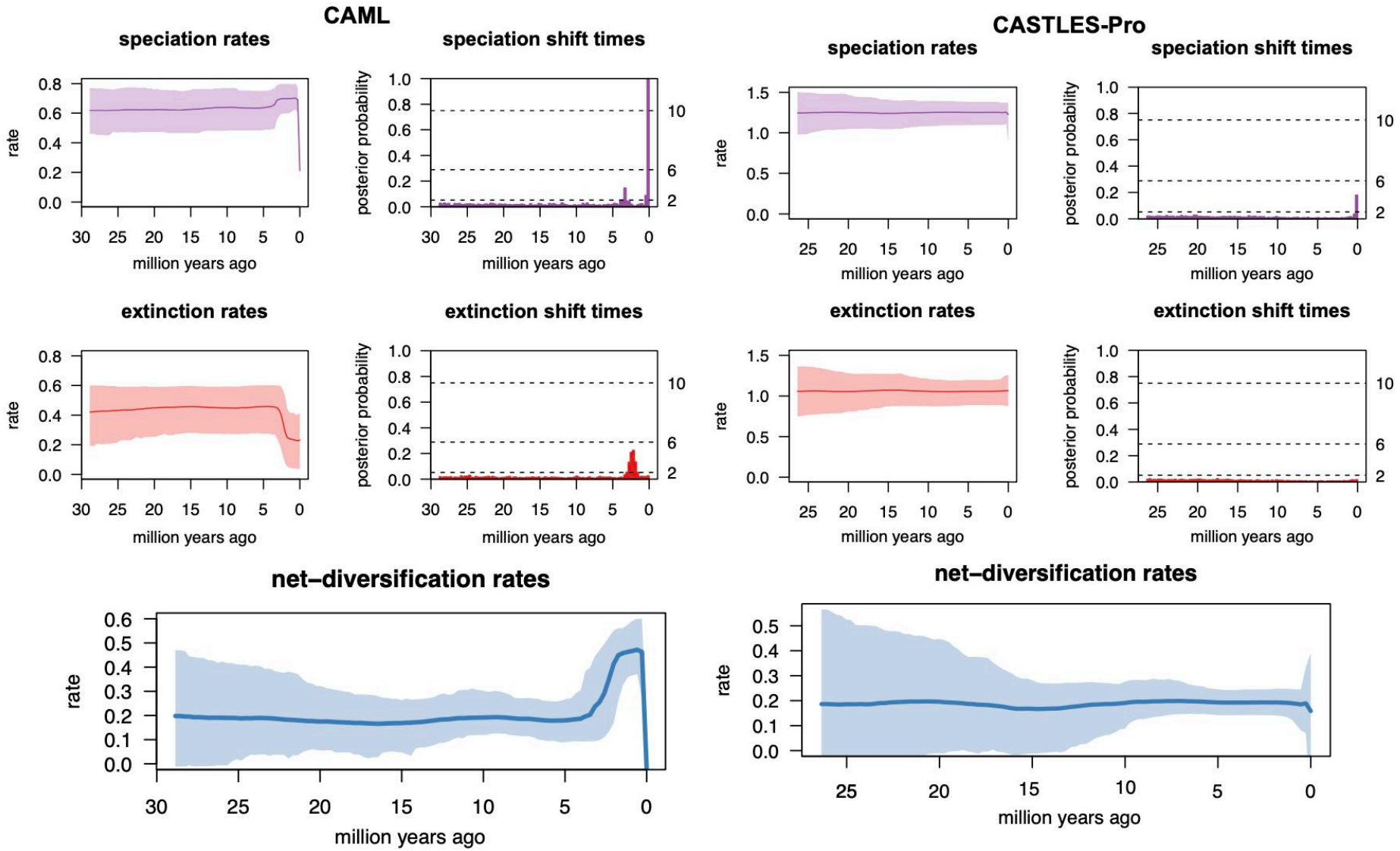
- The dramatic drop in diversification rates can be an artifact of concatenation bias, can we correct it with CASTLES-Pro?

