

# 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

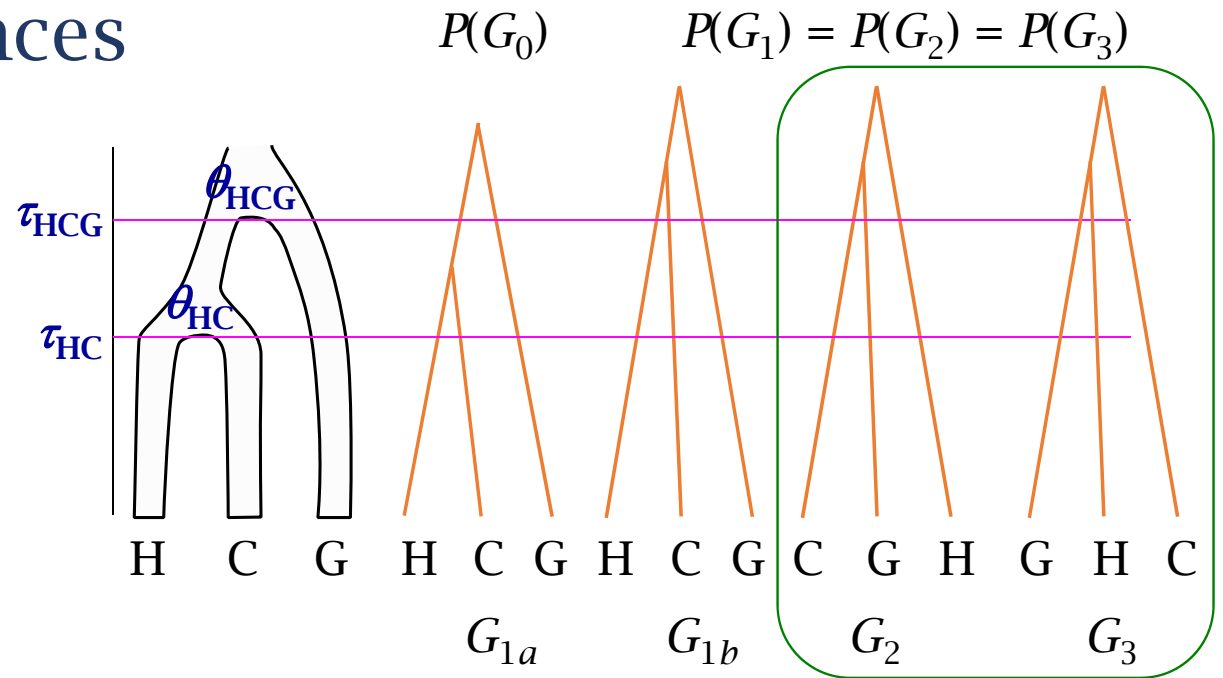
## Parameters:

### Speciation times:

$$\tau_{\text{HC}}, \tau_{\text{HCG}}$$

### Population sizes:

$$\theta_{\text{HC}}, \theta_{\text{HCG}}$$



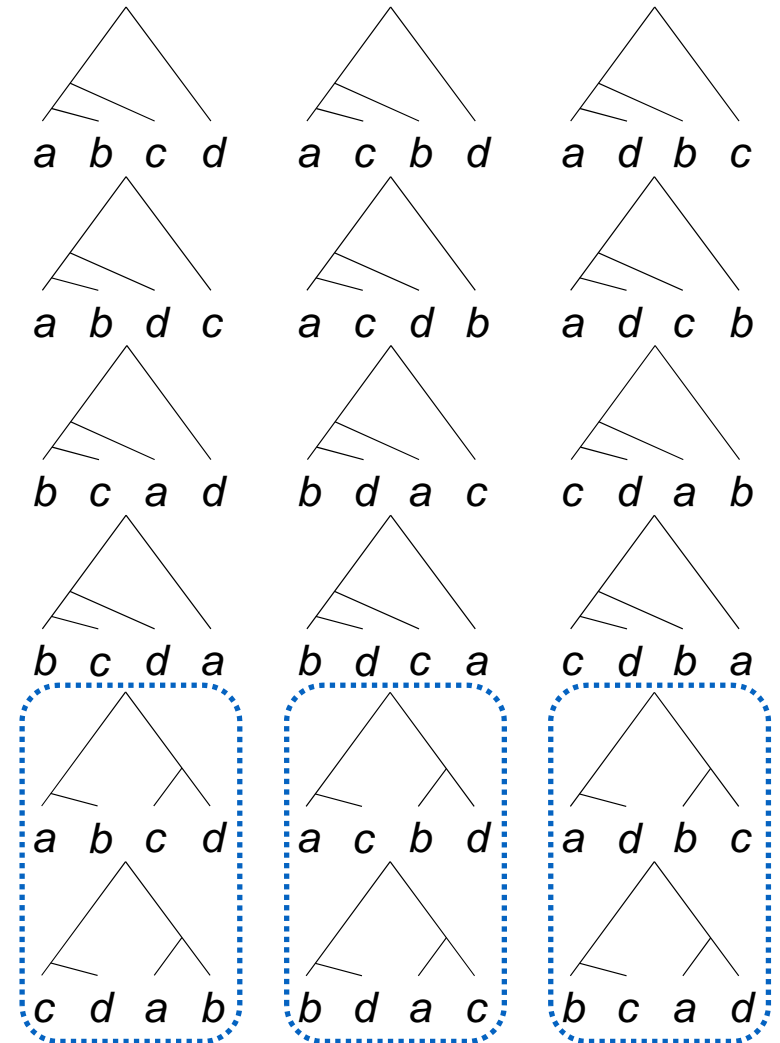
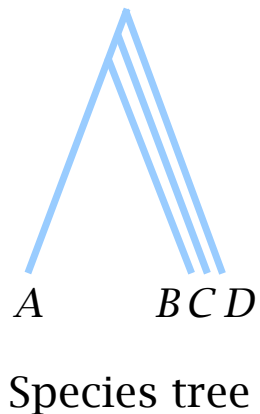
$$P_{\text{mismatch}} = P(G_2) + P(G_3) = \frac{2}{3} [1 - P(G_0)] = \frac{2}{3} e^{-2(\tau_{\text{HCG}} - \tau_{\text{HC}})/\theta_{\text{HC}}}$$

