Tree priors

Chi Zhang

Prior distributions for trees

- Coalescent models
  - Kingman coalescent
  - Coalescent skyline
  - Structured coalescent
- Birth-death models
  - Birth (Yule) process
  - Birth-death (constant rates and skyline)
  - Fossilized birth-death
  - Multitype birth-death

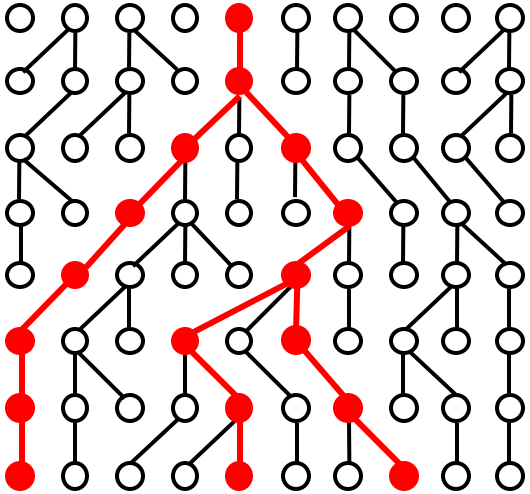
Fisher-Wright model

- Fisher 1930; Wright 1931

Ronald A. Fisher
(1890–1962)

Sewall Wright
(1889–1988)

- Constant population size
- Non-overlapping generations
- Random mating
- Neutral evolution



$2N = 10$

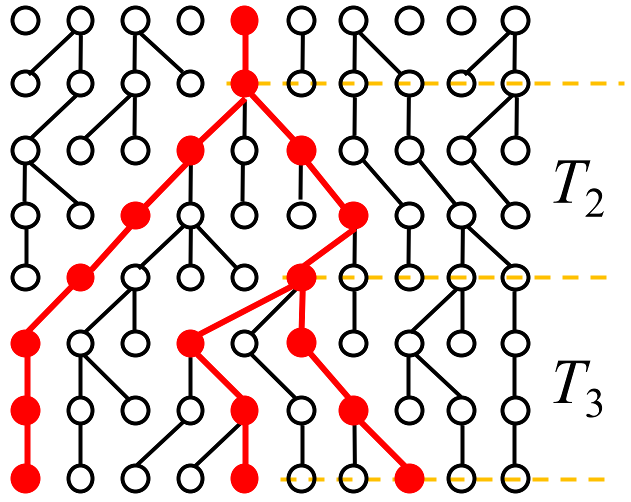
Coalescent process

- For k ($\leq n$) genes, the probability of no coalescent event in the previous generation is

$$\frac{2N-1}{2N}\frac{2N-2}{2N}\dots\frac{2N-k+1}{2N} = \prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right) = 1-\sum_{i=1}^{k-1}\frac{i}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right) \approx 1-\binom{k}{2}\frac{1}{2N}$$

- The probability that the coalescent event (which reduces k genes to k–1 genes) occurs more than j generations ago is

$$\Pr(T_k > j) \approx \left[1-\binom{k}{2}\frac{1}{2N}\right]^j \approx \mathrm{e}^{-\binom{k}{2}\frac{1}{2N}j}$$



$2N = 10$

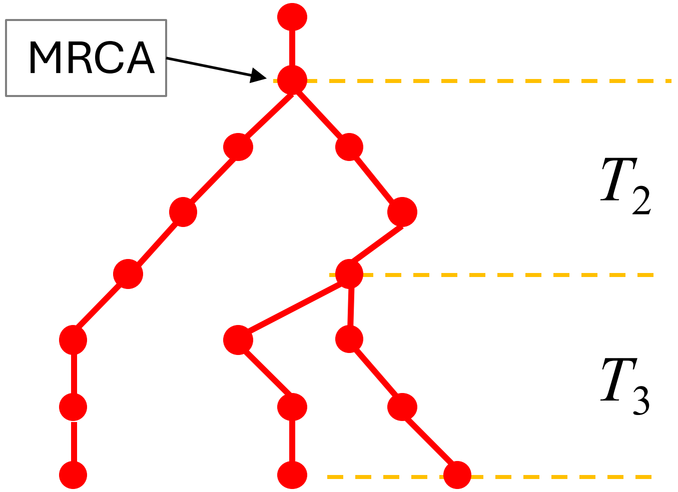
Coalescent process

- The coalescent times, $T_n, \ldots, T_2$, are independent exponential variables

$$f(T_k) \sim \exp\left(\binom{k}{2}\frac{1}{2N}\right)$$

$$\mathrm{E}(T_k) = \frac{4N}{k(k-1)}, \; \mathrm{E}(T_2) = 2N$$

$$\mathrm{E}(T_{\mathrm{MRCA}}) = \sum_{k=2}^{n} \frac{4N}{k(k-1)} = 4N\left(1 - \frac{1}{n}\right)$$

