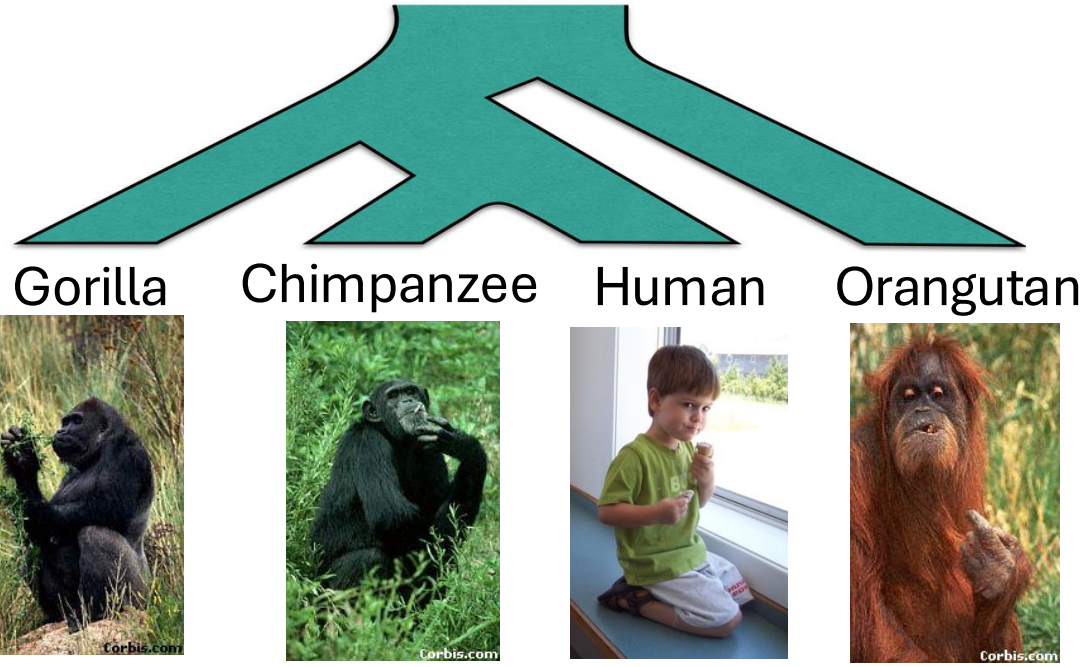
CASTLES-Pro: A tool for species tree branch length estimation despite ILS and GDL

Presenter: Yasamin Tabatabaee

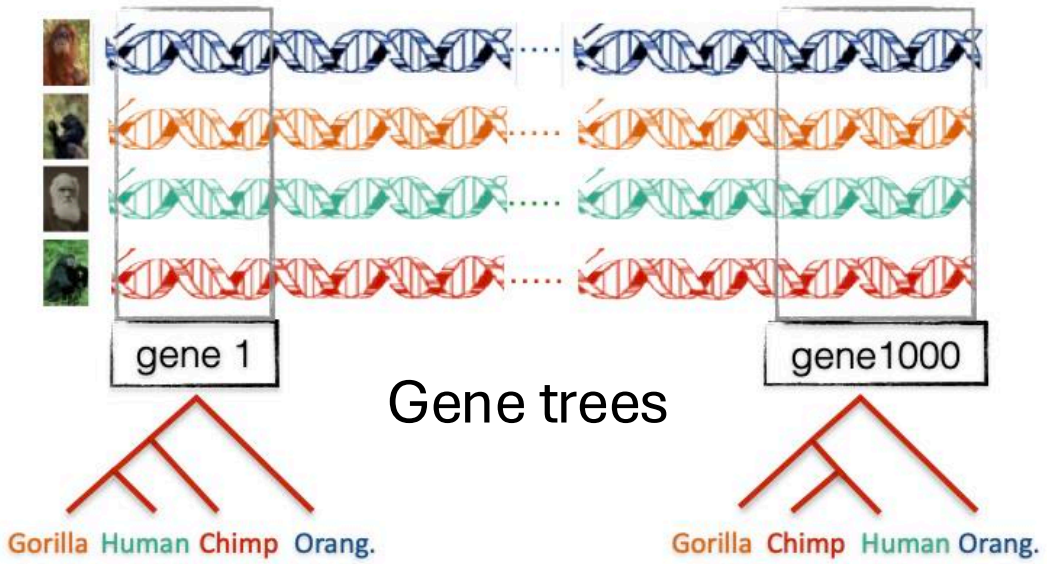
Siebel School of Computing and Data Science

University of Illinois Urbana-Champaign

Phylogenomics and gene tree discordance

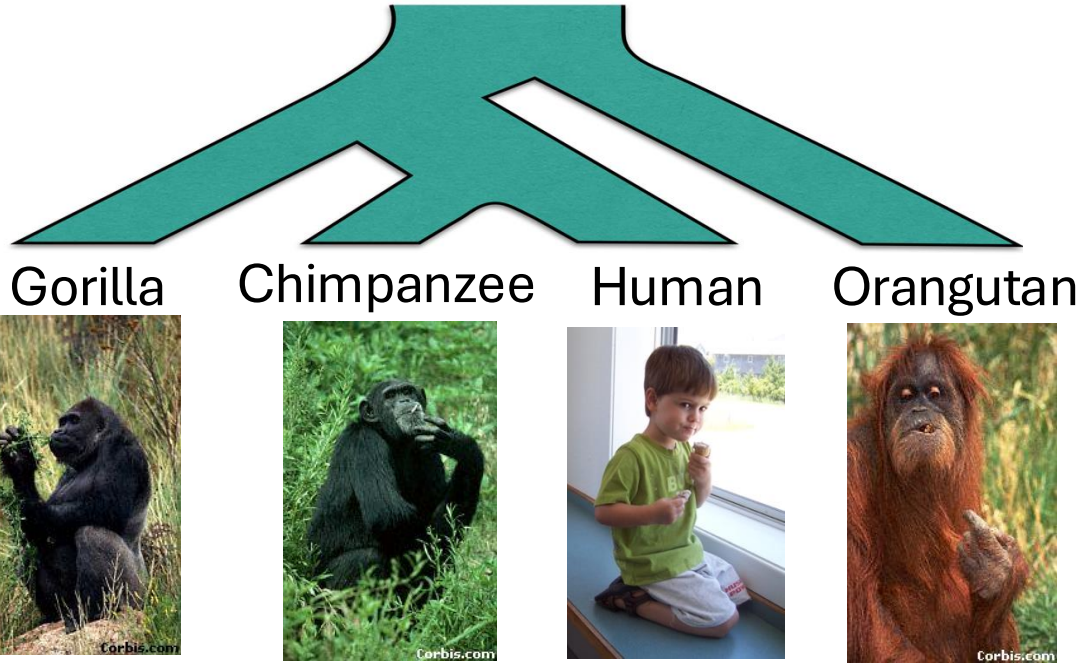


Species tree

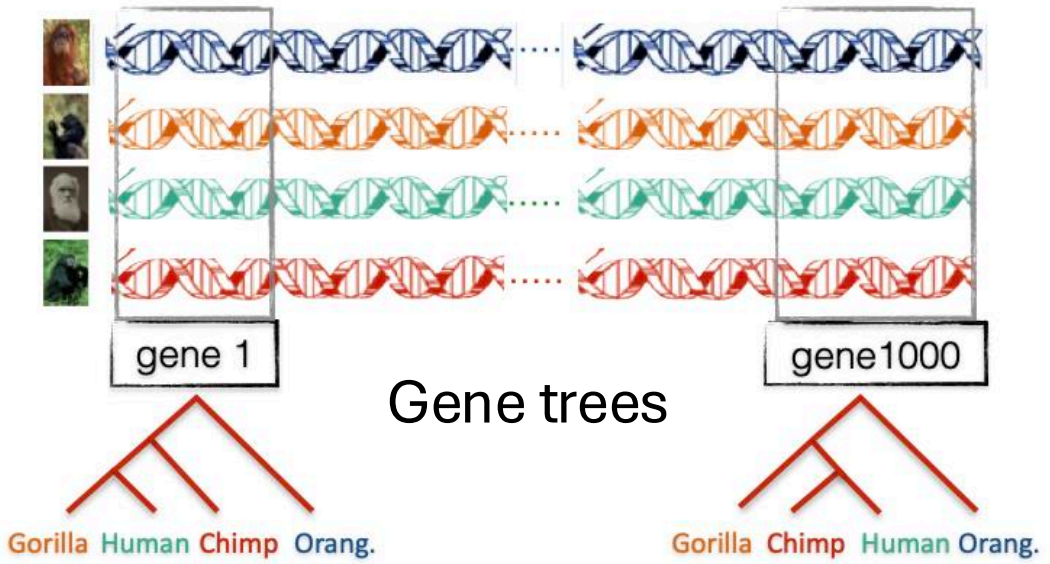


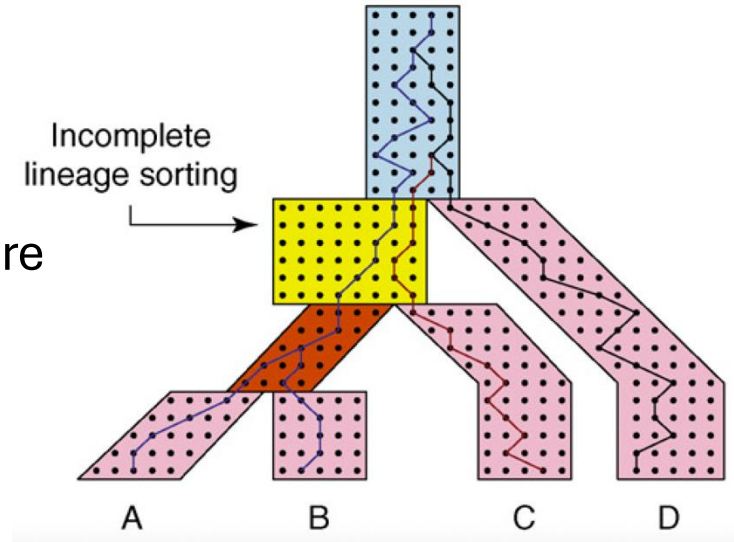
Gene trees

Phylogenomics and gene tree discordance



Species tree

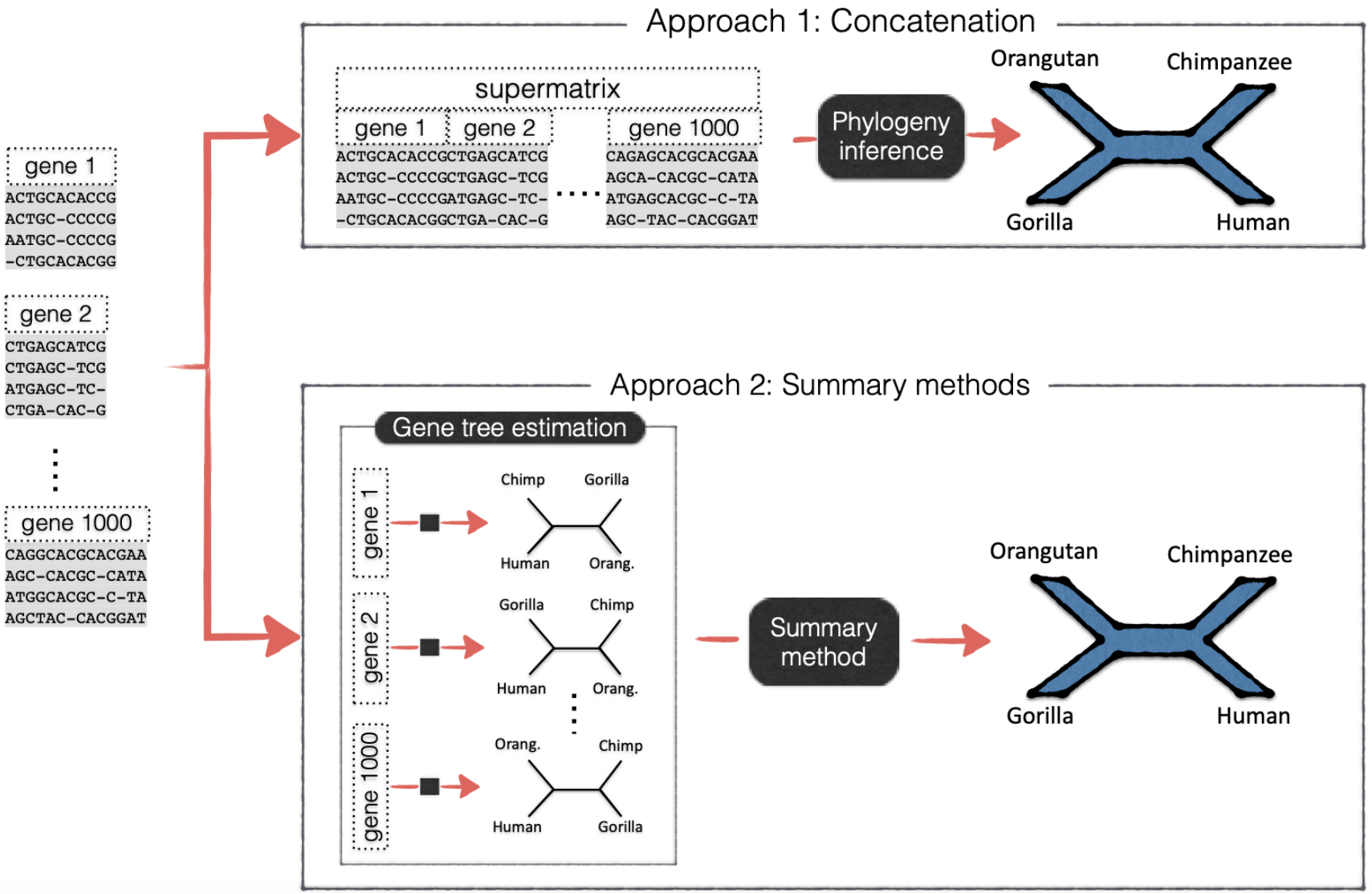




- Incomplete lineage sorting (ILS) and gene duplication and loss (GDL) are major causes of gene tree discordance.
- ILS can be modeled by the multi-species coalescent (MSC) model.