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

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

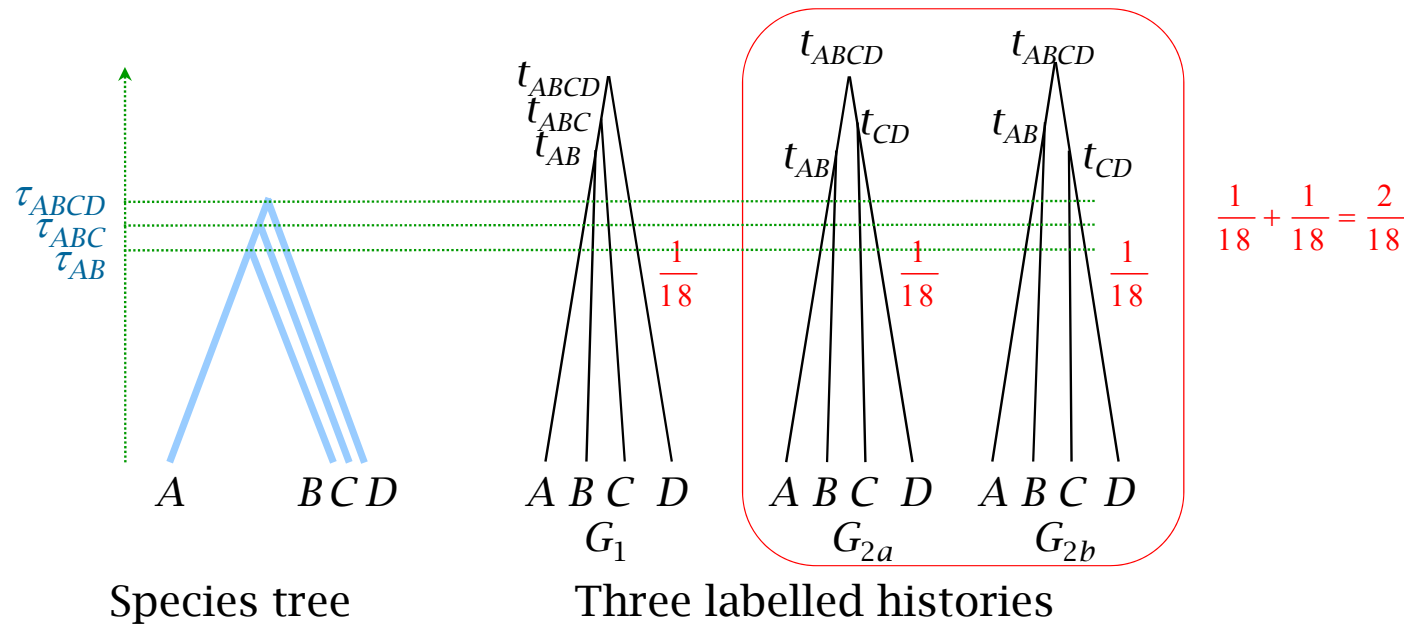
# MSC: 4 sequences from 4 species

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



# MSC, 4 species, anomaly zone

For species tree and parameters ( $\tau$ s &  $\theta$ s) in the **anomaly zone**, the most probable gene tree has a different topology from the species tree.



- 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  $\tau$ s 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  $\geq 5$  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  $S, \{\theta_S, \tau_S\}, \{G_i, t_i\}$ .
2. Iterate
  - Change parameters ( $\theta_S, \tau_S$  in the model).
  - Change gene trees  $\{G_i, t_i\}$ .
  - **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

