Species tree estimation under multispecies coalescent

Ziheng Yang

Department of Genetics, Evolution, and Environment

University College London

Outline

- Species tree estimation in the case of 3 and 4 species
- The anomaly zone
- MCMC moves that change the species tree
- Simulation and empirical results

MSC, 3 species, 3 sequences

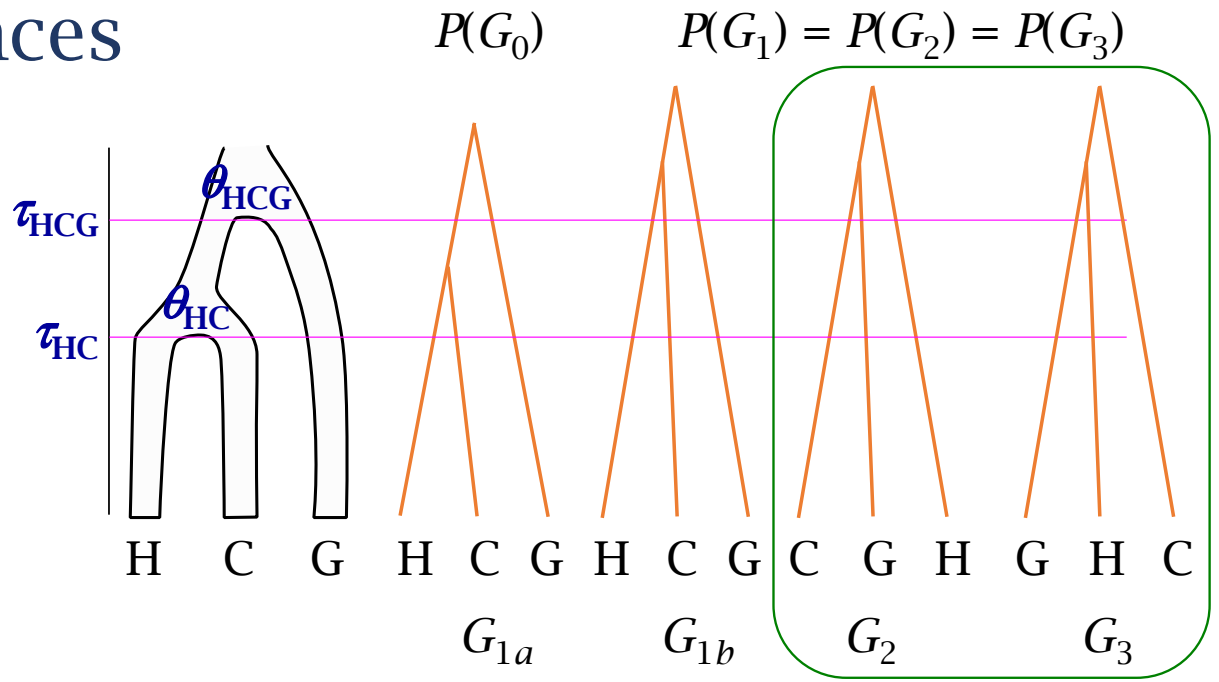
Parameters:

Speciation times:

τ_{HC}, τ_{HCG}

Population sizes:

θ_{HC}, θ_{HCG}



$$P_{\text{mismatch}} = P(G_2) + P(G_3) = \tfrac{2}{3}\left[1 - P(G_0)\right] = \tfrac{2}{3}\mathrm{e}^{-2(\tau_{HCG} - \tau_{HC})/\theta_{HC}}$$

For human-chimpanzee-gorilla, P_{mismatch} = 30%.

Hudson, R. R. 1983. Evolution **37**:203-217.
Chen & Li 2001. Am. J. Hum. Genet. **68**:444-456

