Tree Models

Coalescent Trees, Birth Death Processes, and Beyond...

Will Freyman

Department of Integrative Biology
University of California, Berkeley

freyman@berkeley.edu
http://willfreyman.org

IB290 Grad Seminar in Phylogenetics, Fall 2017

Tree Models

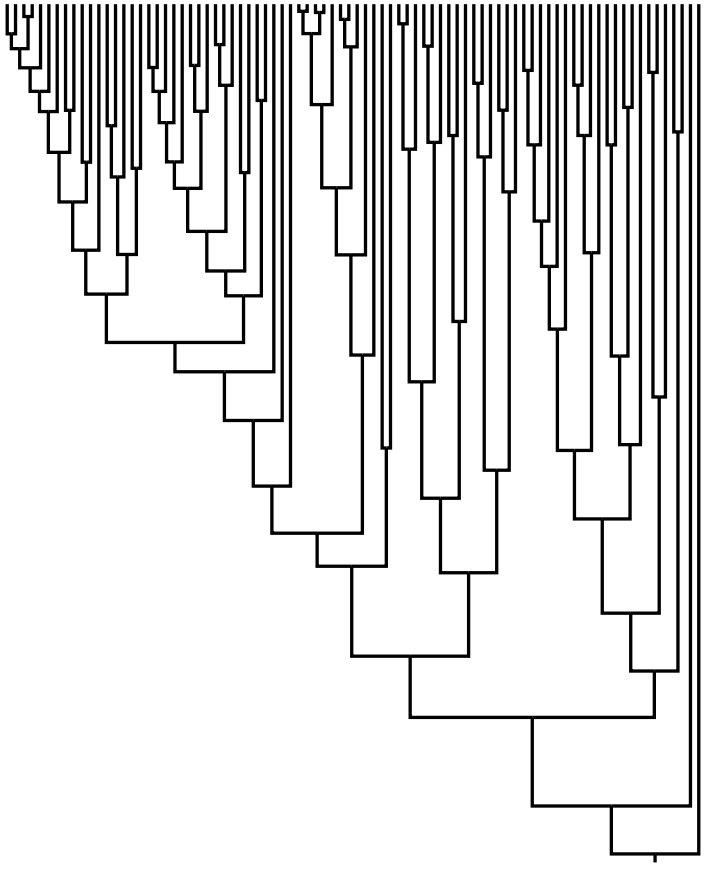
Statistical Distributions of Trees

Priors on evolutionary trees:

- Uniform tree topologies
- Coalescent trees
- Birth death processes

Different priors for different purposes:

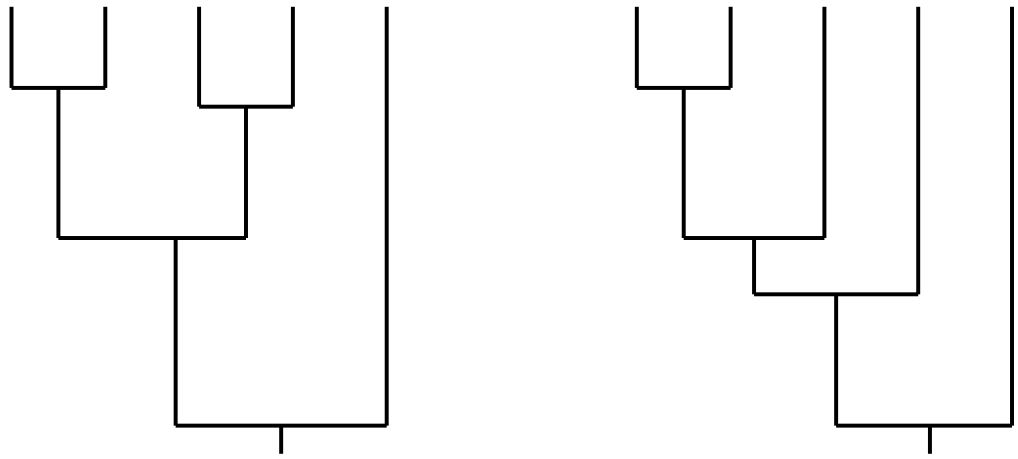
- Branch lengths in substitutions per site?
- In units of time?



Shape, Topology, Labeled History

Three Aspects of Trees

Tree Shape

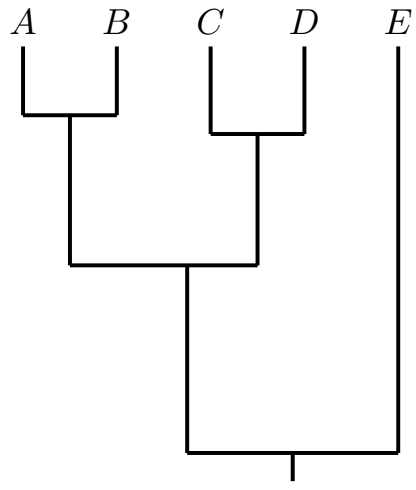


branching diagram with no labels at the tip

Shape, Topology, Labeled History

Three Aspects of Trees

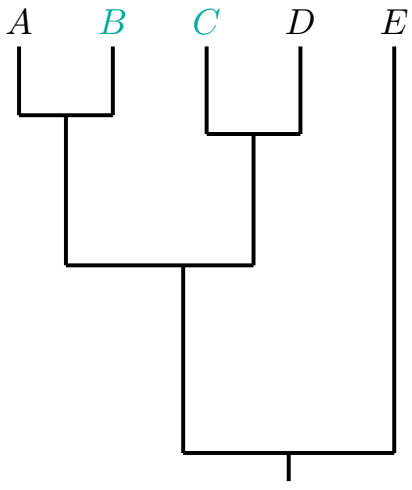
Tree Topology

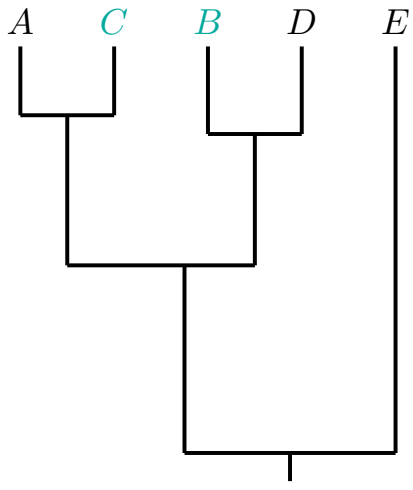


Shape, Topology, Labeled History

Three Aspects of Trees

Tree Topology



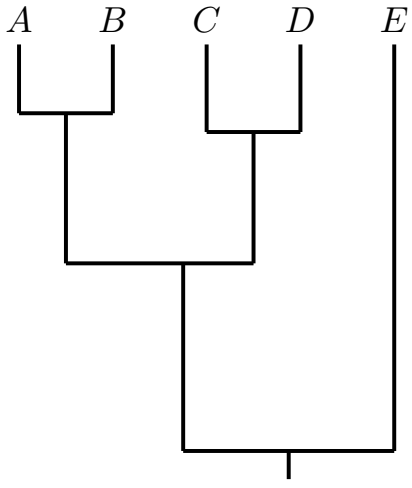


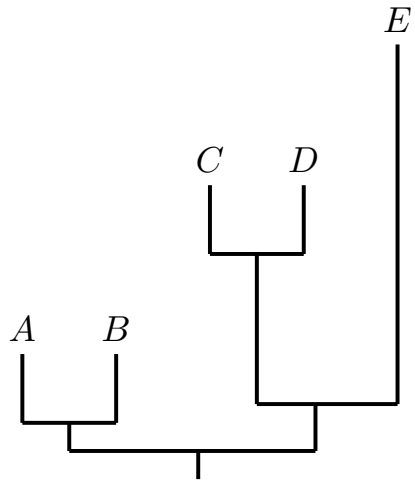
same shape, different topologies...

Shape, Topology, Labeled History

Three Aspects of Trees

Tree Topology



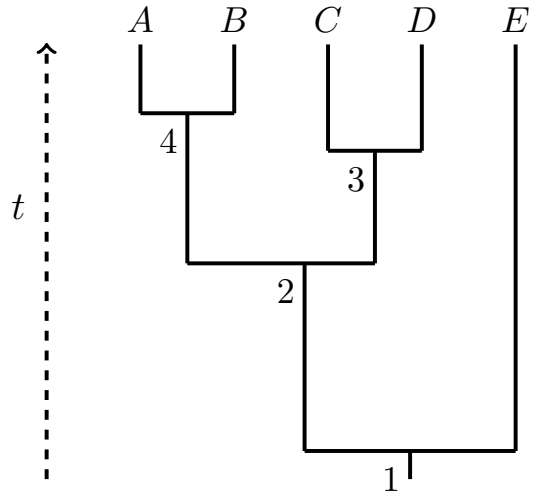


same topology, different roots...

Shape, Topology, Labeled History

Three Aspects of Trees

Labeled History

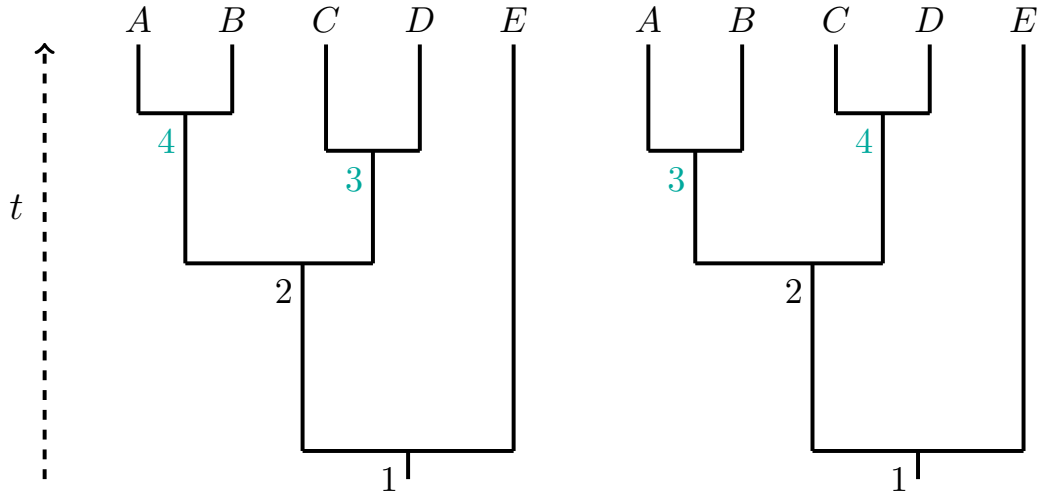


the topology plus a temporal ordering of the nodes

Shape, Topology, Labeled History

Three Aspects of Trees

Labeled History



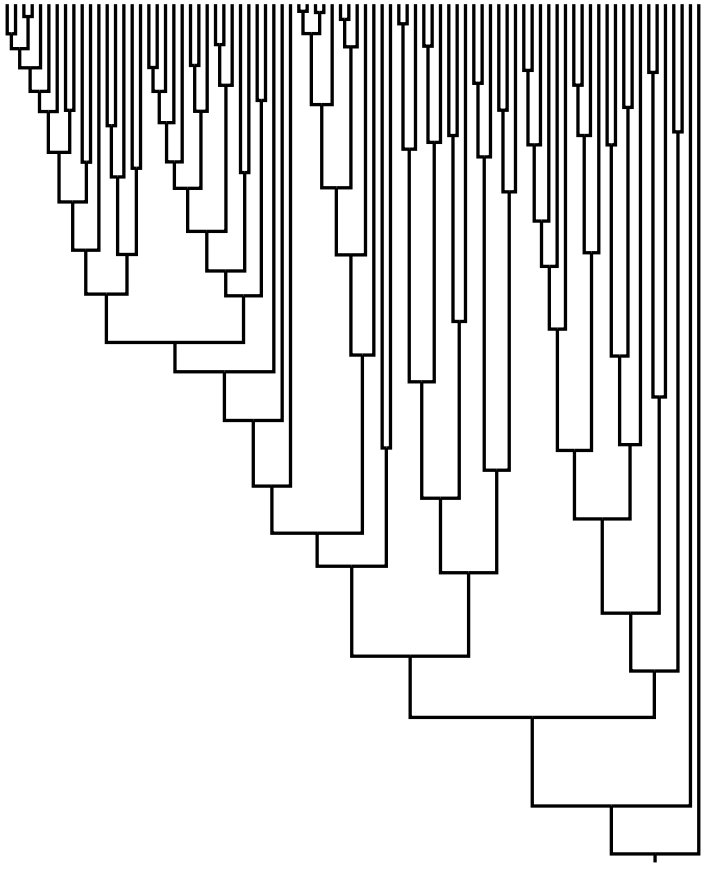
same topology, different labeled histories

Tree Models

Statistical Distributions of Trees

Three tree models we'll introduce today:

- Uniform tree topologies
- Uniform labeled histories
  - Coalescent trees
  - Birth death processes



Uniformly Distributed Tree Topologies

We ignore labeled histories and simply assign each tree topology an equal prior probability:

1. An OK assumption if we don't care about time
2. Branch length in units of the expected # of substitutions per site
3. Not all tree shapes will be equally probable

Uniformly Distributed Tree Topologies

Uniformly distributed tree topologies are:

1. the implicit assumption in `RAxML`, `PAUP*`, etc.
2. the default tree prior in `MrBayes`

In a Bayesian framework we also need to define a prior for branch lengths, something like:

$$v_i \sim Exponential(\lambda = 10.0)$$

Uniformly Distributed Labeled Histories

We often want to disentangle *time* from the *rate* of character change:

- Estimating demographic parameters
- Estimating divergence times
- Estimating diversification rates
  - adaptive radiation
  - key innovations
  - mass extinction

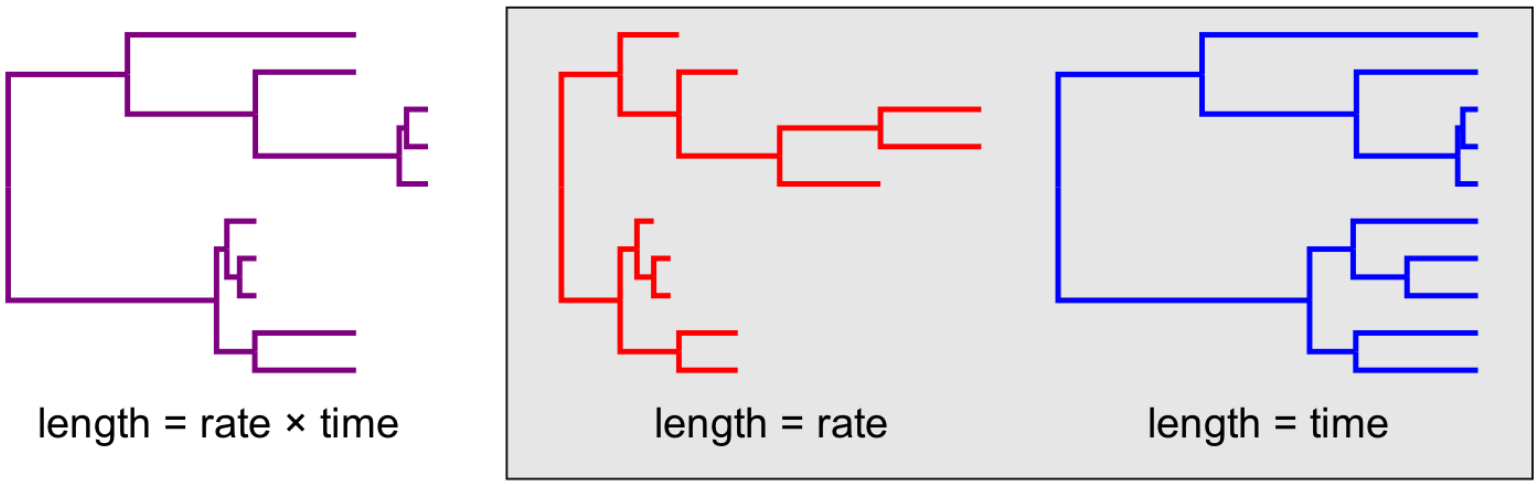
To do this, we must use tree models that account for labeled histories:

- coalescent trees
- birth death processes

Why?

