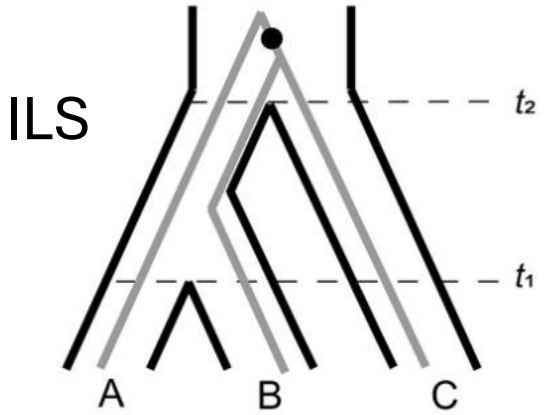
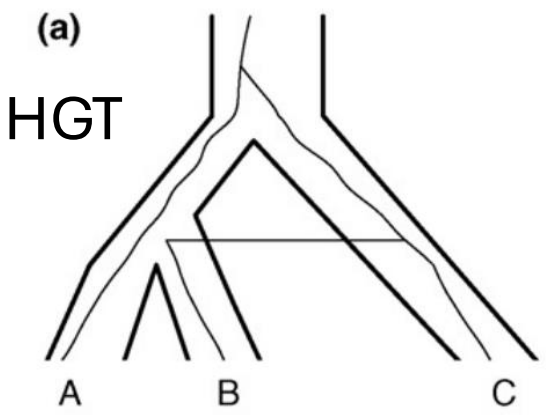
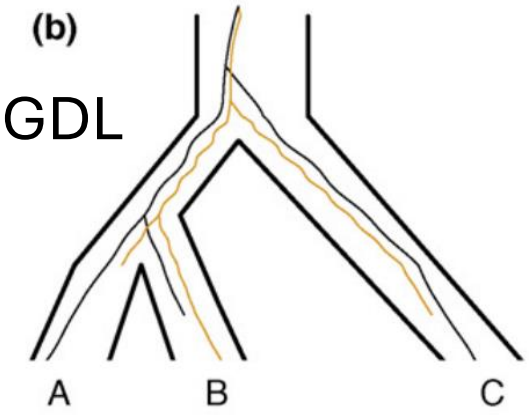
Species tree inference and the multispecies coalescent

