Taming the Beast Workshop

Bayesian inference of species tree and *BEAST

Chi Zhang

June 28, 2016

Species tree

- Species tree — the phylogeny representing the relationships among a group of species

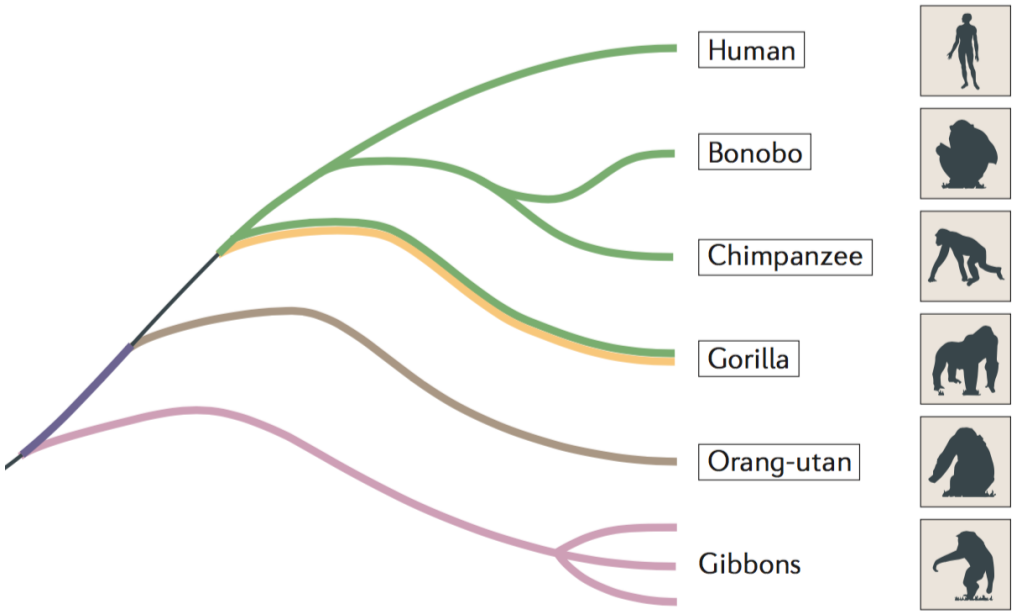


Figure adapted from [Rogers and Gibbs, 2014]

- Gene tree — the phylogeny for sequences at a particular gene locus from those species

Gene tree discordance

- Incomplete lineage sorting

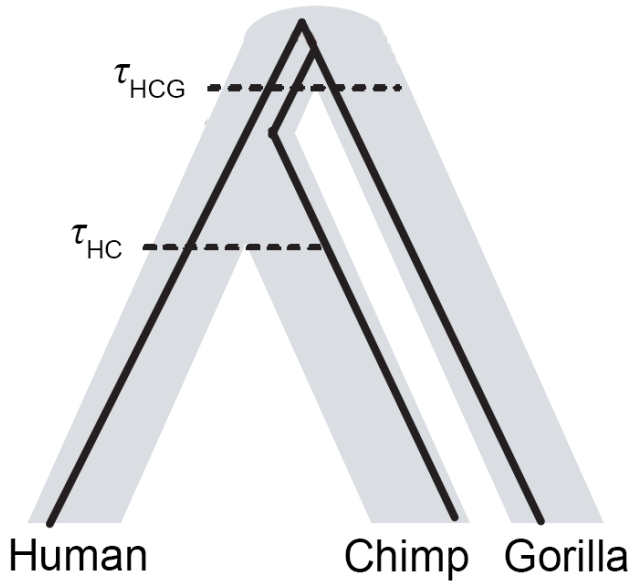


Figure adapted from [Patterson et al., 2006]