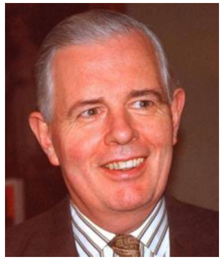


Coalescent process

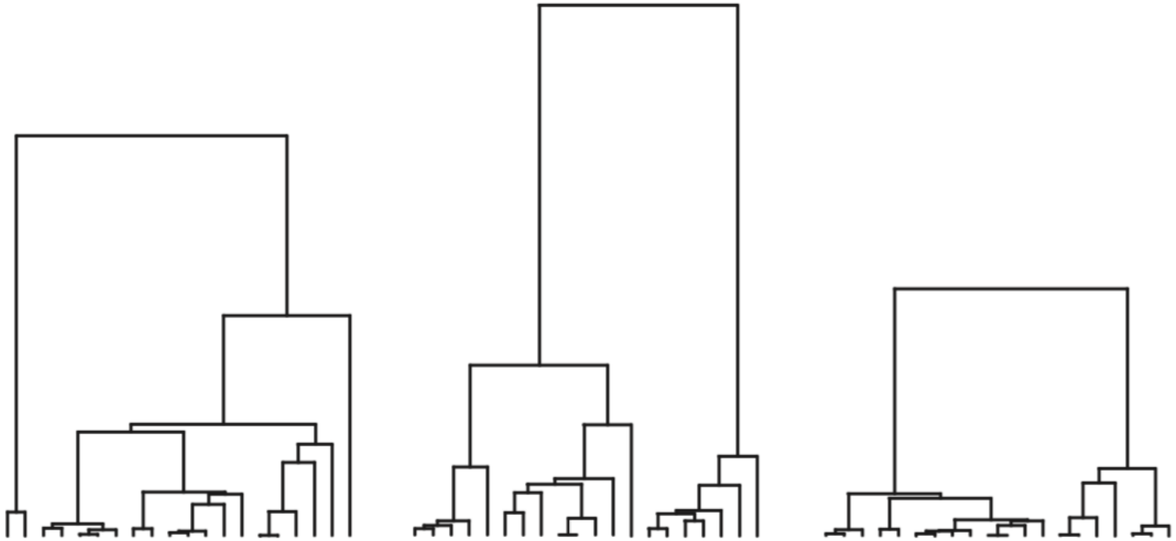
- Limiting condition ($N \to \infty$)

  (Kingman 1982)

- Three coalescent trees for $n = 20$ lineages



John Kingman
(1939–)



Coalescent process

- The joint distribution of tree topology and coalescent times

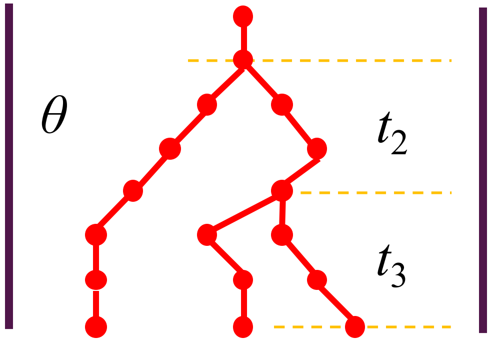
$$f(\tau, T_n, \ldots, T_2) = \prod_{k=2}^{n} \frac{1}{\binom{k}{2}} \binom{k}{2} \frac{1}{2N} \exp\left(-\binom{k}{2} \frac{1}{2N} T_k\right)$$

- For dating time in unit of Myr (or kyr), rescale T_k (in unit of generations) by multiplying the generation time g (in unit of Myr per generation), so that $t_k = T_k g$ is measured by Myr,

$$f(\tau, t_n, \ldots, t_2) = \prod_{k=2}^{n} \frac{1}{2Ng} \exp\left(-\frac{k(k-1)}{4Ng} t_k\right)$$

Coalescent process

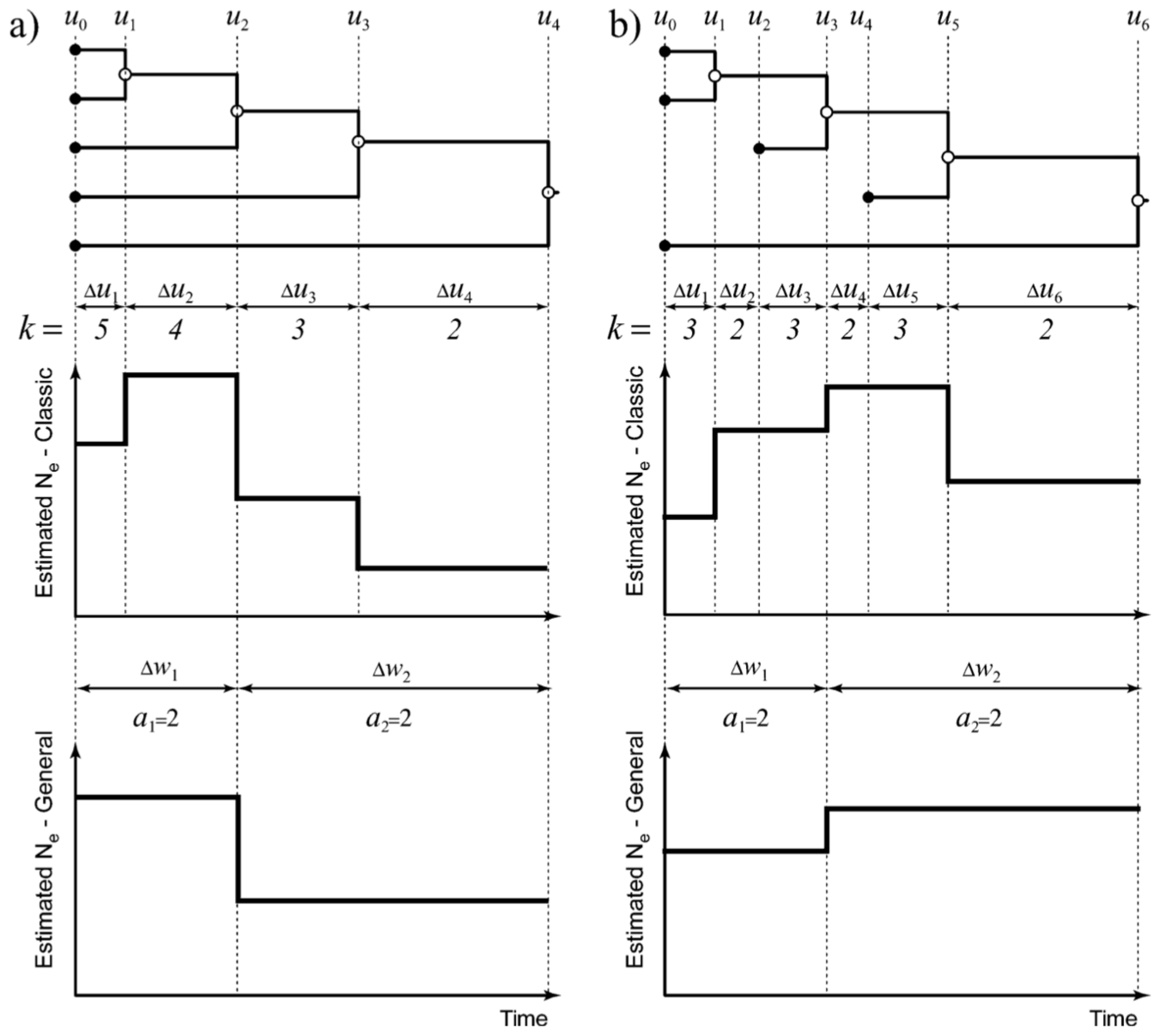
- For inference using sequences from contemporary samples (without time calibration), rescale T_k (in unit of generations) by multiplying the substitution rate μ per site per generation, so that $t_k = T_k\mu$ is measured by expected number of substitutions per site, and $\theta = 4N\mu$, $f(\tau, t_n, \ldots, t_2) = \prod_{k=2}^{n} \frac{2}{\theta} \exp\left(-\frac{k(k-1)}{\theta} t_k\right)$