Image Credit: Harvey et al (2020). "The evolution of a tropical biodiversity hotspot." *Science*

Tabatabaee et al., "Coalescent-based branch length estimation improves dating of species trees". 2025, submitted

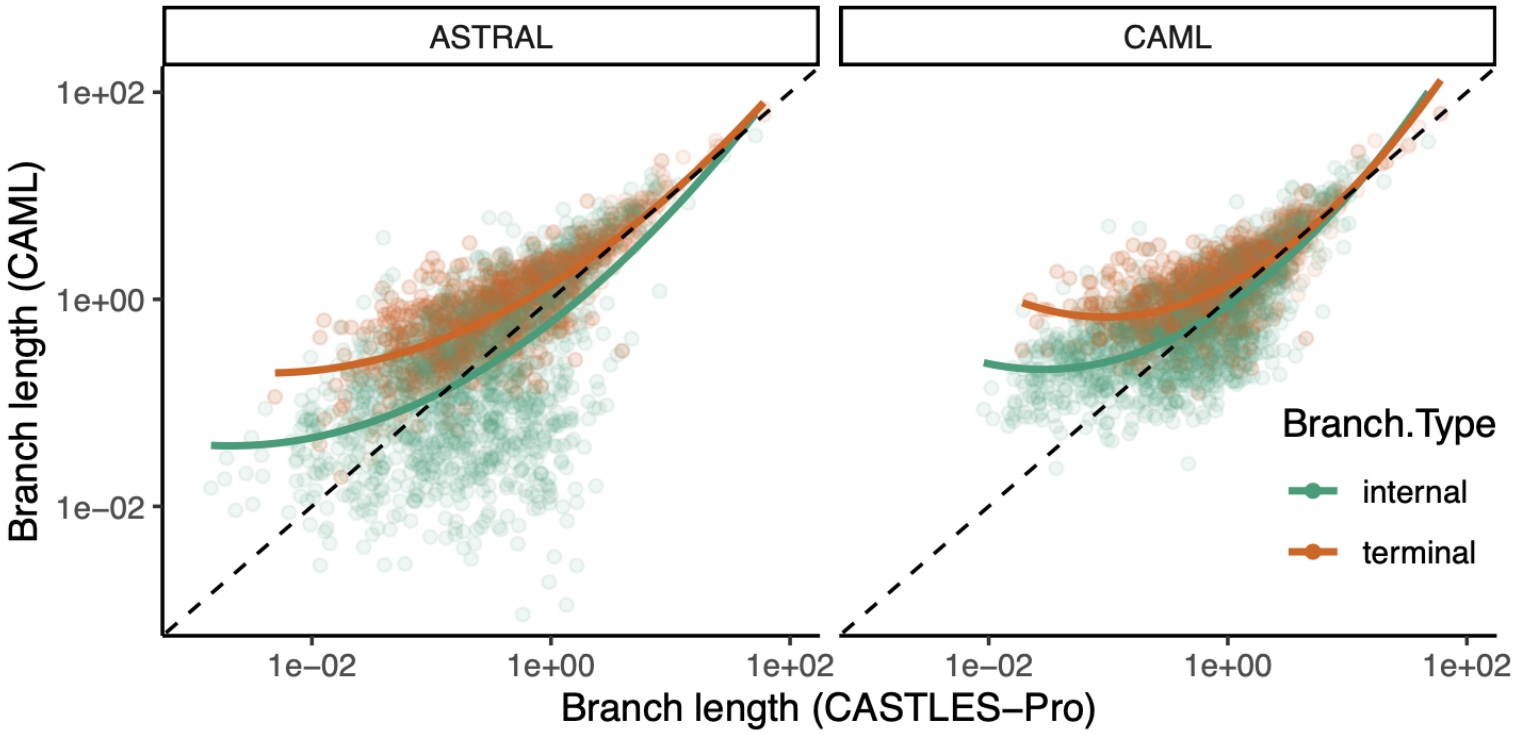
Dating with coalescent-based branch lengths eliminates the shift in diversification rates on the suboscines dataset