Uniformly Distributed Labeled Histories

The *expected # of substitutions/site* occurring along a branch is the product of the *substitution rate* and *time*.



To get branch lengths in unit of time we must estimate *substitution rates* and *time* separately.

Image from Tracy Heath's slides (2017)

Coalescent Trees

Bayesian skyline plot

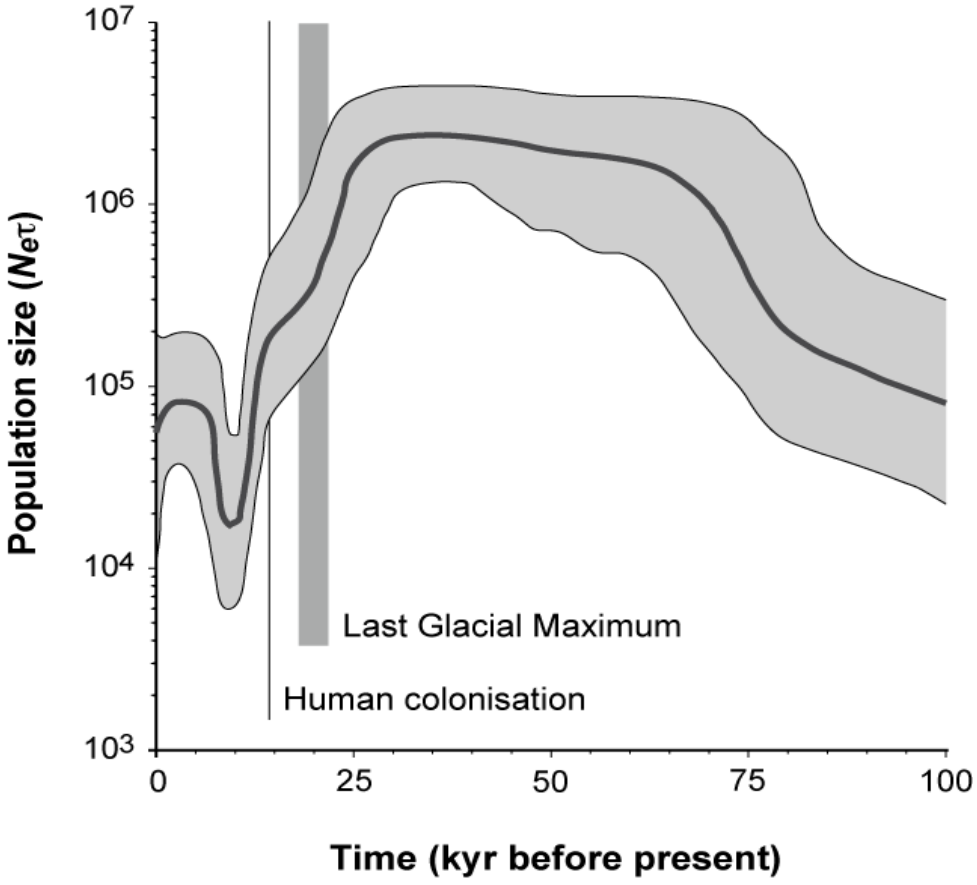
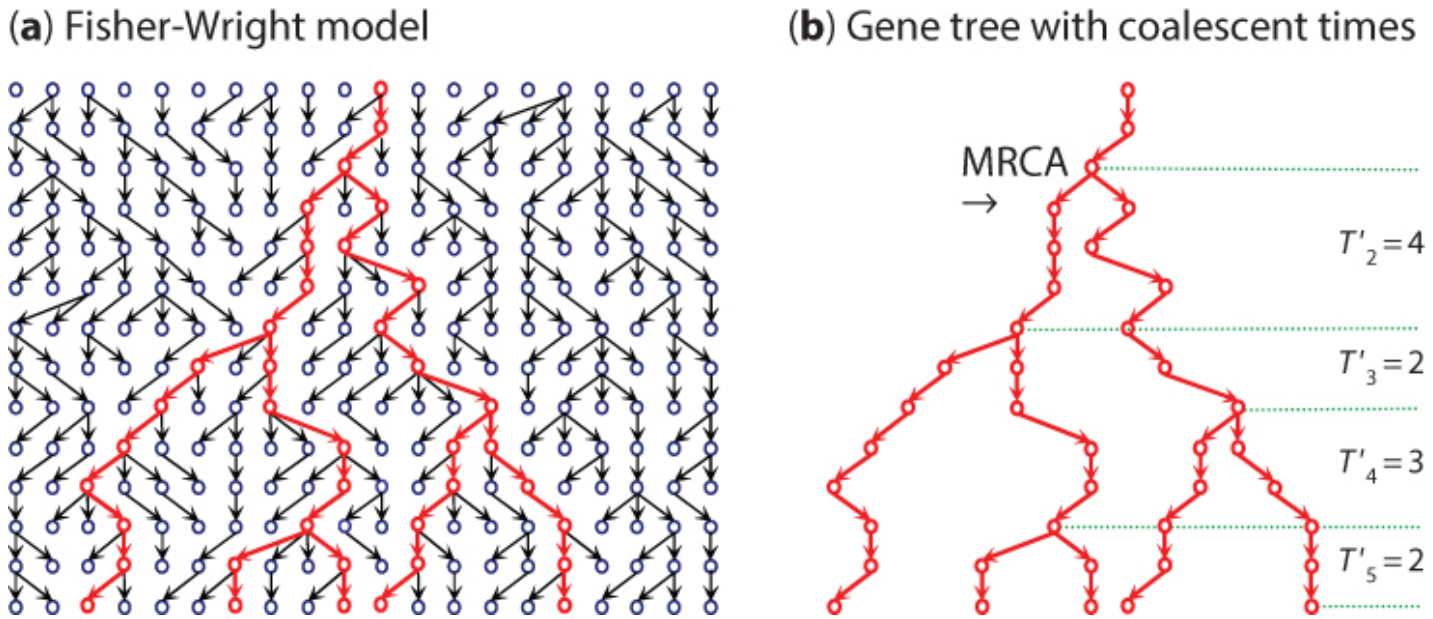


Image from Drummond et al. (2005)

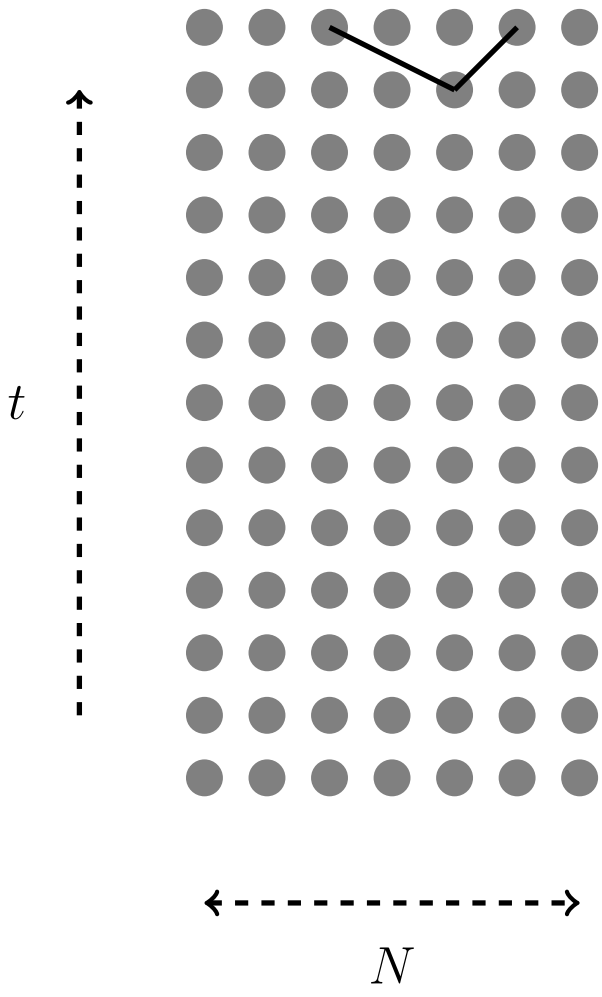
Coalescent Trees



- Fisher-Wright and other classical population genetic models predict allele frequencies in forward time.
- Coalescent theory looks at the same process backwards in time and describes the time until sampled lineages "coalescence".
- These models usually assume random mating, no selection, no structure, no recombination, and no gene flow – but they can be extended to handle these scenarios.

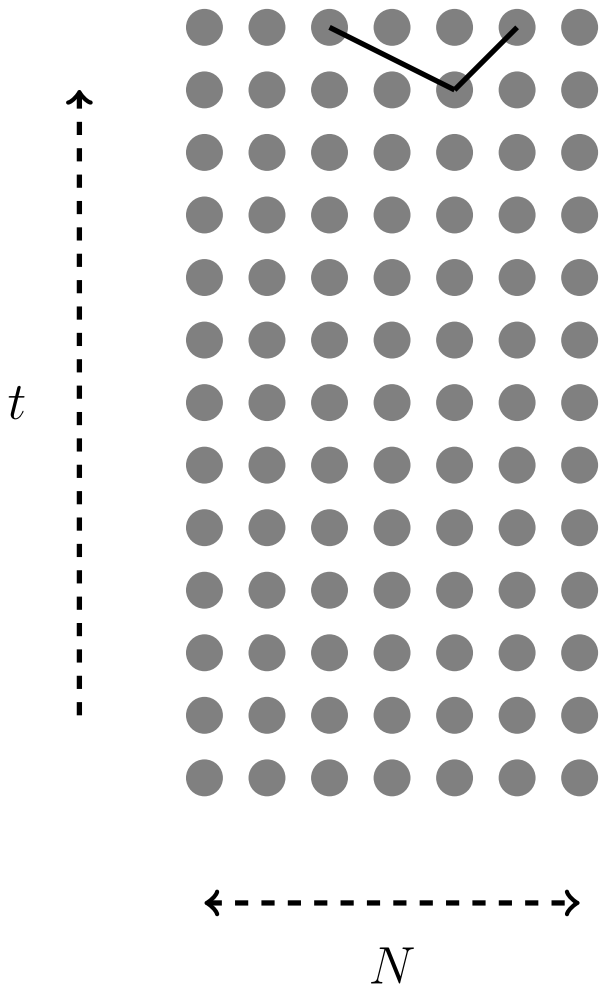
Image from Yang (2014)

Coalescent Trees



What is the probability of two lineages coalescing in a single generation?

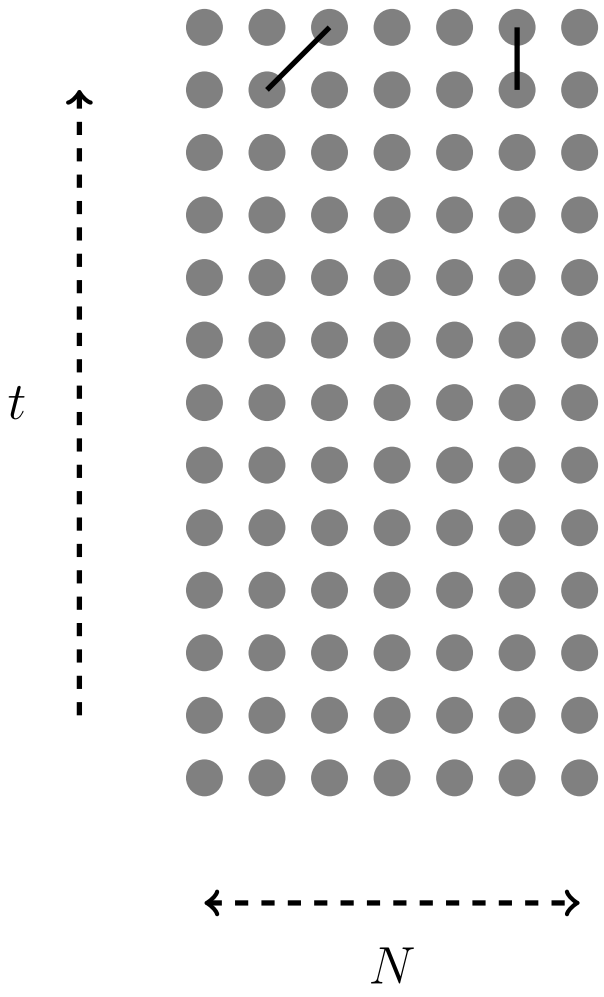
Coalescent Trees



What is the probability of two lineages coalescing in a single generation?