$$f(G|\theta) = \frac{2}{\theta} \mathrm{e}^{-\frac{6t_3}{\theta}} \times \frac{2}{\theta} \mathrm{e}^{-\frac{2t_2}{\theta}}$$

Coalescent skyline

- Population size changes over time (demographic model) (Pybus et al. 2000; Drummond et al. 2005; Heled & Drummond 2008)



Structured coalescent

- Population structure
- Migration among subpopulations (Volz 2012; Vaughan et al. 2014; Müller et al. 2017)

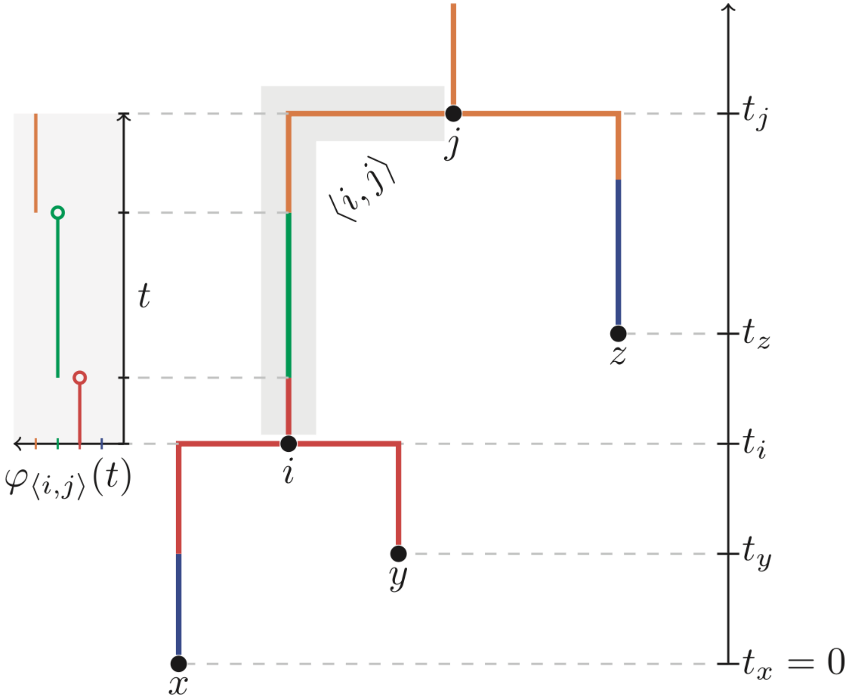


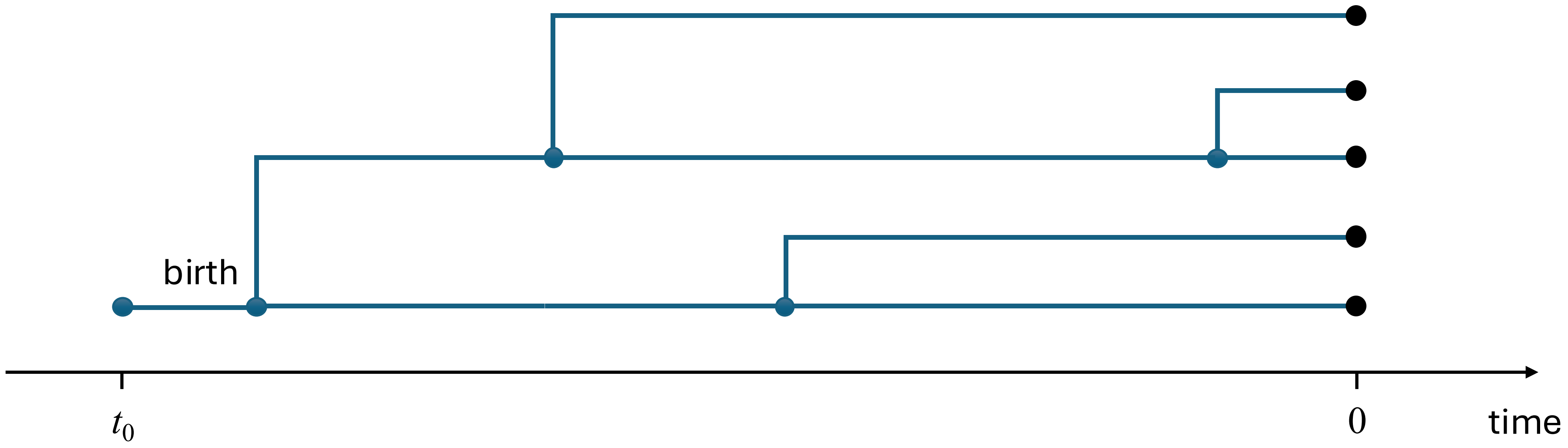
Fig. 1. A structured tree $\mathcal{T} = (V, E, \mathbf{t}, M)$ with $V = I \cup Y$ where $I = \{x, y, z\}$, $Y = \{i, j\}$, $E = \{\langle x, i \rangle, \langle y, i \rangle, \langle i, j \rangle, \langle z, j \rangle\}$ and the coalescence times $\mathbf{t}$ and type mappings M are as shown. Here we have selected the type set $D = \{$**blue**, **red**, **green**, **orange**$\}$, although this can be composed of the values of any discrete trait

Prior distributions for trees

- Coalescent models
    - Kingman coalescent
    - Coalescent skyline
    - Structured coalescent
- Birth-death models
    - Birth (Yule) process
    - Birth-death (constant rates and skyline)
    - Fossilized birth-death
    - Multitype birth-death

