

Coalescent and multispecies coalescent

Ziheng Yang

Department of Genetics, Evolution, and Environment

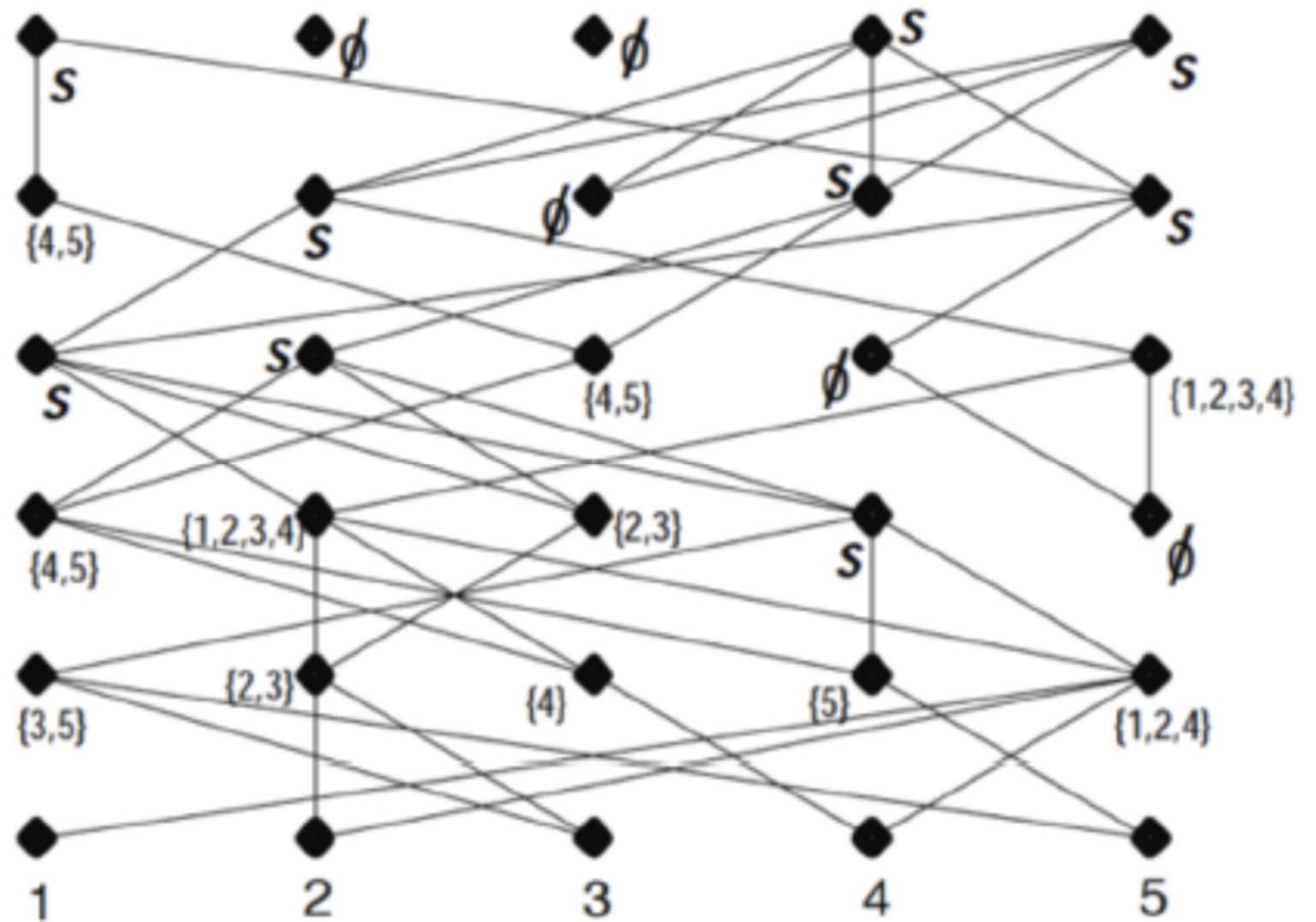
University College London

Outline

- Pedigree, genealogy and coalescent
- Multispecies coalescent (MSC)
- Inference under a fixed model