Gene tree discordance

- Horizontal gene transfer
- Gene duplication and loss

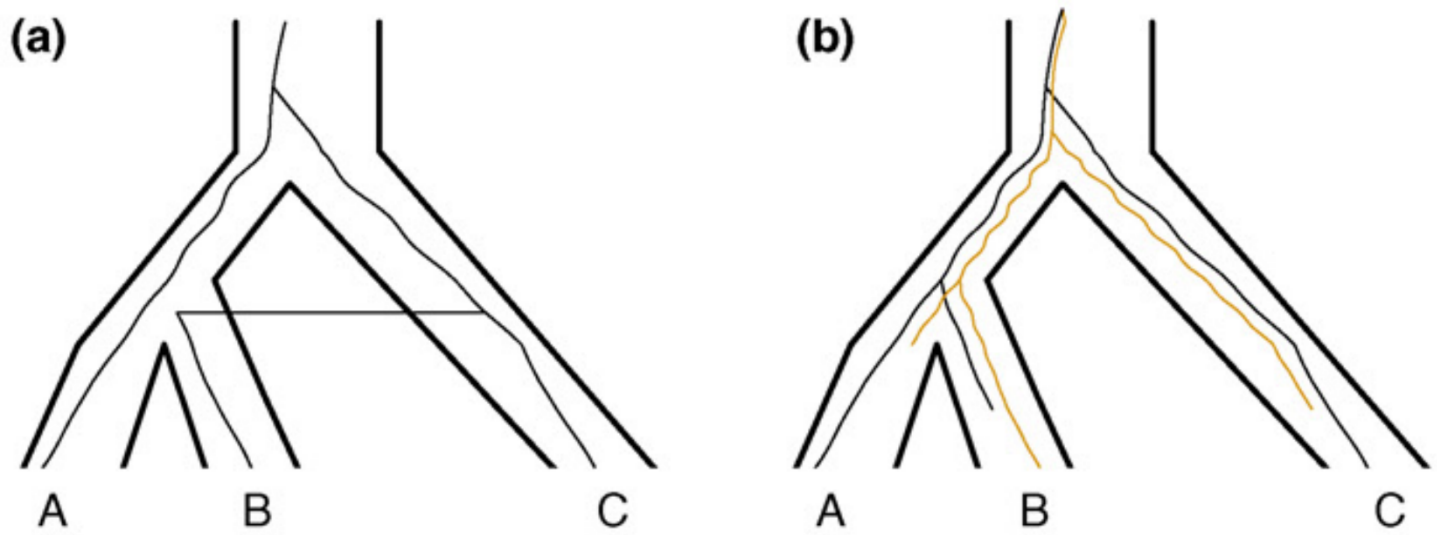


Figure adapted from [Degnan and Rosenberg, 2009]