$\{G_i, t_i\}$ : gene trees and ages

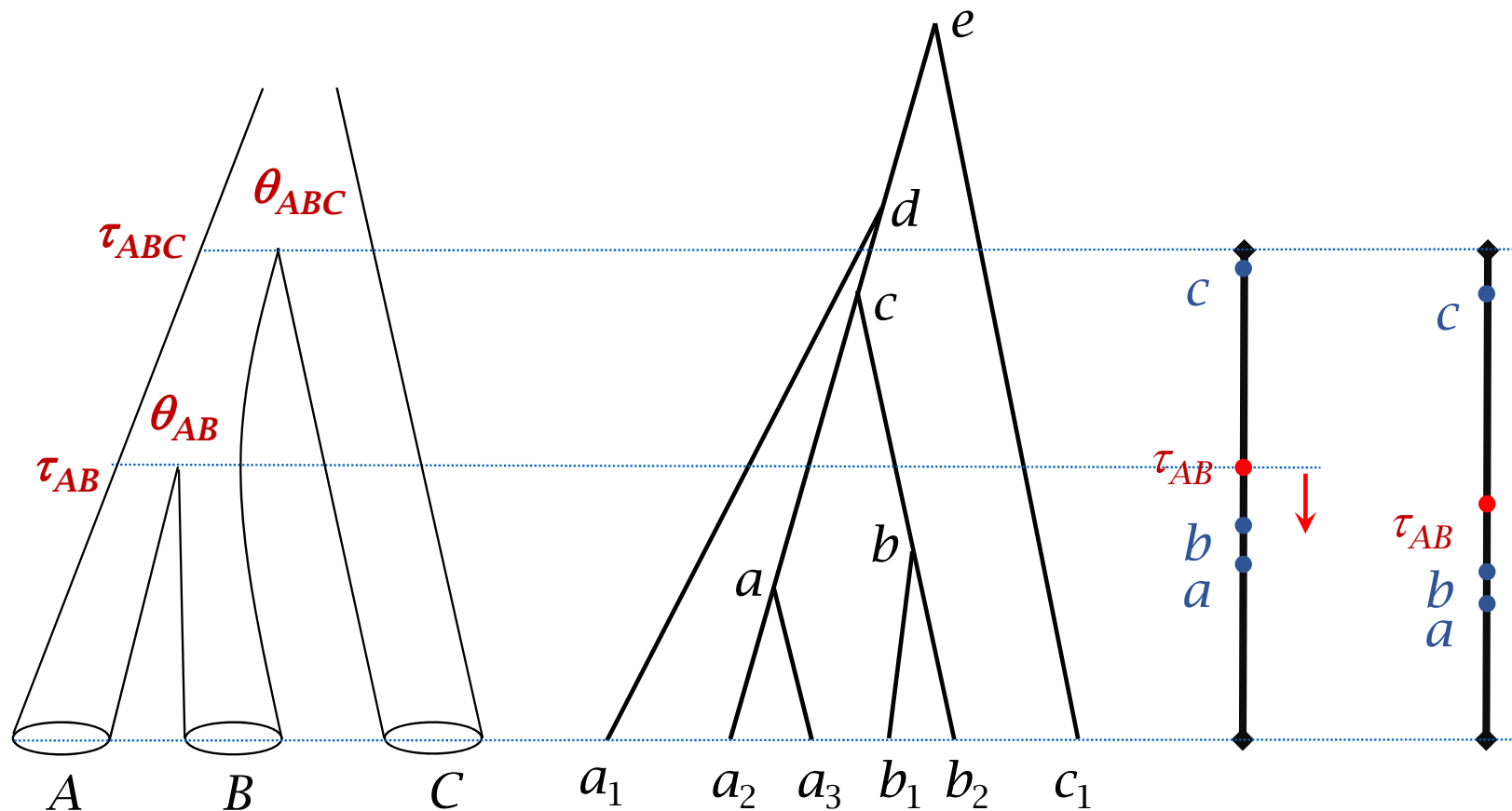
The MCMC algorithm visits the species trees according to their posterior probabilities.

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  $\tau$  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)



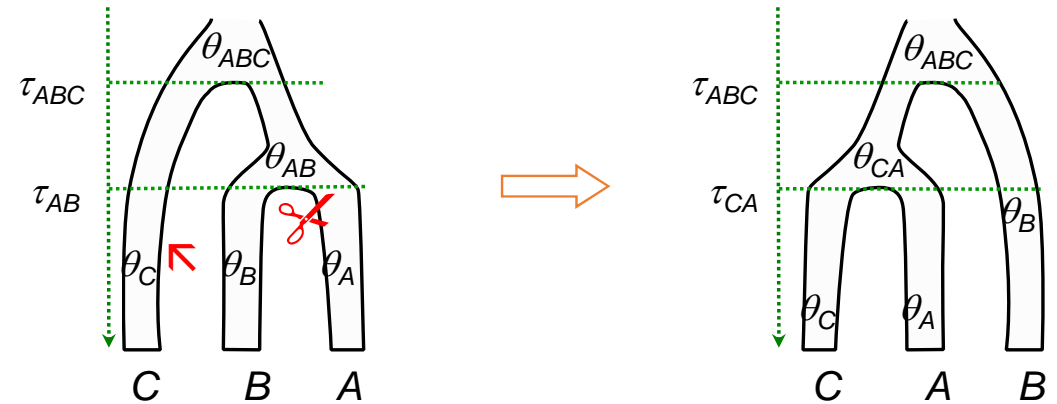
# Rubber-band algorithm to change $\tau$