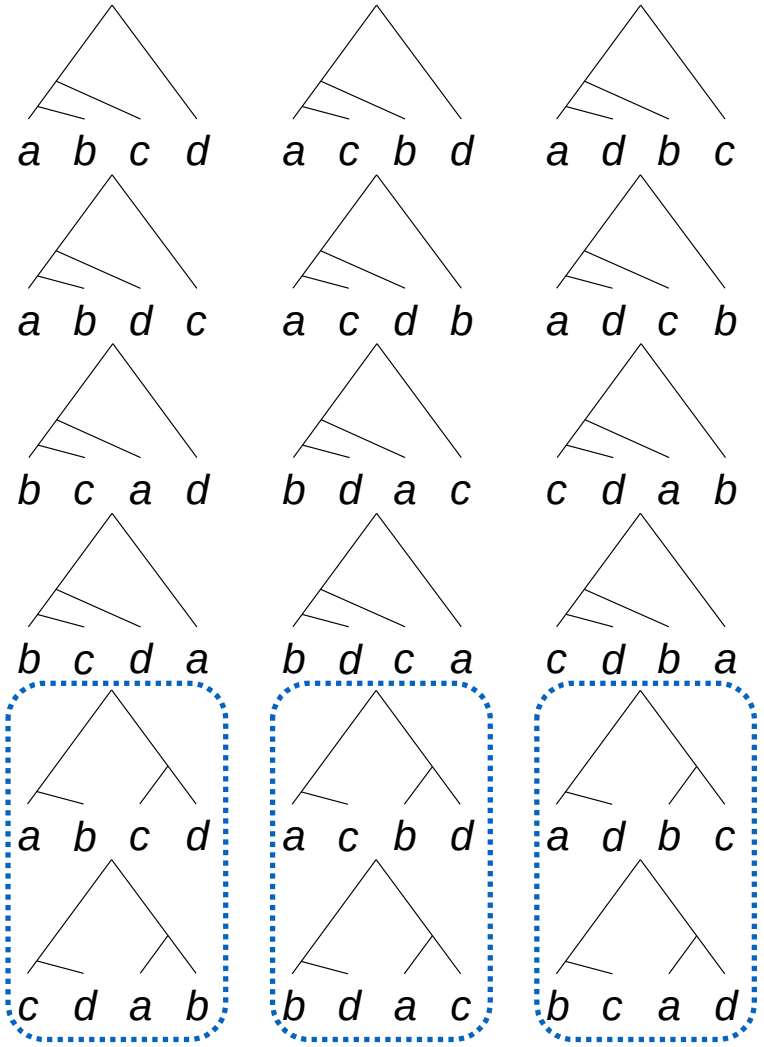
MSC: 4 sequences from 4 species

The probability for each of the 15 (rooted) gene trees is a sum over the compatible labelled histories.

$\tau_{ABCD} > \tau_{ABC} > \tau_{AB}$

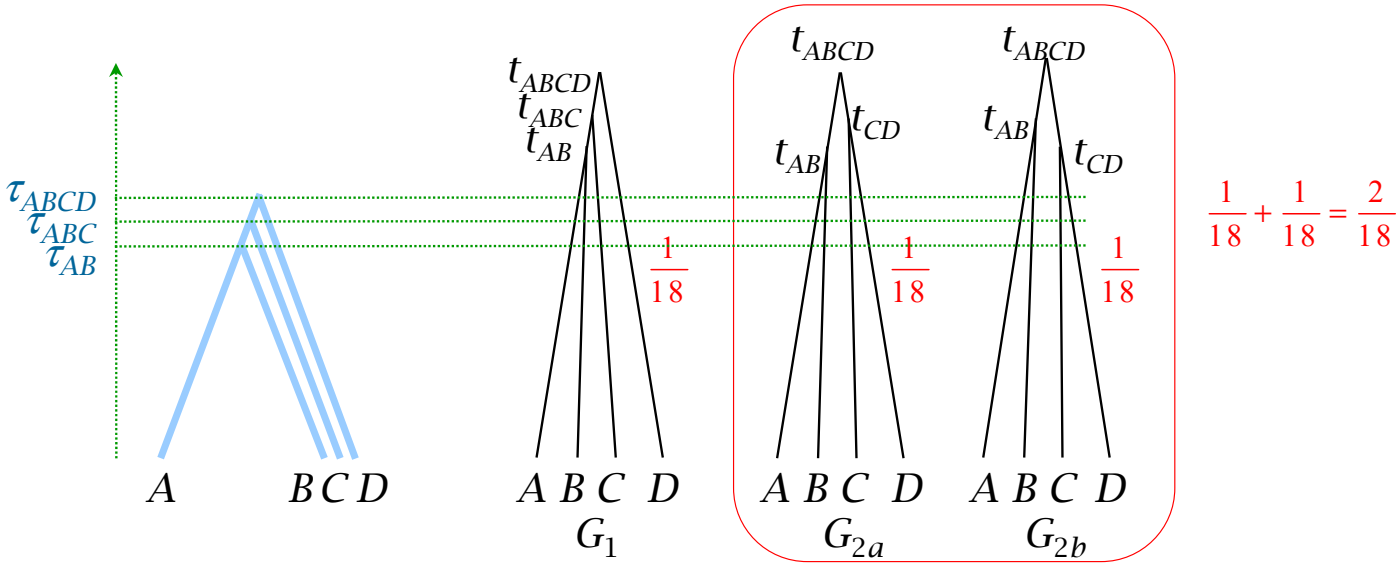
$\tau_{ABCD} \approx \tau_{ABC} \approx \tau_{AB}$