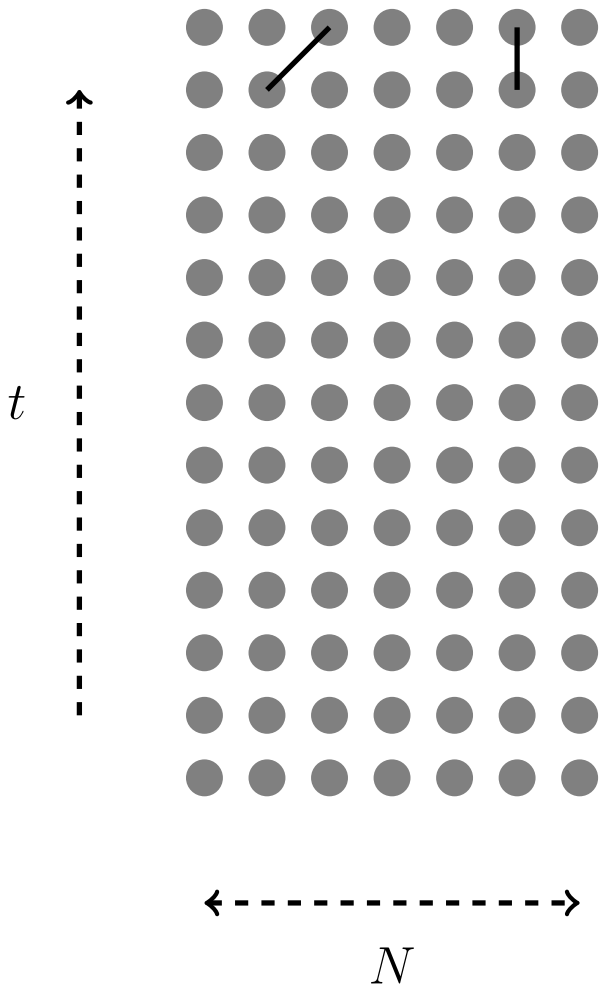
$$P(g = 1|N) = \frac{1}{N}$$

Coalescent Trees



So what is the probability of two lineages *not* coalescing in a single generation?

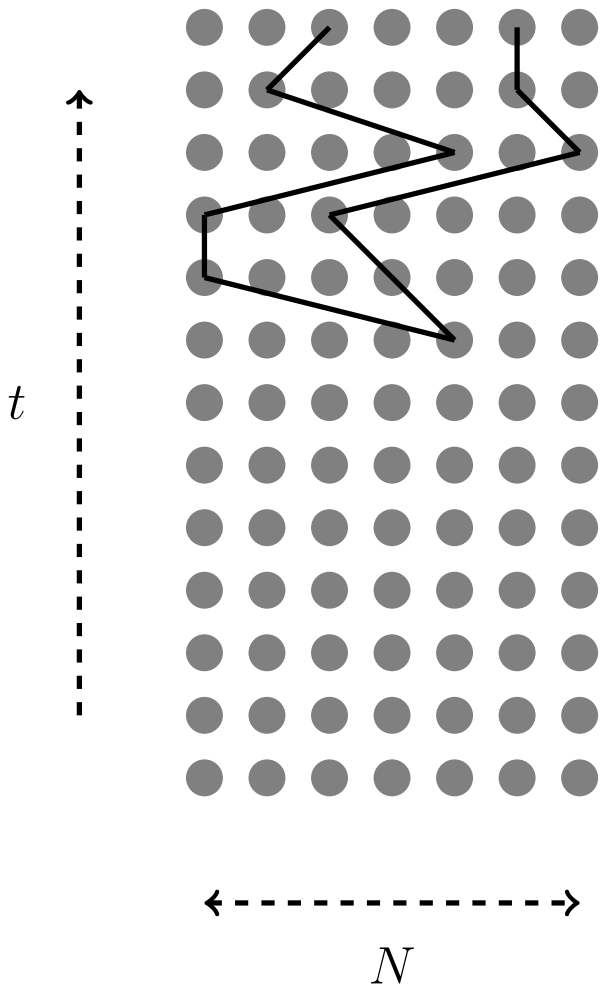
Coalescent Trees



So what is the probability of two lineages *not* coalescing in a single generation?

$$P(g \neq 1|N) = 1 - \frac{1}{N}$$

Coalescent Trees



What is the probability that coalescence occured $g+1$ generations ago?

- Probability of no coalescence for g generations:

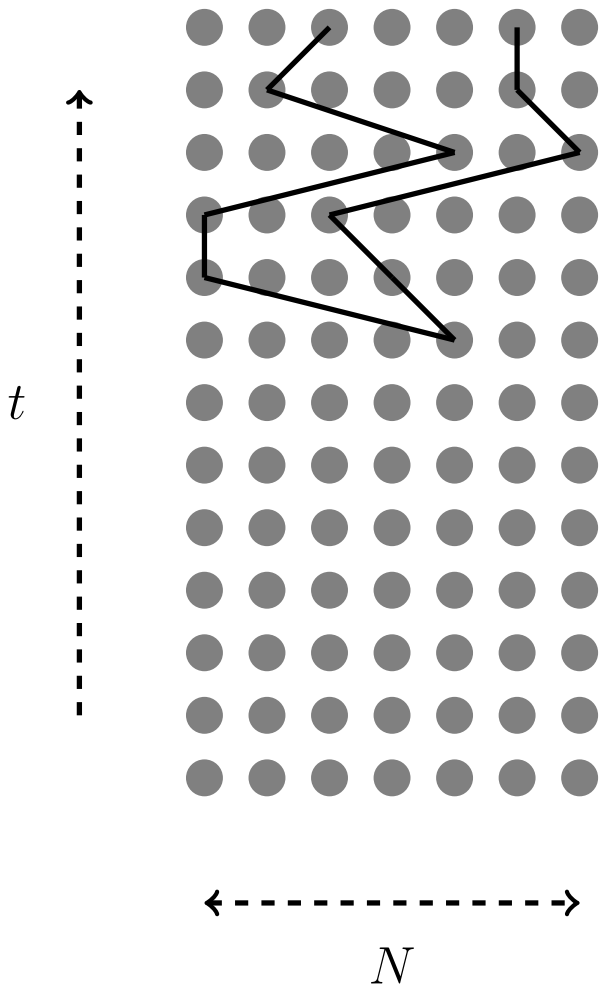
$$\left(1-\frac{1}{N}\right)\times\left(1-\frac{1}{N}\right)\times\cdots=\left(1-\frac{1}{N}\right)^{g}$$

- Followed by probability of coalescence:

$$\frac{1}{N}$$

$$P(g+1|N)=\frac{1}{N}\left(1-\frac{1}{N}\right)^{g}$$

Coalescent Trees



This is the *geometric* distribution:

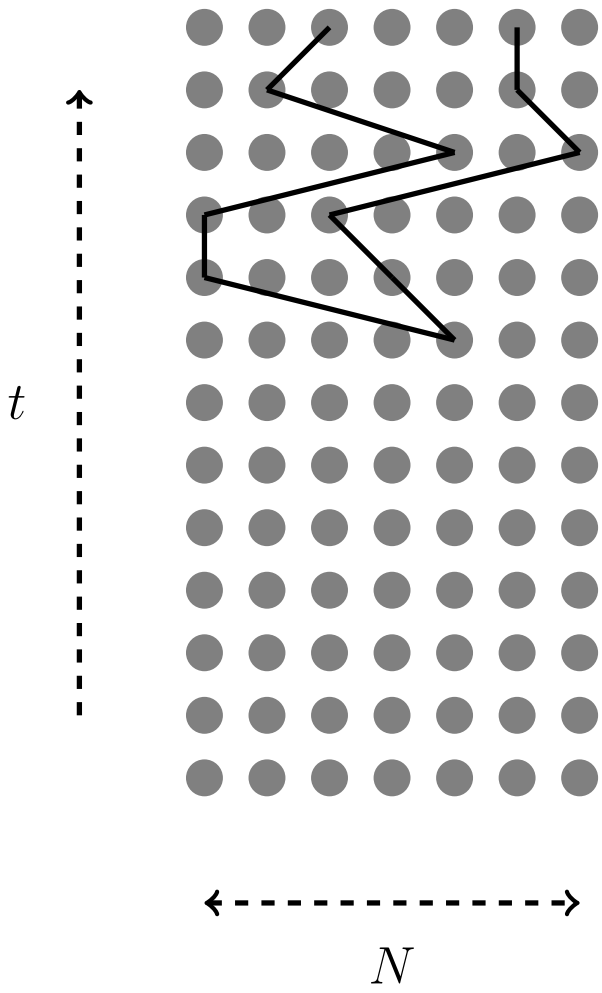
$$= \frac{1}{N}\left(1 - \frac{1}{N}\right)^g$$

It describes the time until the first success in a series of independent trials, where the probability of success is p and the probability of failure is $(1 - p)$.

mean wait time $= 1/p = N$

The expected time until coalescence is N generations!

Coalescent Trees



- The probability of a coalescence among n sampled lineages:

$$\frac{\binom{n}{2}}{N}$$

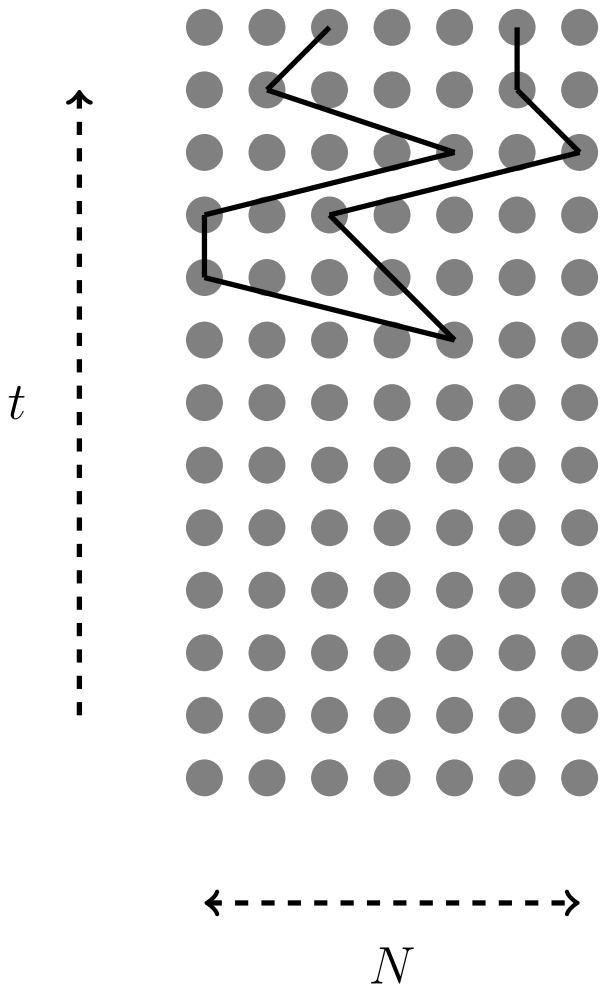
- n choose 2 is the number of ways the coalescent event could have occurred:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!}$$

- So now our probability is:

$$P(g+1|N,n) = \frac{\binom{n}{2}}{N}\left(1 - \frac{\binom{n}{2}}{N}\right)^g$$

Coalescent Trees



- The geometric distribution is a discrete distribution.
- The exponential distribution is the equivalent continuous distribution:

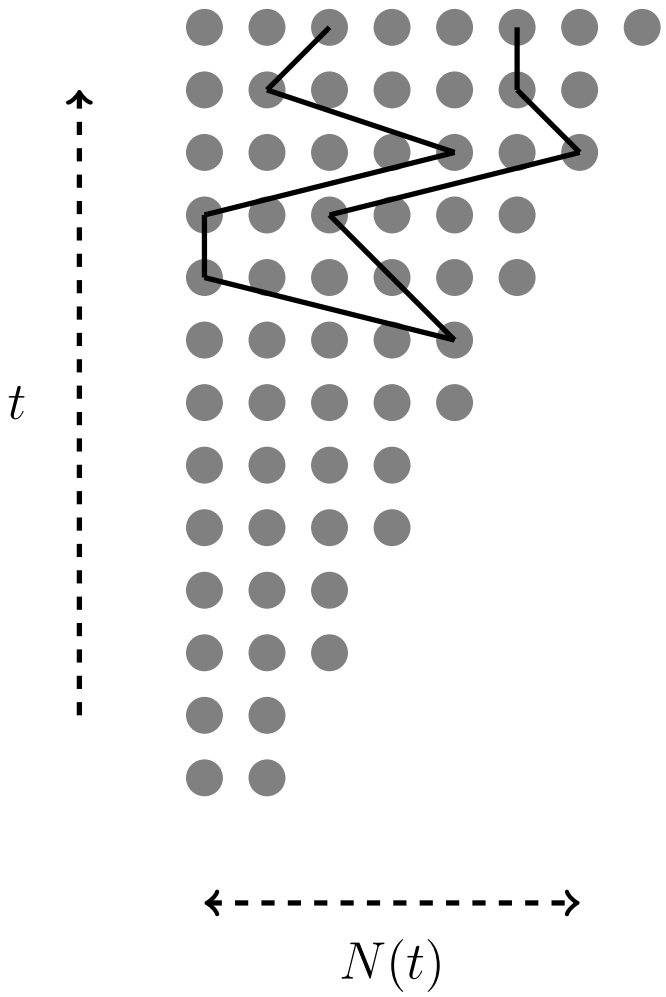
$$\lambda e^{-\lambda t}$$

- Instead of discrete generations, we now use continuous time.
- Now the coalescent process converges to a continuous time Markov process with instantaneous rate of coalescence:

$$\lambda = \frac{\binom{n}{2}}{N}$$

$$f(t|N,n) = \frac{\binom{n}{2}}{N} e^{-\frac{\binom{n}{2}}{N} t}$$

Coalescent Trees



- We have assumed a constant population size N.
- Instead we can specify a function that describes a changing population size through time:

$$N \to N(t)$$

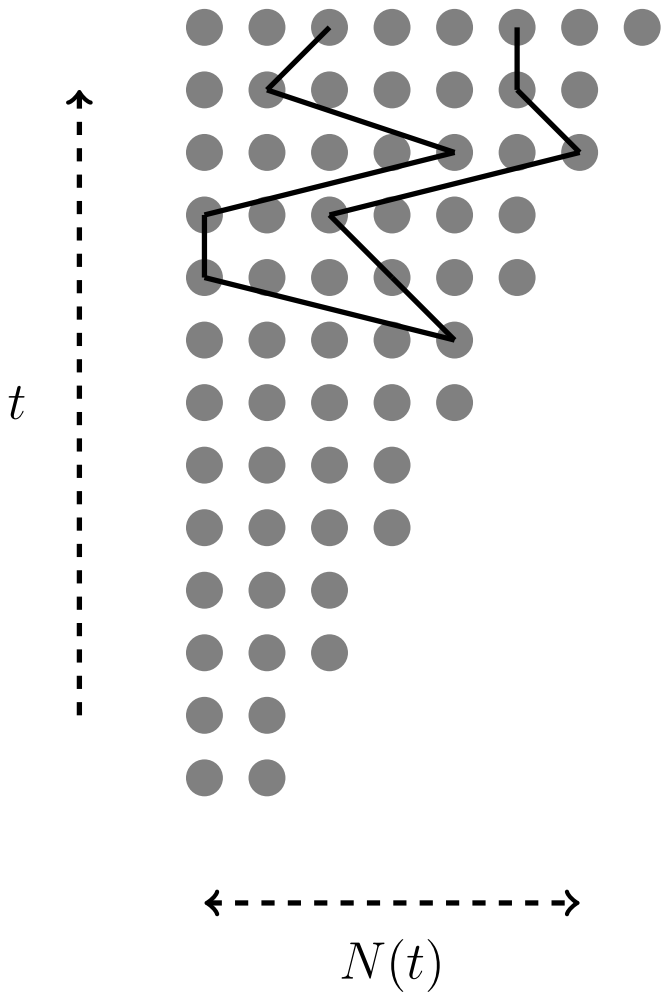
- Now the rate of coalescence is a function of t:

$$\frac{\binom{n}{2}}{N(t)},$$

so we must integrate with respect to t:

$$\frac{\binom{n}{2}}{N} e^{-\frac{\binom{n}{2}}{N}t} \to \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