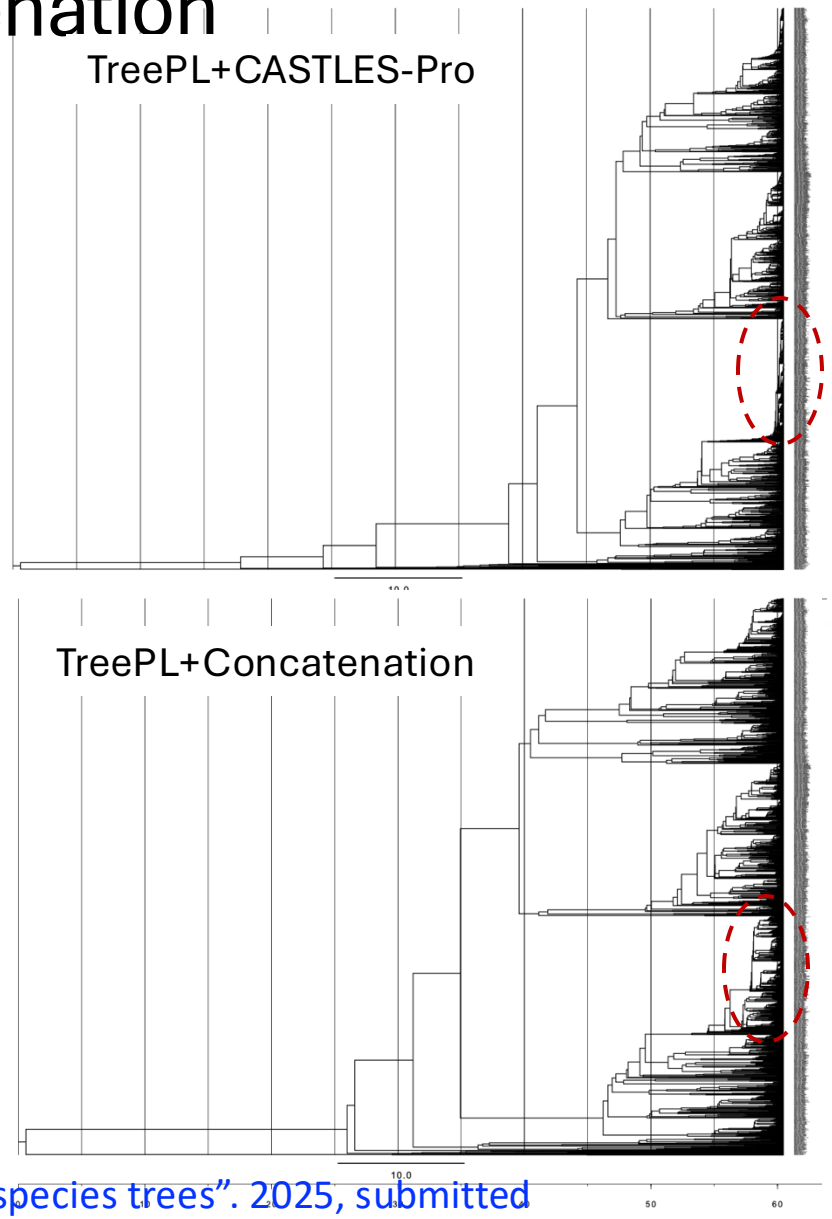
- 1683-taxon suboscines dataset with 2389 genes [Harvey et al (2020), Science]
- ASTRAL topology



Tabatabaee et al., “Coalescent-based branch length estimation improves dating of species trees”. 2025, submitted

Dating with coalescent-based branch lengths produces shorter terminal branches than concatenation





- 1683-taxon suboscines dataset with 2389 genes [Harvey et al (2020), Science]

Tabatabaee et al., "Coalescent-based branch length estimation improves dating of species trees". 2025, submitted

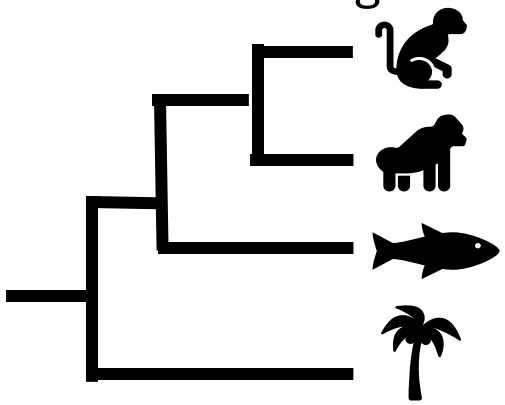
Outline

- Background and Motivation
  - Phylogenomics pipeline
  - Gene tree discordance
  - Species tree estimation
- Overview of Contributions
  - Discordance-aware post-species tree analysis
- Rooting species trees
- Phylogenomic branch length estimation
- Dating species trees and gene trees
- **Conclusions**