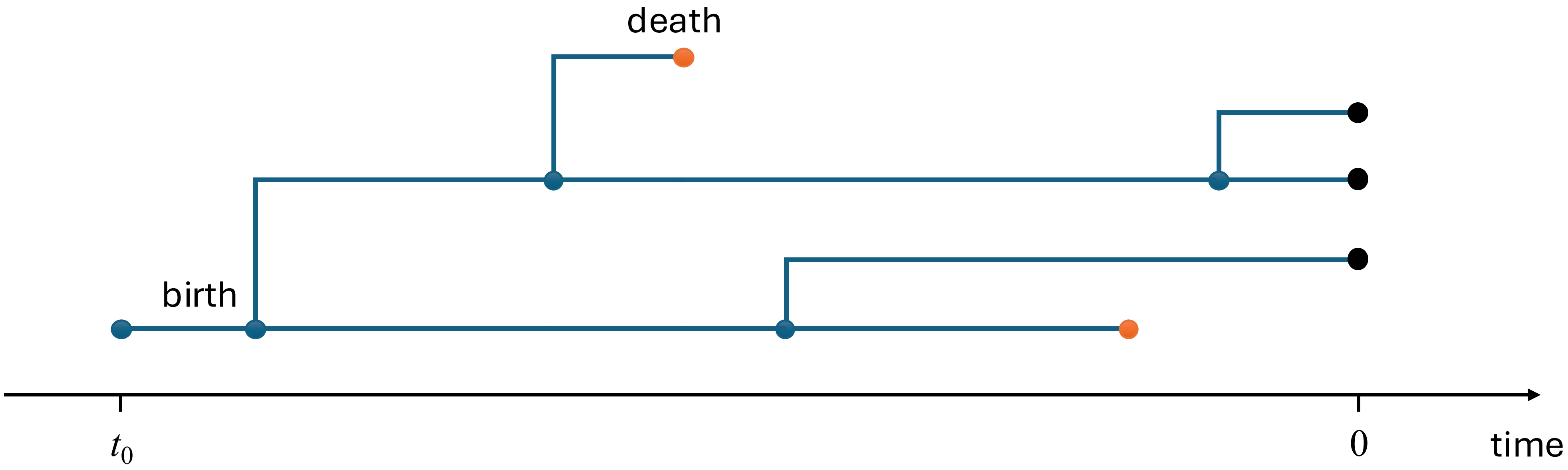
Yule process

- Birth rate, λ



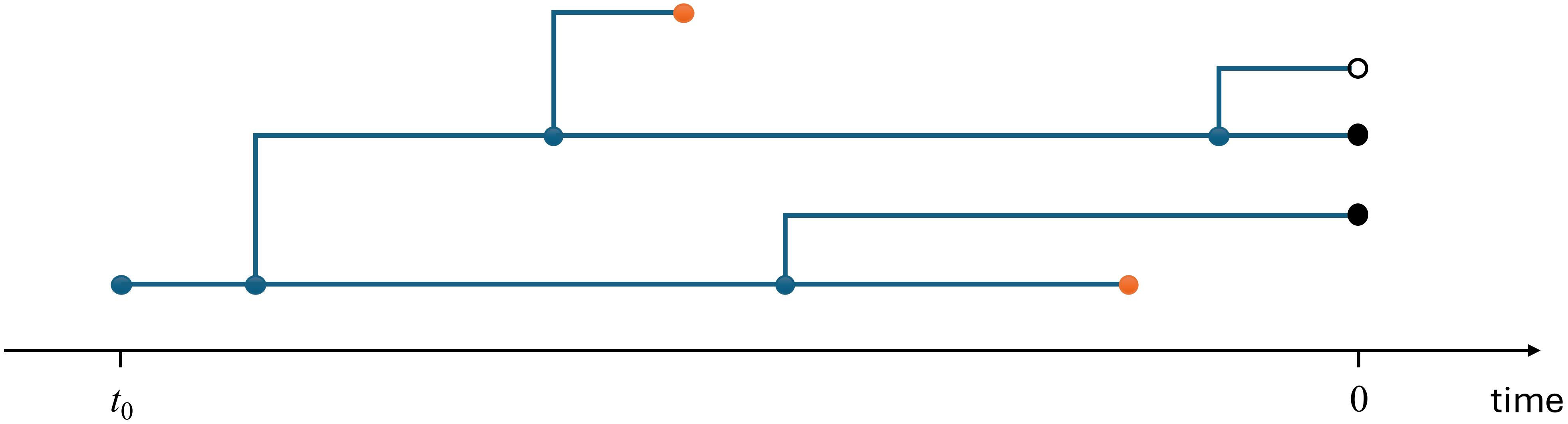
Birth-death process

- Birth rate, λ
- Death rate, μ



Birth-death process

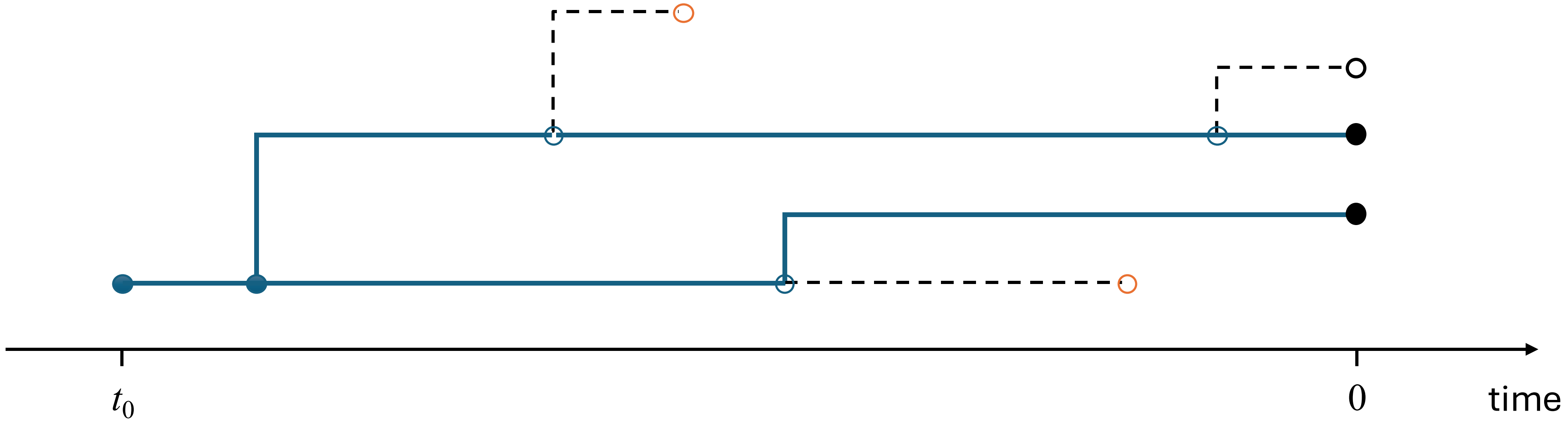
- Birth rate, λ
- Death rate, μ