Species tree



MSC, 4 species, anomaly zone

For species tree and parameters (τs & θs) in the anomaly zone, the most probable gene tree has a different topology from the species tree.



Species tree Three labelled histories

- For 4 seqs, there are 15 rooted tree topologies (3 balanced & 12 unbalanced) and 18 labelled histories.
- If $\tau_{ABCD} = \tau_{ABC} = \tau_{AB}$, $f(G_1) = f(G_{2a}) = f(G_{2b}) = 1/18$.
- If the τs are close, it will be possible for $f(G_1) < f(G_{2a}) + f(G_{2b})$.

The anomaly zone

- If the most probable gene tree differs from the species tree, the species tree is said to be in the **anomaly zone**.
- If the species tree is in the anomaly zone, the method of using the most common gene tree as the species tree estimate (called the *majority-vote* method) will be inconsistent (misleading).
- Anomaly zone does not exist for 3 species.
  It exists for the unbalanced species tree of 4 species, and exists for any species tree of ≥5 species.

Degnan JH, Salter LA (2005 Evolution 59:24-37)
Degnan JH, Rosenberg NA (2006 PLoS Genet. 2:e68)

MCMC samples from the posterior:
$f(S, \{\tau s, \theta s\}, \{G_i, t_i\} \mid \text{Data})$

1. Initialize S, $\{\theta s, \tau s\}$, $\{G_i, \boldsymbol{t}_i\}$.
2. Iterate
   - Change parameters (θs, τs in the model).
   - Change gene trees $\{G_i, \boldsymbol{t}_i\}$.
   - Change species tree S (by NNI, SPR, NodeSlider).
   - Save on the disk every k iterations.

S: species tree
$\{\theta s, \tau s\}$: parameters in the MSC
$\{G_i, \boldsymbol{t}_i\}$: gene trees and ages

The MCMC algorithm visits the species trees according to their posterior probabilities.