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

Species tree estimation

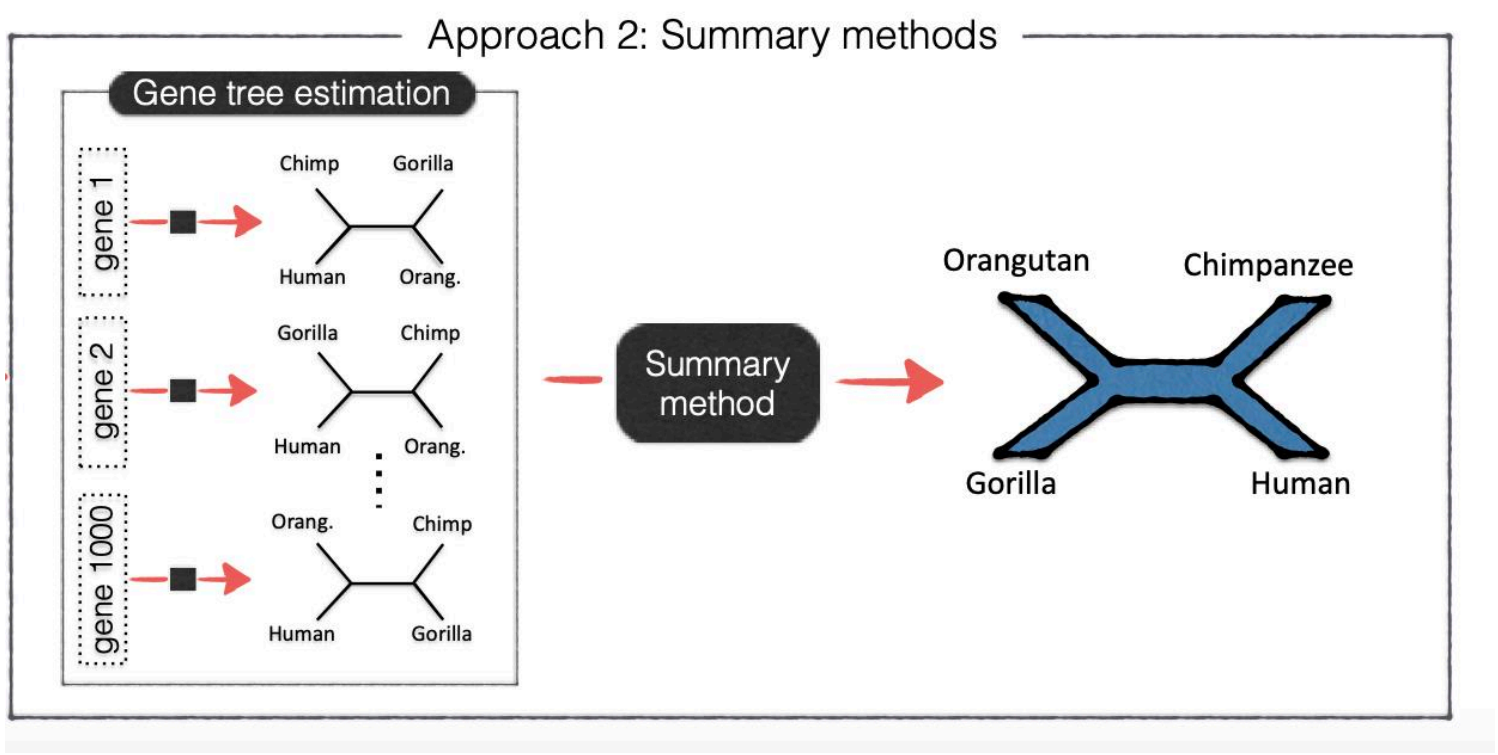


Maximum Likelihood, e.g.
RAxML [Stamatakis, 2014]
FastTree [Price et al, 2010]

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

Image Credit: Siavash Mirarab

Branch length estimation and dating



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

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

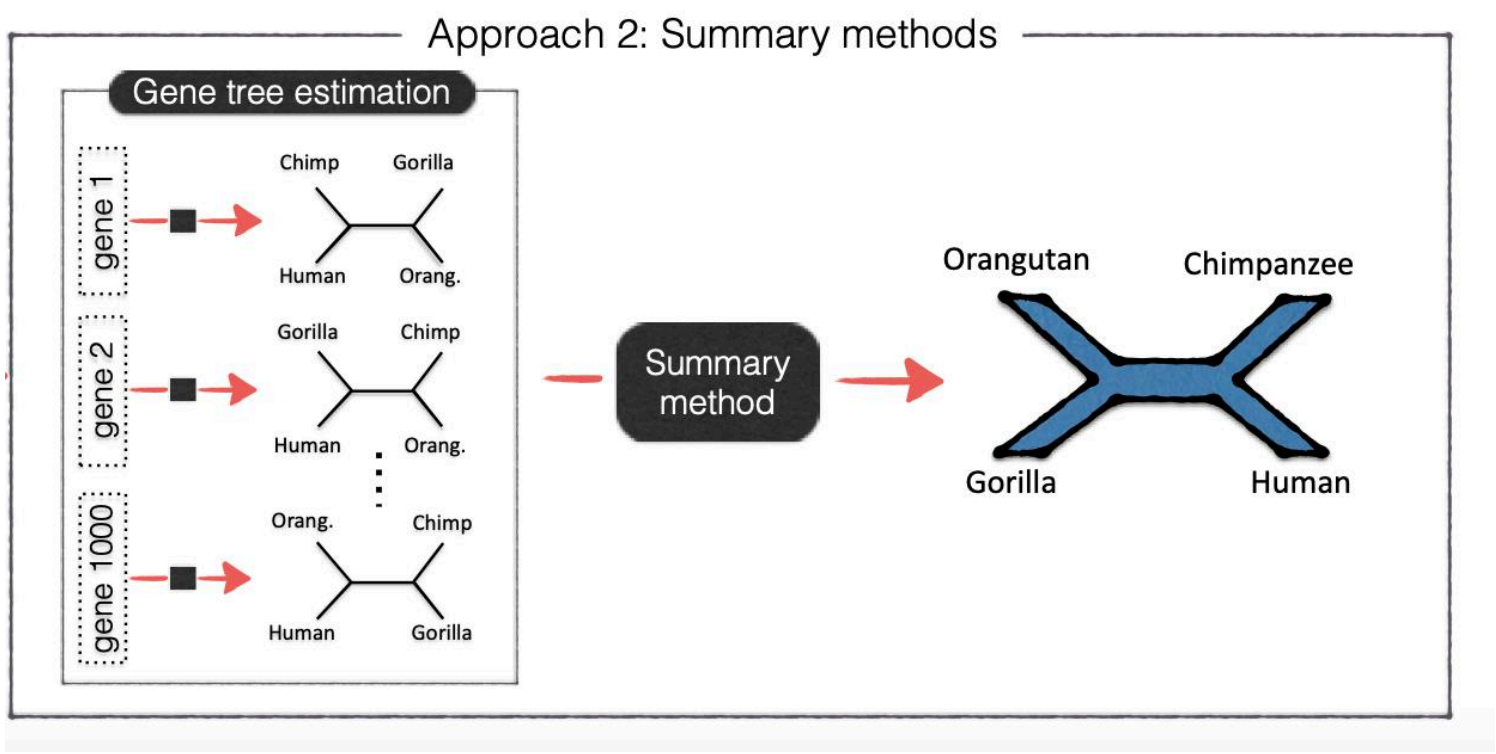
- **Two-step approach:**

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

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

Image Credit: Siavash Mirarab

Branch length estimation and dating



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

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

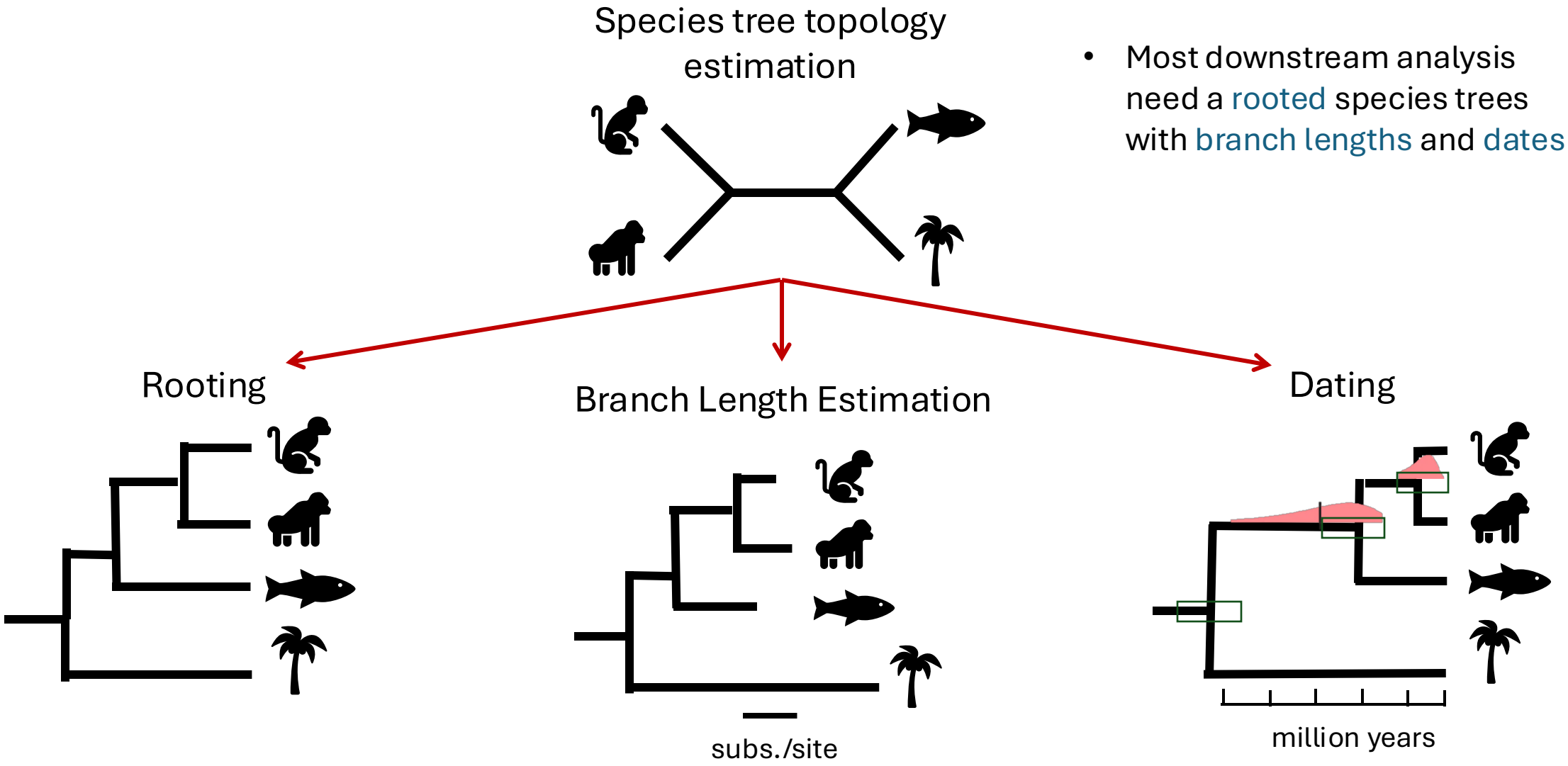
- **Two-step approach:**

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

  - Based on some form of concatenation analysis
  - Ignores heterogeneity

Image Credit: Siavash Mirarab

Discordance-aware post-species tree analysis



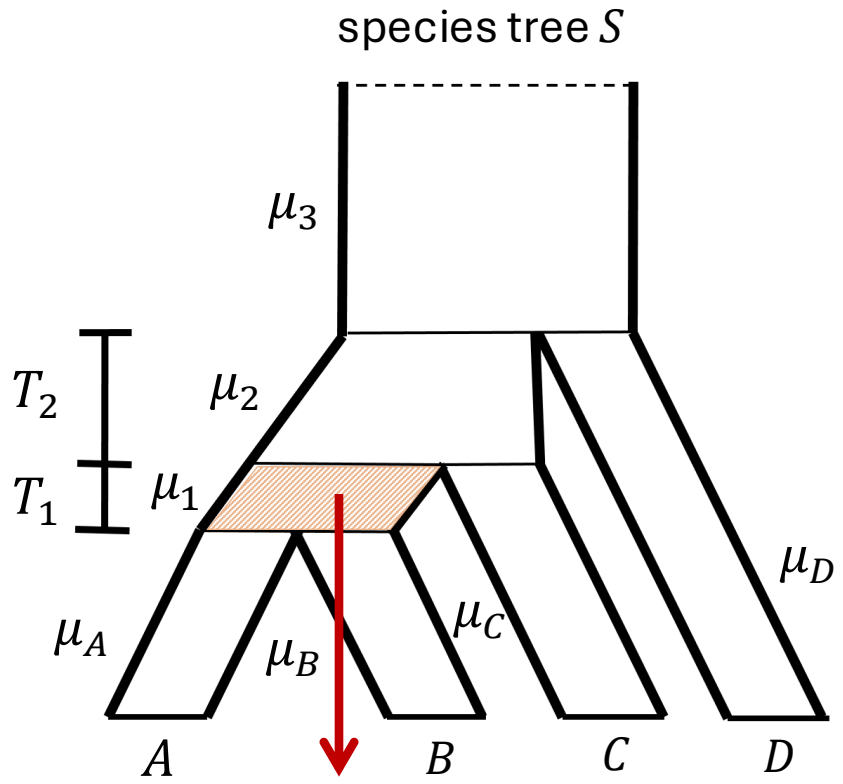
- Most downstream analysis need a rooted species trees with branch lengths and dates

CASTLES / CASTLES-Pro

A branch length estimation method that…

- estimates branch lengths in substitution units (SU)
- addresses gene tree heterogeneity due to ILS and GDL and variation in mutation rates
- is scalable to large genome-wide datasets with hundreds to thousands of genes and species