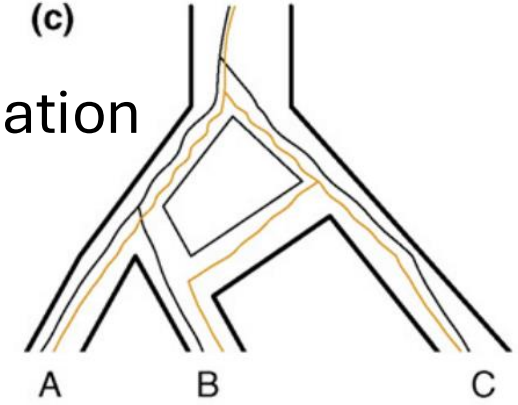
Chi Zhang

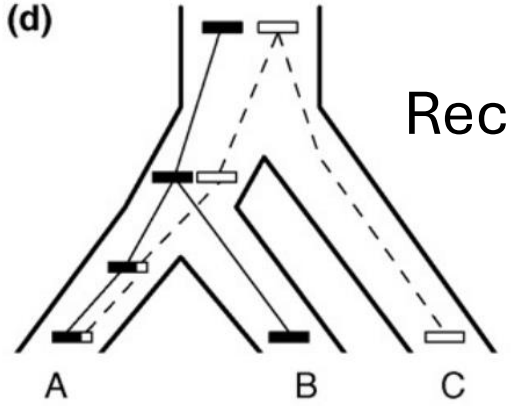
Gene tree–species tree discordance





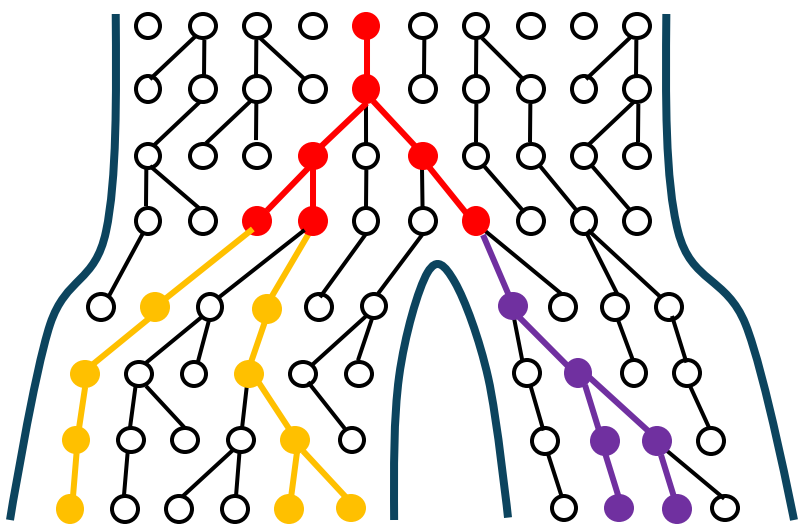


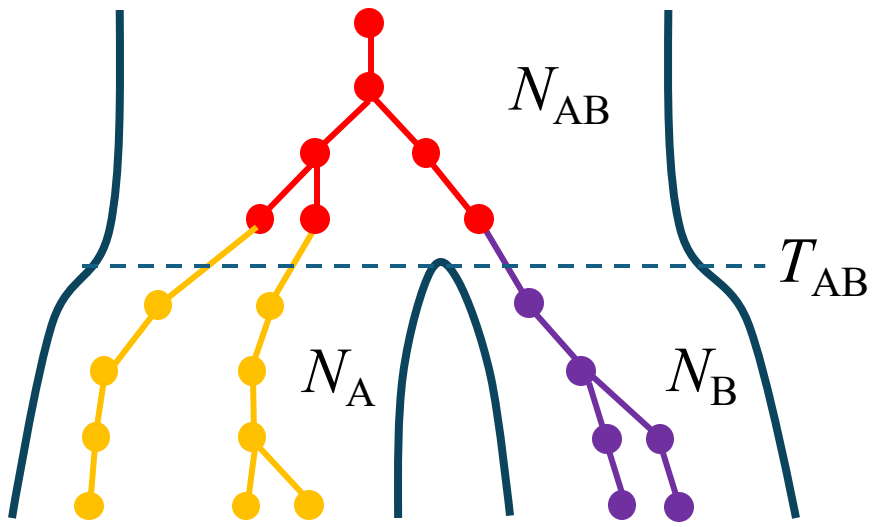




(Degnan & Rosenberg 2009)

Multispecies coalescent (MSC)



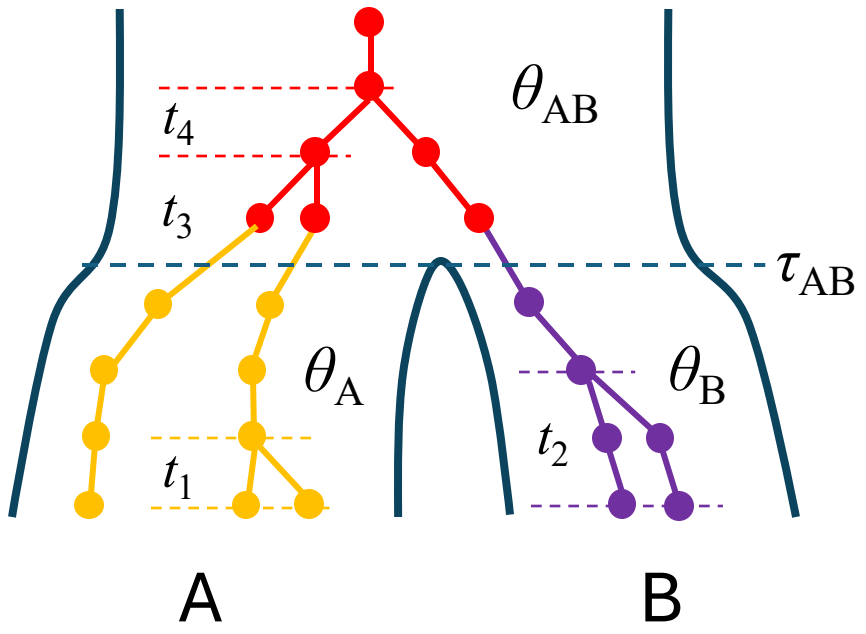


inter-specific coalescent (Takahata 1989)
censored coalescent (Rannala & Yang 2003)
multispecies coalescent (Liu et al. 2009)