Gene tree discordance

- Hybridization

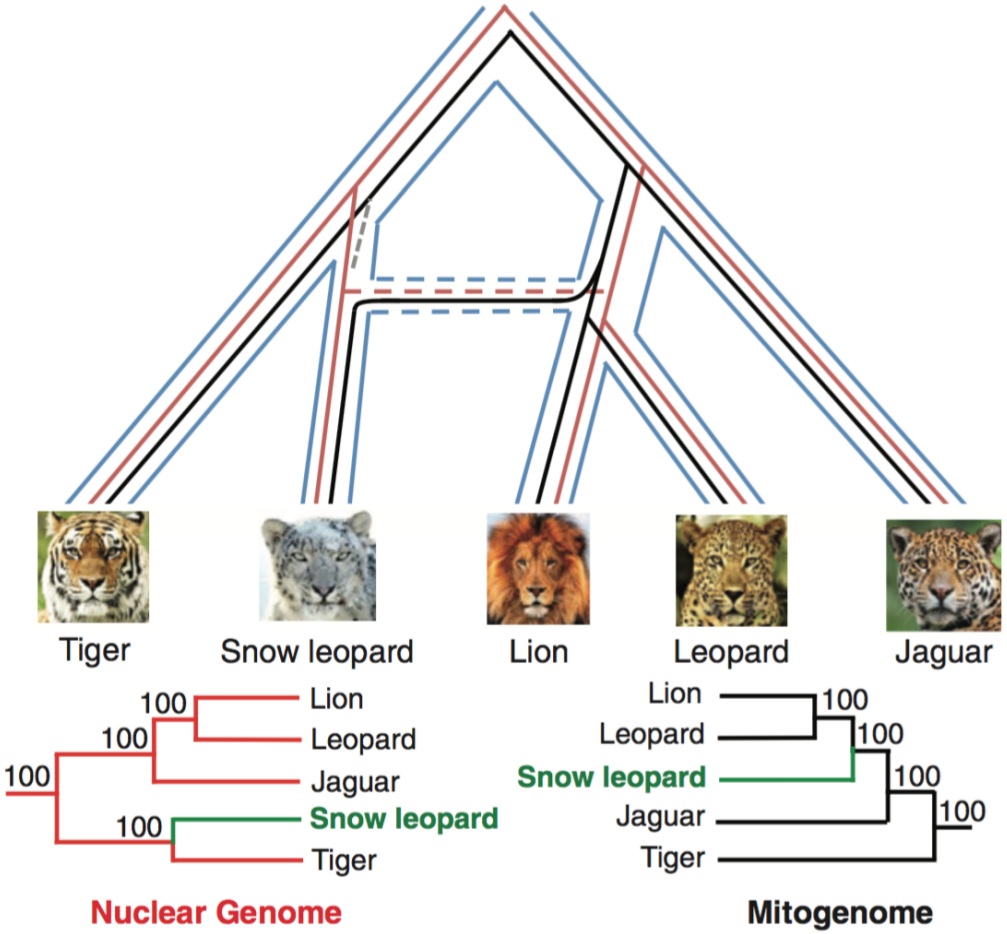
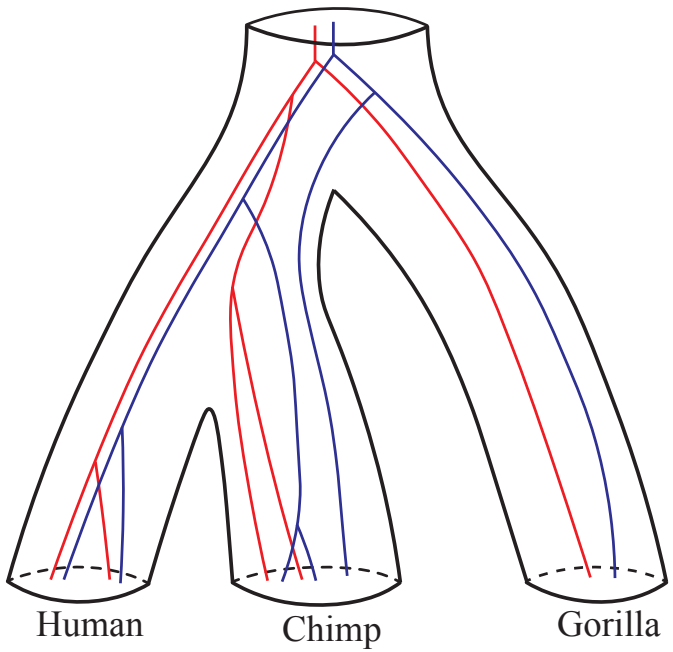


Figure adapted from [Li et al., 2016]

Species tree inference and *BEAST

- A Bayesian method to infer species tree from multilocus sequence data [Heled and Drummond, 2010]
- *BEAST, a functionality of BEAST2

- Gene trees are embedded in the species tree under the multispecies coalescent model [Rannala and Yang, 2003]
  - incomplete lineage sorting
- Gene trees are independent among loci



Species tree prior

- The prior for species tree S has two parts:

$$P(S) = P(S_T)P(N)$$

  - S_T — species time tree
  - N — population size functions
- $P(S_T)$ — typically a Yule (pure-birth) or birth-death prior
  - we can assign a hyperprior for the speciation (birth) rate (and extinction (death) rate, if birth-death)
- $P(N)$ — constant or continuous-linear