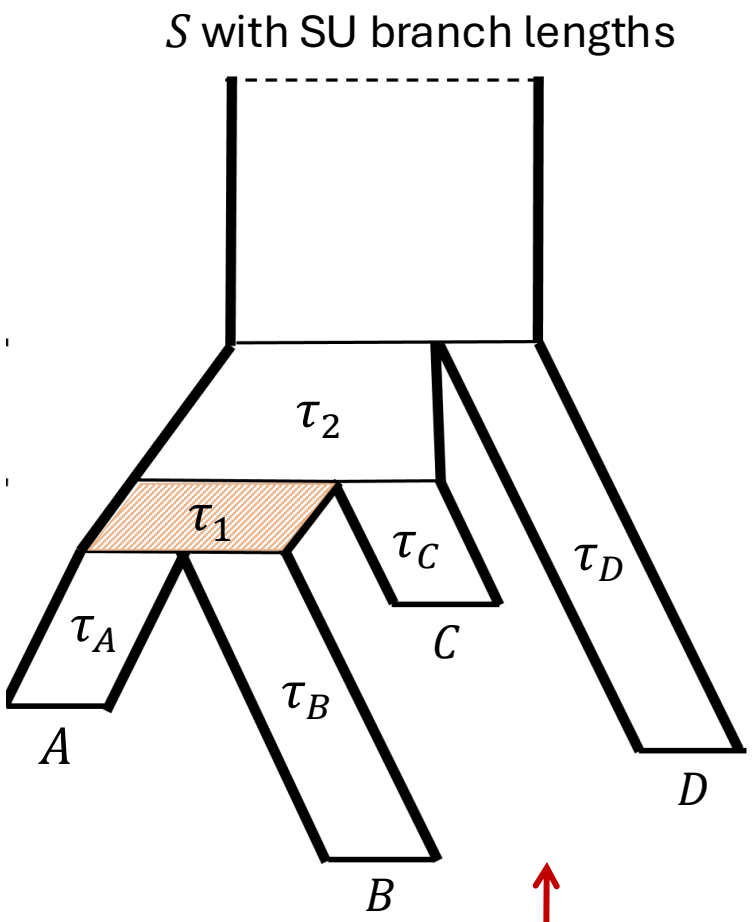
MSC+Substitution model

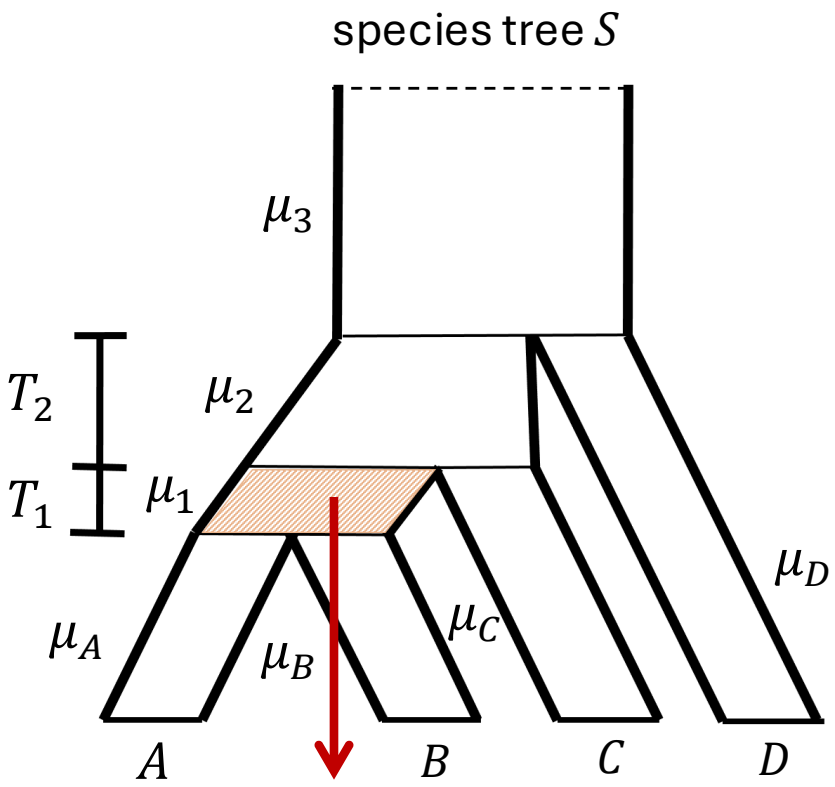


Parameters of branch 1:

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

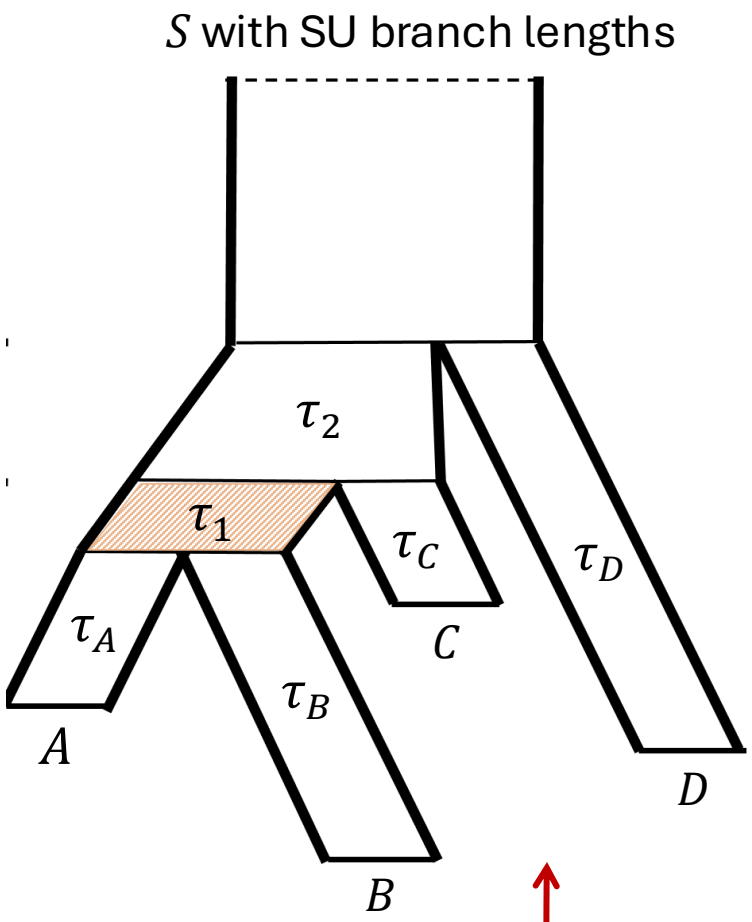


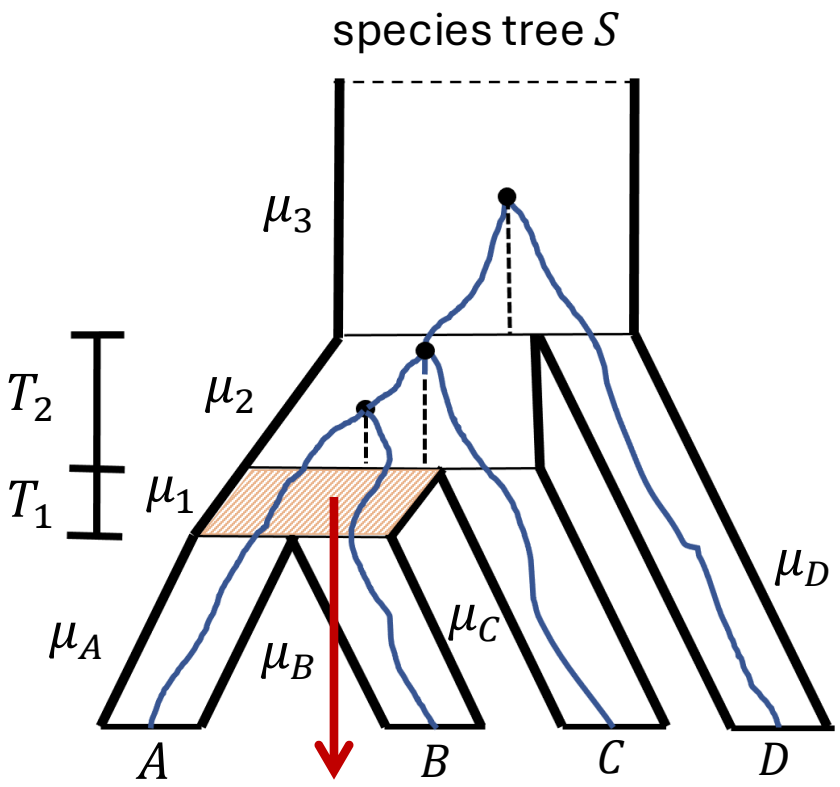


Parameters of branch 1:

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

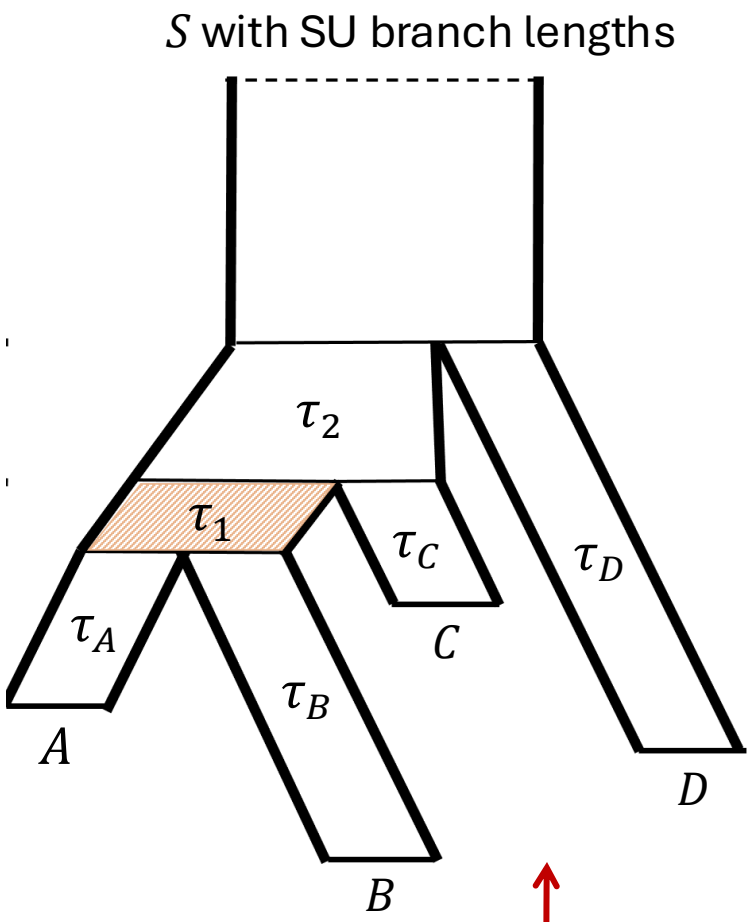


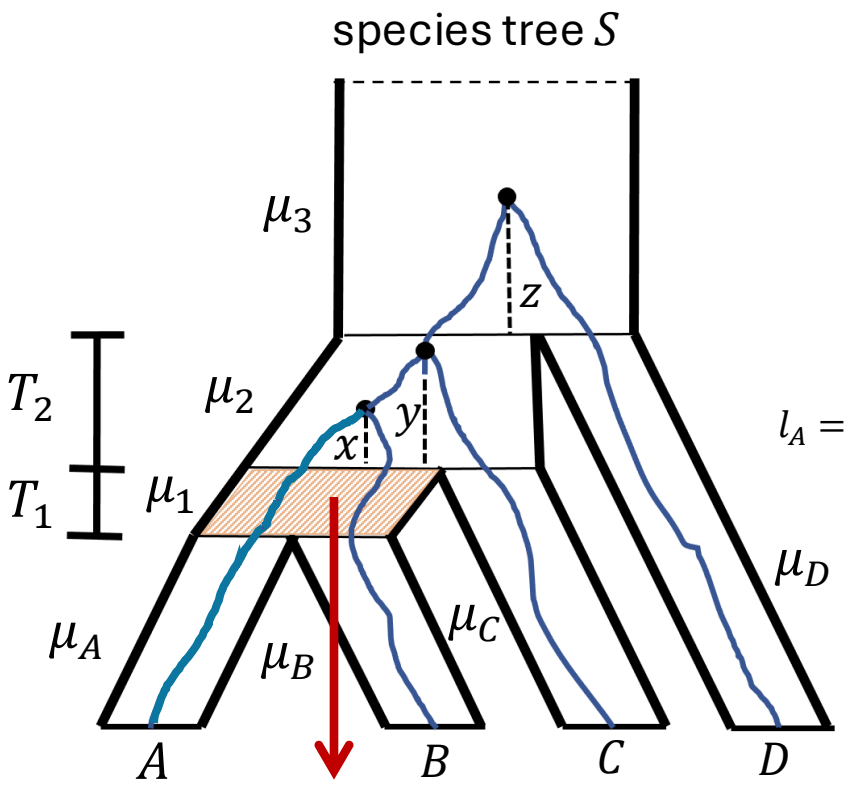


Parameters of branch 1:

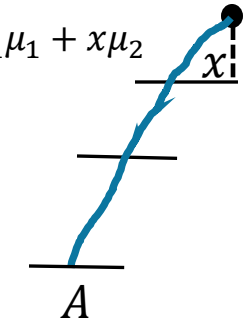
μ_1	Substitution rate
T_1	Branch length in coalescent units (CU)
$\tau_1 = T_1 . \mu_1$	Branch length in substitution units (SU)