Multispecies coalescent (MSC)

- Incomplete lineage sorting

$$f(G|\theta_A,\theta_B,\theta_{AB},\tau_{AB}) = \frac{2}{\theta_A}\mathrm{e}^{-\frac{6t_1}{\theta_A}} \times \mathrm{e}^{-\frac{2(\tau_{AB}-t_1)}{\theta_A}} \times \frac{2}{\theta_B}\mathrm{e}^{-\frac{2t_2}{\theta_B}} \times \frac{2}{\theta_{AB}}\mathrm{e}^{-\frac{6t_3}{\theta_{AB}}} \times \frac{2}{\theta_{AB}}\mathrm{e}^{-\frac{2t_4}{\theta_{AB}}}$$



- The multispecies coalescent (MSC) provides the probability distribution of a gene tree G given the species tree S

Multispecies coalescent (MSC)

- Complete isolation after speciation
  - Coalescent events happen in ancestral populations for lineages from different species
- Complete linkage with locus and free recombination among loci
  - Gene trees are independent among loci
- Gene trees are embedded in the species tree
  - Their distributions are given by the multispecies coalescent process

Implementations of MSC in *BEAST*

- *BEAST (Heled & Drummond 2010)
  - built-in functionality of BEAST2
- StarBEAST2 (Ogilvie et al. 2017)
  - population sizes can be integrated out analytically (Jones 2015)
  - relaxed molecular clock per species branch (instead of per gene branch)
  - more efficient proposals (coordinated operators, Rannala & Yang 2015)
- StarBEAST3 (Douglas et al. 2022)
  - more efficient proposals
  - parallelization (requires estimating population sizes)

StarBEAST3

$$f(S, \mathbf{G}, \Theta|D) \propto f(S|\Theta)f(\Theta)\prod_{i=1}^{k} f(g_i|S,\Theta)f(D_i|g_i,\Theta)$$

(Douglas et al. 2022)