Species tree prior

- Constant population sizes

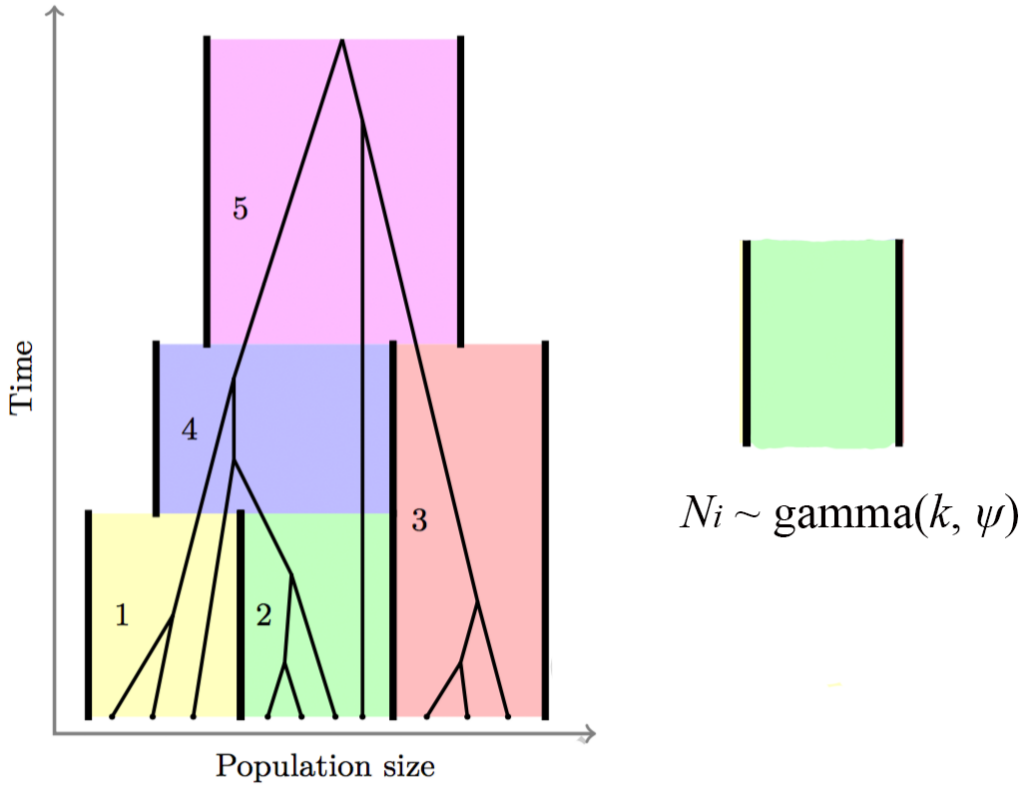


Figure adapted from [Drummond and Bouckaert, 2015]