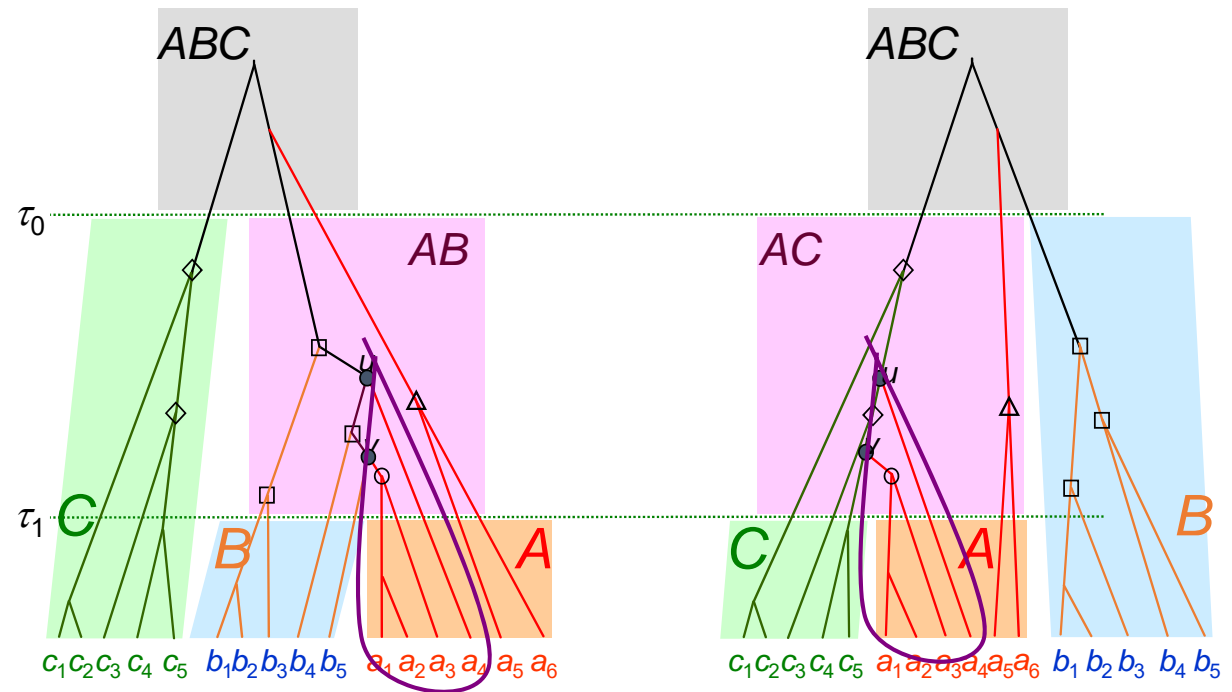
(Rannala & Yang, 2003 Genetics)

# Species tree change by NNI

Species tree



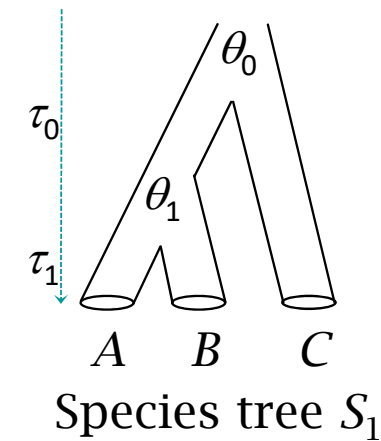
Gene trees



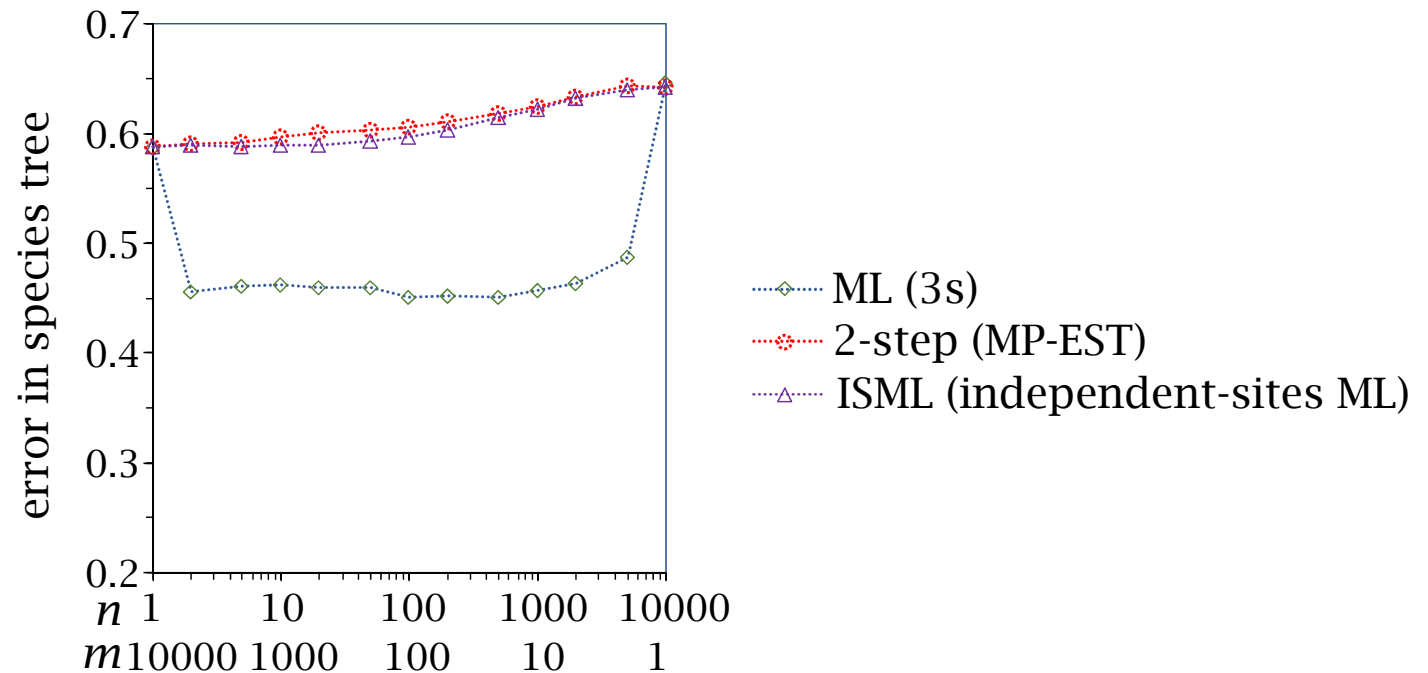
(Yang & Rannala 2014 MBE)