Bruce Rannala's pedigree



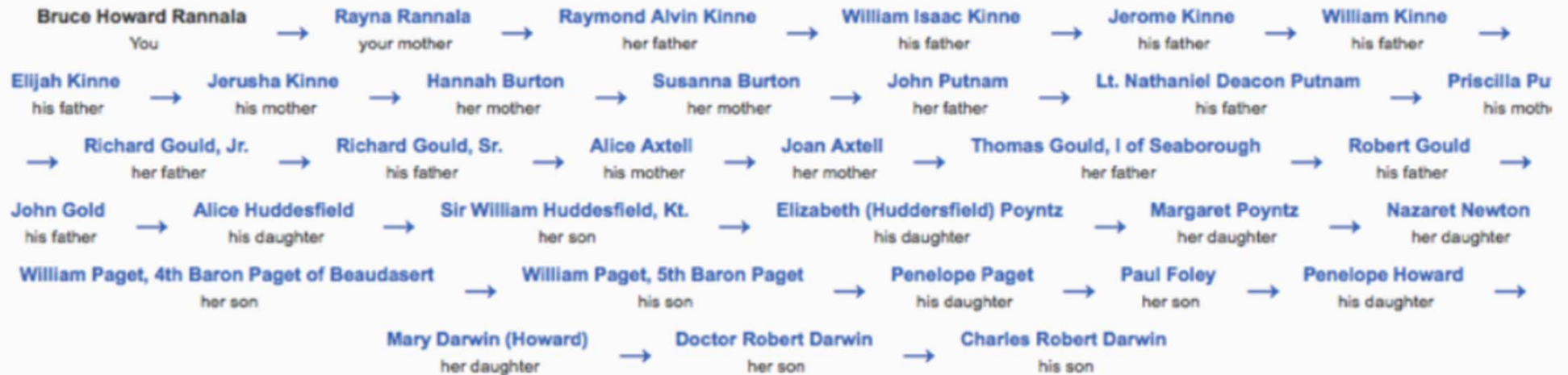


Bruce Howard Rannala

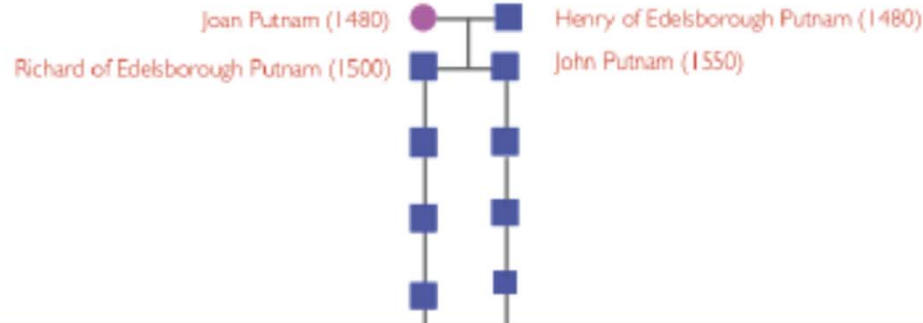


Charles Robert Darwin

Charles Robert Darwin is Bruce Howard Rannala's 12th cousin 6 times removed!

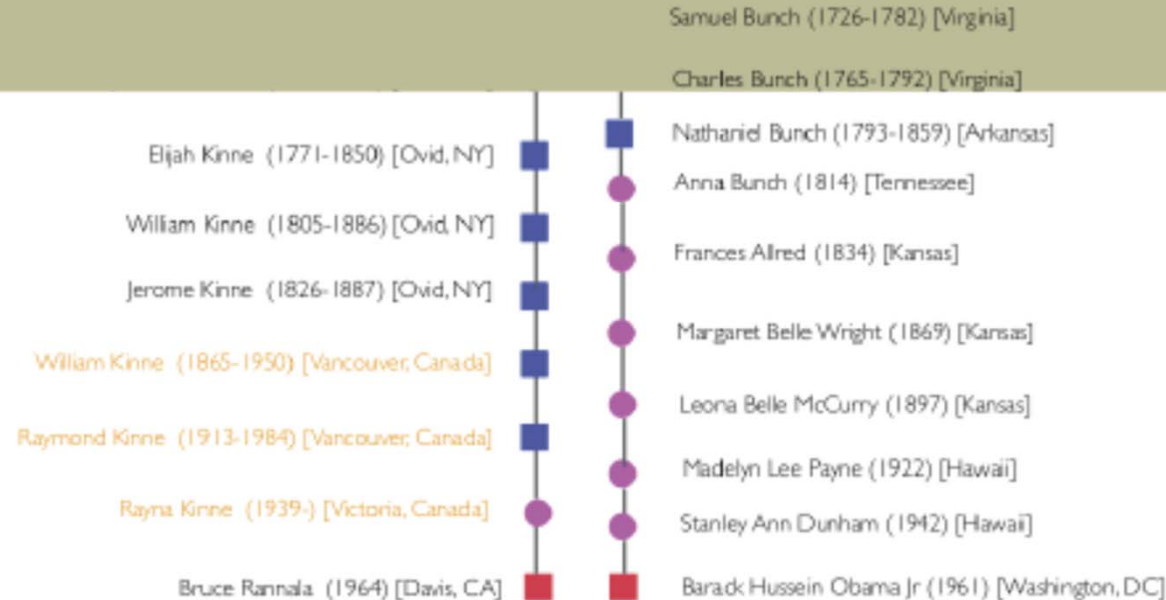


John Gould

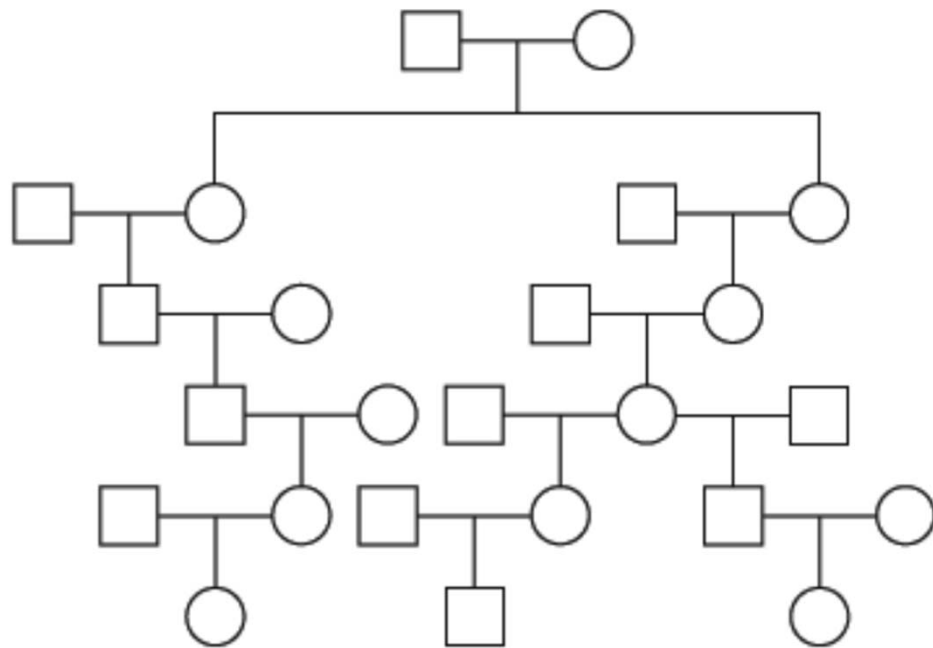


Kinship coefficient of Bruce Rannala and Barack Obama:

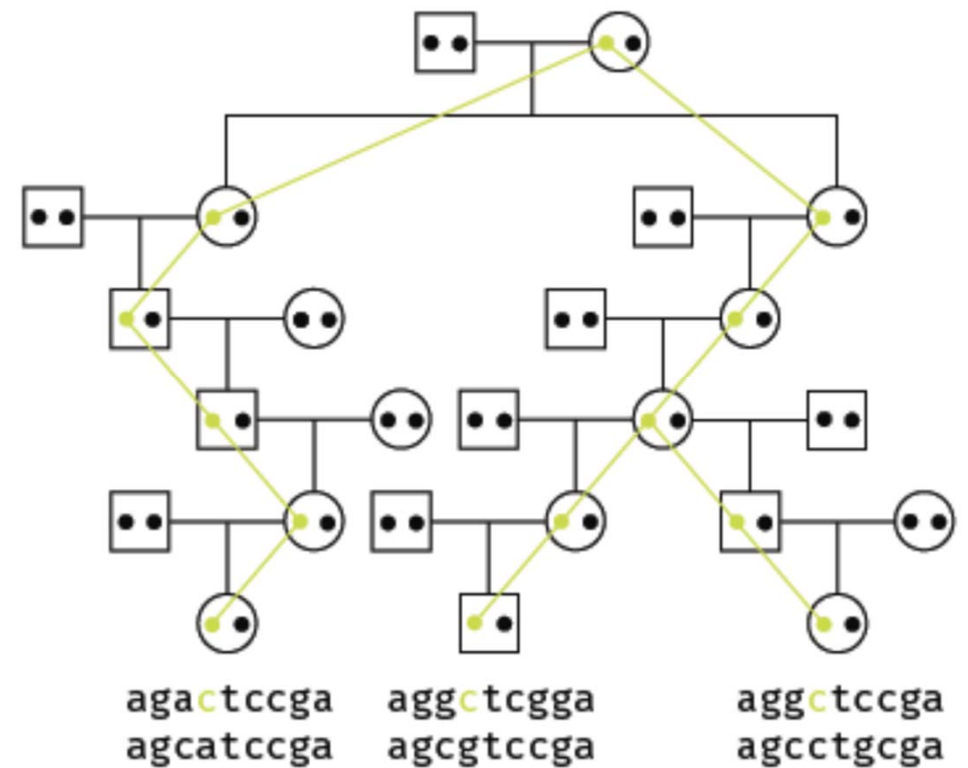
$$(1/2)^{34} = 5.8 \times 10^{(-11)}$$



Pedigree



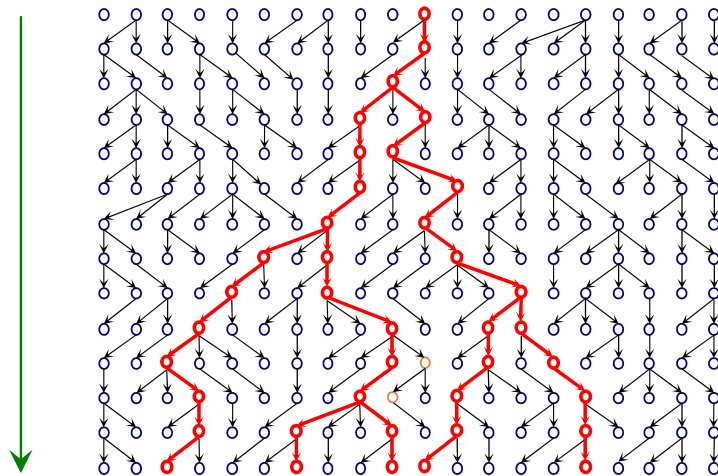
Gene tree within pedigree



Fisher-Wright model & coalescent

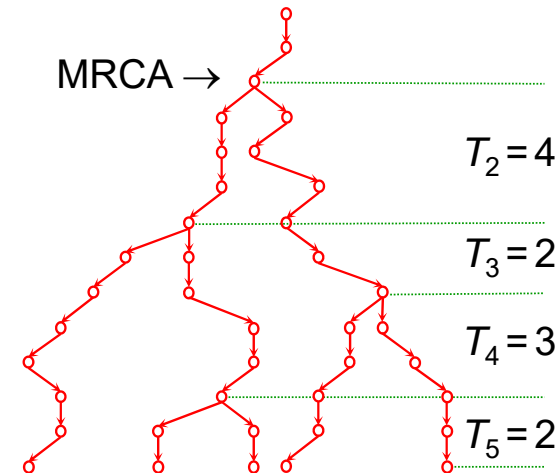
(a) Fisher-Wright model
(Constant population size, non-overlapping generations, random mating)

Time



$$N = 10 \quad (2N = 20)$$

(b) Coalescent process.
The process of lineage joining when one traces the genealogical history of the sample backwards in time.

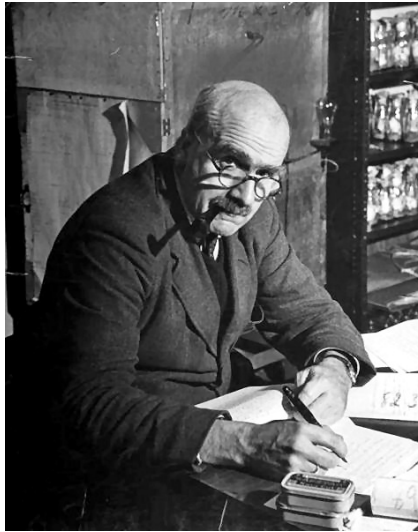


$$T_j \sim \exp\left(\frac{1}{2N} \times \frac{j(j-1)}{2}\right)$$

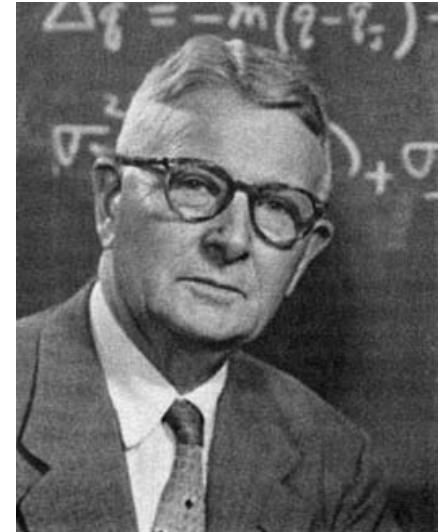
Classic population genetics studies the changes of allele frequencies in a population with time running forward (e.g., *diffusion approximation*)



Ronald A. Fisher (1890-1962)
Galton Professor, UCL (1933-1943)



JBS Haldane (1892-1964)
Weldon Professor, UCL (1933-1956)