MSC+Substitution model





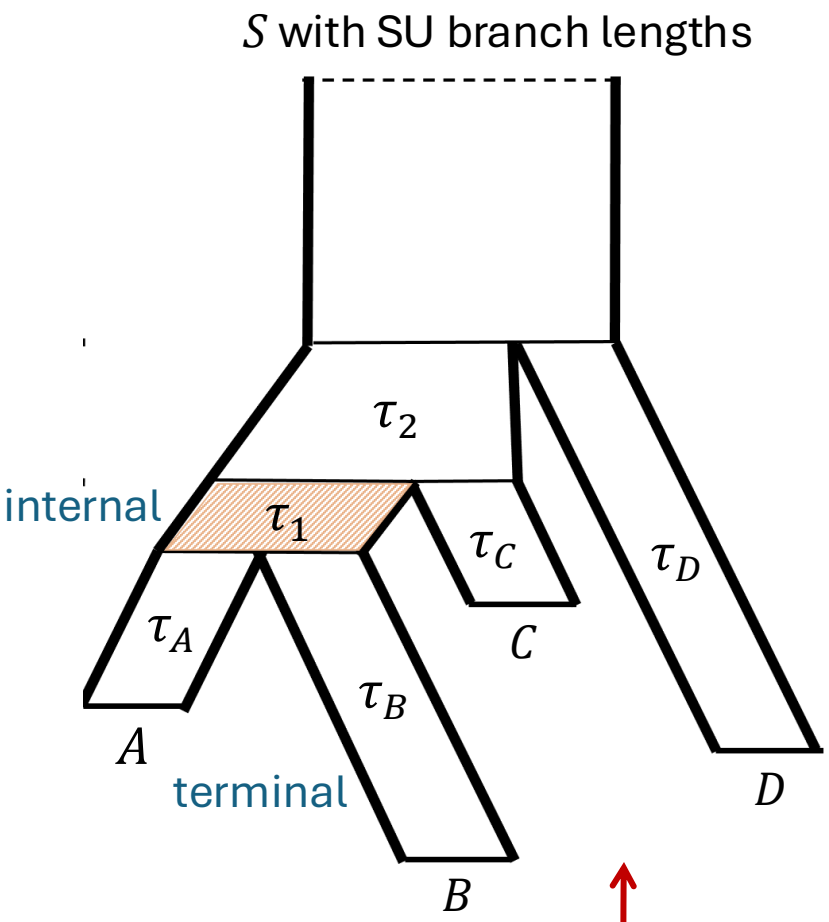
$l_A = T_A\mu_A + T_1\mu_1 + x\mu_2$

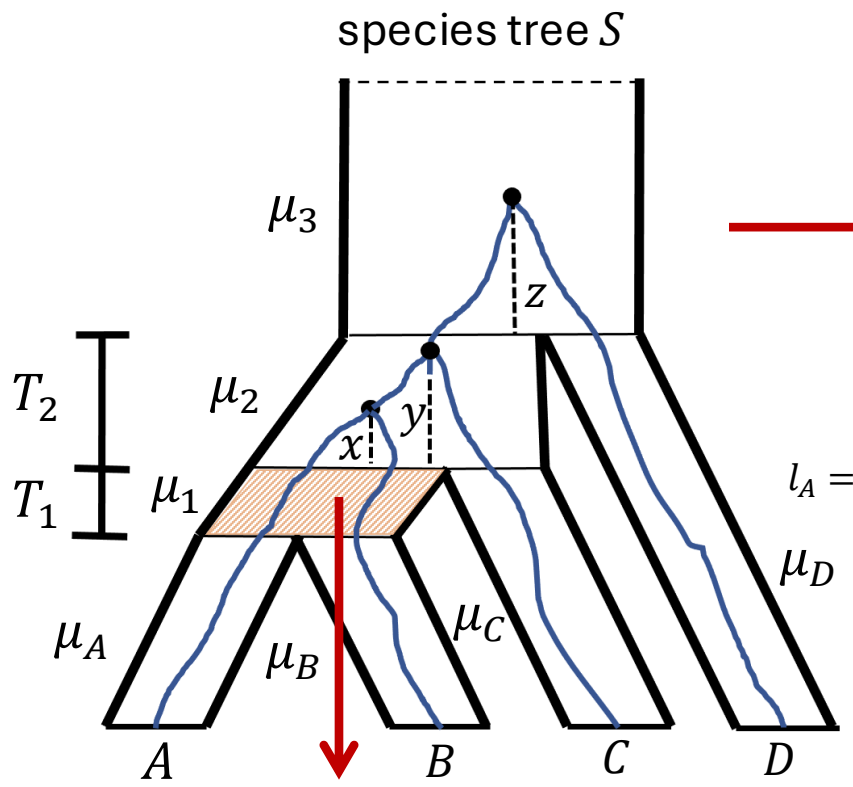


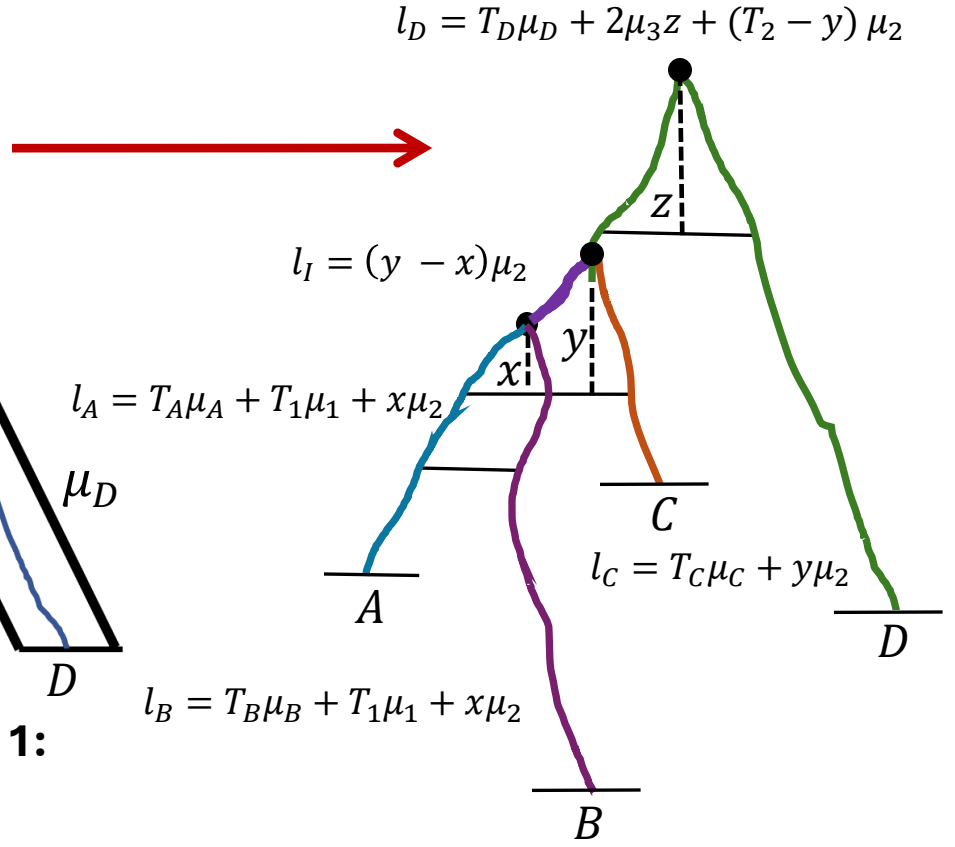
Parameters of branch 1:

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



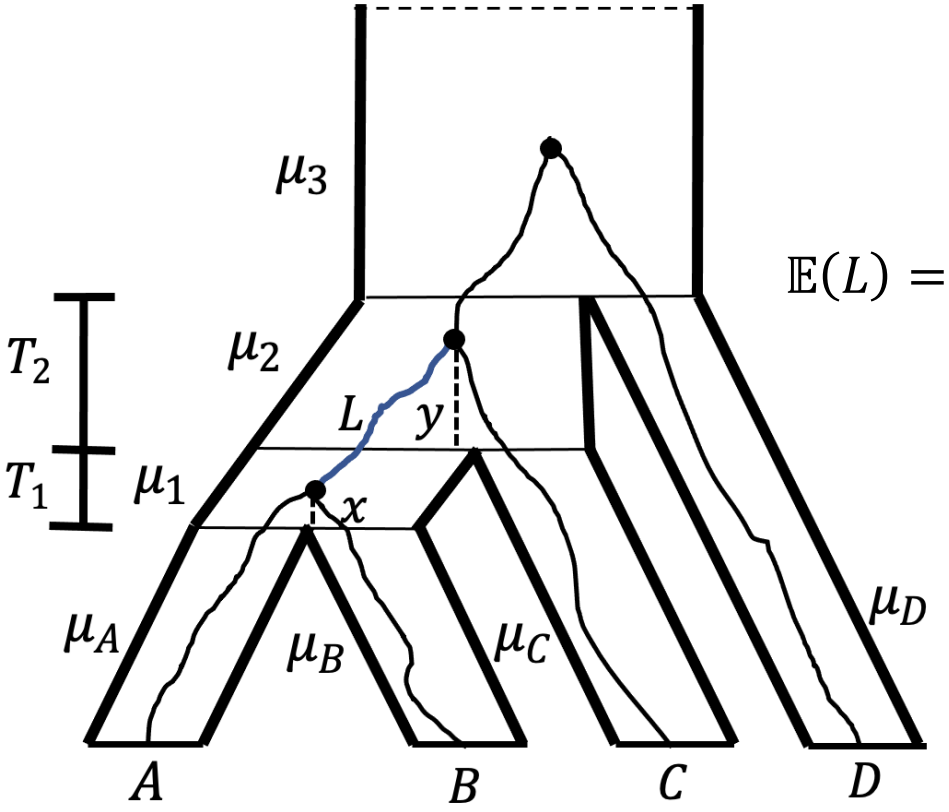




Parameters of branch 1:

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

Expected quartet branch lengths under MSC



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

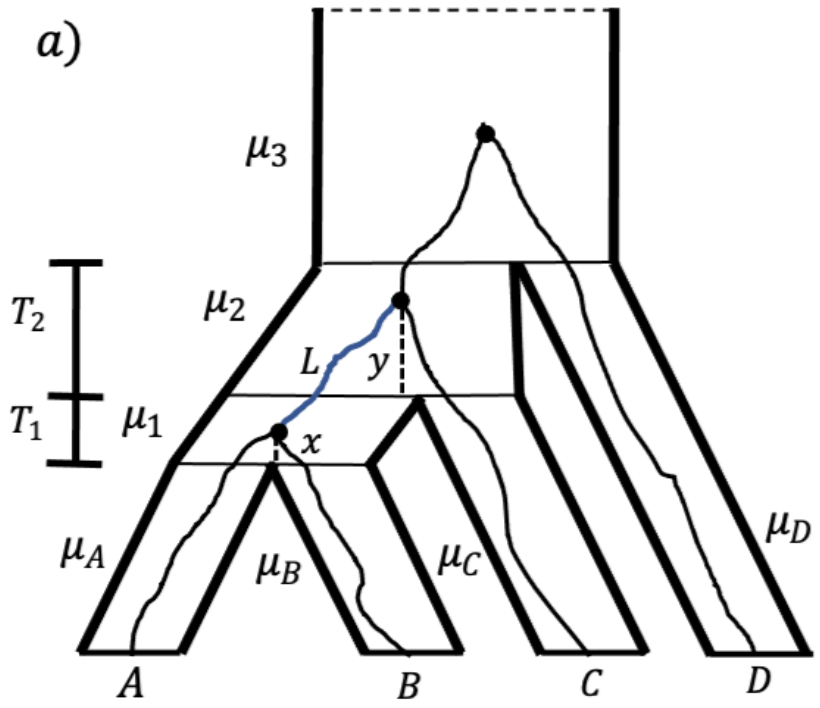
$$\mathbb{E}(L) = \int_0^{T_1}\int_0^{T_2} e^{-x}\, e^{-y}\big((T_1 - x)\mu_1 + y\mu_2\big)\,dy\,dx$$

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

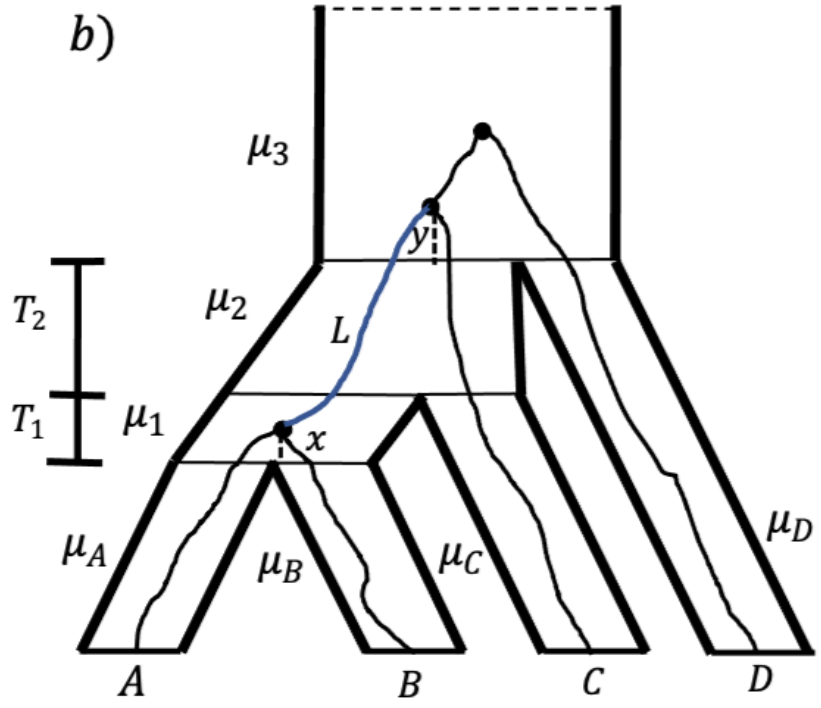
- Under MSC, waiting times before coalescent events are exponential random variables with rate $\lambda = \binom{k}{2}$ where k is the number of lineages entering an interval