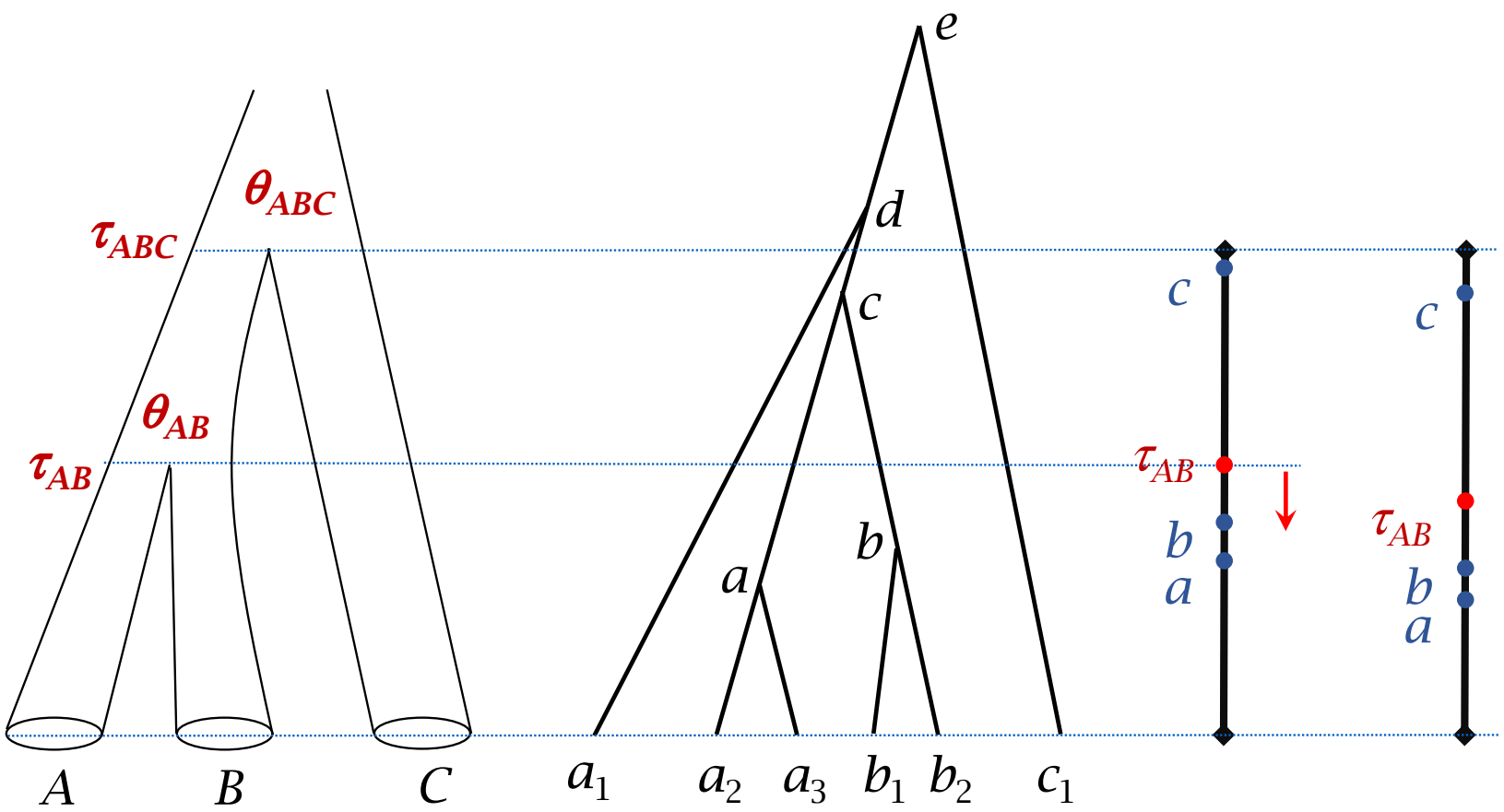
The constraint that gene trees fit inside species tree is a major issue when we design MCMC moves to change the species tree.

(a) Rubber-bound algorithm to change the species divergence time τ in the multispecies coalescent model in MCMC (Rannala & Yang 2003 Genetics)

(b) Rubber-bound with proportional scaling to change the species delimitation model in rjMCMC (Rannala & Yang 2013 Genetics)