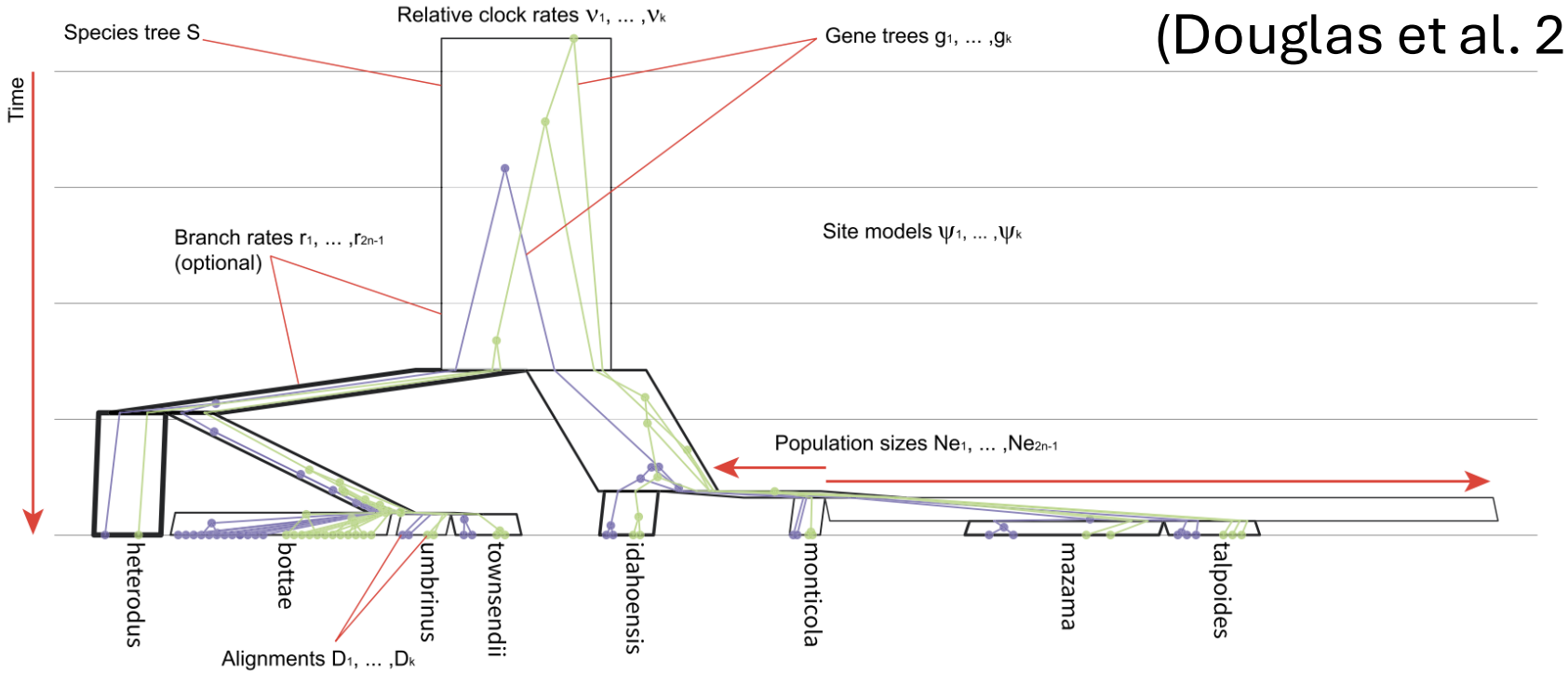
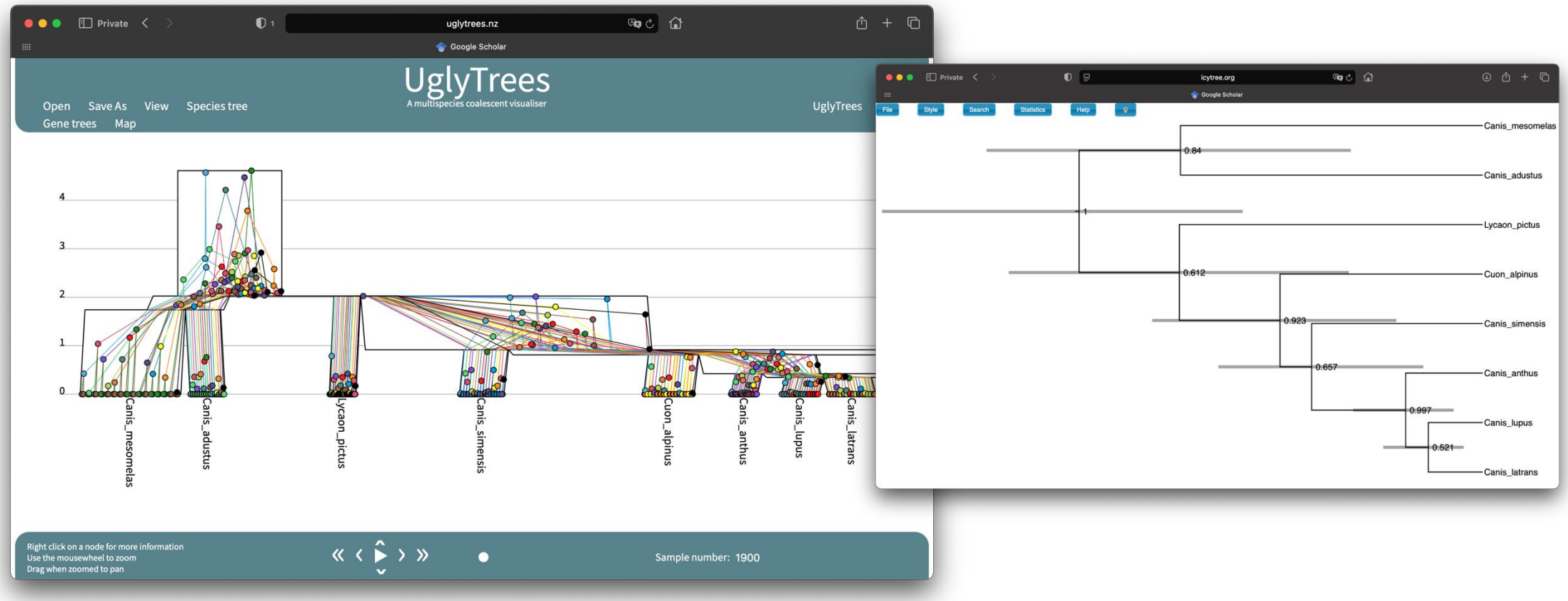


FIGURE 1. Depiction of the multispecies coalescent model, with $k=2$ gene trees constrained within a single species tree S with $n=8$ species. In this depiction, node heights (age) run along the y-axis and species-tree node widths are proportional to effective population sizes (arbitrary units). The relative molecular substitution rate of each species-tree branch is proportional to line thickness. Tree was built from a Gopher data set (Belfiore et al. 2008) and visualized using UglyTrees (Douglas 2020).