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

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

CASTLES

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

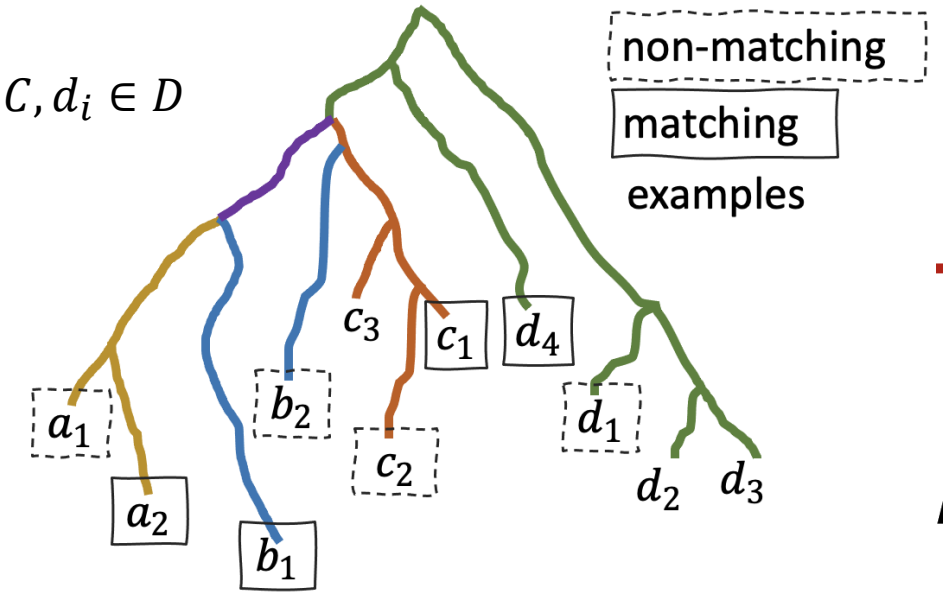
Input:
- Rooted species tree *topology* S
- A set of gene trees $\mathcal{G}$ with SU branch lengths

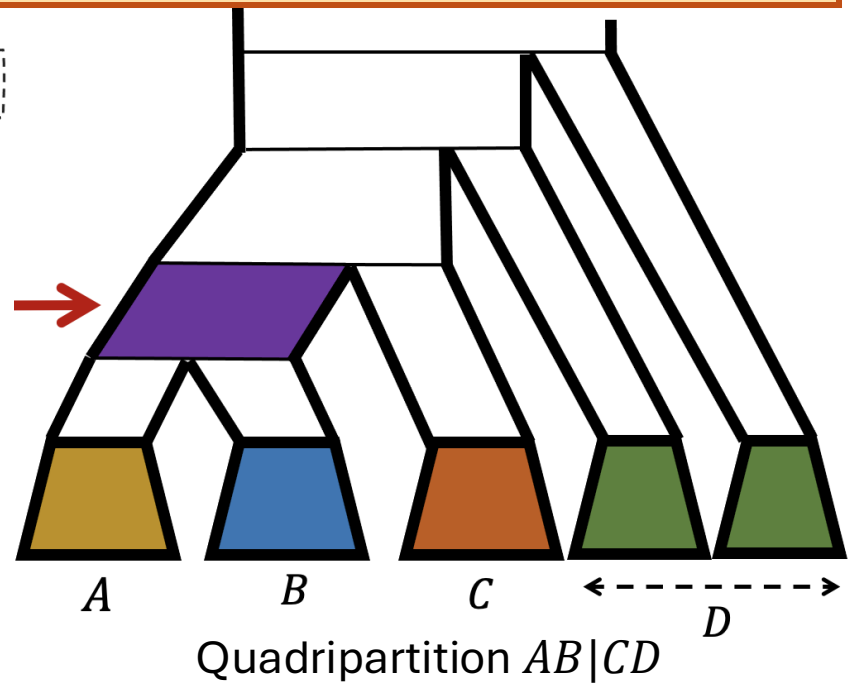
Output:
- Species tree S with SU branch lengths

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



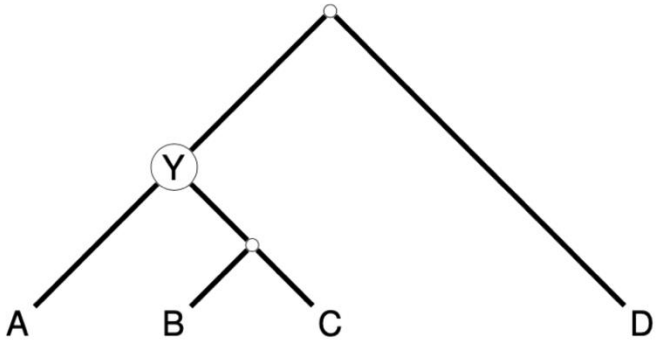


Quadripartition $AB|CD$

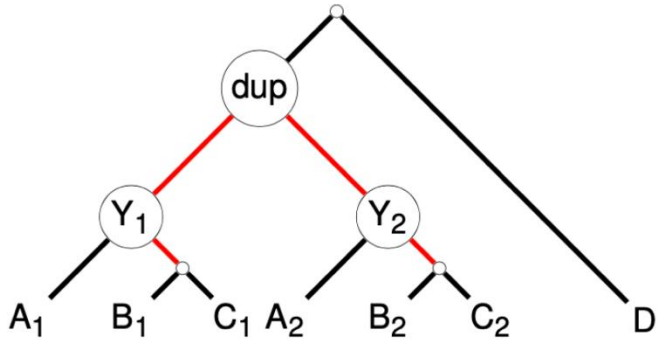
CASTLES-Pro

"Pro" stands for PaRalogs and Orthologs…

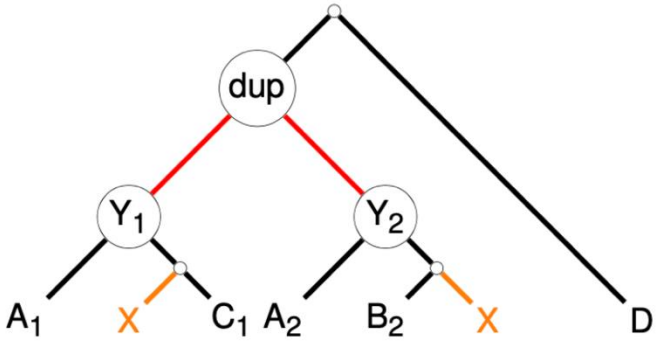
- Extends CASTLES to work with multi-copy genes
- Only considers quartets devoid of paralogous genes
- Improves upon CASTLES by modifying some of its equations
- Relaxes some of the approximations used in CASTLES and is more theoretically appealing
- Implements a weighting scheme to account for uneven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



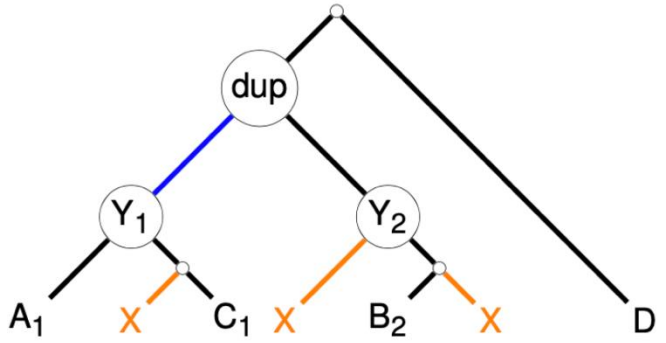
(a) Species tree T^*



(b) Gene tree M_1 with one duplication.



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



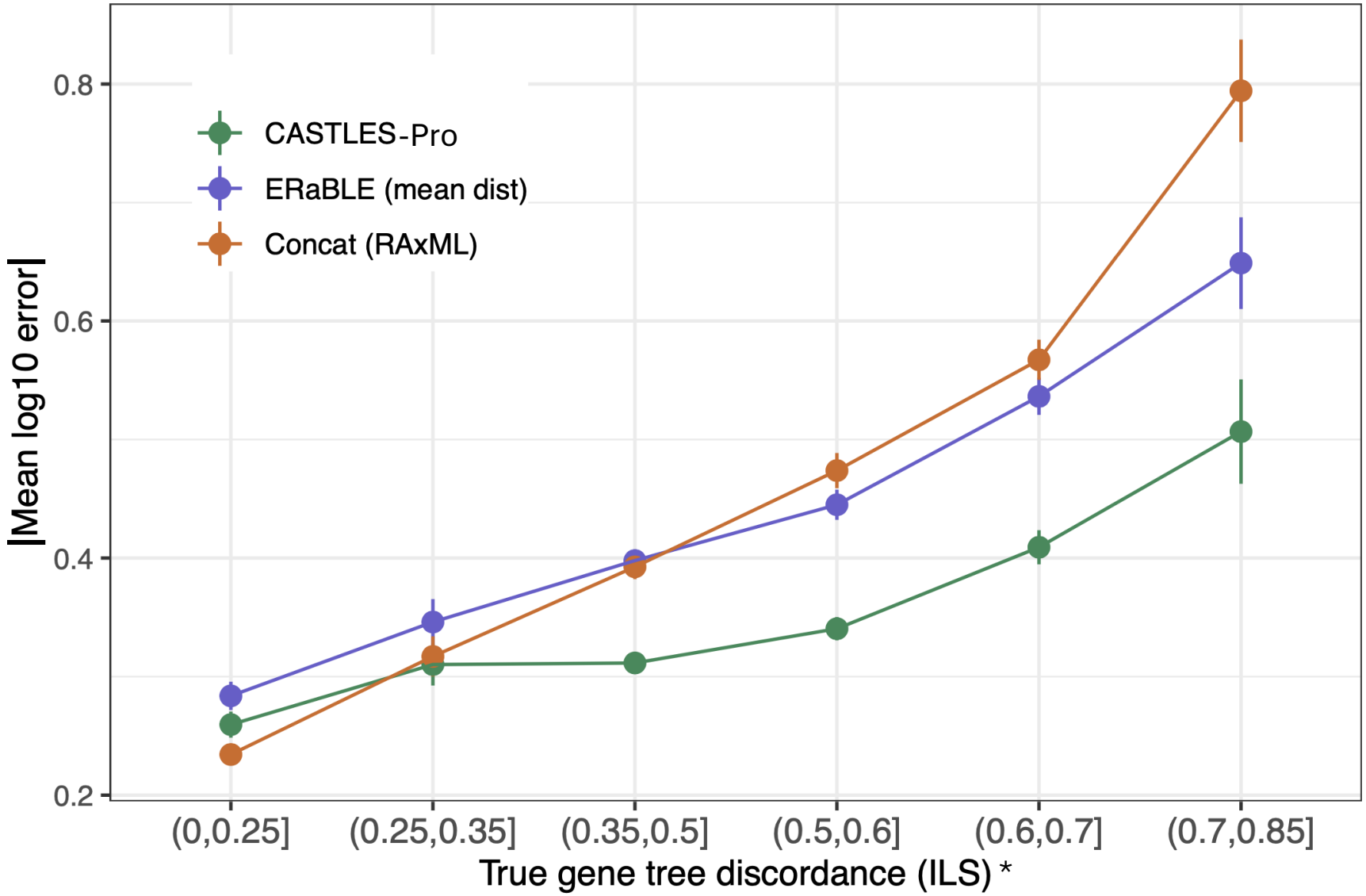
(d) Gene tree with one duplication and three losses.

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

CASTLES-Pro's advantage increases with ILS

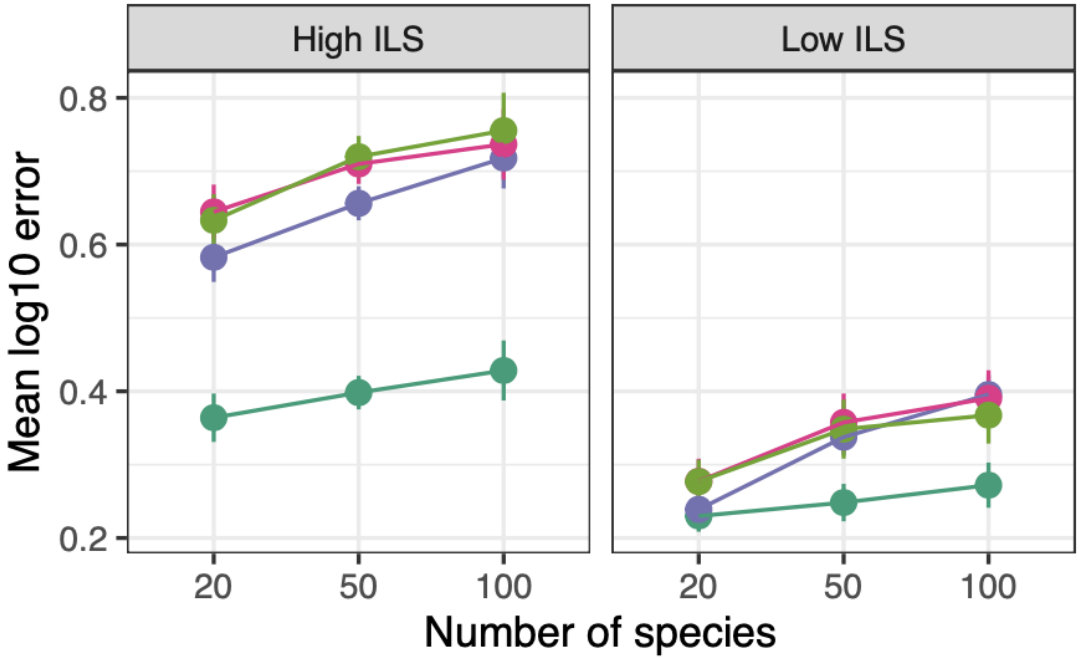
- 30-taxon ILS simulated dataset with 500 estimated gene trees [Mai et al (2017)]