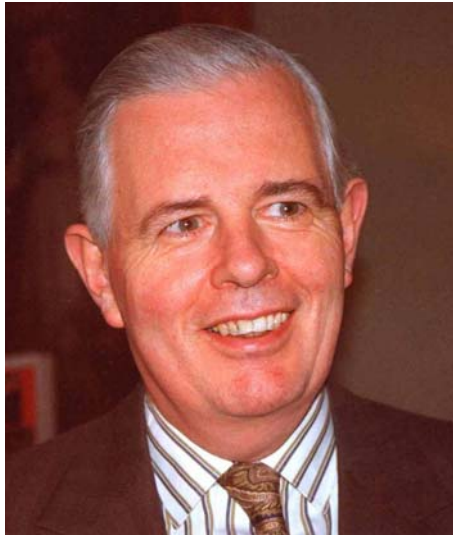


Sewall Wright (1889-1988)

“Three giants in population genetics, two in UCL”

Coalescent runs the time machine backwards

John Kingman (1939-)



The coalescent: 2 genes

The probability that two genes share a common ancestor (parent) in the previous generation is $1/(2N)$. The probability that two genes share a common ancestor j generations back is

$$\Pr\{T_2 = j\} = \left(1 - \frac{1}{2N}\right)^{j-1} \times \frac{1}{2N}$$

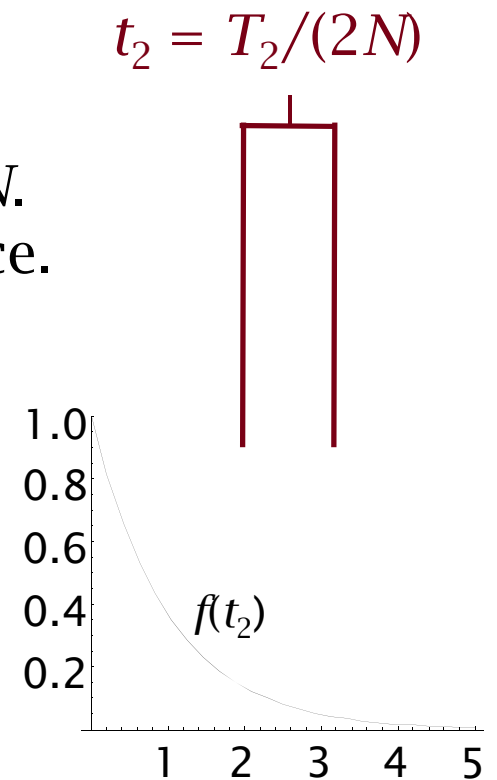
This is known as a geometric distribution and has mean $2N$. It takes on average $2N$ generations for two genes to coalesce.

Let $t_2 = T_2/(2N)$ so that one time unit is $2N$ generations. Then t_2 is exponential with mean 1:

$$f(t_2) = e^{-t_2}$$

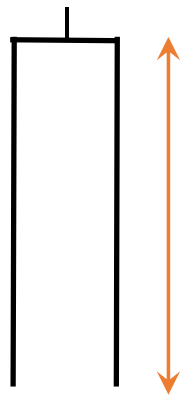
N : population size of a diploid species

$2N$: number of sequences at any locus in the population



Genetic diversity in a population is measured by $\theta = 4N\mu$, where μ is the mutation rate, which is the average difference per site between two sequences.

For the human, $\theta_H = 0.0006$: two sequences taken at random from the human population are different at 0.06% of sites. This means $N \sim 10,000$ (using $g = 15y$, $\mu = 10^{-9}/\text{site}/\text{year}$).



Average coalescent time is $2N$ generations.

Average sequence distance is $\theta = 2N \times \mu \times 2$.

Coalescent time scale, Poisson & exponential

If an event occurs as a *Poisson process* at the rate λ , the waiting time has an *exponential distribution* with probability density function

$$f(t) = \lambda e^{-\lambda t}$$

and mean $1/\lambda$. The probability for no event before time t is

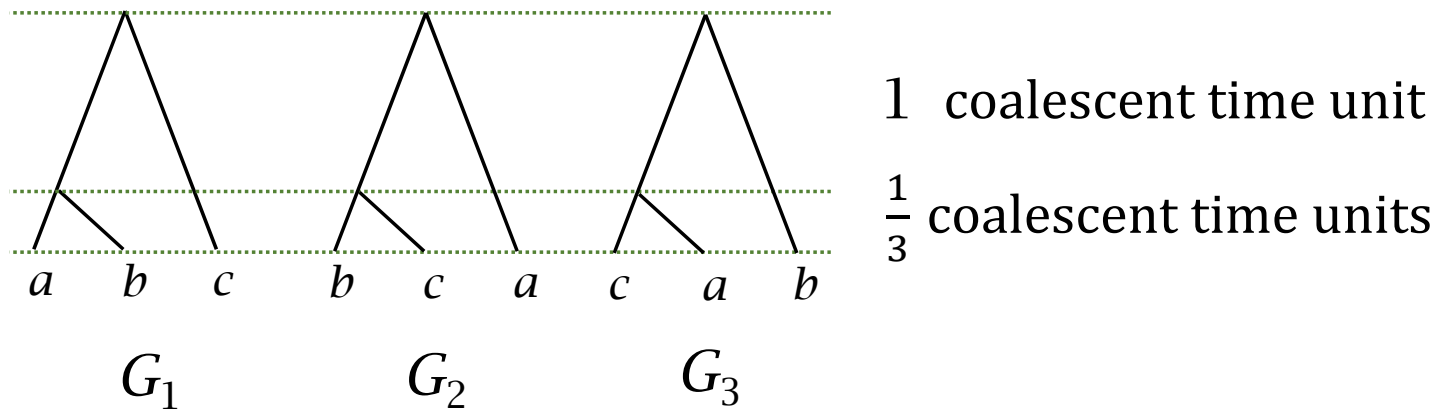
$$\Pr(X > t) = e^{-\lambda t}.$$

Any 2 sequences coalesce like a Poisson process with rate λ .