# Species tree estimation in the case of 3 species

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



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



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.

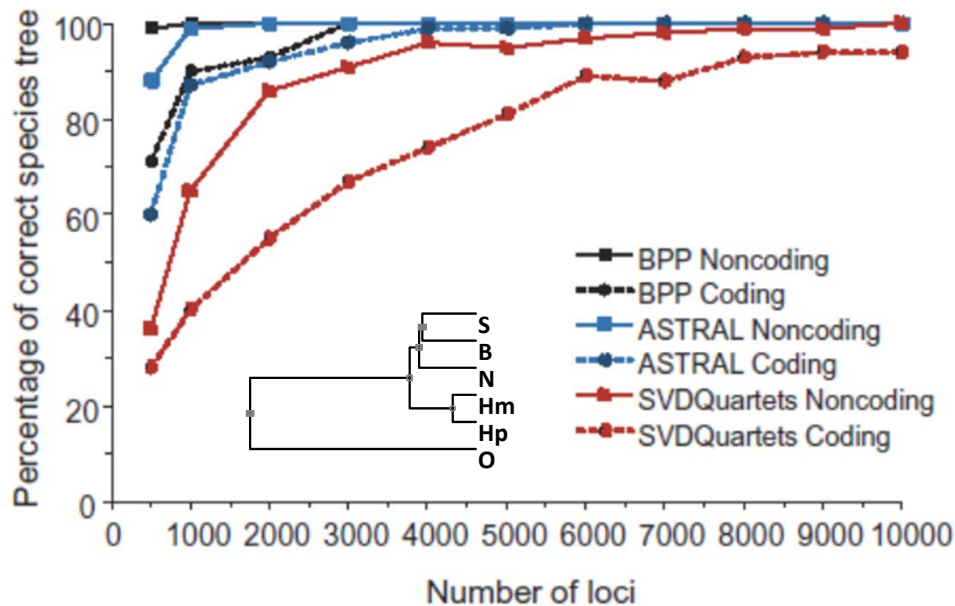


Fig. 6. The percentage of correct species trees in simulated data sets

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

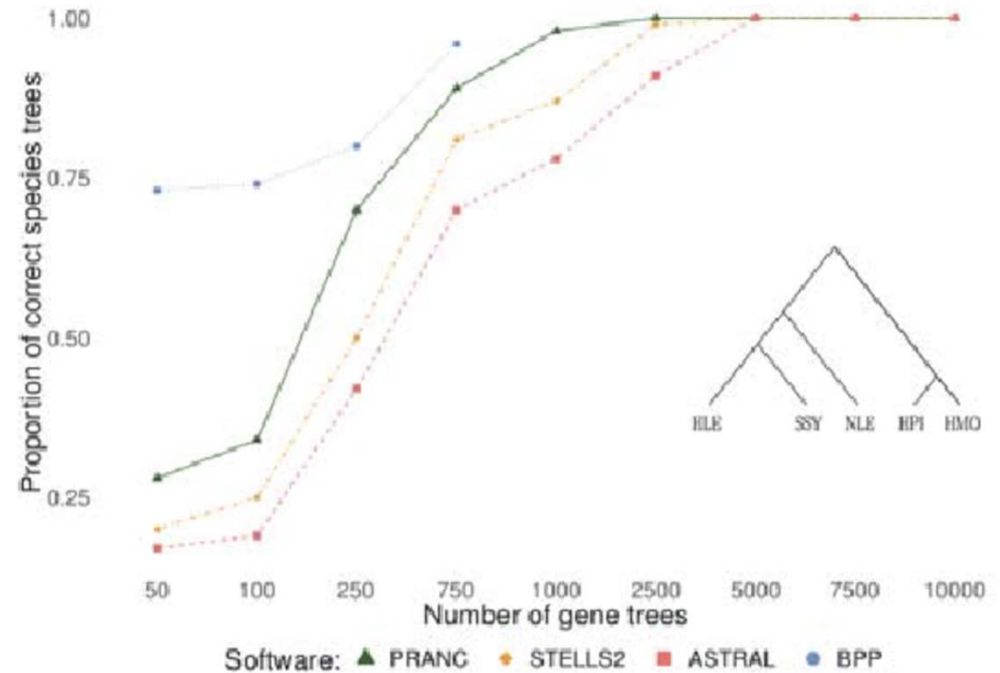
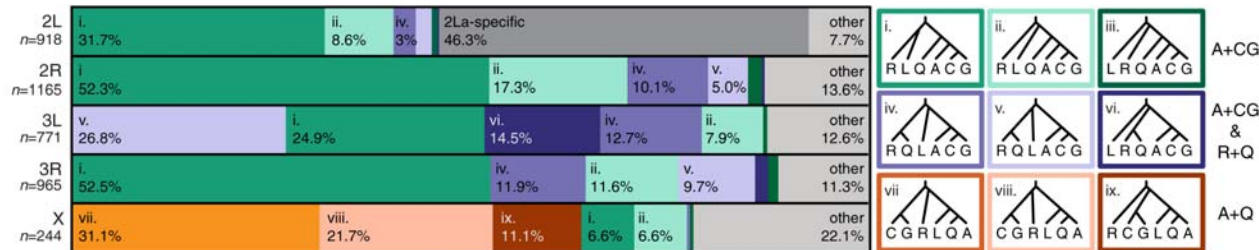