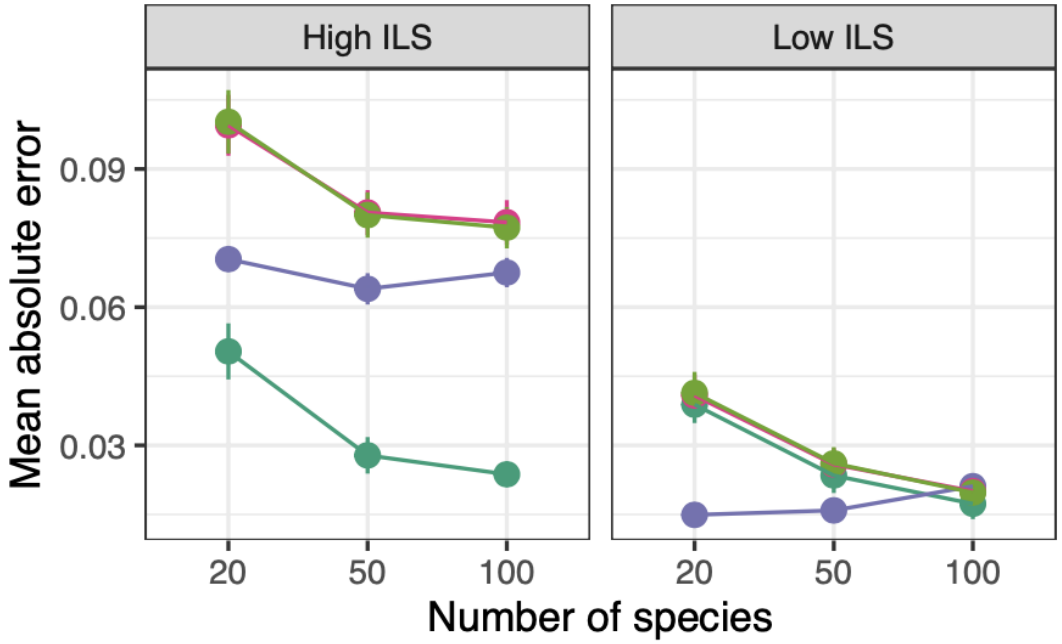
|log10 (est. length / true length)|



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

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





CASTLES–Pro

CA–DISCO(RAxML)

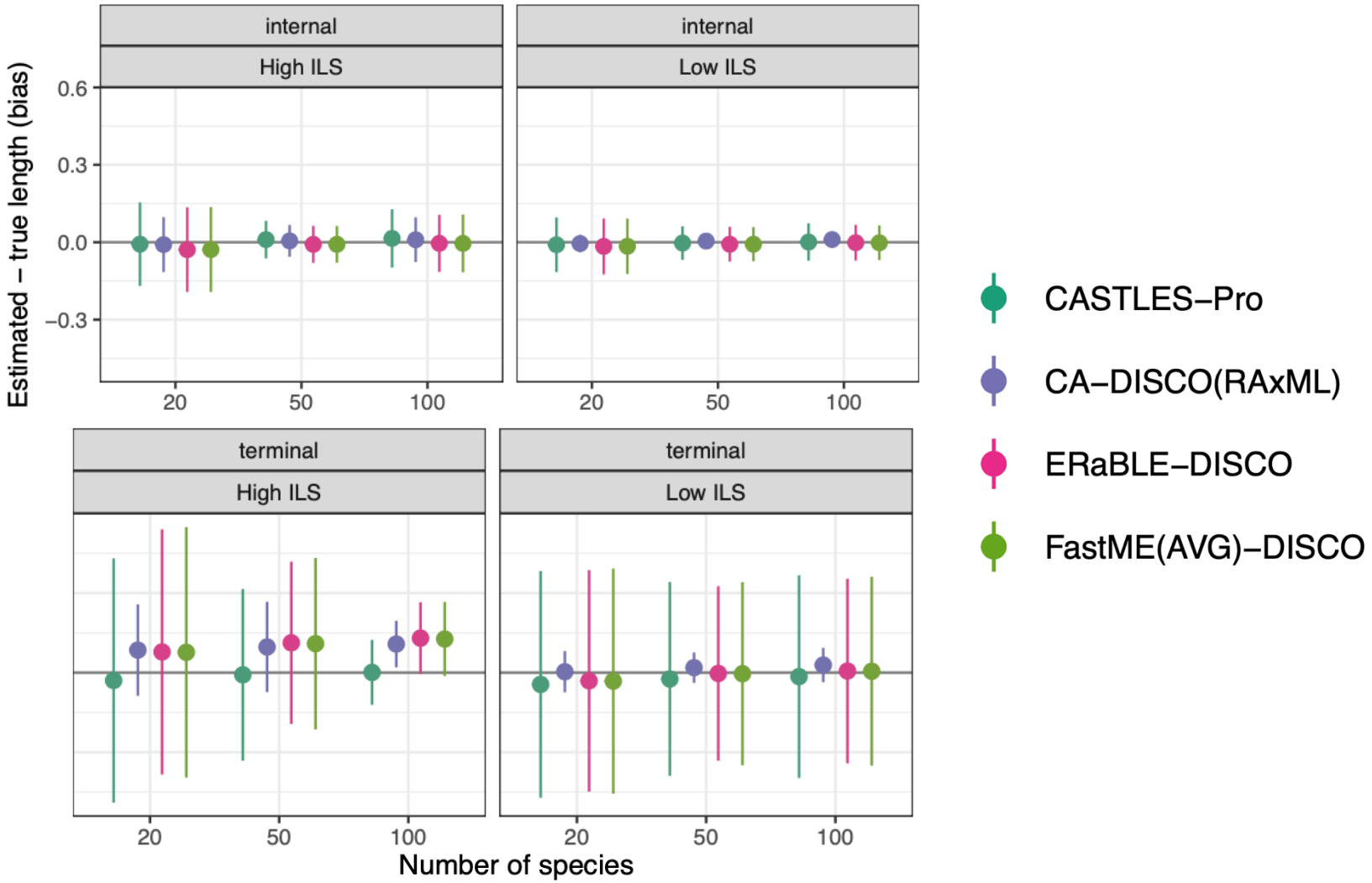
ERaBLE–DISCO

FastME(AVG)–DISCO

- DISCO [Willson et al (2022)] decomposes multi-copy gene trees into single-copy ones
- GDL+ILS simulated dataset with 1000 estimated gene trees [Willson et al (2022,2023)]

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

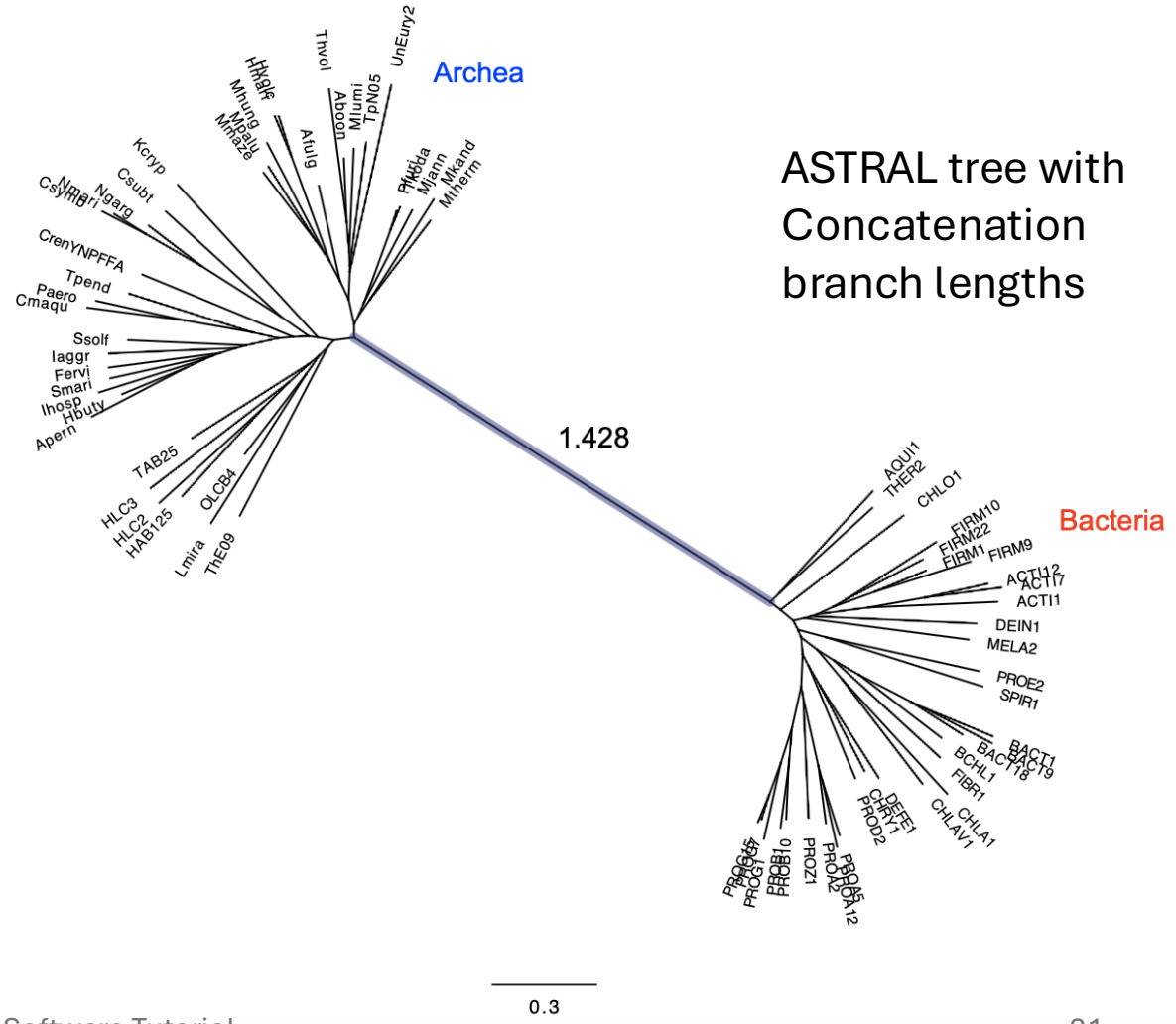
- DISCO [Willson et al (2022)] decomposes multi-copy gene trees into single-copy ones
- GDL+ILS simulated dataset with 1000 estimated gene trees [Willson et al (2022,2023)]



Biological 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)



Biological 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

- Can CASTLES-Pro correct this?