Time unit	Rate (λ)	Mean waiting time
(i) Generation	$1/(2N)$	$2N$
(ii) $2N$ generations	1	1
(iii) 1 mutation per site	$2/\theta$	$\theta/2$

The coalescent: $n=3$ sequences

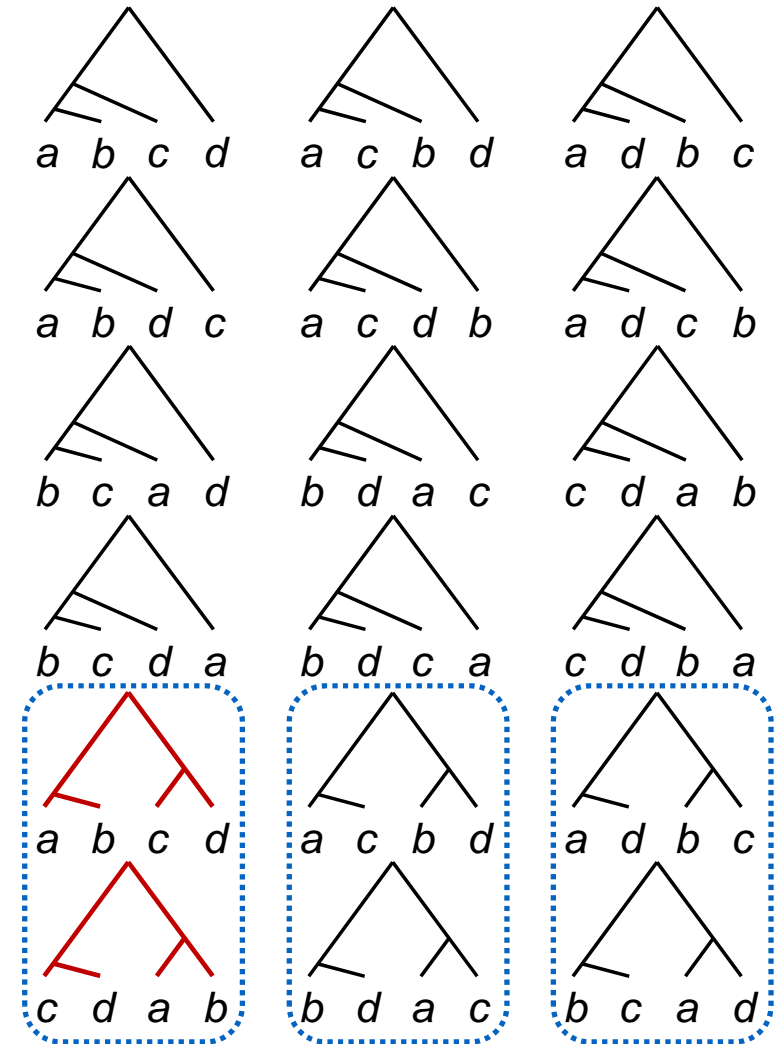
- There are 3 possible gene trees for a sample of 3 sequences, each with probability $\frac{1}{3}$.
- The first waiting time has mean $\frac{1}{3}$ while the second has mean 1 (One time unit is $2N$ generations).



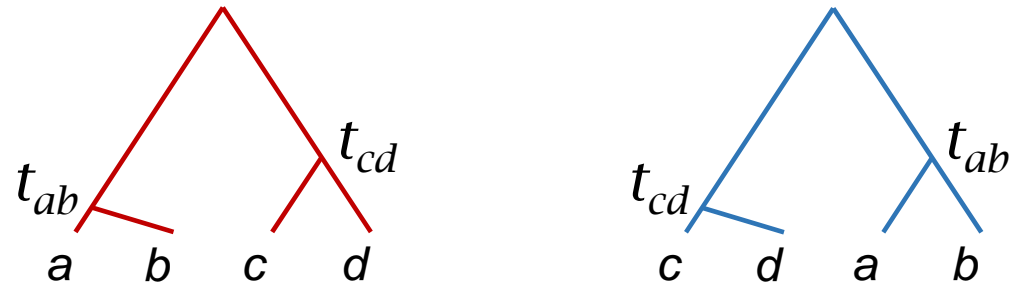
The coalescent: $n=4$ sequences

- There are 18 possible **labelled histories** (ranked gene trees) for $n=4$, each with probability $\frac{1}{18}$.
- There are 15 **gene trees**, with probability $\frac{1}{18}$ for each unbalanced tree or $\frac{2}{18}$ for each balanced tree.
Each balanced tree is compatible with two labelled histories while each unbalanced tree is compatible with one.
- Coalescent waiting times are independent exponential variables

$$\binom{4}{2}\binom{3}{2}\binom{2}{2}=18$$



A **labelled history** (or **ranked gene tree**) is a (rooted) gene tree with interior nodes ranked by age



In **left** tree: $t_{ab} < t_{cd}$. Sequences *a* & *b* coalesce first.

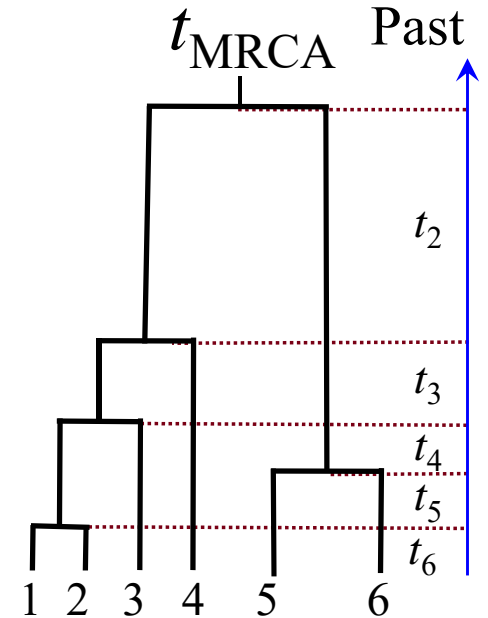
In **right** tree: $t_{cd} < t_{ab}$. Sequences *c* & *d* coalesce first.

The balanced gene tree ((*a*, *b*), (*c*, *d*)) is compatible with two labelled histories and so has probability $\frac{2}{18}$.