- Extant sampling probability, ρ



Birth-death process

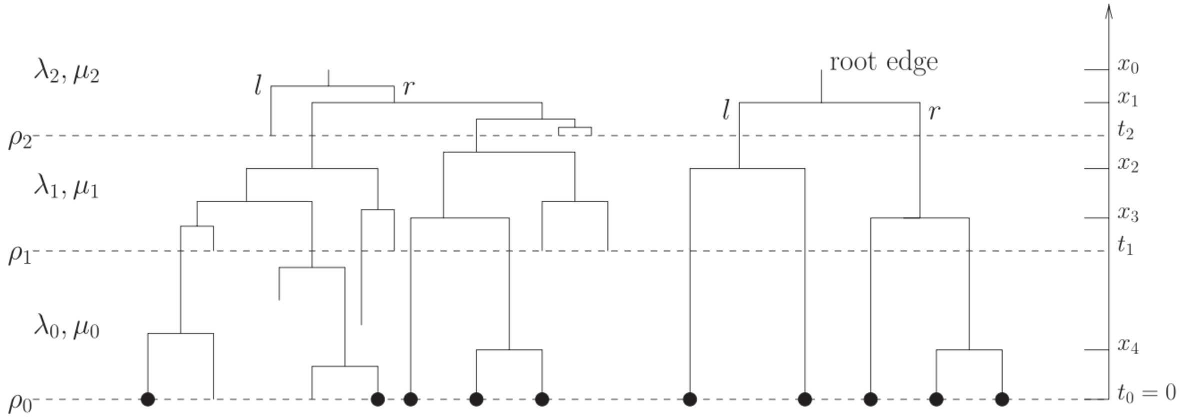
- Sampled tree
- Joint distribution on topology (τ, labelled history) and times ($\mathbf{t}$)

  (Yang & Rannala 1997, Stadler 2009)



Birth-death skyline

(Stadler 2011 PNAS)