Species tree prior

- Continuous-linear population sizes

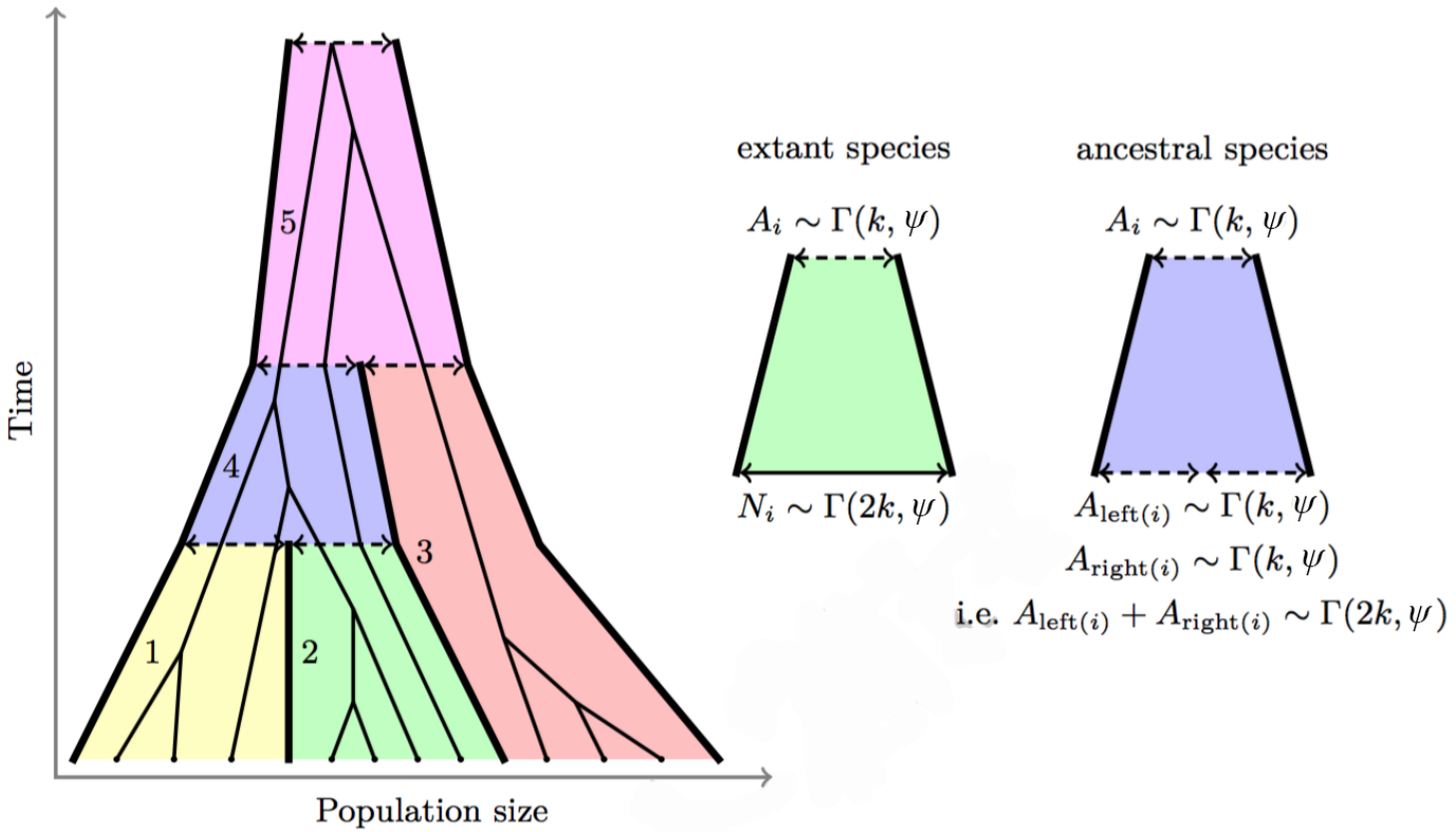


Figure adapted from [Drummond and Bouckaert, 2015]

Species tree prior

- In *BEAST, the prior type for $\mathbb{N}$ is fixed to gamma
- The gamma shape parameter $\mathbb{k}$ is fixed to 2, but we can assign a hyperprior for ψ, the scale parameter of the gamma
- (This ψ parameter is called "population mean" in Beauti, but the prior mean is actually 2ψ when the population sizes are constant)