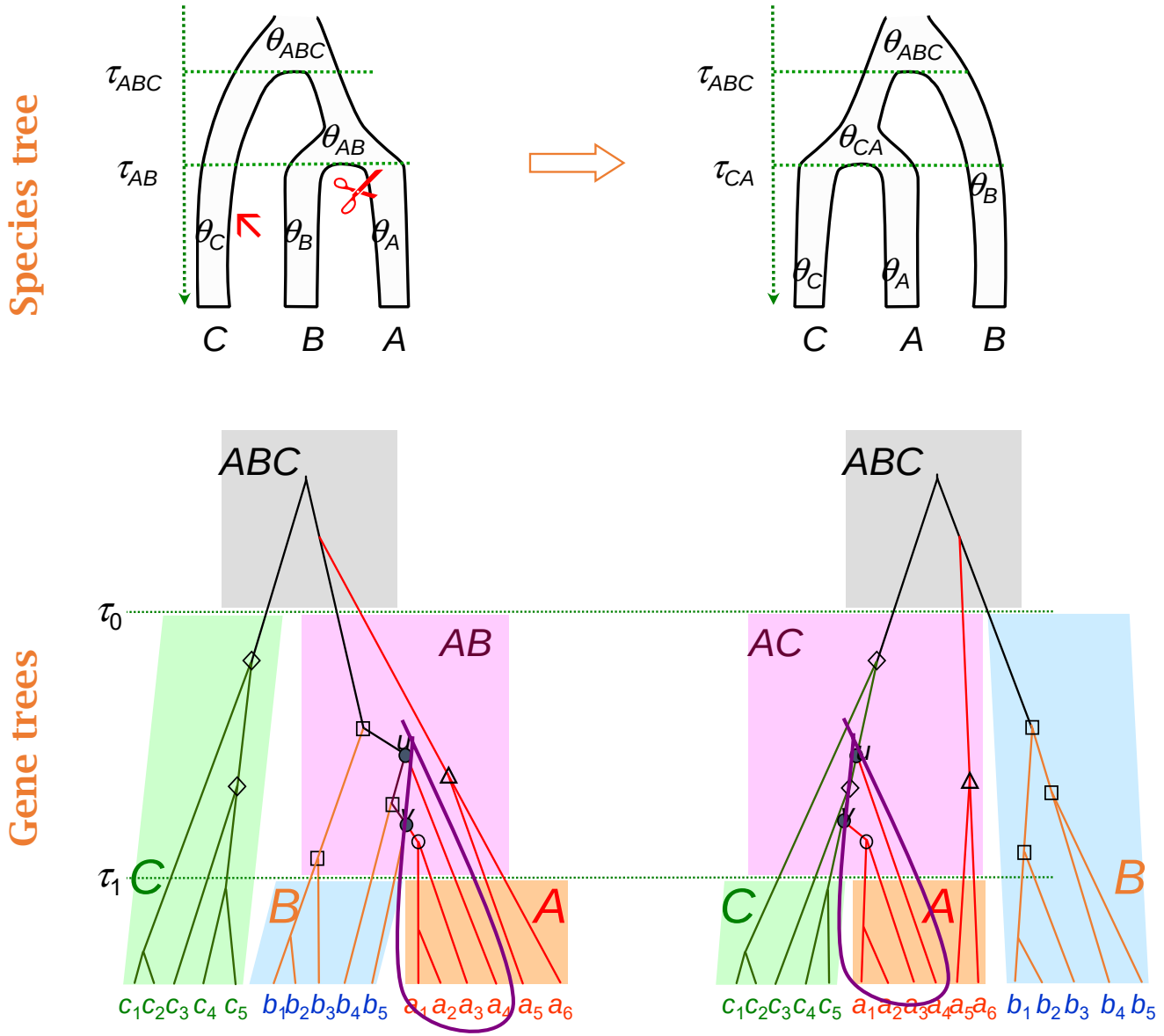
(c) NNI, SPR, & NodeSlider to change the species tree in transmodel MCMC (Yang & Rannala 2014 MBE; Rannala & Yang 2017 Syst Biol)

Rubber-band algorithm to change τ



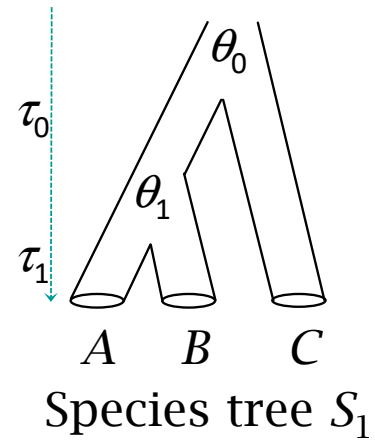
(Rannala & Yang, 2003 Genetics)

Species tree change by NNI



(Yang & Rannala 2014 MBE)