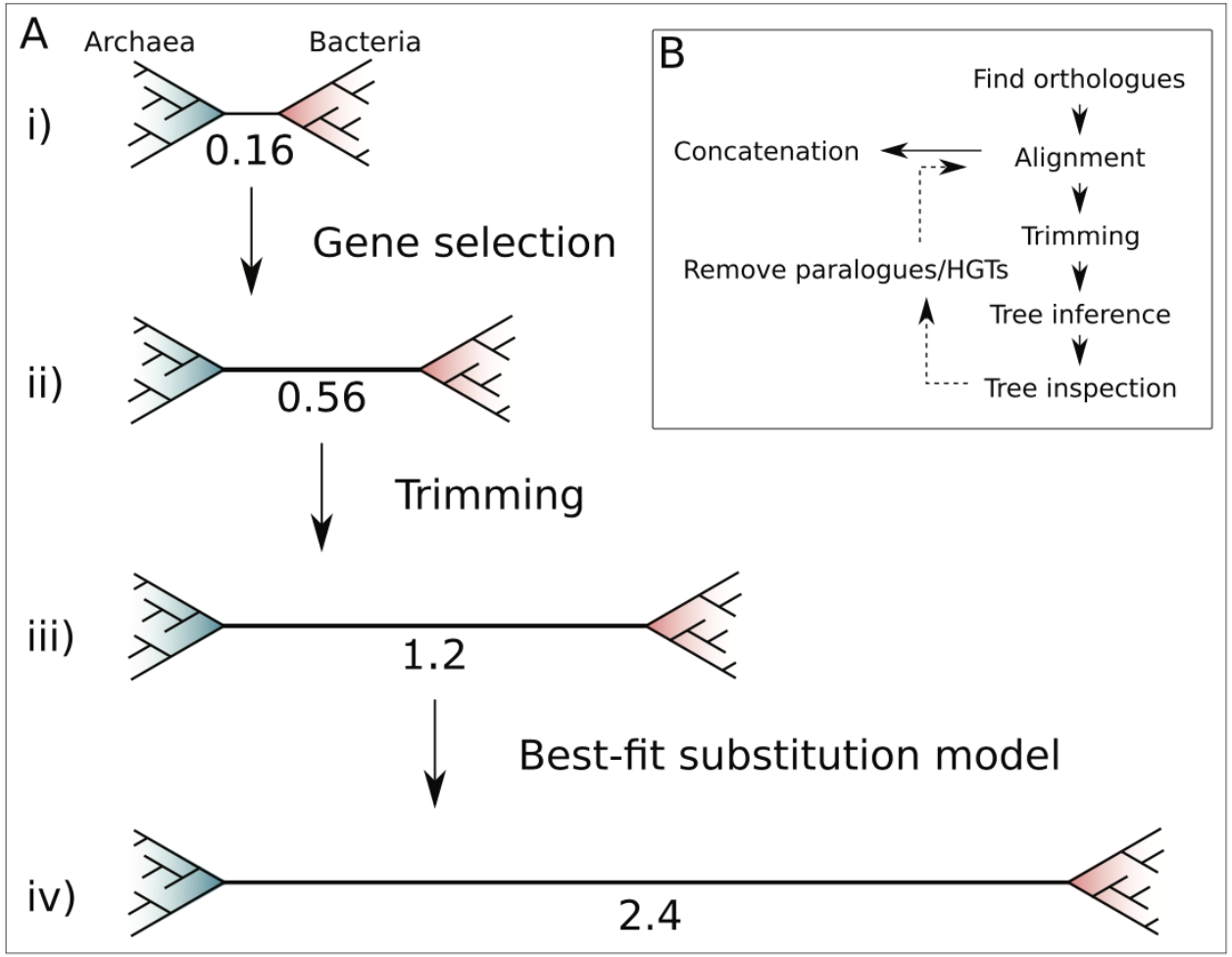
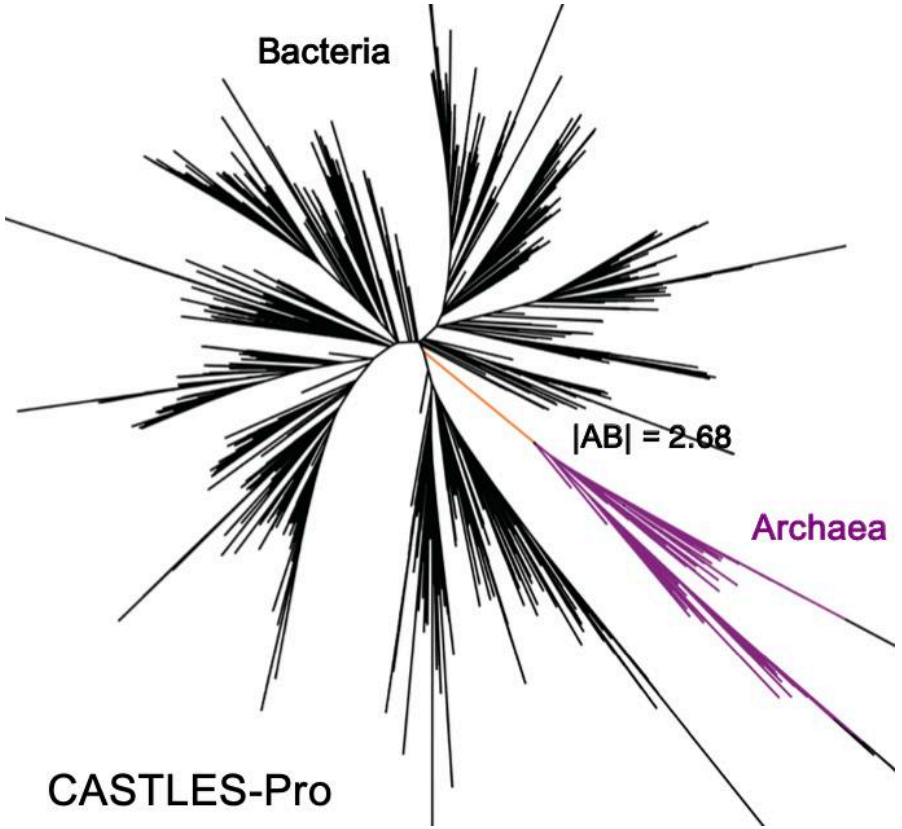
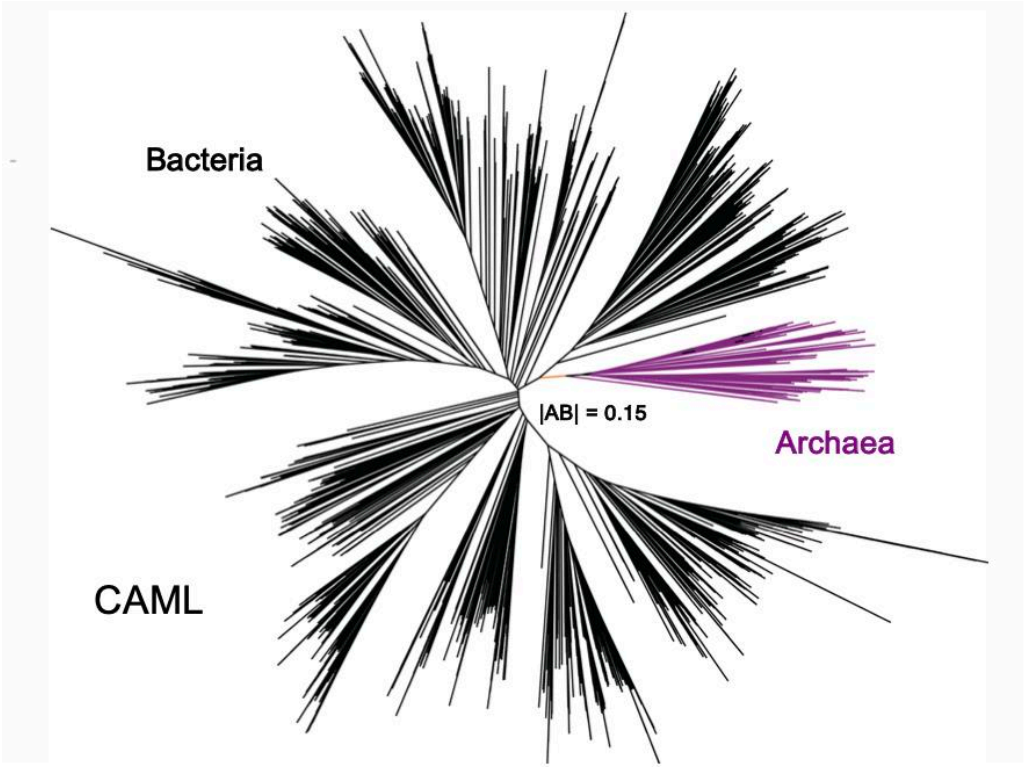


Image Credit: Moody et al., 2022, "An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *Elife*
"

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

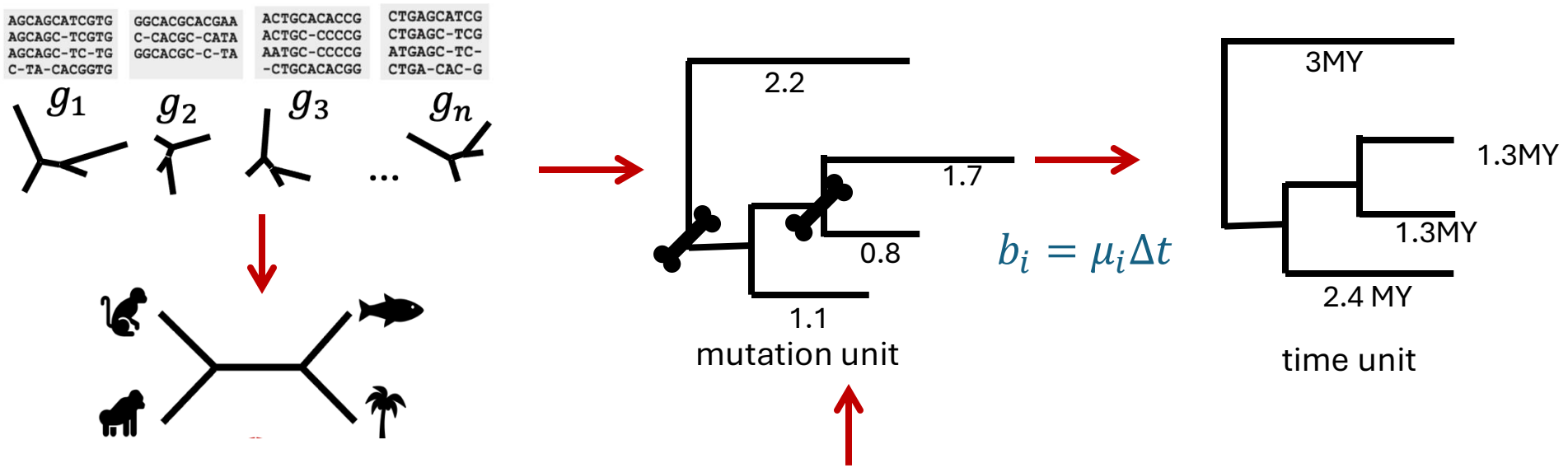




- Bacterial dataset with 10,575 species and 381 marker genes [Zhu et al. (2019)].

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

Typical likelihood-based dating pipeline



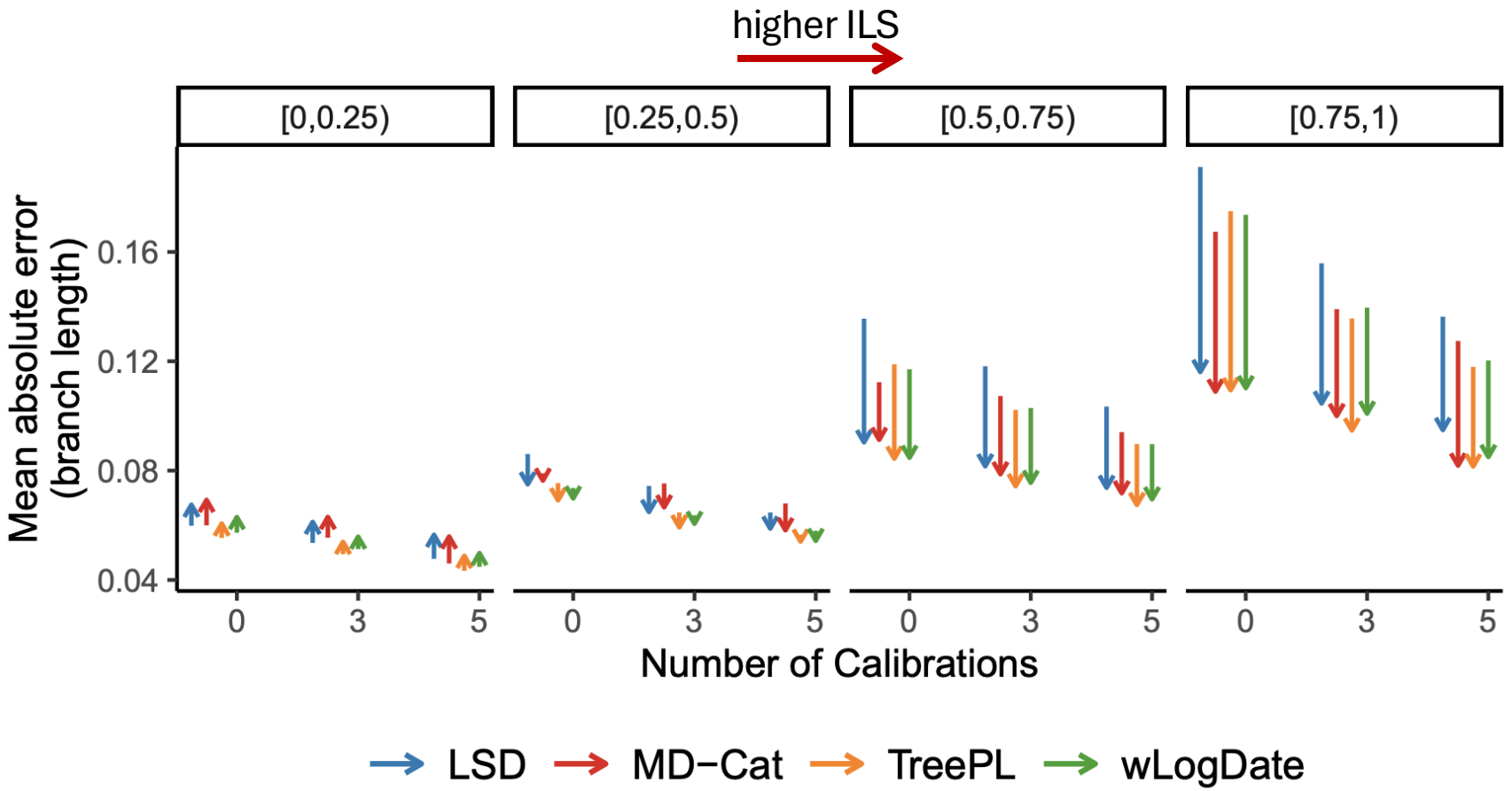
- Can accounting for gene tree discordance in the branch length estimation step improve the dating pipeline?

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:

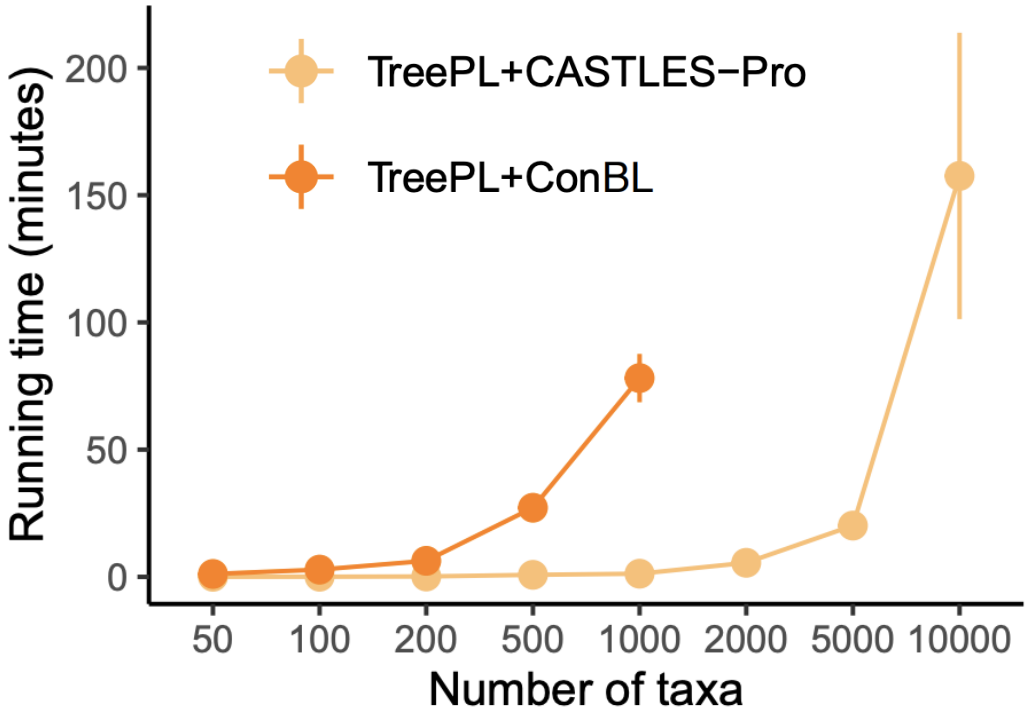
- Least-square dating (LSD)
- wLogDate
- treePL
- MD-Cat