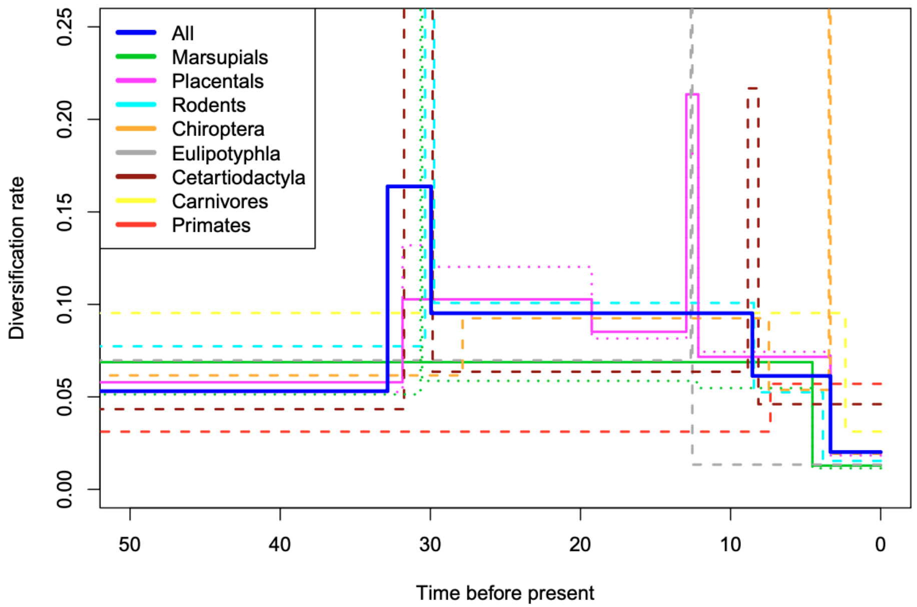
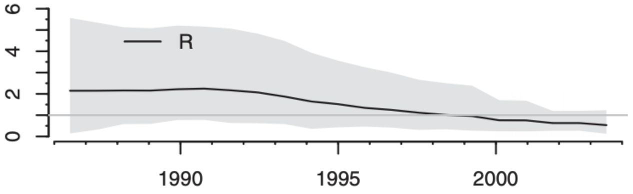


Fig. 3. Maximum-likelihood diversification rate estimates (per million years)

(Stadler et al. 2013 PNAS)

Fig. 2. HIV-1 in the United Kingdom: Birth–death skyline plot



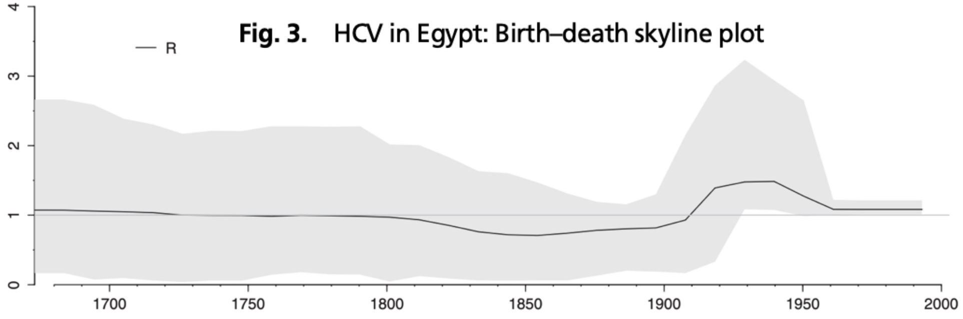
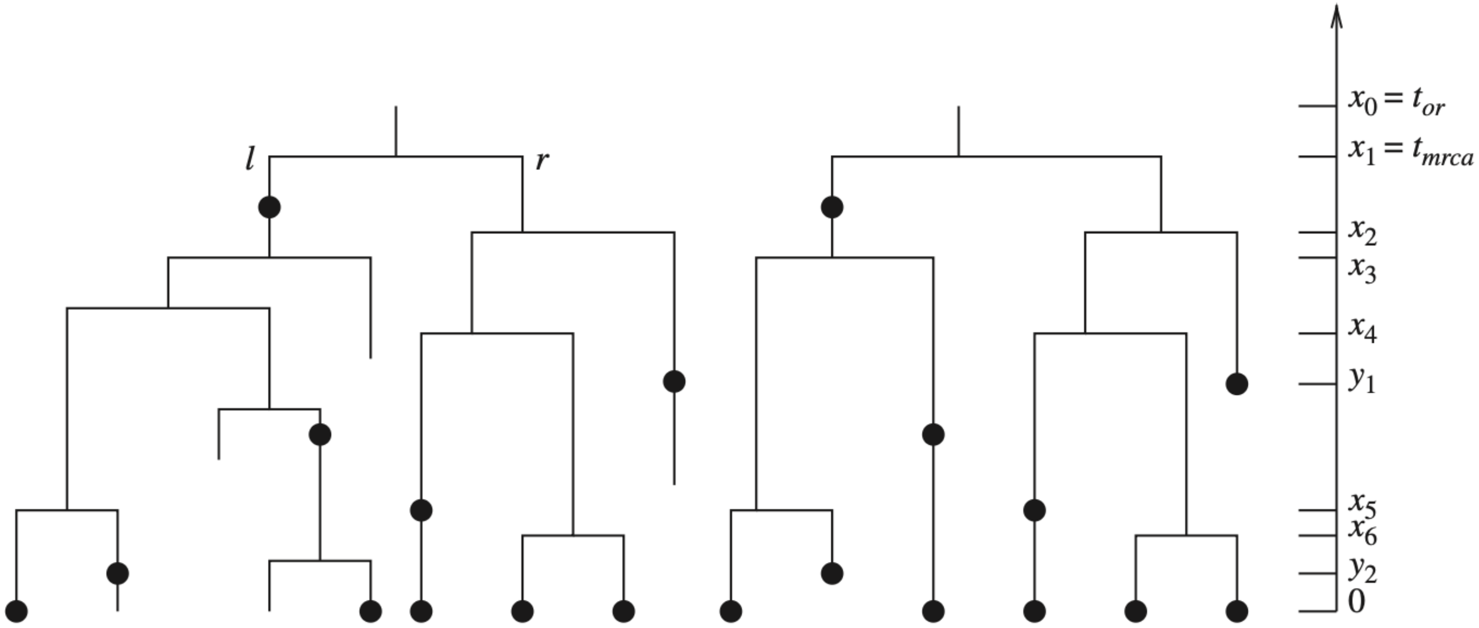