The coalescent: n sequences

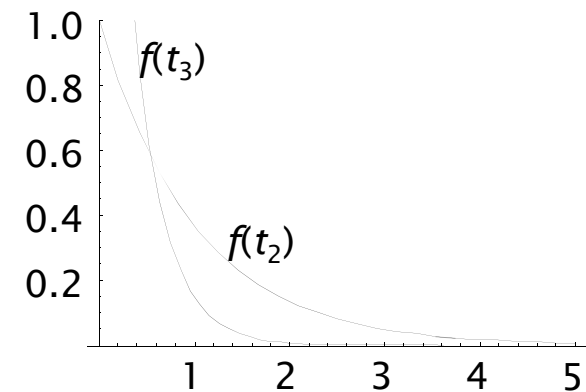
(i) Each of the $H_n = \binom{n}{2} \binom{n-1}{2} \dots \binom{2}{2}$ **labelled histories** (G_i) has equal probability, $P(G_i) = \frac{1}{H_n}$.

(ii) Coalescent rate is $\frac{1}{2N}$ for each pair of sequences.

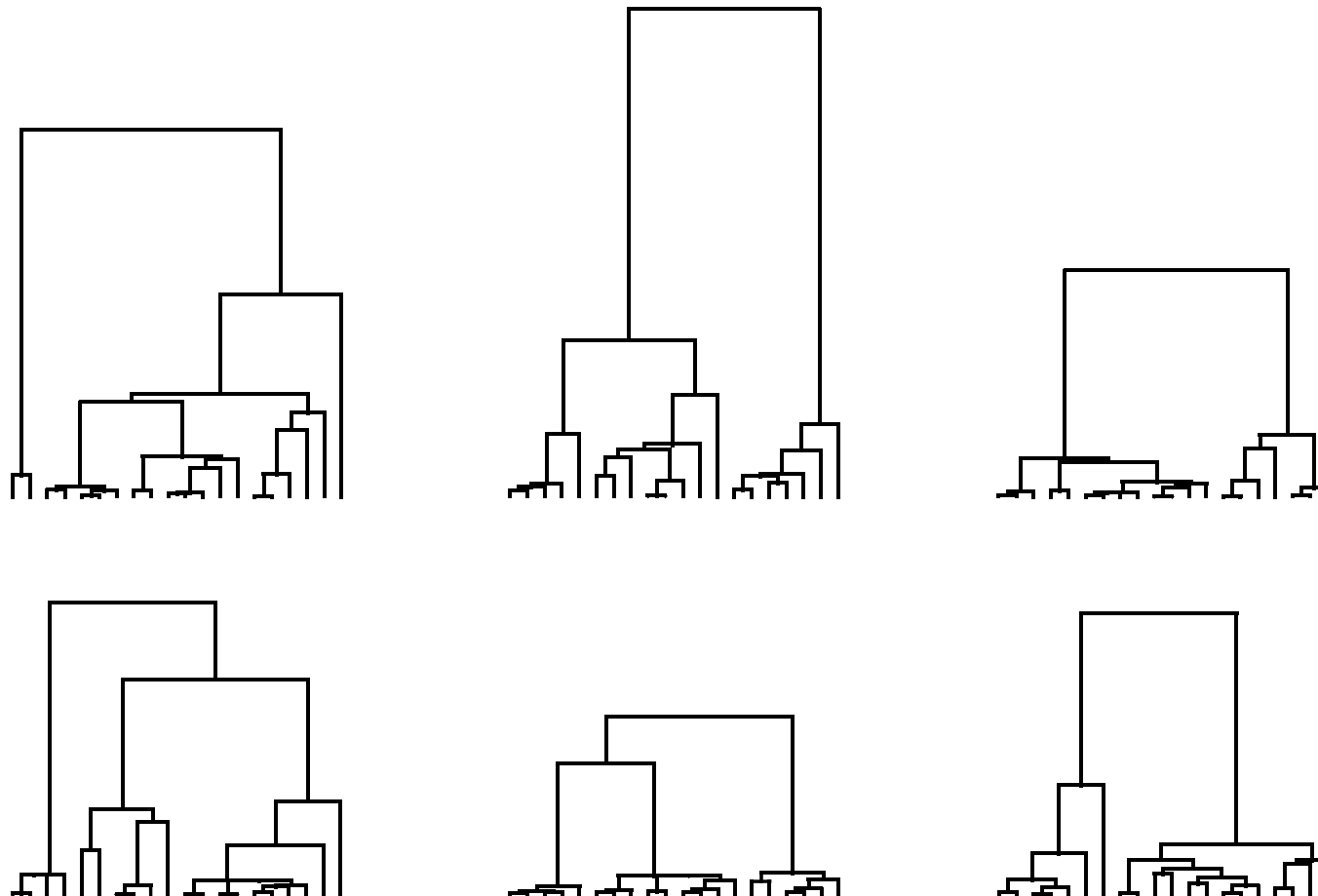


$$f(G_i, t_n, \dots, t_2) = \left[\frac{1}{2N} \dots \frac{1}{2N} \right] \times \exp \left\{ -\binom{n}{2} \frac{1}{2N} t_n - \binom{n-1}{2} \frac{1}{2N} t_{n-1} - \dots - \frac{1}{2N} t_2 \right\}$$