Coalescent Trees

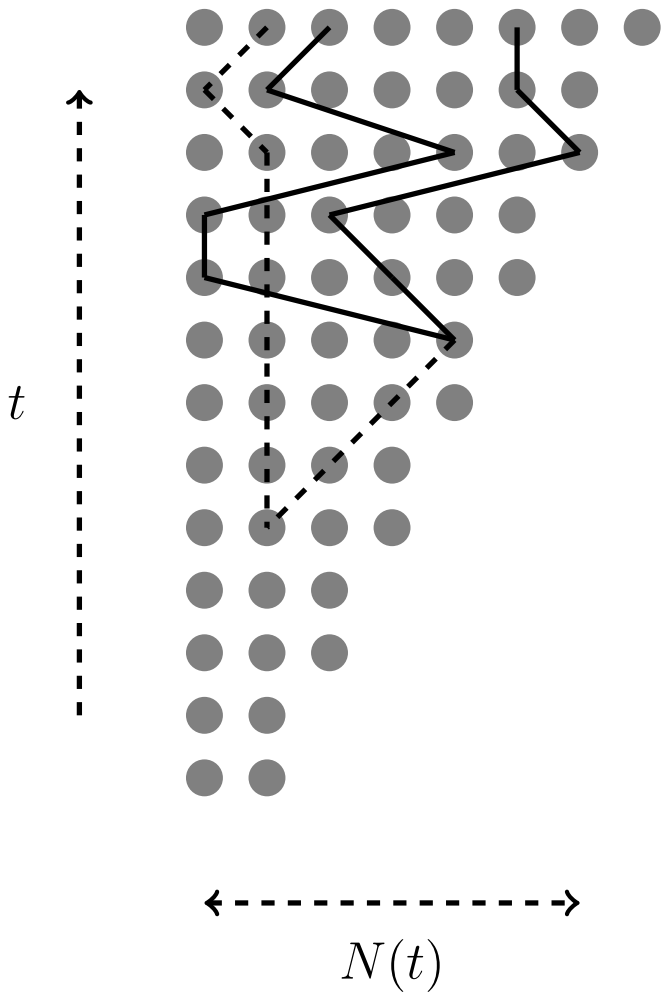


- So given a set of samples n and a demographic function $N(t)$ we know the time t of a coalescent event occurring has the distribution:

$$f(t|N(t),n) = \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

- But what about more than one coalescent event?

Coalescent Trees



- Define a list of i coalescent times:

$$\mathcal{C} = t_1, t_2, \ldots, t_i$$

- And finally:

$$f(\mathcal{C}|N(t), n) = \prod_{j=1}^{i} \frac{\binom{n}{2}}{N(t_j)} \exp\left(-\int_0^{t_j} \frac{\binom{n}{2}}{N(t)} dt\right)$$

- This gives us the probability density of a coalescent tree (a labeled history) within a lineage.
- It relates:
  1. the population size, to the
  2. the times of coalescent events

Coalescent Trees

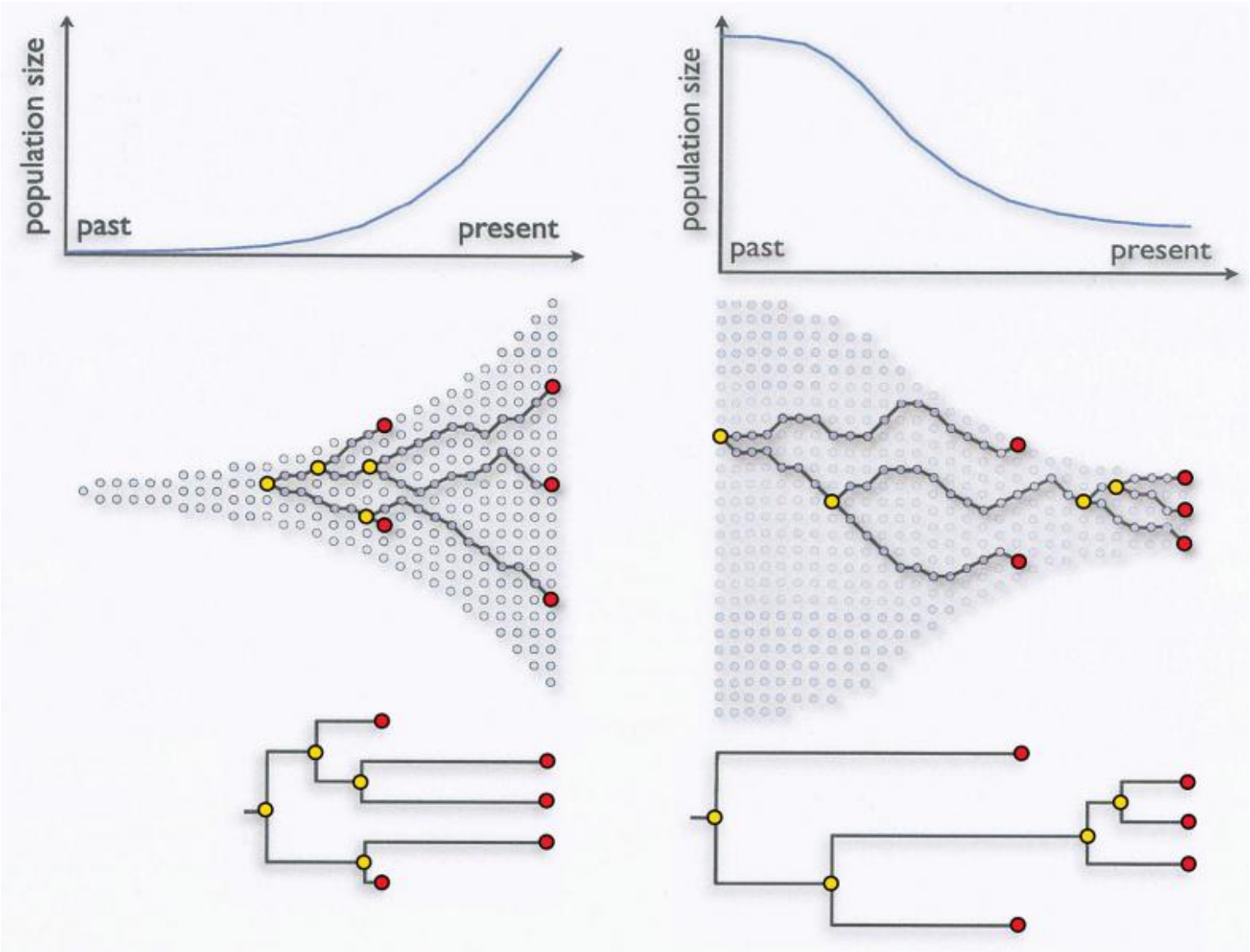
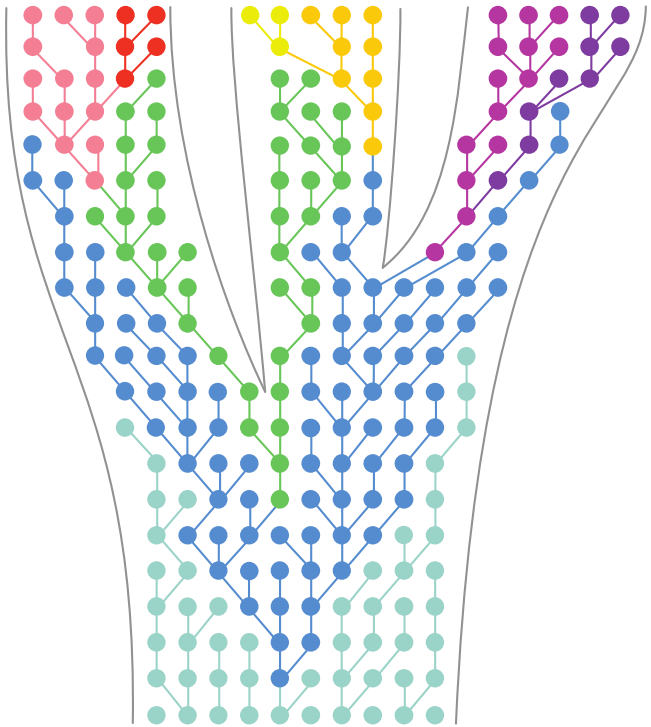


Image from Sainani (2009)

Coalescent Trees



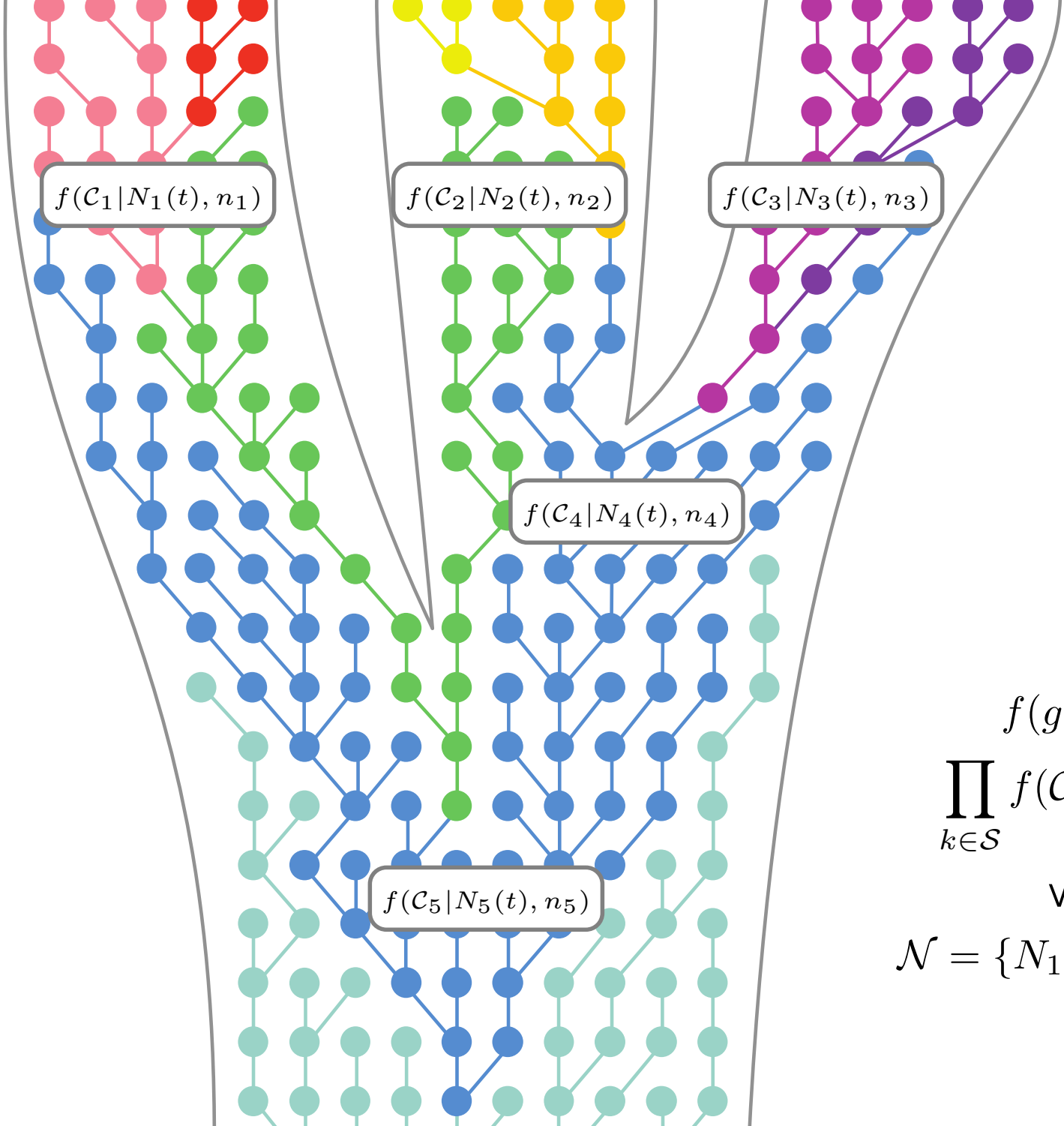
How can we link coalescent theory and phylogenetic theory?

- Each branch of the phylogeny is a lineage.
- We already derived the probability of a coalescent history within a single branch:

$$f(\mathcal{C}|N(t),n) = \prod_{j=1}^{i} \frac{\binom{n}{2}}{N(t_j)} \exp\left(-\int_0^{t_j} \frac{\binom{n}{2}}{N(t)} dt\right)$$

- The probability density of the coalescent history of a "gene tree" embedded within a "species tree" is the product of the coalescent probabilities for each branch...

Image from Leliaert et al. (2014)



$$f(g|\mathcal{S}, \mathcal{N}) = \prod_{k \in \mathcal{S}} f(\mathcal{C}_k|N_k(t), n_k)$$

where

$$\mathcal{N} = \{N_1(t), \ldots, N_k(t)\}$$

Coalescent Trees

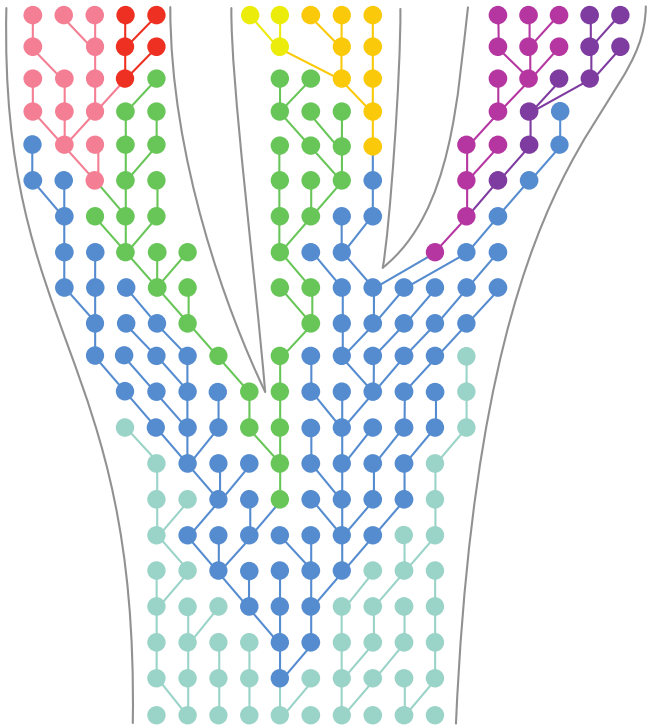


Image from Leliaert et al. (2014)

Now we have everything we need to describe the *multispecies coalescent*:

$$f(\mathcal{S},\mathcal{N}|\mathcal{D}) = \frac{\prod_{i=1}^{n} f(d_i|g_i) f(g_i|\mathcal{S},\mathcal{N}) f(\mathcal{S}) f(\mathcal{N})}{f(\mathcal{D})}$$

This is the fully parameterized model as implemented in software like:

- `RevBayes`
- `*BEAST`
- `BPP`

Since the model is computationally intensive there are many methods that approximate it like `SVDQuartets` and `ASTRAL`.