Fig. 3. HCV in Egypt: Birth–death skyline plot

Fossilized birth-death (FBD) process

- Birth rate, λ
- Death rate, μ
- Fossil sampling rate, ψ
- Extant sampling probability, ρ (Stadler 2010; Heath et al. 2014)

Joint distribution of $\{\tau, \mathbf{t}\}$: $f(\tau, \mathbf{t} | \lambda, \mu, \psi, \rho, x_0)$



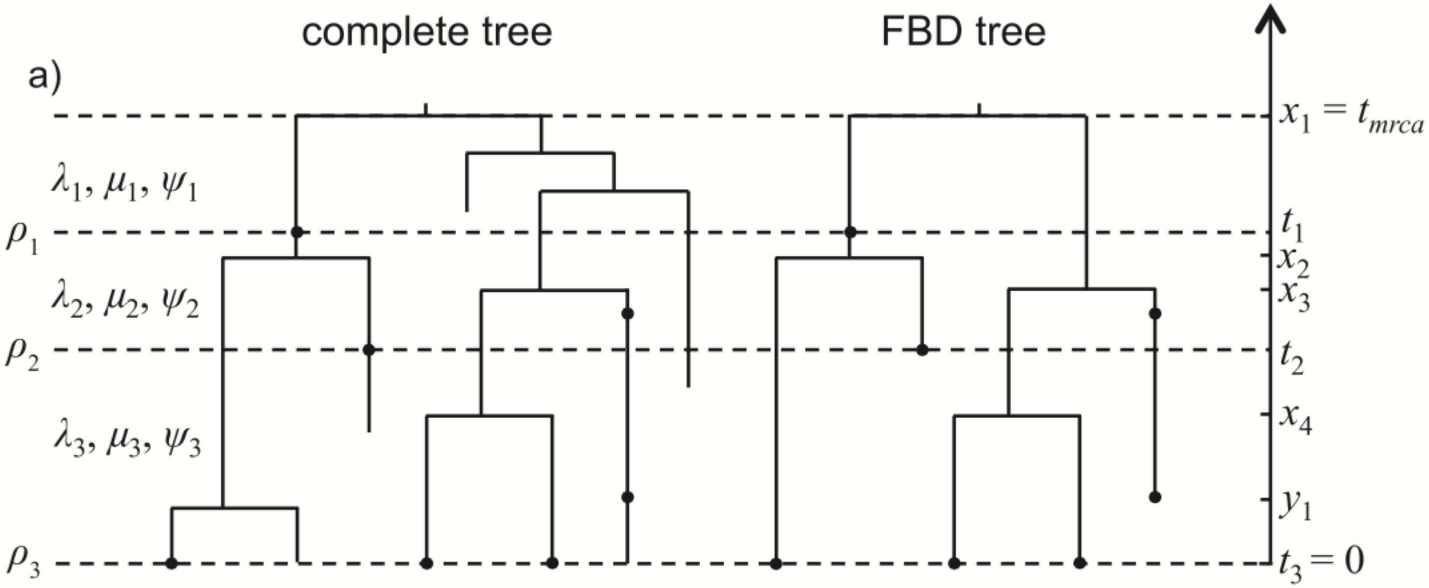
Reparameterization

- Fossilized BD (macroevolution)
  - $d = \lambda - \mu$
  - $v = \mu / \lambda$
  - $s = \psi / (\mu + \psi)$
  - $r = 0$
- Serially-sampled BD (epidemiology)
  - $R = \lambda / (\mu + \psi r)$
  - $\delta = \mu + \psi r$
  - $s = \psi r / \delta$
  - $\rho = 0$

FBD skyline

- Birth rates, $\boldsymbol{\lambda}$
- Death rates, $\boldsymbol{\mu}$
- Fossil sampling rates, $\boldsymbol{\psi}$
- Extant sampling probability, ρ