Multispecies coalescent model

- The prior for gene tree $\mathbf{g}$, given species tree S

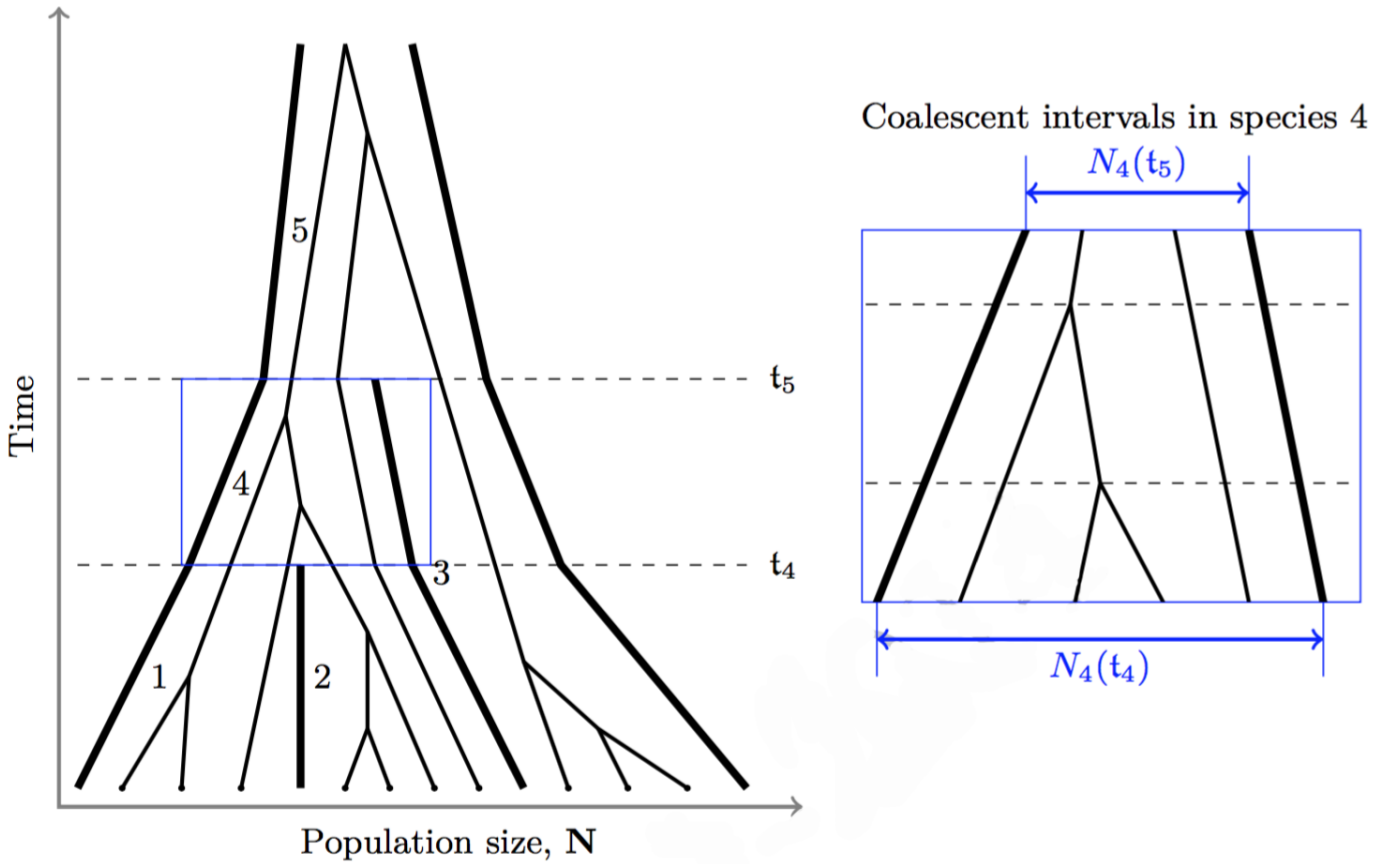


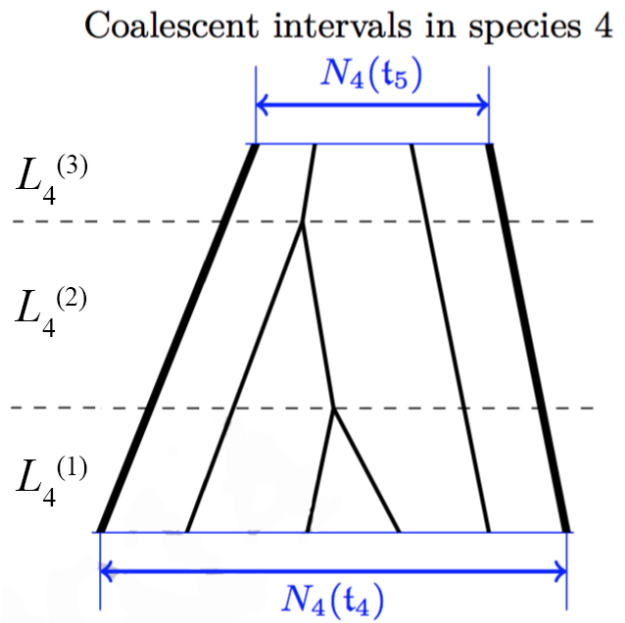
Figure adapted from [Drummond and Bouckaert, 2015]