Coalescent Trees

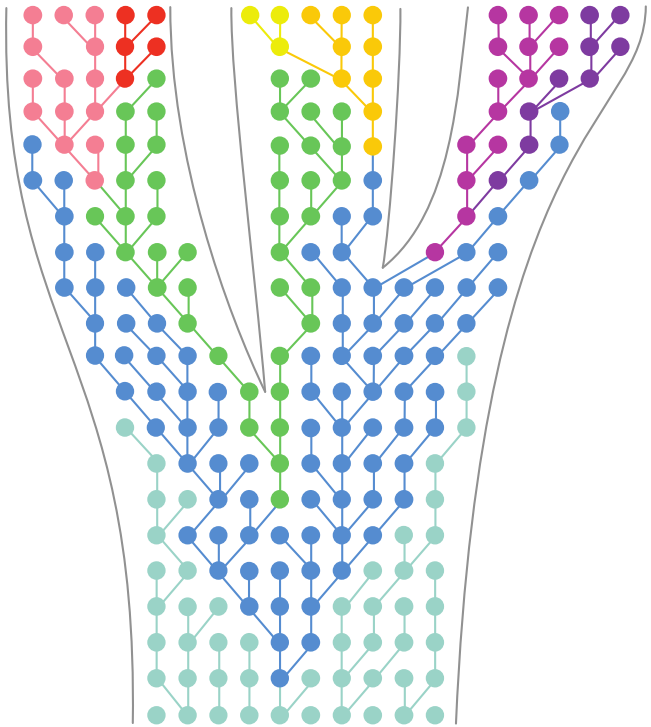


Image from Leliaert et al. (2014)

The multispecies coalescent:

$$f(\mathcal{S},\mathcal{N}|\mathcal{D}) = \frac{\prod_{i=1}^{n} f(d_i|g_i) f(g_i|\mathcal{S},\mathcal{N}) f(\mathcal{S}) f(\mathcal{N})}{f(\mathcal{D})}$$

$f(d_i|g_i) =$ Felsenstein likelihood for gene alignment given a gene tree

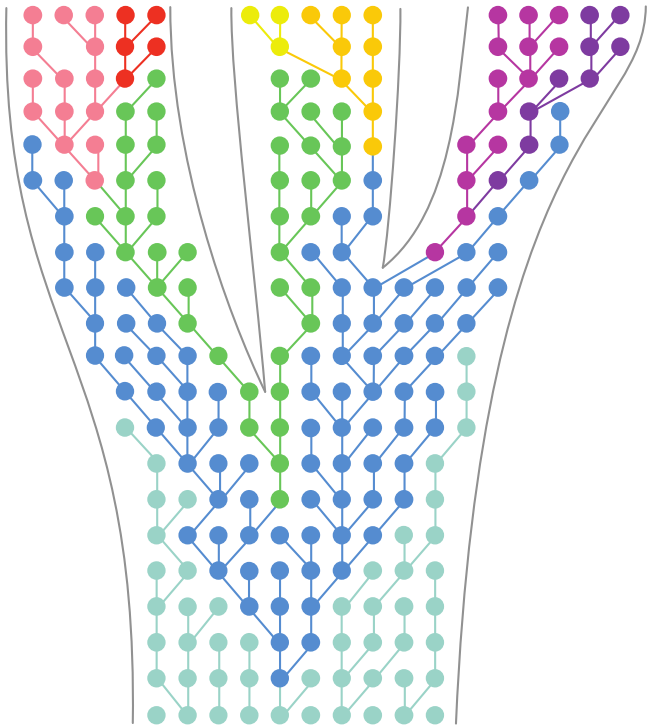
$f(g_i|\mathcal{S},\mathcal{N}) =$ coalescent probability of gene tree given species tree

$f(\mathcal{S}) =$ prior probability of species tree

$f(\mathcal{N}) =$ prior probability of population sizes

$f(\mathcal{D}) =$ marginal likelihood

Coalescent Trees



What sort of prior could we use for the species tree?

$$f(\mathcal{S}) = ?$$

Birth-death process!

Image from Leliaert et al. (2014)

Birth-Death Processes

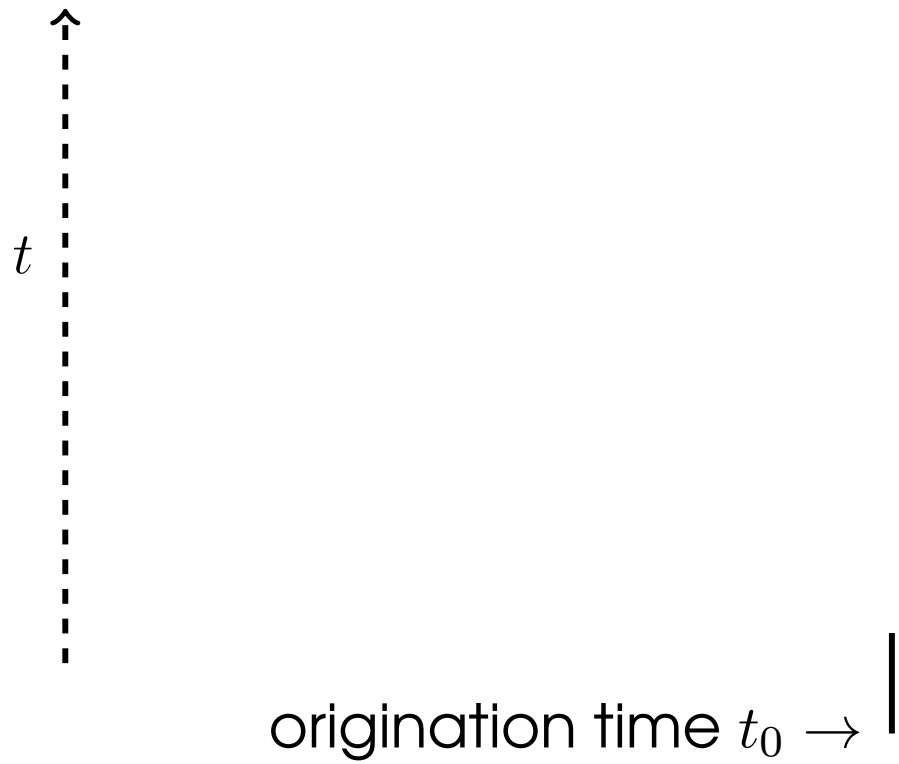
- A species gives birth to a new species with rate λ
- A species goes extinct with rate μ
- This is a continuous-time Markov process with the rate matrix:

$$Q_{ij} = \begin{cases} i\lambda & j = i+1, i \geq 1, \\ i\mu & j = i-1, i \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

- How many states does it have?
- How are the times between events distributed in a Markov process?

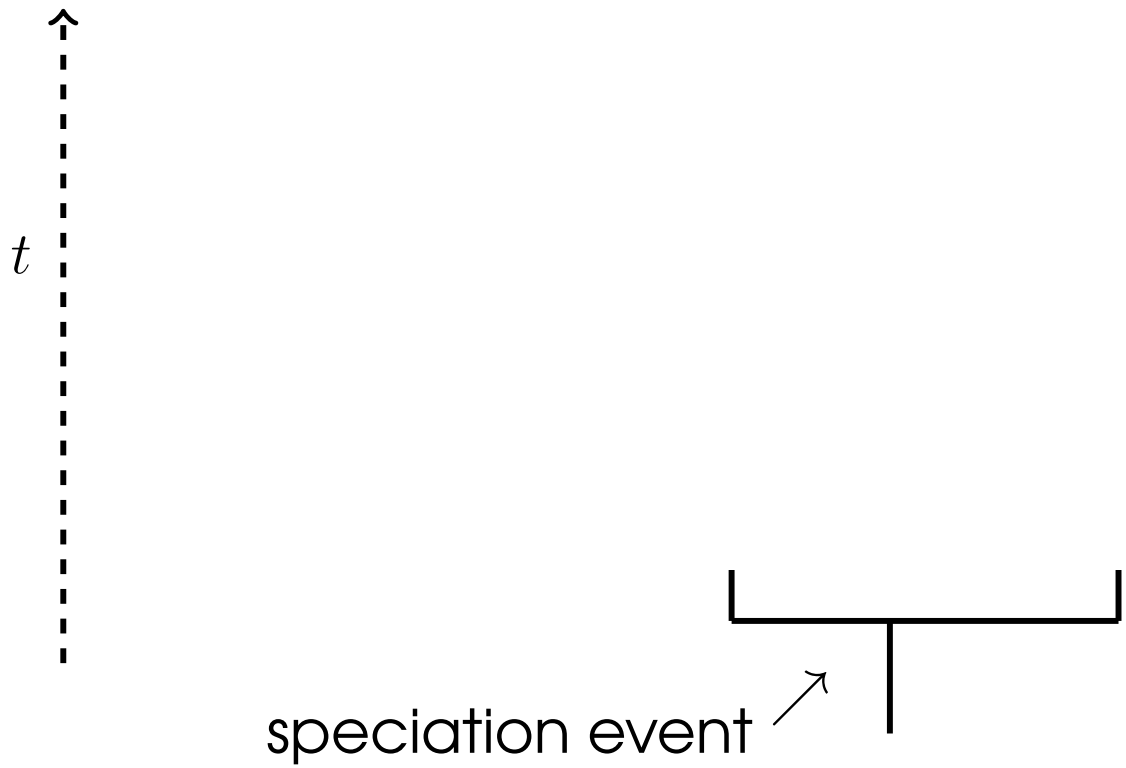
Birth-Death Processes

Now we can simulate a tree using the birth-death process:



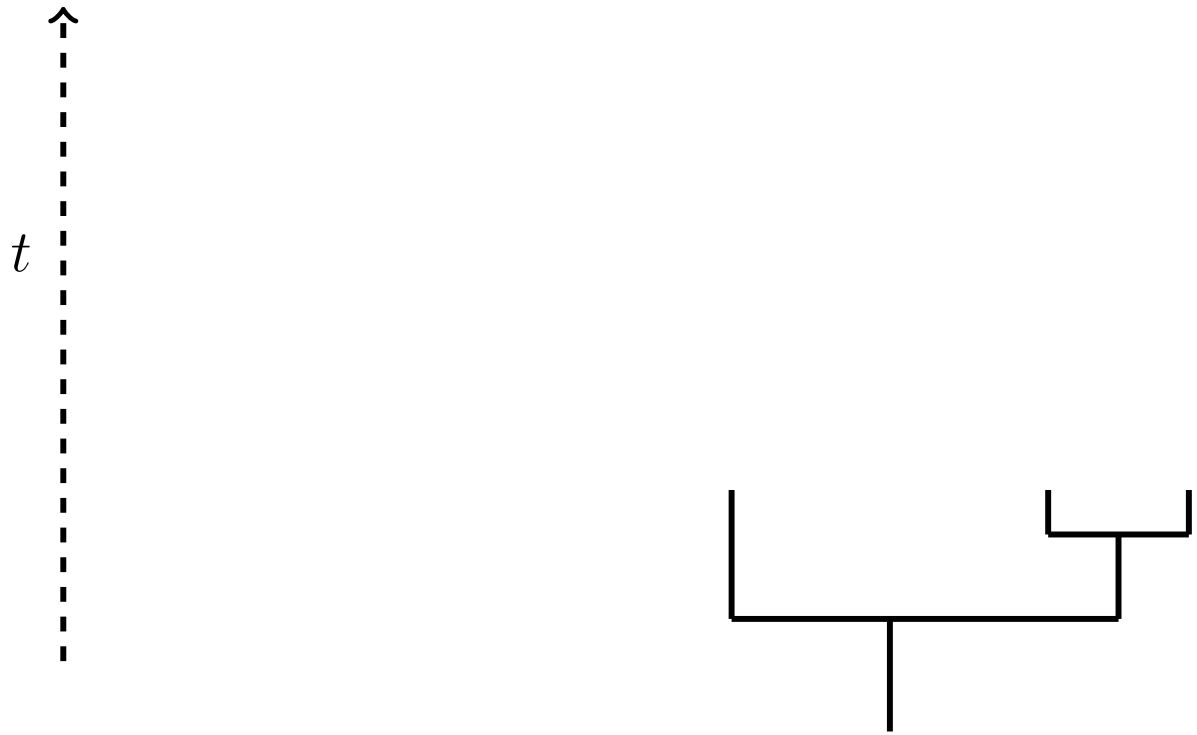
Birth-Death Processes

Now we can simulate a tree using the birth-death process:



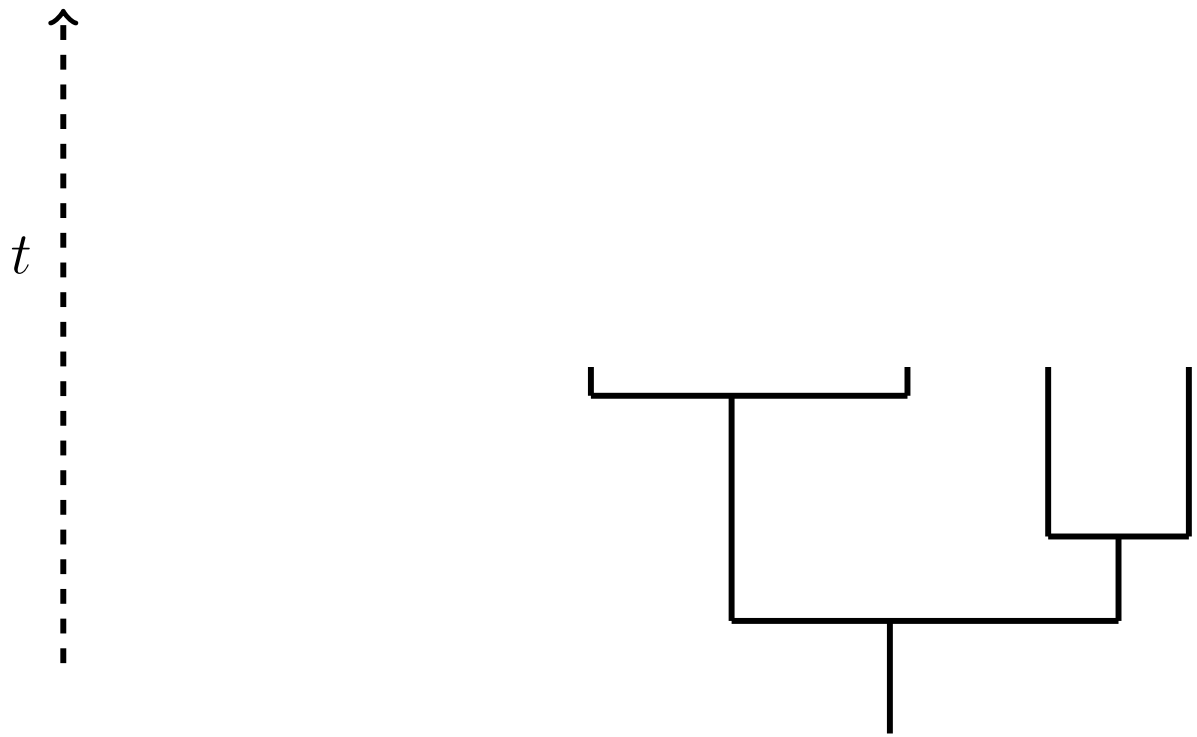
Birth-Death Processes

Now we can simulate a tree using the birth-death process:



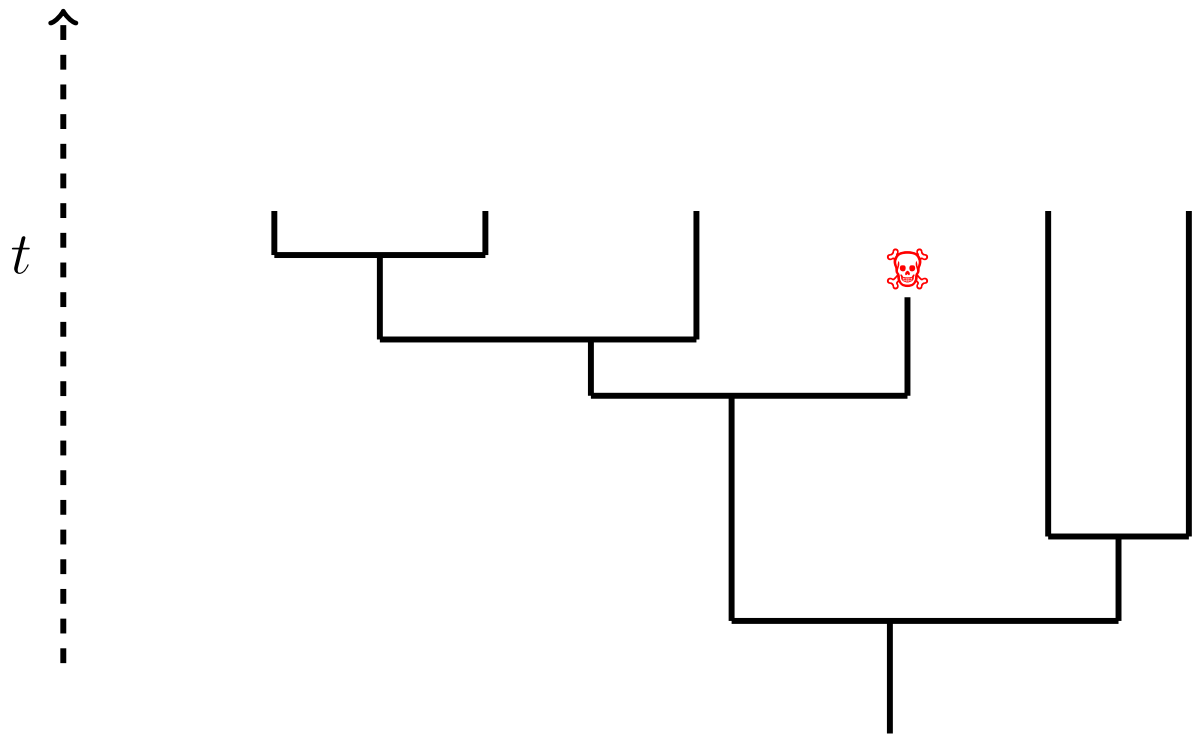
Birth-Death Processes

Now we can simulate a tree using the birth-death process:



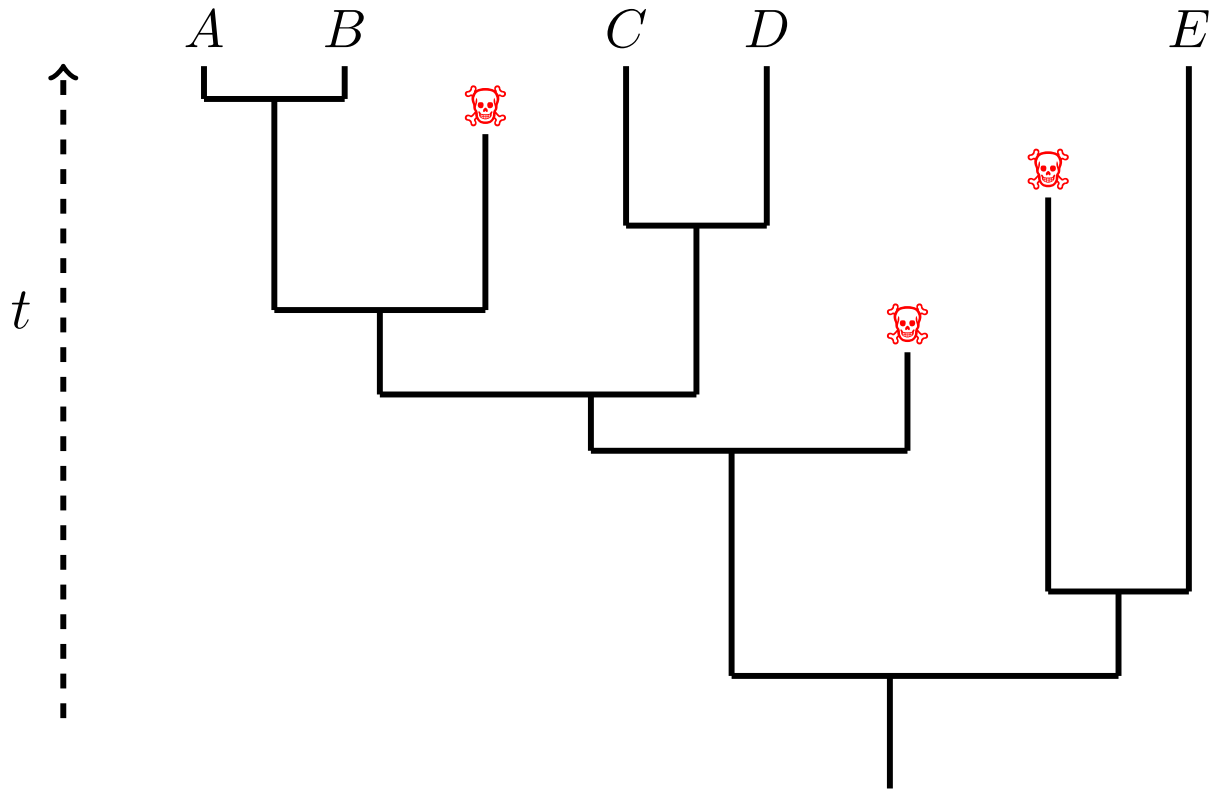
Birth-Death Processes

Now we can simulate a tree using the birth-death process:



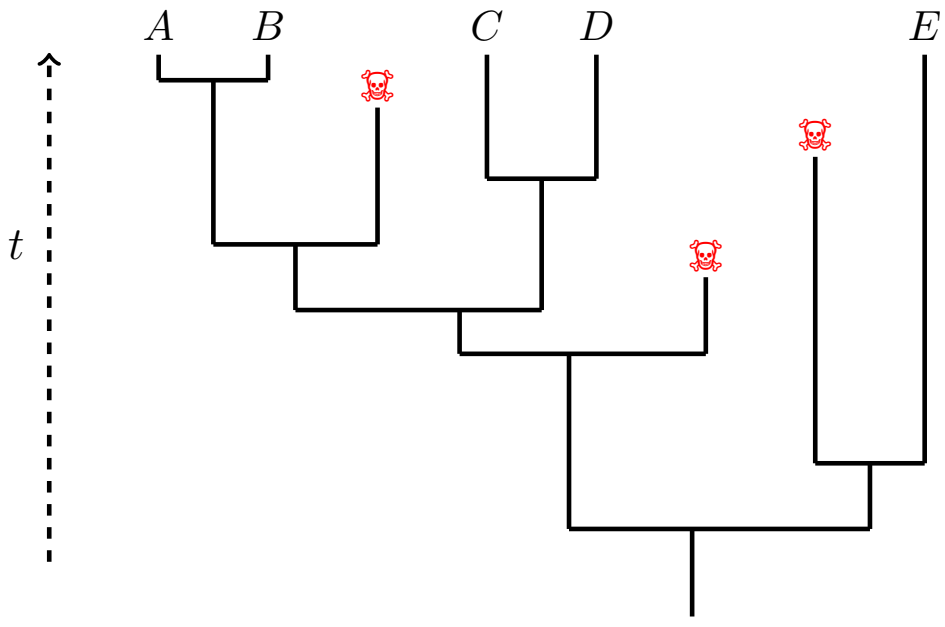
Birth-Death Processes

Complete simulated tree with 5 extant lineages:

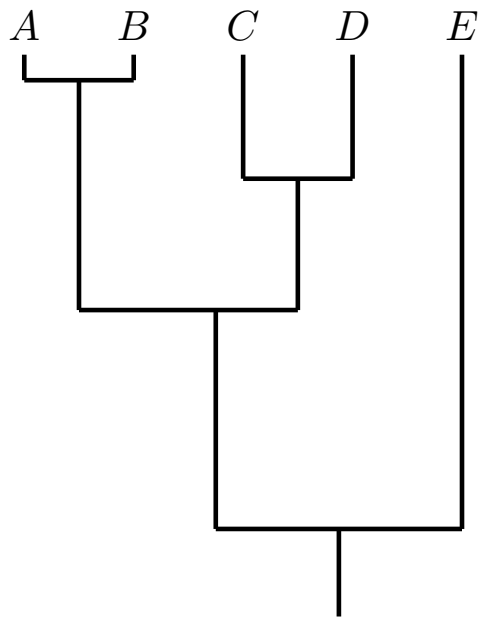


Birth-Death Processes

Complete Tree

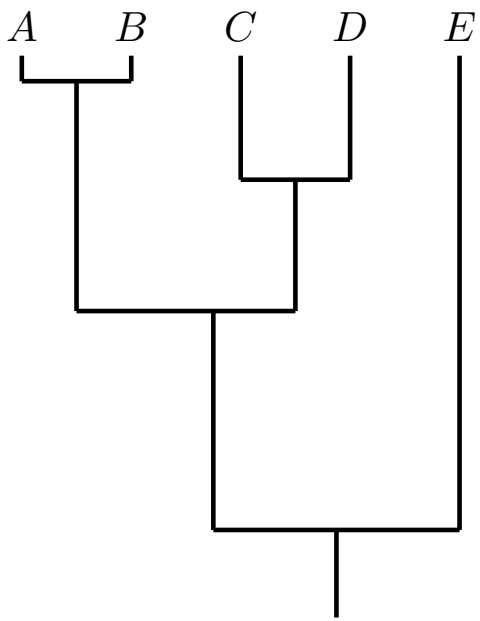


Reconstructed Tree



Birth-Death Processes

Reconstructed Tree



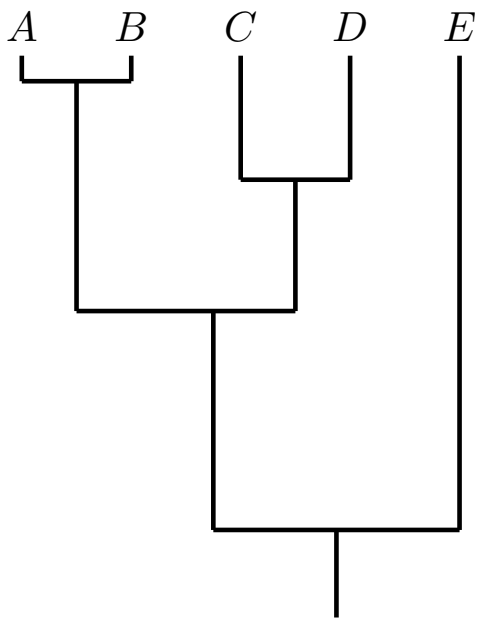
The probability density function of the set of speciation times $\mathcal{T}$ of a reconstructed tree under the constant-rate birth-death process is:

$$f(\mathcal{T}|N(t_0)=1,\lambda,\mu) = (n_p-1)!\lambda^{n_p-1}\frac{r^3 e^{-r(t_p-t_0)}}{(re^{-r(t_p-t_0)})^3} \times \prod_{i=1}^{n_p-1}\frac{r^2 e^{-r(t_p-t_i)}}{(re^{-r(t_p-t_i)})^2}$$

where $r=\lambda-\mu$, n_p is the number of lineages that survived to the present, t_p is the time at the present, and conditioned on there being one lineage at the origination time t_0.

Birth-Death Processes

Reconstructed Tree



Now we can use the probability density function $f(\mathcal{T}|N(t_0) = 1, \lambda, \mu)$ to estimate divergence times and speciation/extinction rates.