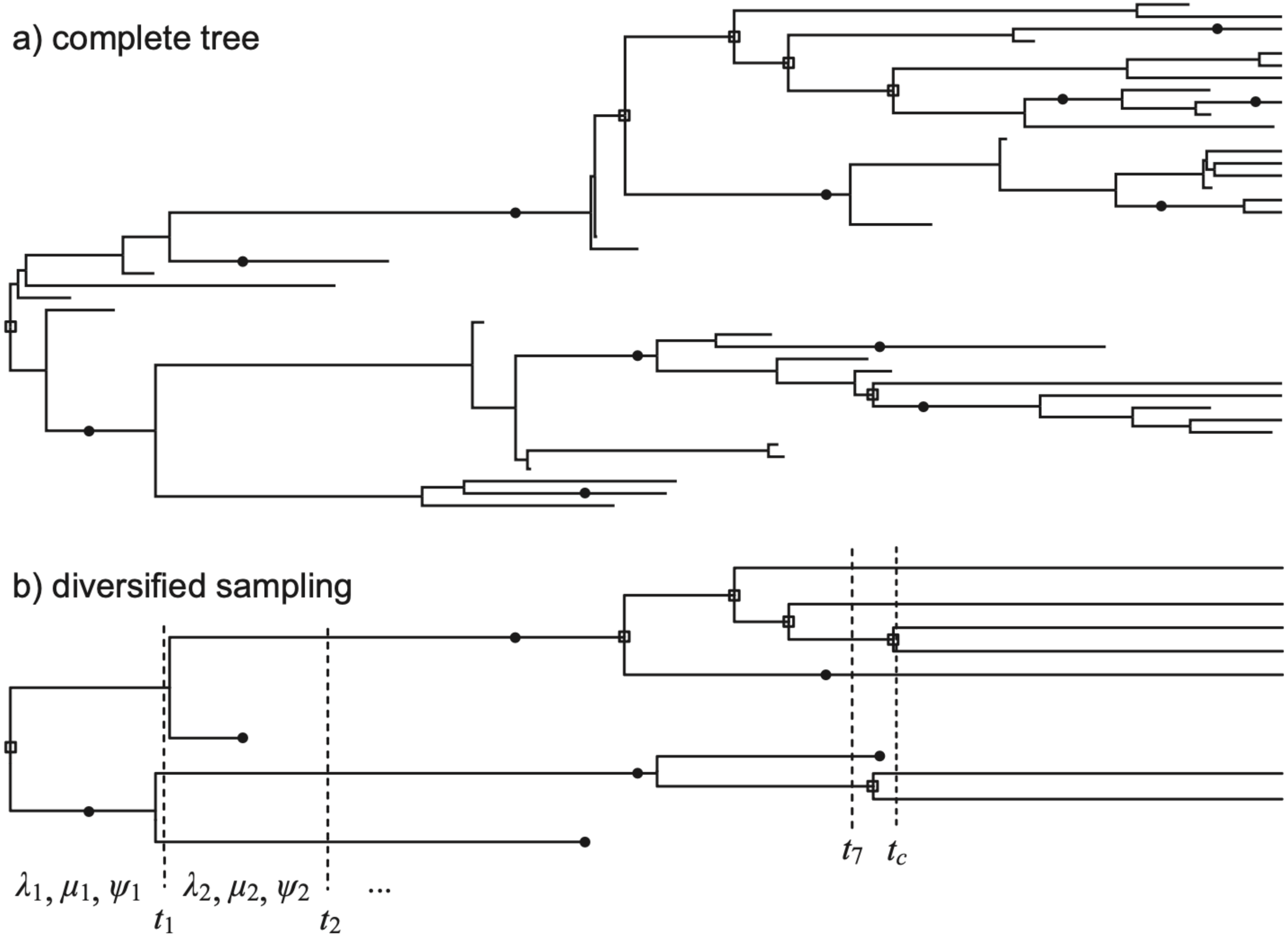
(Gavryushkina et al. 2014; Zhang et al. 2016)



FBD skyline

- Diversified extant sampling

(Zhang et al. 2016)



Multitype birth-death

(Barido-Sottani et al. 2020; Barido-Sottani & Morlon 2023)

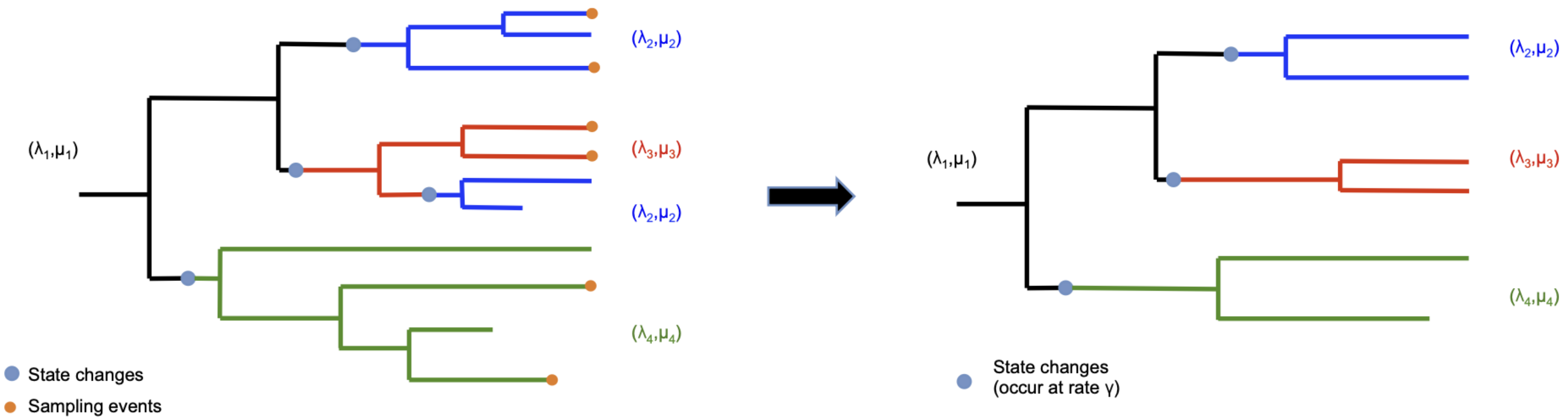


FIGURE 1. Visual representation of the MTBD model on a complete tree (left) with sampling events indicated in orange, and on the corresponding reconstructed tree (right). Each type is represented by a color: the ancestral type, in black, starts at the root. The other types, in blue, red, and green, start at change points along the tree. The same type can be present in multiple clades along the tree, such as the blue type in the complete tree.

Coalescent vs. birth-death

- Coalescent
  - backward in time
  - samples are from a population or closely-related populations
  - population size and its dynamics
  - epidemiology
- Brith-death
  - forward in time
  - samples can be cells, individuals, or species
  - diversification and sampling rates
  - epidemiology and macroevolution

(Stadler 2009;
Stadler et al. 2015
Cornuault et al. 2025)