Multispecies coalescent model

- The prob. distribution of gene time tree g given species tree S, is:

$$P(g|S) = \prod_{j=1}^{2s-1} P(L_j(g)|N_j(t))$$

- s — number of extant species ($2s-1$ branches totally)
- $N_j(t)$ — population size function (linear)
- $L_j(g)$ — coalescent intervals for genealogy g that are contained in the j'th branch of species tree S



Molecular clock model

- $P(\mathbf{c})$ — prior for the molecular clock model of genealogy $\mathbf{g}$
  - strict clock — typically fix to 1.0 for the first locus, and infer the relative clock rates for the rest loci
  - relaxed clock
- $P(\theta)$ — prior for the substitution model parameters
- e.g. HKY85,
  - Prior for transition/transversion rate ratio (κ), e.g. gamma(2,1)
  - Prior for base frequencies ($\pi_T, \pi_C, \pi_A, \pi_G$), e.g. Dirichlet(1,1,1,1)

Felsenstein likelihood

- The probability (likelihood) of data $\mathbf{d_i}$ (alignment at locus $\mathbf{i}$), given the gene time tree $\mathbf{g_i}$, molecular clock $\mathbf{c_i}$, and substitution model $\boldsymbol{\theta_i}$, is:

$$P(\mathbf{d_i}|\mathbf{g_i}, \mathbf{c_i}, \boldsymbol{\theta_i})$$

Priors and likelihood

- $P(S)$ — prior for species tree
- $P(g_i|S)$ — prior for gene tree i (multispecies coalescent)
- $P(c_i)$ — prior for clock rate of locus i
- $P(\theta_i)$ — prior for substitution parameters of locus i
- $P(d_i|g_i, c_i, \theta_i)$ — likelihood of data at locus i

Posterior

- The posterior distribution of the species tree S and other paremeters given data D is:

$$P(S, \mathbf{g}, \mathbf{c}, \Theta|D) \propto P(S) \prod_{i=1}^{n} P(g_i|S) P(c_i) P(\theta_i) P(d_i|g_i, c_i, \theta_i)$$

- The data $D = \{d_1, d_2, \ldots, d_n\}$ is composed of n alignments, one per locus.

Integrating out population sizes

- Assume constant population sizes
- Assign i.i.d inverse-gamma(α, β) prior for N_j
  - mean $= \beta/(\alpha - 1)$
- The population sizes $\mathbf{N}$ can be integrated out from $P(g|S)$ [Jones, 2015]

- Specify α and β in the invgamma prior (instead of ψ in the gamma prior)

starBEAST2

- A more efficient implementation and an upgrade of *BEAST
  - Population sizes integrated out [Jones, 2015]
  - Relaxed molecular clock per species tree branch (instead of per gene tree branch)
  - More efficient MCMC proposals for the species tree and gene trees (coordinated operators)
    [Jones, 2015, Rannala and Yang, 2015]
- Available at `github.com/genomescale/starbeast2`, will be released soon (as a BEAST2 add-on)

References I

- Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340.
- Drummond, A. J. and Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press.
- Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580.
- Jones, G. R. (2015). Species delimitation and phylogeny estimation under the multispecies coalescent. *bioRxiv*.
- Li, G., Davis, B. W., Eizirik, E., and Murphy, W. J. (2016). Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Research*, 26(1):1–11.
- Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103–1108.
- Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656.
- Rannala, B. and Yang, Z. (2015). Efficient Bayesian species tree inference under the multi-species coalescent. *arXiv.org*.
- Rogers, J. and Gibbs, R. A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature Reviews Genetics*, 15(5):347–359.