In a Bayesian setting we must specify our priors. A few possible parameterizations:

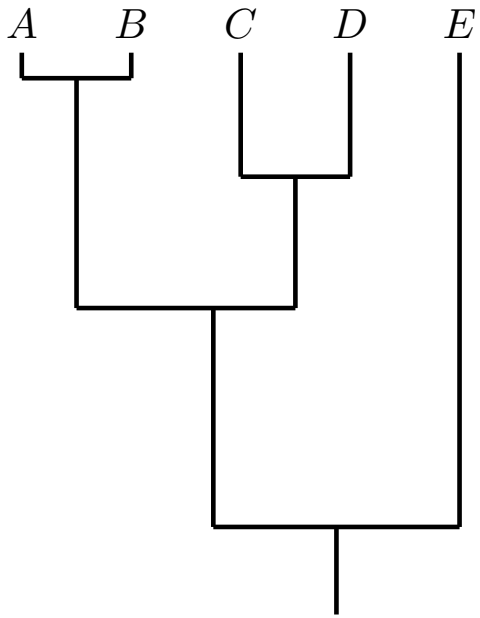
1.
   - Speciation rate: λ
   - Extinction rate: μ

2.
   - Speciation rate: λ
   - Turnover rate: μ/λ

3.
   - Net-diversification rate: $\lambda - \mu$
   - Turnover rate: μ/λ

Birth-Death Processes

Reconstructed Tree



- What are reasonable values for the priors?
- We have good prior information about net-diversification:

$$E[\lambda - \mu] = \ln\left(\frac{n_p}{n_0}\right)/t_0$$

- If we assume speciation is greater than extinction (not always a good assumption):

$$\mu/\lambda \sim \text{Beta}(1, 1)$$

Birth-Death Processes

Divergence Time Estimation

Node calibrations:

- Normal distribution
- Lognormal distribution
- Exponential distribution
- Uniform distribution w/ hard min & soft max
- Uniform distribution w/ hard min & hard max
- Point value
- Fossilized birth-death

Tip calibrations:

- Empirical calibrated radiocarbon sampler
- Normal distribution
- Uniform distribution w/ hard min & max
- Point value

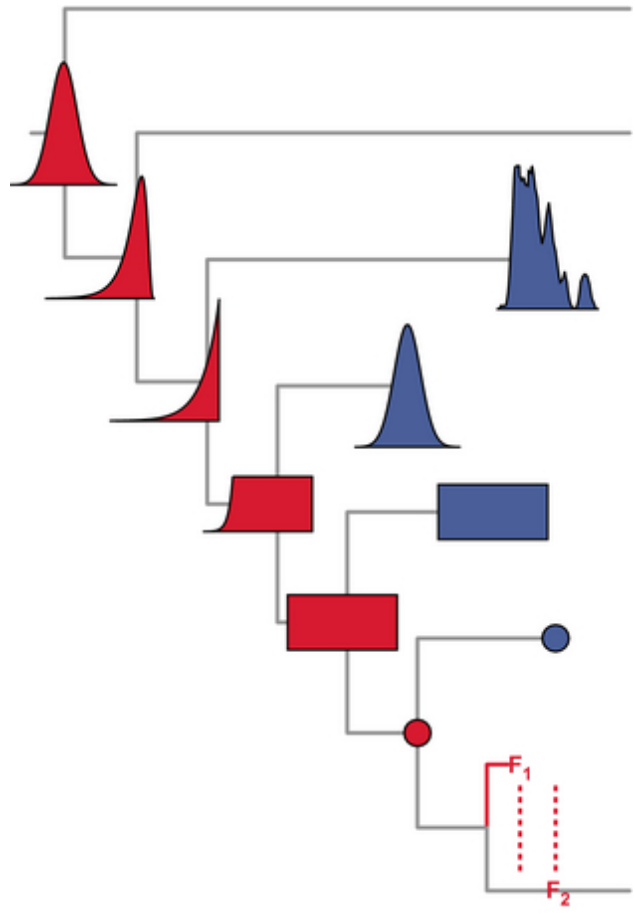


Image from Ho & Duchêne (2014)

Birth-Death Processes

Diversification rate estimation:

1. Constant diversification rates
2. Diversification rates through time
3. Character-dependent diversification rates
4. Branch-specific diversification rates

Birth-Death Processes

Diversification rate estimation:

1. ~~Constant diversification rates~~
2. Diversification rates through time
3. Character-dependent diversification rates
4. Branch-specific diversification rates

Birth-Death Processes

Diversification rates through time

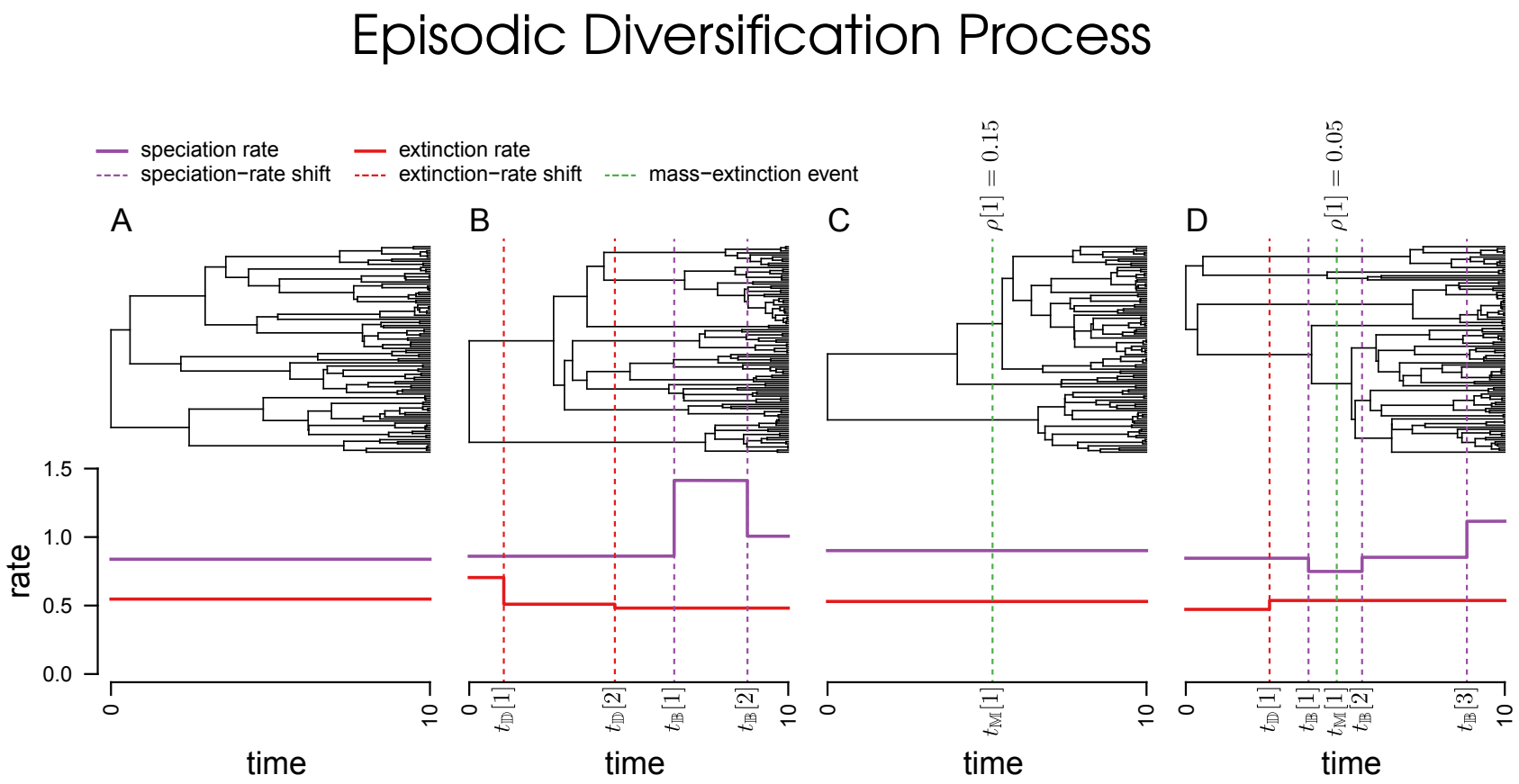
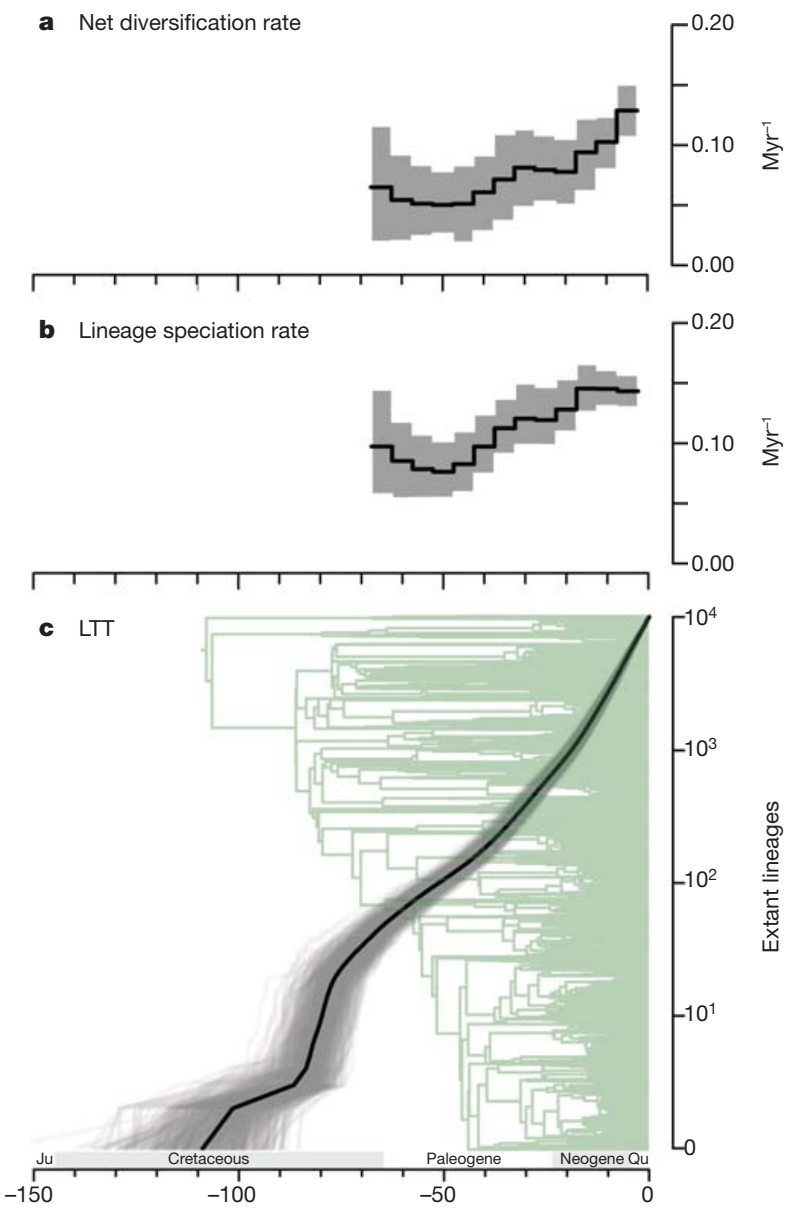


Image from Sebastian Höhna's slides (2017)

Birth-Death Processes

Diversification rates through time



Bird diversification by Jetz et al. (2012)

Diversification rates estimated in 5 million year intervals

Birth-Death Processes

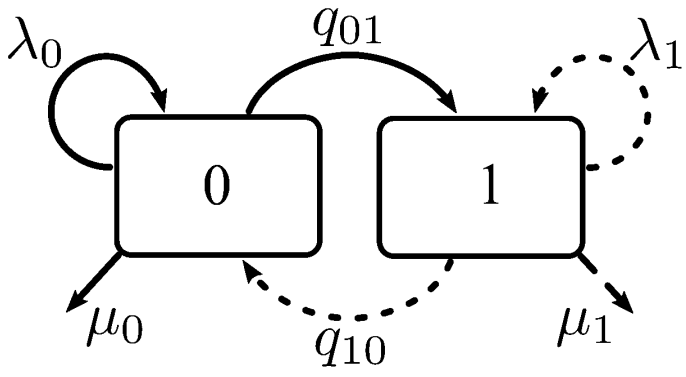
Diversification rate estimation:

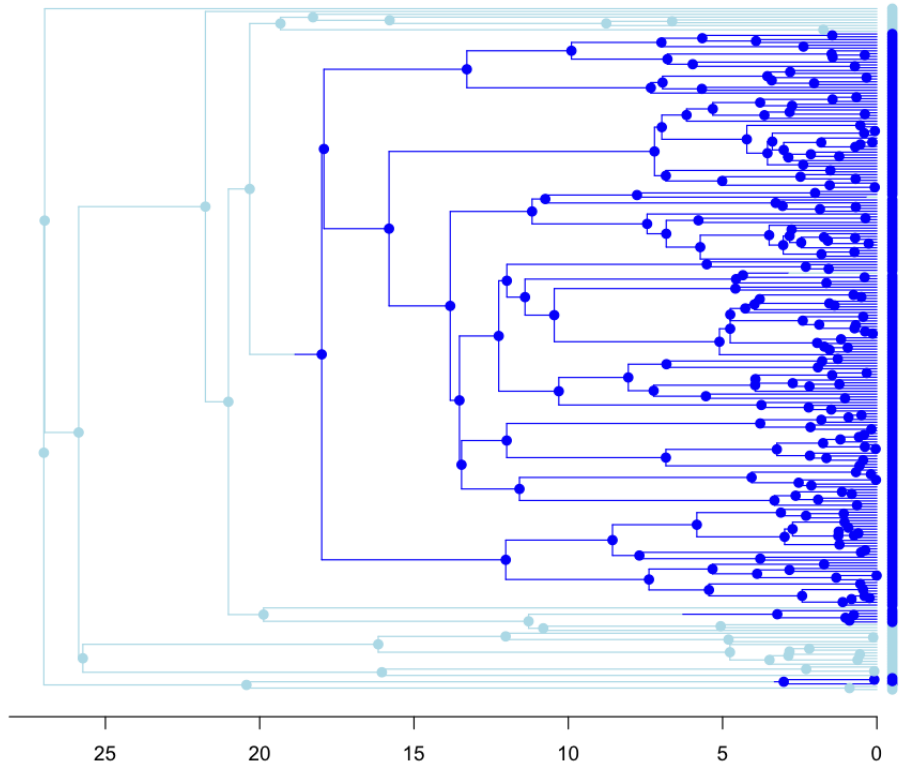
1. ~~Constant diversification rates~~
2. ~~Diversification rates through time~~
3. Character-dependent diversification rates
4. Branch-specific diversification rates

Birth-Death Processes

Character-dependent diversification rates

Joint Models of the Tree and Character Evolution





Binary State Speciation and Extinction (BiSSE) Model

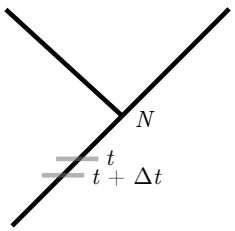
Birth-Death Processes

Character-dependent diversification rates

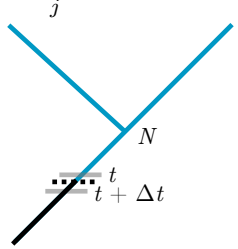
BiSSE, MuSSE, HiSSE, GeoSSE, ChromoSSE are all special cases of ClaSSE

a

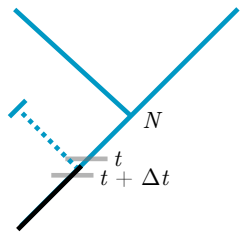
$$\frac{dD_{Ni}(t)}{dt} = -\left(\sum_j \sum_k \lambda_{ijk} + \sum_j Q_{ij} + \mu\right) D_{Ni}(t) + \sum_j Q_{ij} D_{Nj}(t) + \sum_j \sum_k \lambda_{ijk}\left(D_{Ni}(t)E_j(t) + D_{Nj}(t)E_i(t)\right)$$



no event occurred



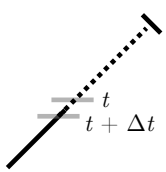
anagenetic change



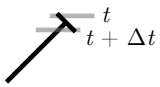
speciation followed by extinction w/ possible cladogenetic change

b

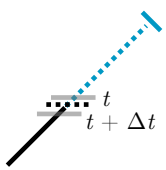
$$\frac{dE_i(t)}{dt} = -\left(\sum_j \sum_k \lambda_{ijk} + \sum_j Q_{ij} + \mu\right) E_i(t) + \mu + \sum_j Q_{ij} E_j(t) + \sum_j \sum_k \lambda_{ijk} E_j(t) E_k(t)$$