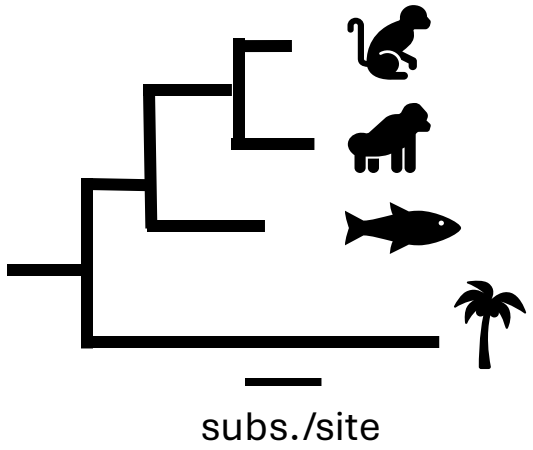
Summary of contributions

This dissertation introduces scalable and accurate methods for phylogenomic analysis, especially post-species tree analysis, that explicitly model gene tree discordance.

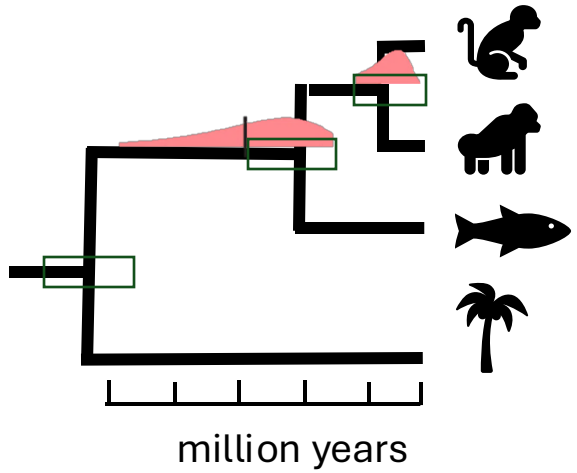
Rooting



Branch Length Estimation



Dating



- **QR** and **QR-STAR** are methods for rooting species trees that address incomplete lineage sorting, and QR-STAR comes with guarantees of consistency.
- **CASTLES** and **CASTLES-Pro** are methods for phylogenomic branch length estimation designed to work with ILS and GDL+ILS respectively.
- Ongoing research focuses on dating species trees and gene trees while addressing gene tree discordance.

Publications

Rooting species trees

- Tabatabaee, Y., Sarker, K., and Warnow, T. (2022). Quintet Rooting: rooting species trees under the multi-species coalescent model. *Bioinformatics, 38(Supplement 1):*i109–i117. (special issue for ISMB 2022)
- Tabatabaee, Y., Roch, S., and Warnow, T. (2023). Statistically consistent rooting of species trees under the multispecies coalescent model. *International Conference on Research in Computational Molecular Biology (RECOMB 2023),* pages 41–57. Springer.
- Tabatabaee, Y., Roch, S., and Warnow, T. (2023). QR-STAR: A polynomial-time statistically consistent method for rooting species trees under the coalescent. *Journal of Computational Biology*, 30(11):1146–1181. (extended version of RECOMB 2023 paper)
- Willson, J., Tabatabaee, Y., Liu, B., and Warnow, T. (2023). DISCO+QR: rooting species trees in the presence of GDL and ILS. *Bioinformatics Advances,* 3(1):vbad015.

Phylogenomic branch length estimation and dating

- Tabatabaee, Y., Zhang, C., Warnow, T., and Mirarab, S. (2023). Phylogenomic branch length estimation using quartets. *Bioinformatics, 39(Supplement 1):*i185–i193. (special issue for ISMB/ECCB 2023)
- Arasti, S., Tabaghi, P., Tabatabaee, Y., and Mirarab, S. (2024). Optimal tree metric matching enables phylogenomic branch length reconciliation. *RECOMB* 2024
- Tabatabaee, Y., Zhang, C., Arasti, S. and Mirarab, S. (2025). Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss. Submitted.
- Tabatabaee, Y., Claramunt, S., and Mirarab, S. (2025). Coalescent-based branch length estimation improves dating of species trees. Submitted.

Acknowledgements

Thank you!

Funding:

NSF

NIH National Institutes of Health

Computing Resources:
UIUC Campus Cluster


Tandy Warnow


Siavash Mirarab


Sebastien Roch


Santiago Claramunt


Chao Zhang


Kowshika Sarker


James Willson


Puoya Tabaghi


Shayesteh Arasti

Members of Warnow Lab

Members of Mirarab Lab