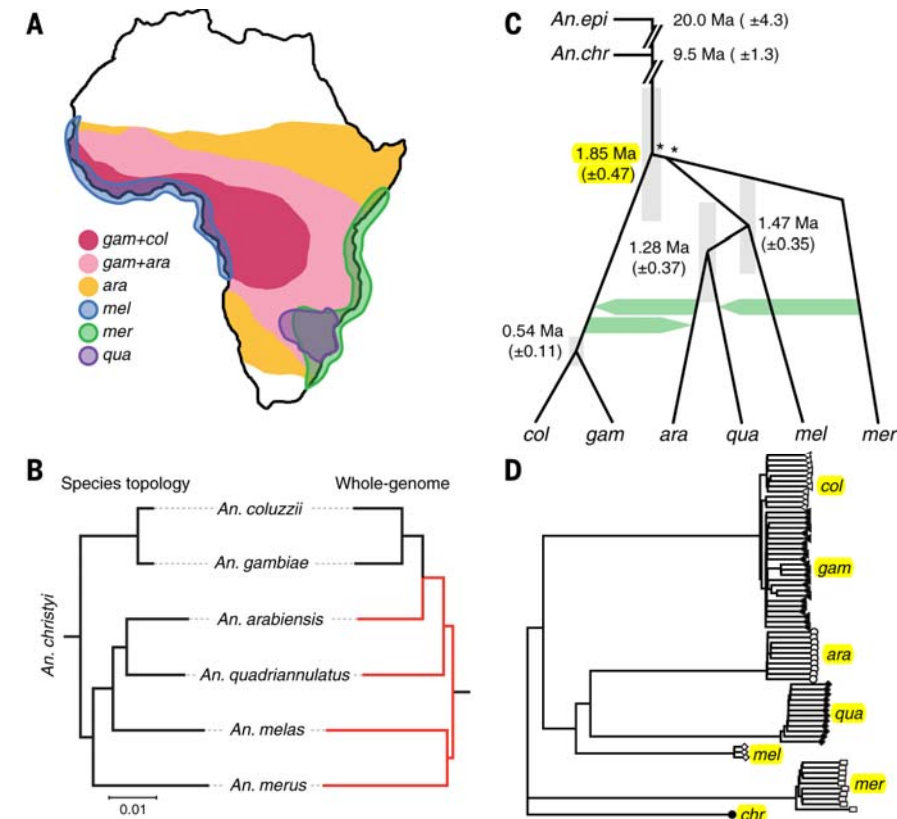
Fig. 1. The proportion of correct species trees in a gibbon dataset obtained by four different

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

# Species tree estimation for the *Anopheles gambiae* species complex



**Fig. 2. Phylogenies inferred from regions on the autosomes differ strongly from phylogenies inferred from regions on the X chromosome.** ML-rooted taxa (trees vii to ix), whereas the autosomes indicate that A and C+G are sister groups (trees i to vi). Large portions of the autosomes (particularly 3L and 3R)



**Fig. 1. Distribution and phylogenetic relationships of sequenced members of the *An. gambiae* complex.** (A) Schematic geographic distribution of *An. gambiae* (gam, formerly *An. gambiae* S form).

(Fontaine MC, *et al.* 2015. Science 347:1258524)