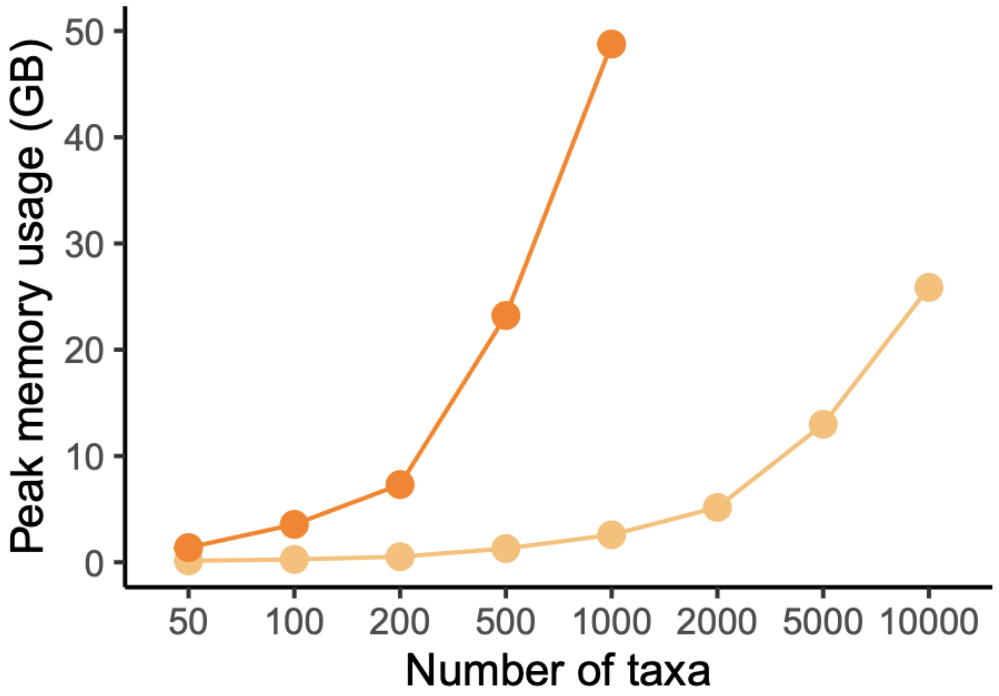


- CASTLES-Pro improves the accuracy of dating when ILS is at least moderate

Coalescent-based dating scales to tens of thousands of species and genes

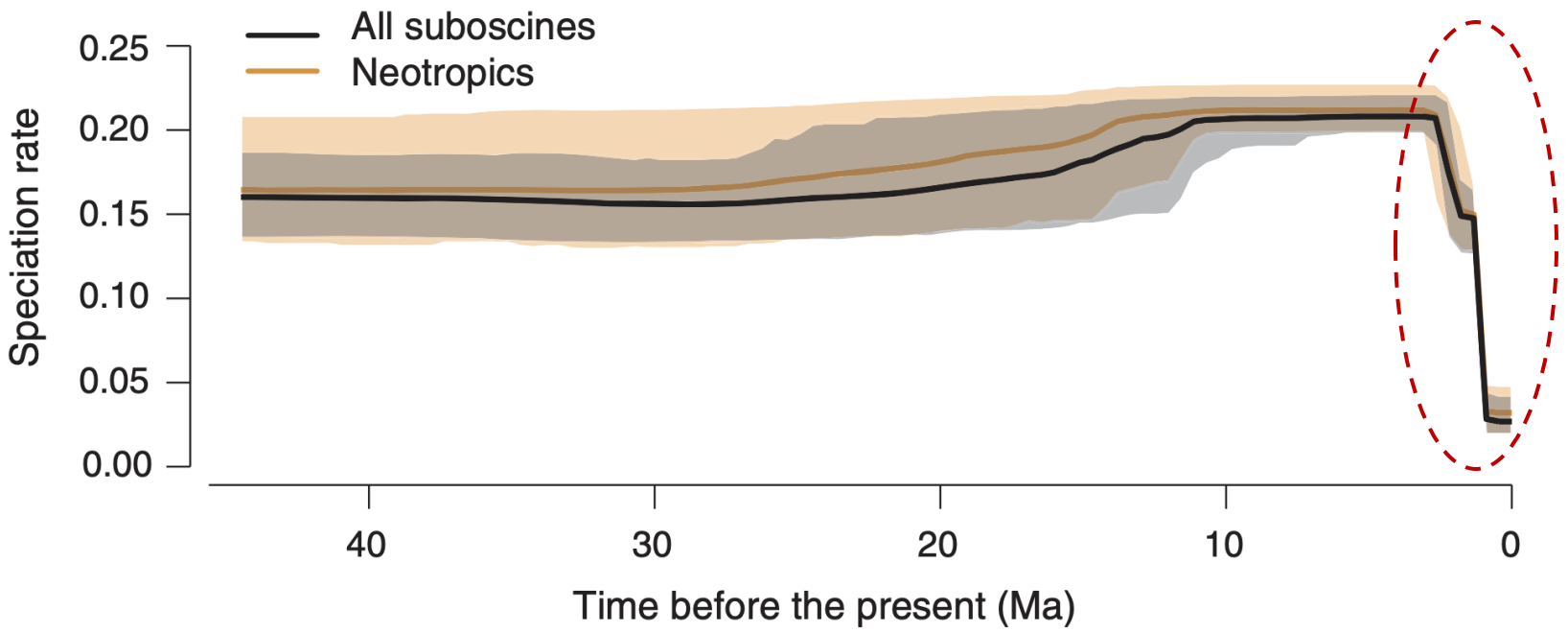
- Simulated dataset with 1000 genes, moderate ILS





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

Overestimation bias of concatenation can impact diversification analysis



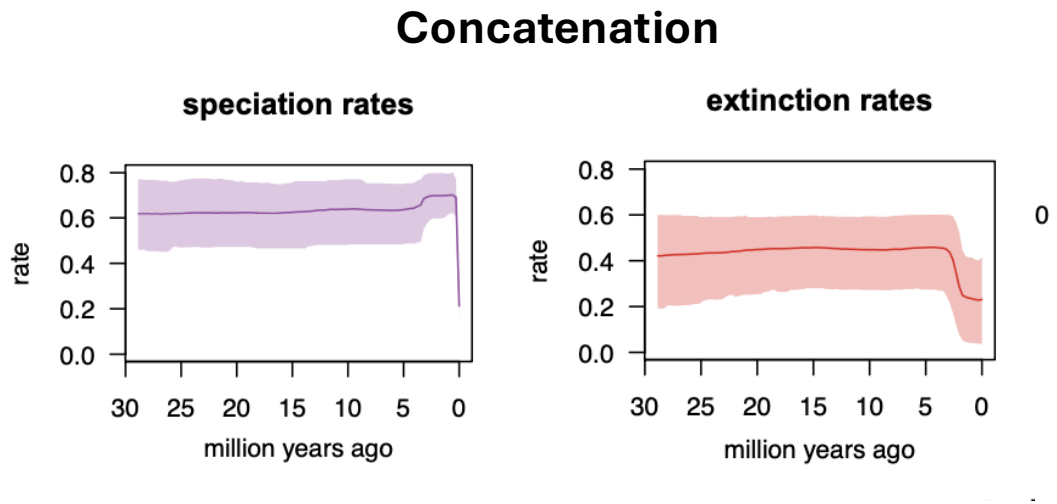
- Harvey et al., (2020) studied the diversification of suboscines, the largest tropical bird radiation.
- Diversification rates are stable over most of the history of the group aside from a drop within the past 2 Ma
- 1683 species and 2389 genes

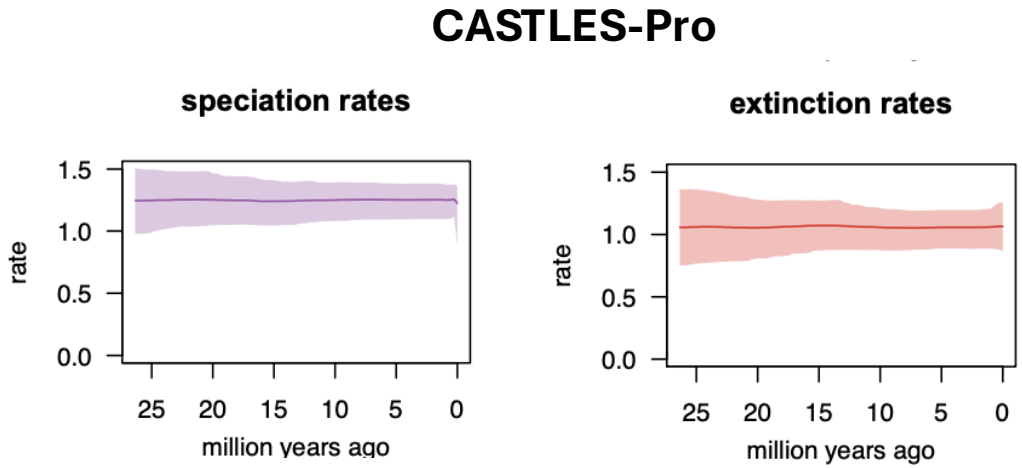
- The dramatic drop in diversification rates can be an artifact of concatenation bias, can we correct it with CASTLES-Pro?

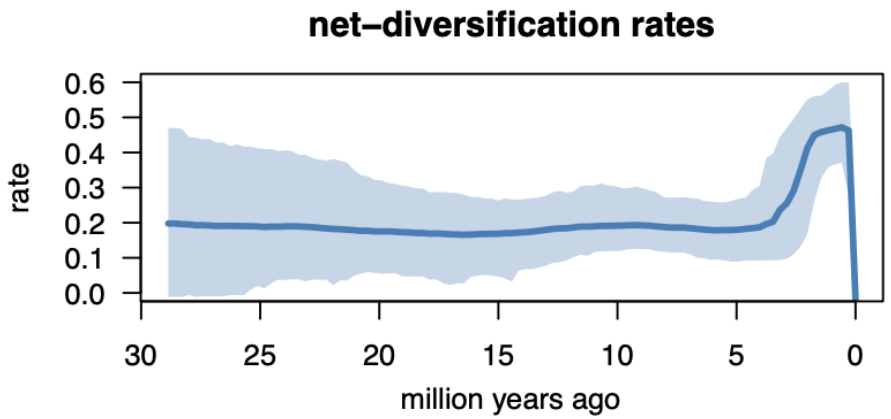
Image Credit: Harvey et al (2020). "The evolution of a tropical biodiversity hotspot." *Science*

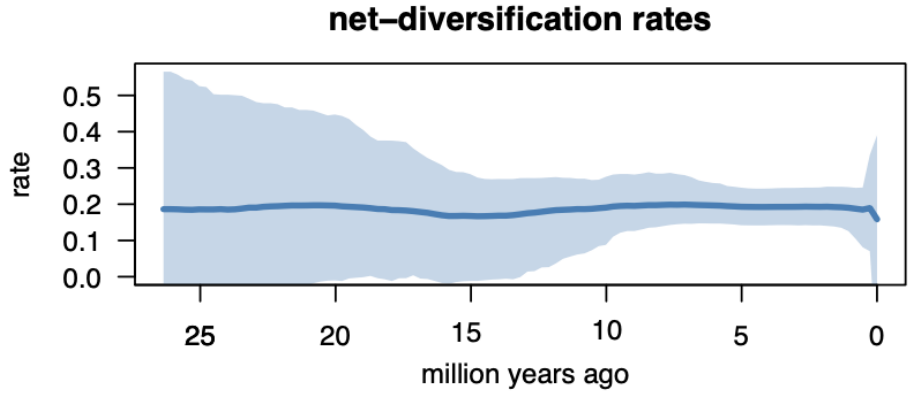
Dating with CASTLES-Pro branch lengths eliminates the shift in diversification rates on the suboscines dataset

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









Installation

- CASTLES-Pro is available as part of the ASTER package of tools.

- Option 1: Download executables from https://github.com/chaoszhang/ASTER/tree/master

- Option 2: Use conda: conda install aster

Execution (branch length estimation)

- Estimating the topology and branch lengths simultaneously

```
astral4 -i INPUT_FILE –o OUTPUT_FILE
```

- Estimating branch lengths on an existing topology

```
astral4 -C -c EXISTING_TREE -i INPUT_FILE –o OUTPUT_FILE
```

Execution (dating and diversification analysis)

- Estimating dates using a maximum likelihood method, e.g., LSD

```
lsd2 -i CASTLES_PRO_TREE -d CALIBRATIONS -s SEQ_LENGTH
```

- Diversification rates can be estimated using CoMet analysis with the TESS R library

```
tess.analysis(tree,
        empiricalHyperPriors = TRUE,
        MRCA = TRUE,
        estimateNumberMassExtinctions = FALSE,
        MAX_ITERATIONS = 1000000,
        dir = ”OUTPUT_TREE")
```

Summary

- CASTLES-Pro is a scalable method for estimating branch lengths of a species tree in the presence of ILS and GDL and it is also robust to HGT
- Its generally more accurate than traditional concatenation-based pipelines for branch length estimation
- It can be used in pipelines for dating species trees and diversification analysis.

Related Publications and Tools

- Tabatabaee, Y., Zhang, C., Arasti, S. and Mirarab, S. (2025). Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss. bioRxiv.
- Tabatabaee, Y., Claramunt, S., and Mirarab, S. (2025). Coalescent-based branch length estimation improves dating of species trees. bioRxiv.
- Zhang, C., Nielsen, R., Mirarab, S. (2025). ASTER: A Package for Large-scale Phylogenomic Reconstructions, Molecular Biology and Evolution, 2025, msaf172
- 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., Warnow, T., and Mirarab, S. (2023). Phylogenomic branch length estimation using quartets. *Bioinformatics, 39(Supplement 1):*i185–i193.

- CASTLES-Pro (implemented in ASTER): https://github.com/chaoszhang/ASTER/
- Least-square-dating: https://github.com/tothuhien/lsd2
- Newick Utilities: https://gensoft.pasteur.fr/docs/newick-utils/1.6/nwutils_tutorial.pdf

Acknowledgements

Thank you!

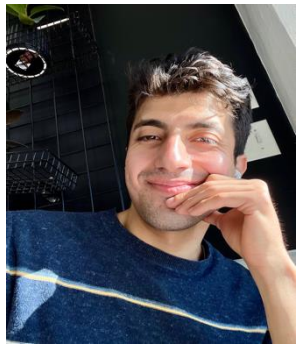

Tandy Warnow


Siavash Mirarab


Chao Zhang


Santiago Claramunt


Shayesteh Arasti


Puoya Tabaghi

Funding

NSF

NIH National Institutes of Health

COMPUTER SCIENCE

Computing Resources

- UIUC Campus Cluster
- Expanse at San Diego Supercomputing Center