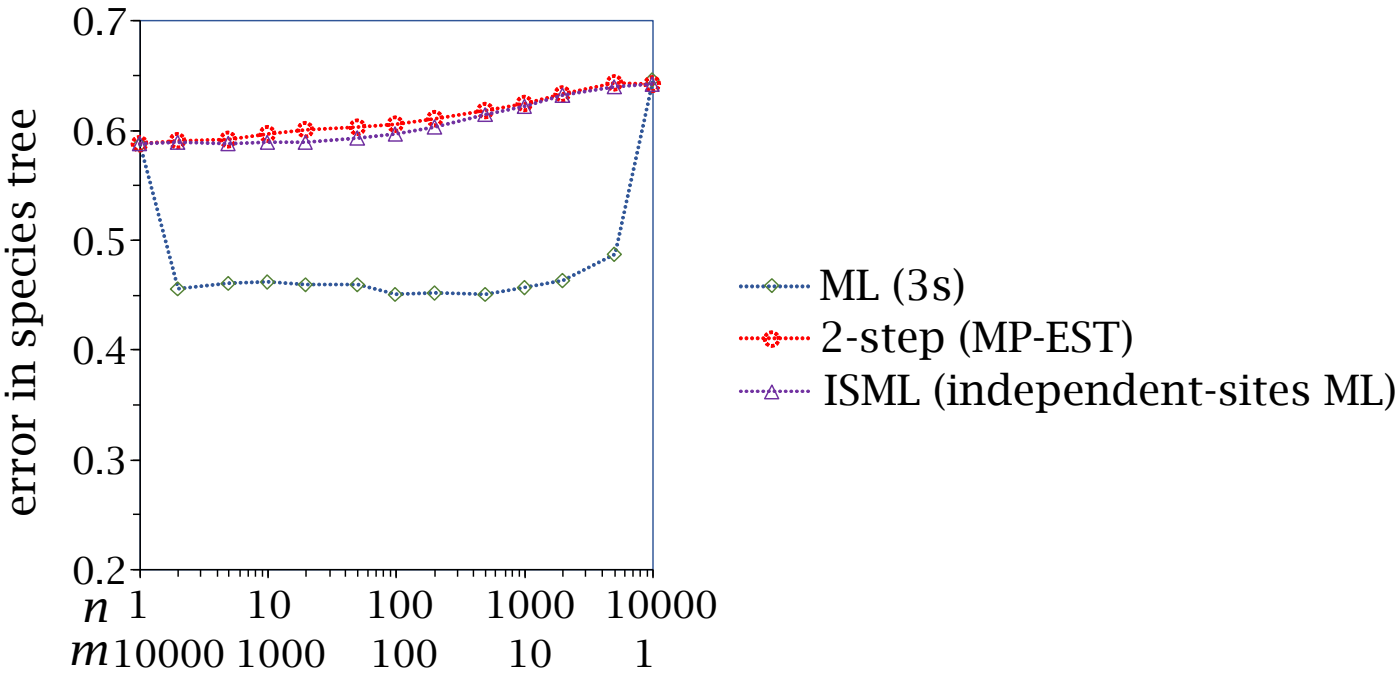
Species tree estimation in the case of 3 species



Species tree S_1

$\tau_0 = 0.02$, $\tau_1 = 0.019$,
$\theta_0 = 0.01$, $\theta_1 = 0.05$

(c) $nm = 10^4$ fixed, n sites m loci



There is more information in 2 genes each of 5000 sites than in 10000 independent sites.

Zhu T, Yang Z. 2021. Complexity of the simplest species tree problem. Mol Biol Evol. 10.1093/molbev/msab009

Full likelihood methods are much more efficient when the species tree is hard.

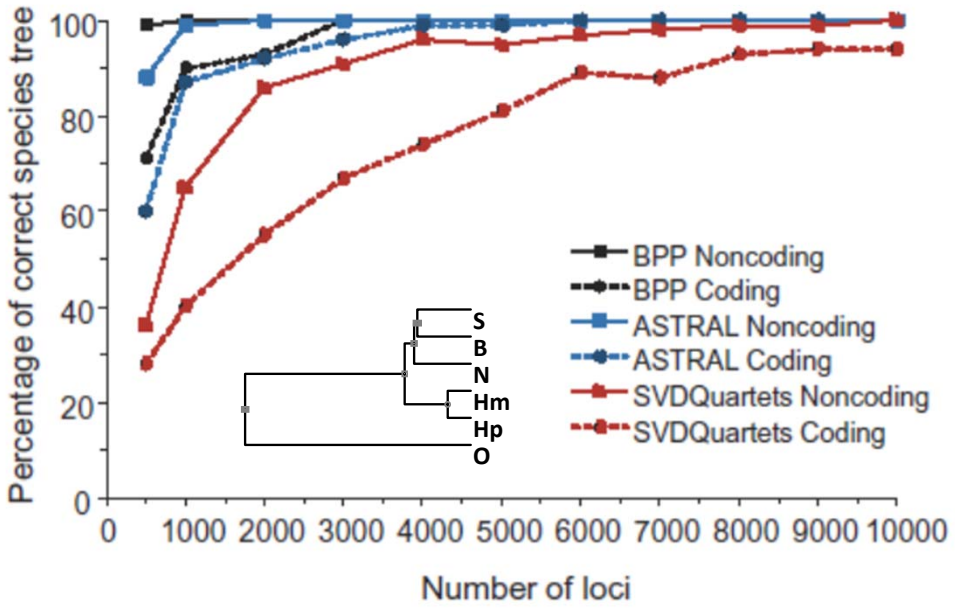


FIG. 6. The percentage of correct species trees in simulated data sets

(Shi & Yang 2018. Mol Biol Evol 35:159-179)