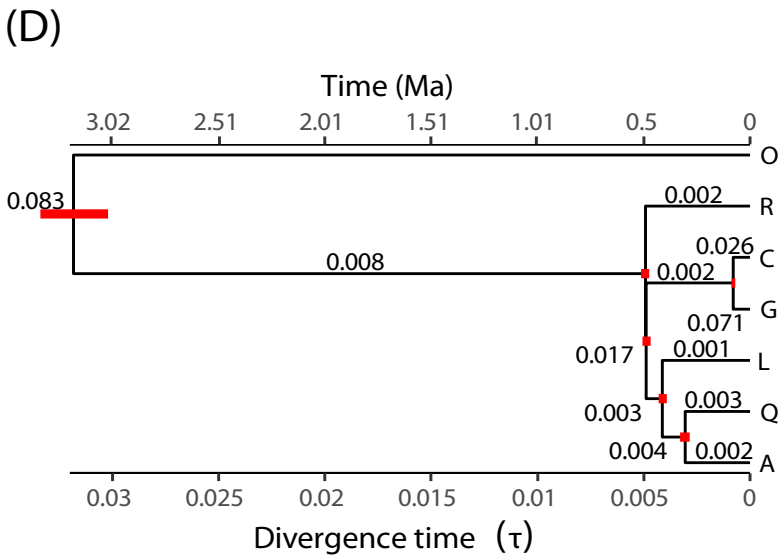
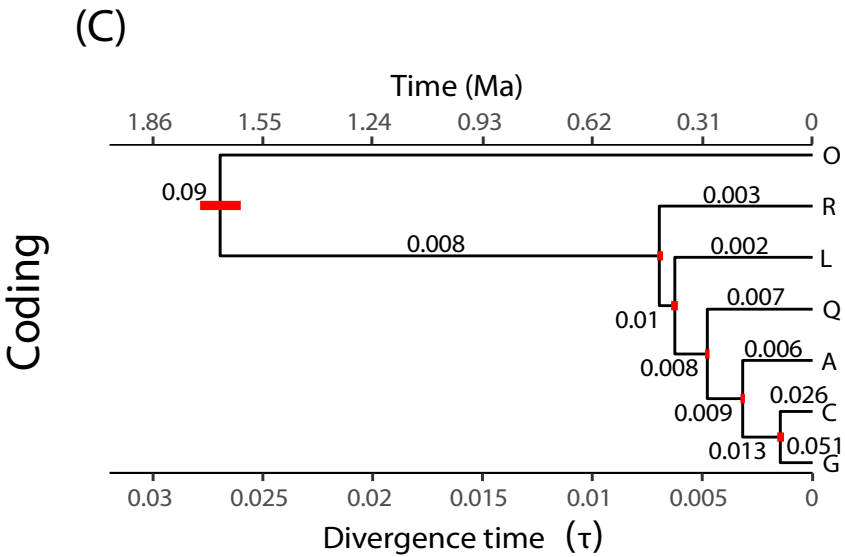
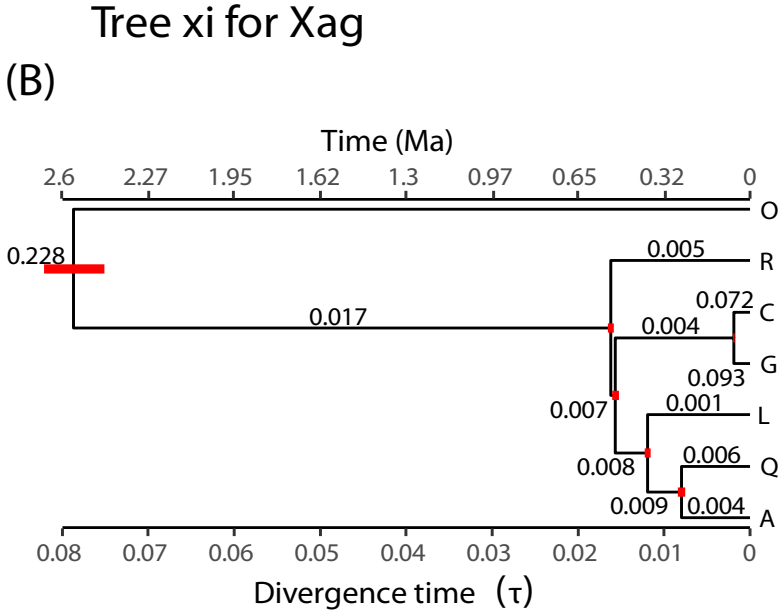
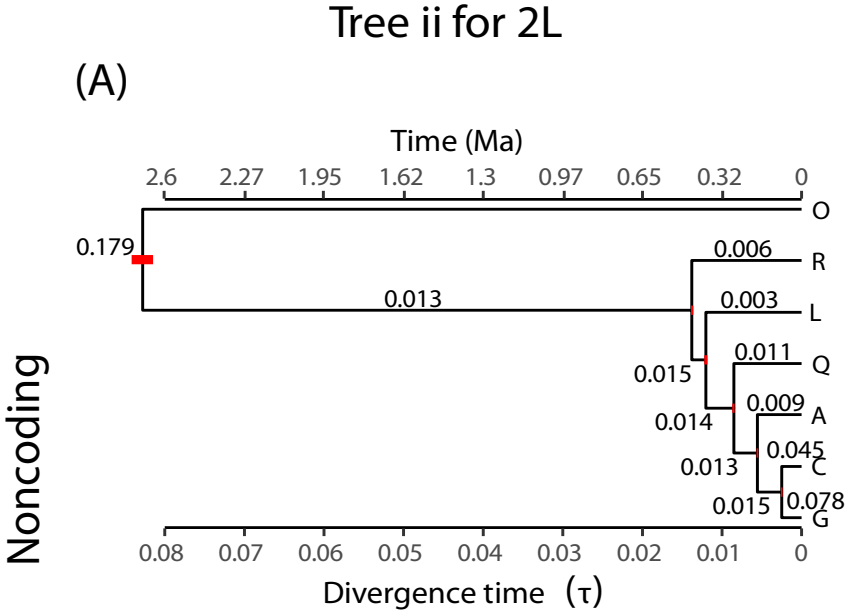
# BPP analysis of 100-loci blocks

Assumptions:

JC, clock,  
no gene flow



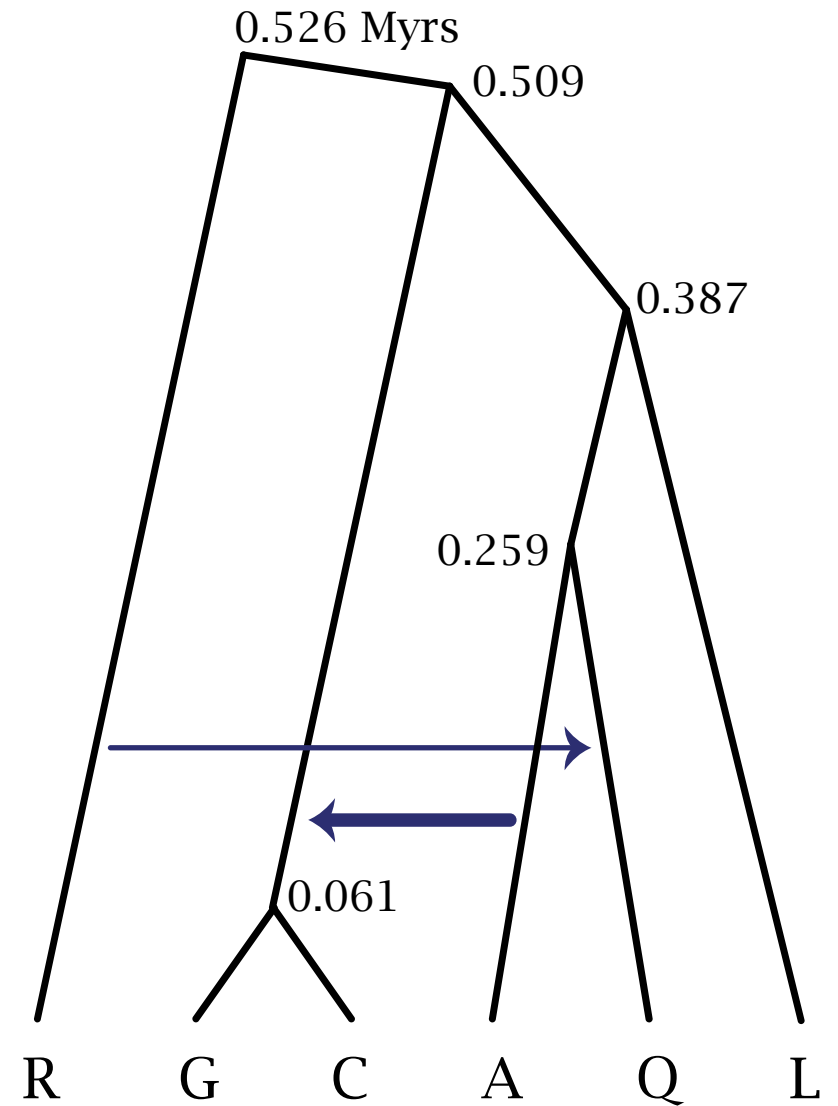
# BPP trees



# *A. Gambiae* phylogeny

Estimates of  $\tau_s$  are converted to divergence times using the *Drosophila* mutation rate of  $2.8 \times 10^{-9}$  /site/generation, and 11 generations/year.

Thawornwattana Y, Dalquen DA, Yang Z. 2018.  
Mol Biol Evol 35: 2512-2527.



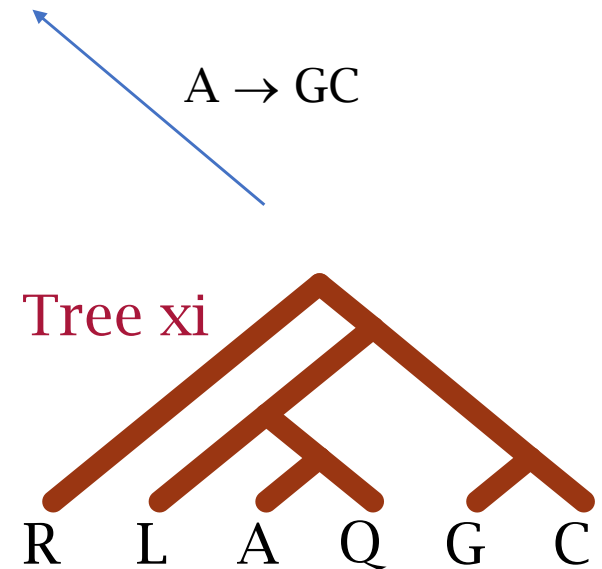


A  $\rightarrow$  GC introgression in tree xi leads to tree ii.



The X chromosome tree (**tree xi**) is the true species tree.

The autosomes tree (**tree ii**) is the result of **tree xi** and A  $\rightarrow$  GC gene flow.



Thawornwattana Y, et al. 2018. Mol Biol Evol 35:2512-2527.

# Comparison between Fontaine *et al.* and our study

Fontaine, <i>et al.</i> 2015	Thawornwattana et al. (2018)
Different genomic regions support different species relationships. Xag phylogeny reflects the true species relationship, while the autosome trees are results of introgression.	Yes.
Xag tree is tree ix.	tree xi
Autosomal tree is tree i.	tree ii
<ul style="list-style-type: none"><li>• Tree xi from the Xag is the most parsimonious <b>chromosomal inversion</b> phylogeny. It is consistent with tree ii for autosomes plus A→G introgression.</li><li>• Simulation shows that the trees inferred in Fontaine, <i>et al.</i> (2015) are artifactual. Our concatenation/RAxML results mimic their sliding-windows analysis.</li></ul>	

# 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