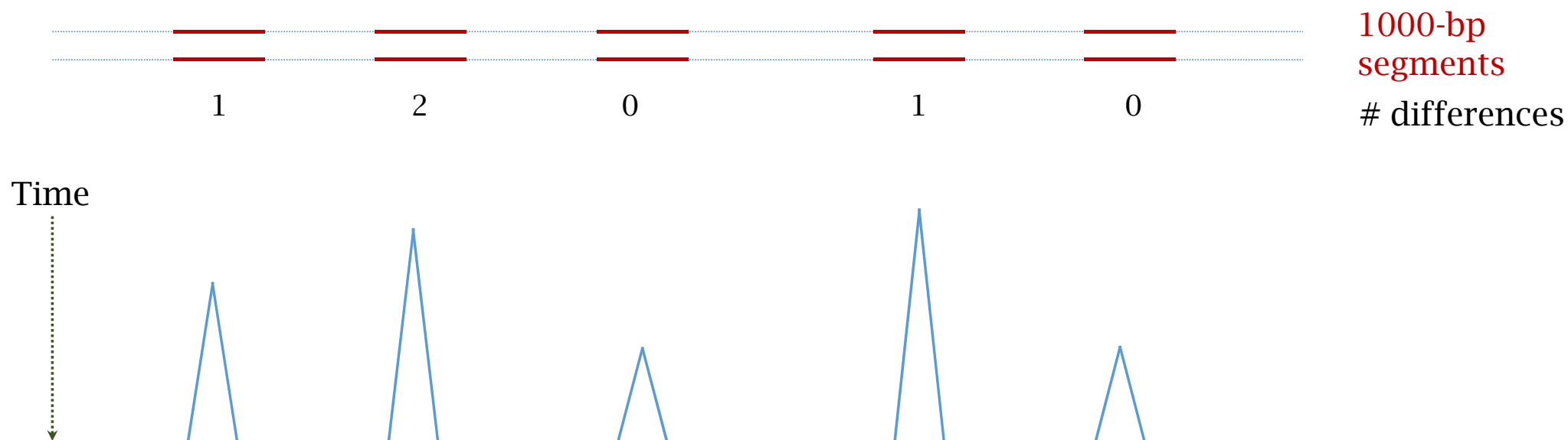
It takes on average $\sim 2 \times 2N$ ($\pm 2.15N$) generations for the whole sample to coalesce, and $2N$ generations for the last two lineages to coalesce.



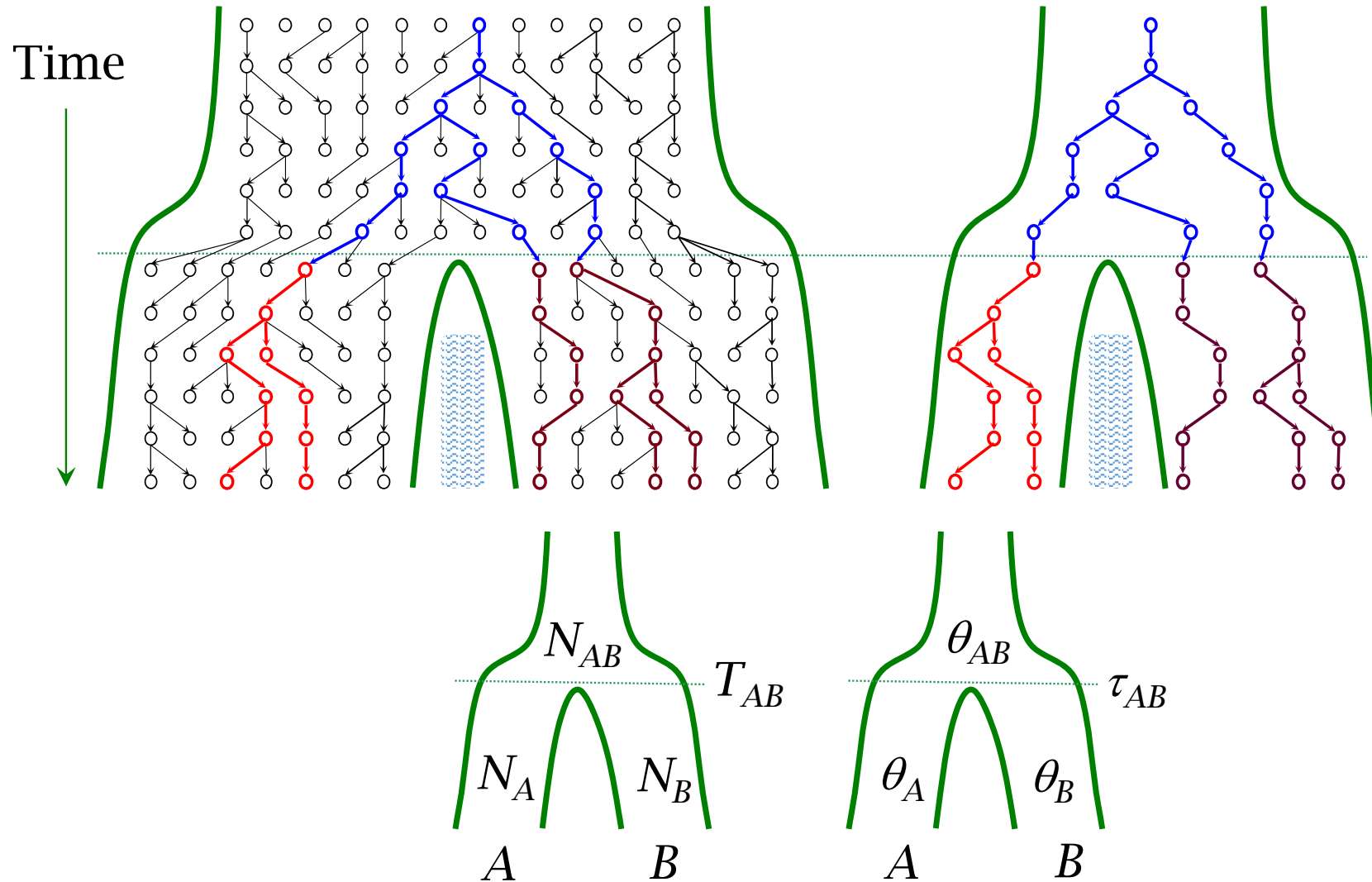
The coalescent ($n = 20$)



Coalescent time fluctuates across the genome according to an exponential distribution, with mean $2N$ (generations).