Priors

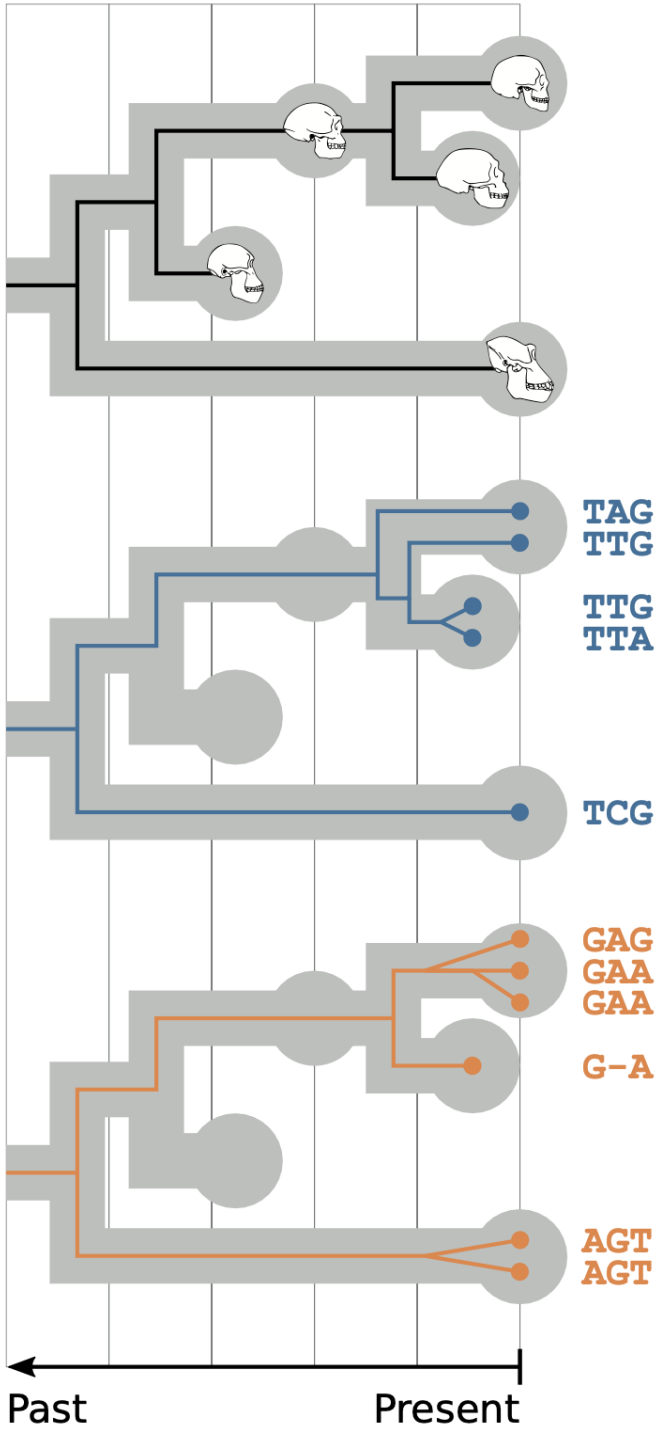
- Prior for the species tree
  - Yule or birth-death for topology and branch lengths
  - InvGamma(2, β_N) for population sizes (prior mean = β_N)
- Prior for the gene trees
  - MSC
- Prior for the evolutionary rate per species branch
  - strict (fixed rate or assign a prior)
  - relaxed ($r_b \sim$ Lognormal; fixed mean or assign a hyperprior)
- Prior for the relative rate per gene
  - Lognormal / Gamma (prior mean = 1.0)

Exploring and summarizing the posterior trees



Total Evidence Dating

- A unified model integrating morphological characters and multilocus molecular sequences
- FBD-MSC (Ogilvie et al. 2021)
  - FBD for the species tree
  - MSC for the gene trees
- Trait evolution along the species tree
- Molecular evolution along the gene trees



Total Evidence Dating

- MSC vs.

  Concatenation
- Concatenation
  - all genes (and traits) evolve along the same tree
- Higher level taxa

(Gavryushkina et al. 2017)