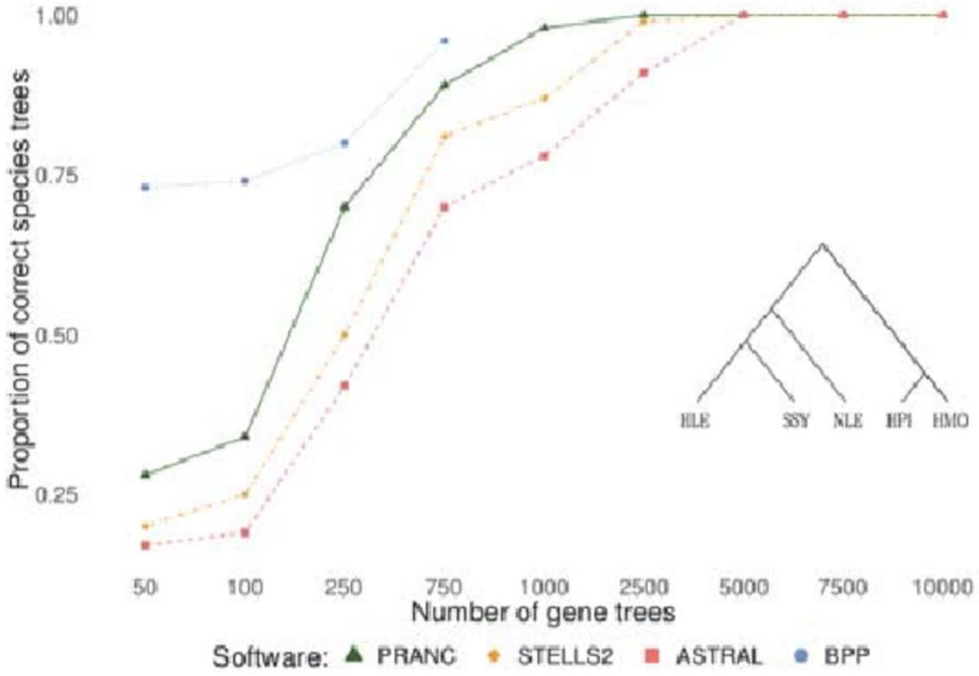


Fig. 1. The proportion of correct species trees in a gibbon dataset obtained by four different

(Kim A, Degnan J. 2020. *Bioinformatics* 10.1093/bioinformatics/btaa605)

Summary

- MSC (with and without introgression) is a powerful framework for analysis of genomic sequence data from closely related species.
- Full likelihood methods account for genealogical variations across the genome (among loci), as well as phylogenetic uncertainties in gene trees.
- In challenging species tree problems, summary methods have far poorer statistical properties but they run much faster.
- Deep coalescence and introgression are major challenges to species tree inference for shallow phylogenies. Deep trees are even harder.

References

Flouri T, Jiao X, Rannala B, Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol Biol Evol 35:2585-2593.

Flouri T, Rannala B, Yang Z. 2020. A tutorial on the use of BPP for species tree estimation and species delimitation. Pp. 5.6.1-16 *in Scornavacca C, Delsuc F, and Galtier N, eds. Phylogenetics in the Genomic Era.*

Flouri T, Jiao X, Rannala B, Yang Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol Biol Evol 37:1211-1223.

Jiao X, Yang Z. 2021. Defining species when there is gene flow. Syst Biol 70:108–119.

Jiao X, Flouri T, Rannala B, Yang Z. 2020. The impact of cross-species gene flow on species tree estimation. Syst Biol 69:830-847.

Thawornwattana Y, Dalquen DA, Yang Z. 2018. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the Anopheles gambiae species complex. Mol Biol Evol 35:2512-2527.

Zhu T, Yang Z. 2021. Complexity of the simplest species tree problem. Mol Biol Evol. 10.1093/molbev/msab009