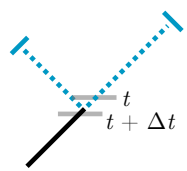
no event followed by extinction



extinction



anagenetic change followed by extinction

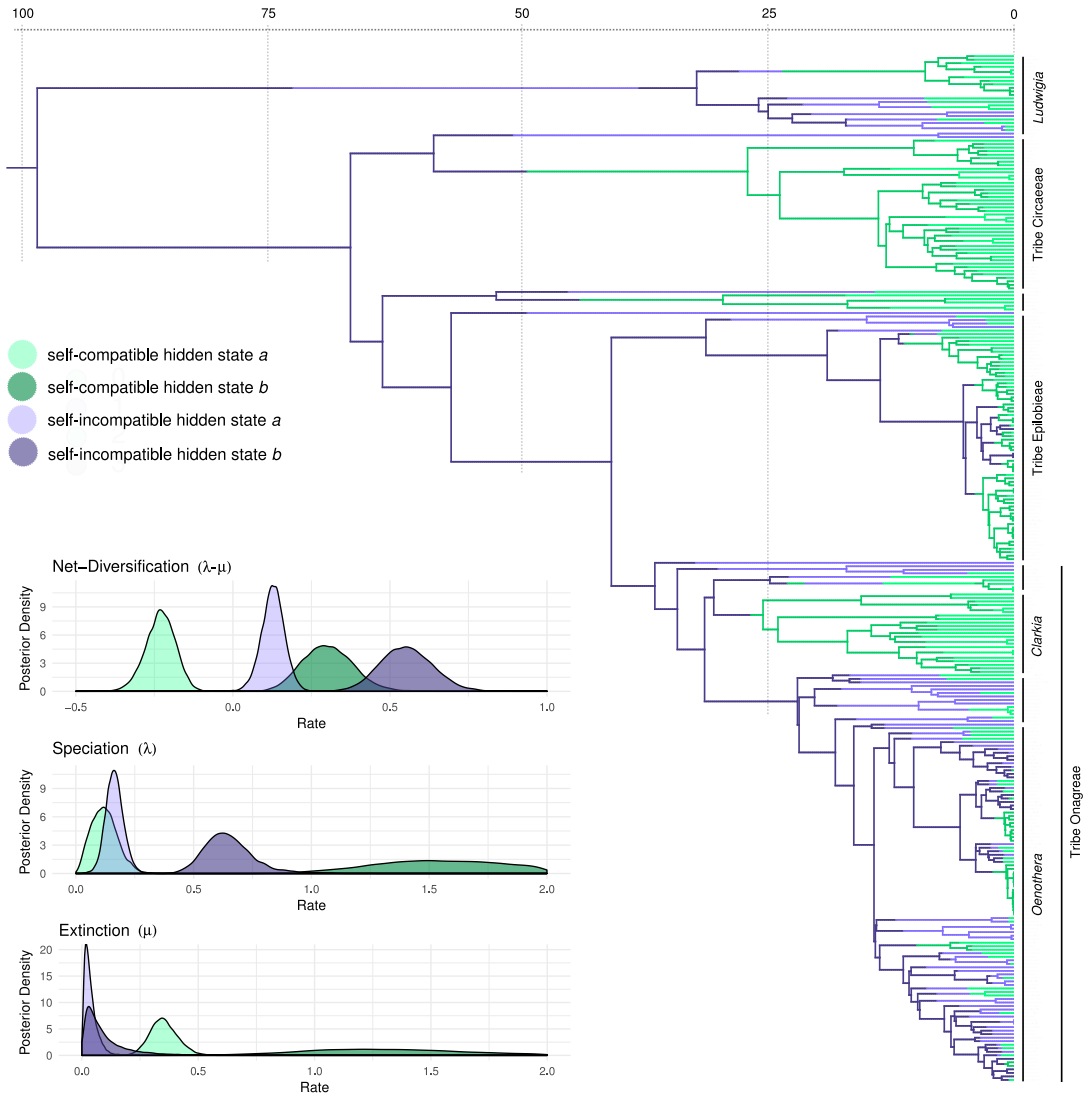


speciation followed by extinction w/ possible cladogenetic change

Image from Freyman & Höhna (2017)

Birth-Death Processes

Character-dependent diversification rates



Changes in mating system have different long and short term evolutionary consequences.

The time lag from the loss of self-incompatibility until the onset of evolutionary decline:

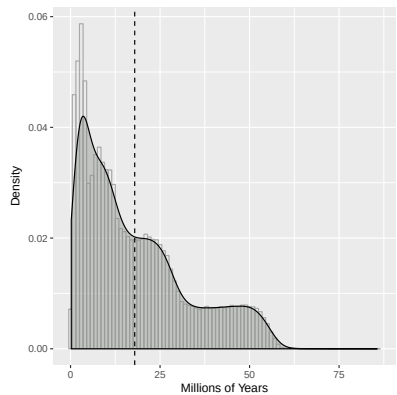


Image from Freyman & Höhna (in prep)

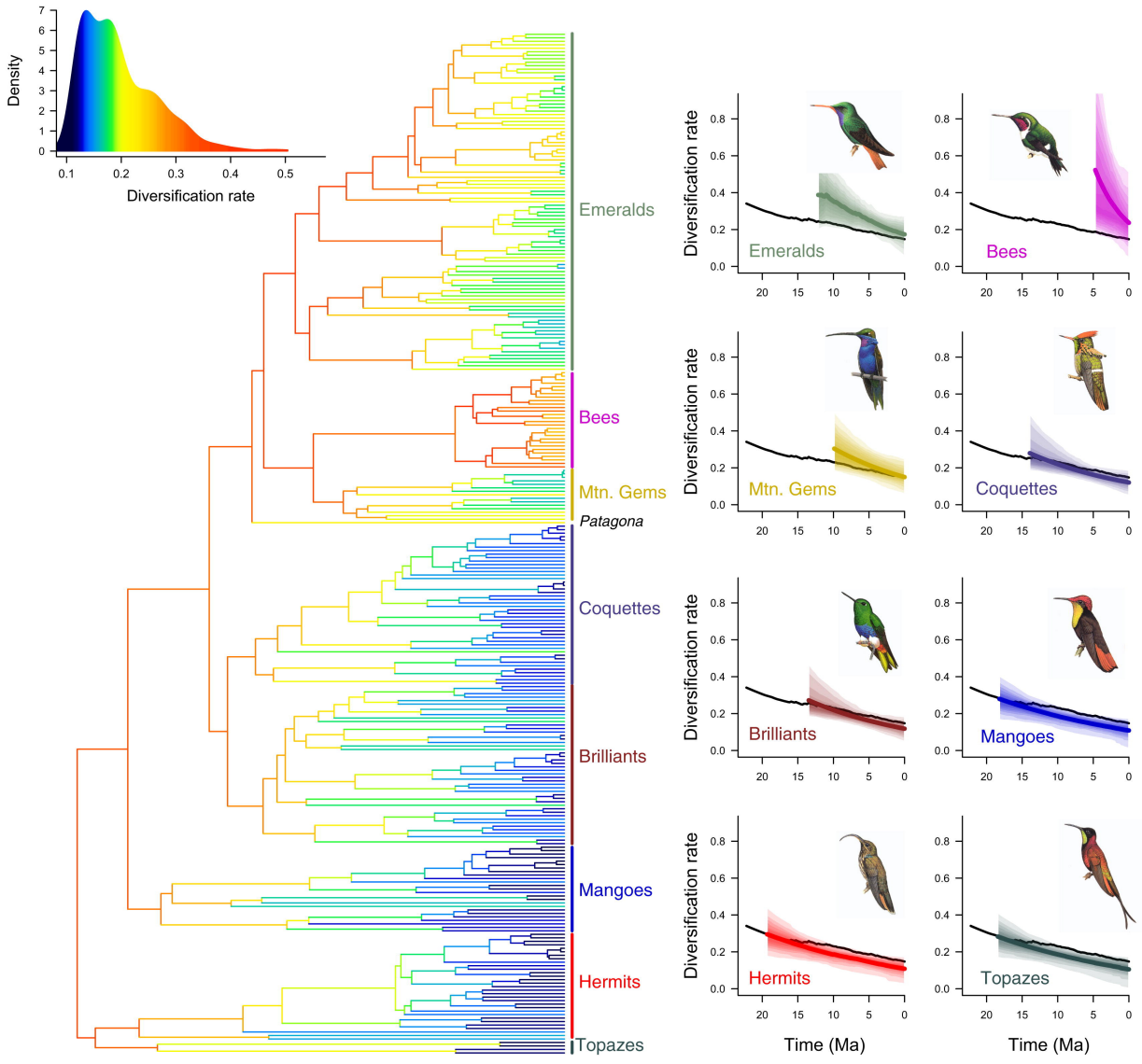
Birth-Death Processes

Diversification rate estimation:

1. ~~Constant diversification rates~~
2. ~~Diversification rates through time~~
3. ~~Character-dependent diversification rates~~
4. Branch-specific diversification rates

Birth-Death Processes

Branch-specific diversification rates



Hummingbird diversification by McGuire et al. (2014)

Diversification rates estimated using BAMM

Birth-Death Processes

Branch-specific diversification rates

Modeling issues in BAMM

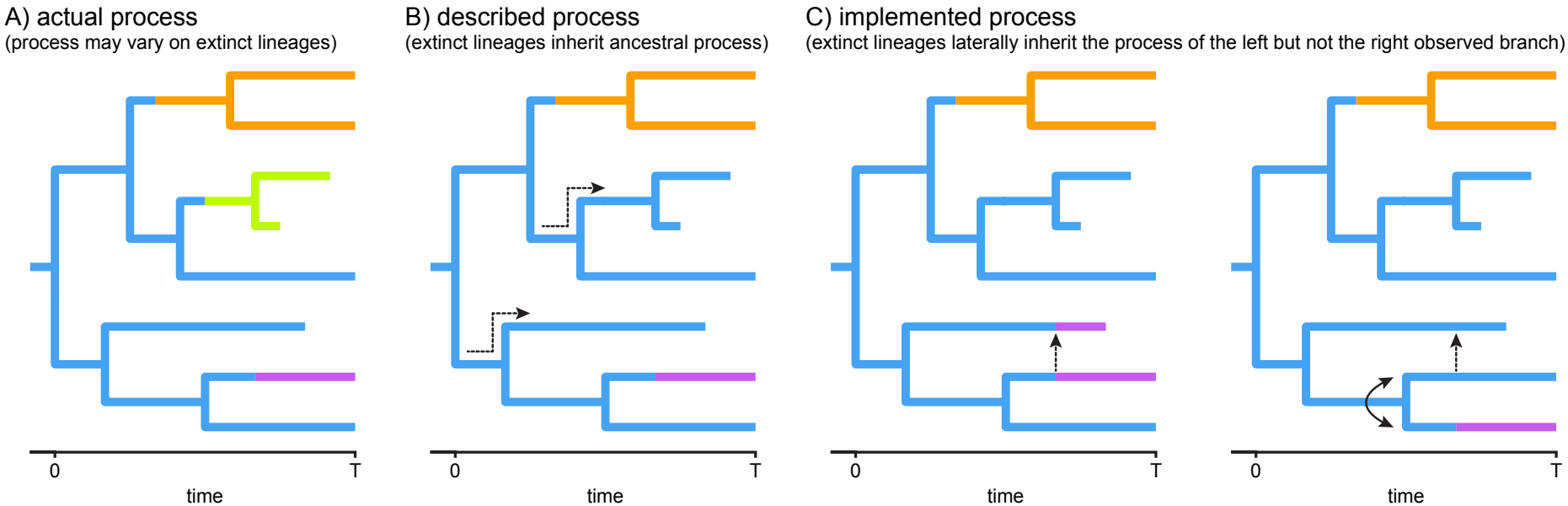
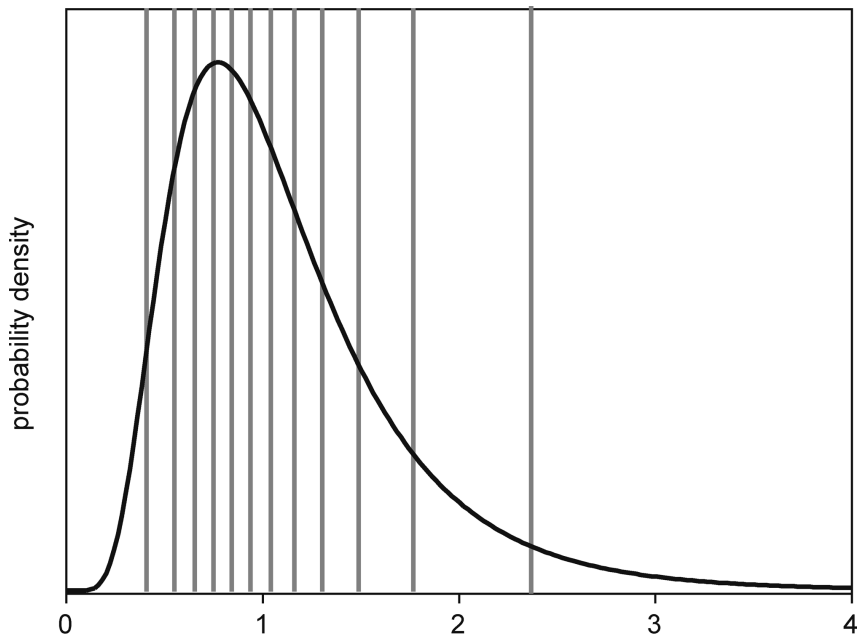


Image from Sebastian Höhna's slides (2017)

Birth-Death Processes

Branch-specific diversification rates



The PERSEUS solution:

Discretize speciation and extinction rates

Use MuSSE with all tip states (rate categories) unknown

Tree Models

Three approaches covered today:

1. Uniform Tree Topologies
2. Coalescent Trees
3. Birth-Death Processes

What about phylogenetic networks?!