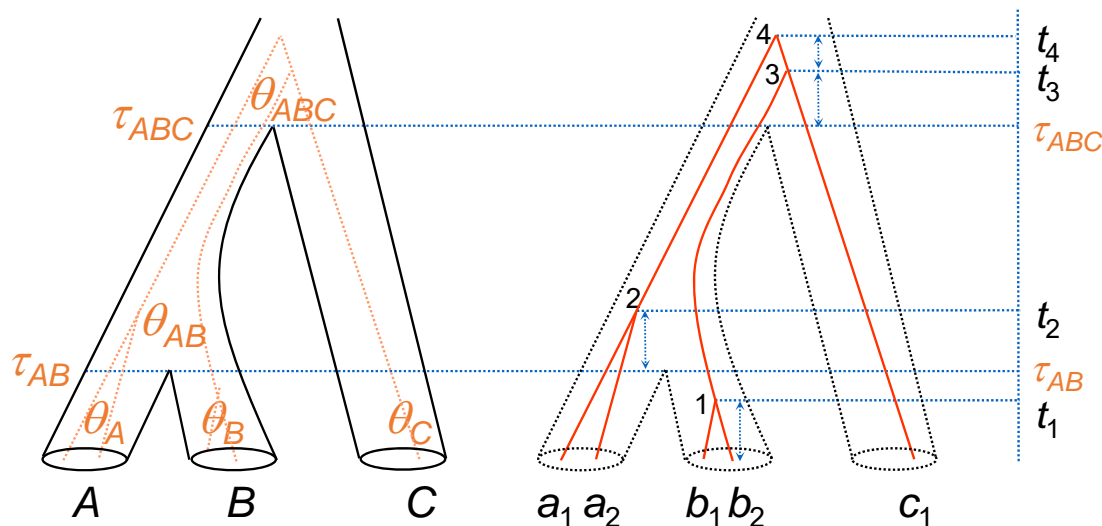


Multispecies coalescent (MSC)



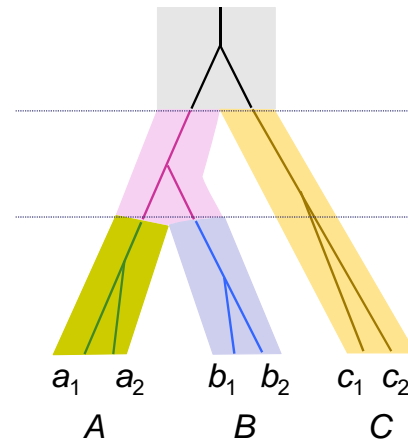
Multispecies coalescent (MSC, Liu *et al.* 2009) or **censored** coalescent (Rannala and Yang 2003) or **inter-specific** coalescent (Takahata 1989)

- Parameters: divergence times (τ) and population sizes (θ).
- Lineages join independently in different populations.
- Coalescent rate is reset when lineages enter a new species.
- Genes split before species (**gene trees fit inside species tree**).

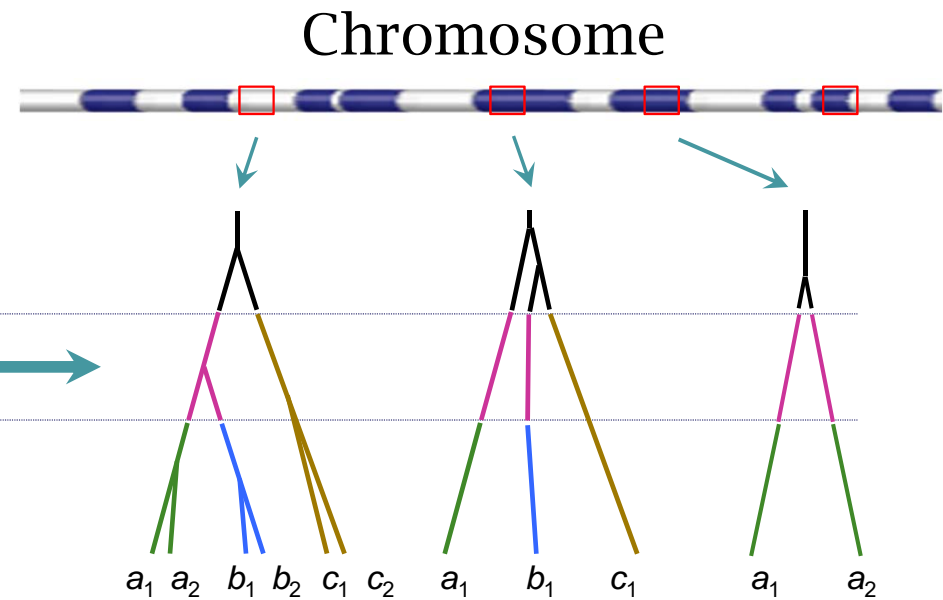


Rannala & Yang (2003 *Genetics* 164:1645-1656)

Model, data, and inference



(a) Species tree



(b) Gene trees

a1 AAGCTTCACC ... a1 GCTATCACCA ... a1 TAGCATCACC ...
a2 AAGCTTCACC ... b1 GCTATCACCA ... a2 TAGCATCACC ...
b1 AAGCTTCACC ... c1 GCTATCACC^G ...
b2 AAGCTTCACC ...
c1 AAGCTTCAT^C ...
c2 AAGCTTCAT^C ...

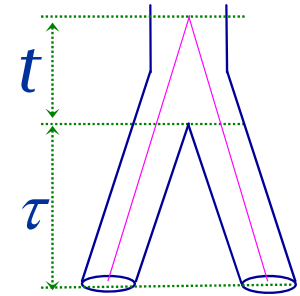
(c) Multi-locus sequence data

MSC provides a framework for analyzing genomic data from different species

Multispecies coalescent

Two species

- Gillespie, J. H., and C. H. Langley (1979. *J. Mol. Evol.* 13:27-34)
The number of substitutions [between 2 species] is the sum of a Poisson and a geometric random variable.
- Takahata, N. (1986. *Genet. Res.* 48:187-190)
The variance in H-C sequence divergence among loci was used to estimate the ancestral population size θ_{HC}



Three species

- Hudson R.R. (1983 *Evolution* 37:203-217)
derived the gene tree-species tree mismatch probability for 3 species.
- Chen & Li (2001 *AJHG* 68:444-456) used it to estimate θ_{HC} .
- Takahata, N., et al. (1995 *TPB* 48:198-221): ML for 3 species

Structured coalescent

- Li, W.-H. 1976. *TPB* 10, 303-308.
- Griffiths, R. C. 1981. *J. Math. Biol.* 12:251-261.
- Slatkin, M. 1987. *TPB* 32:42-49.
- Notohara, M. 1990. *J. Math. Biol.* 29:59-75.

Multispecies coalescent, incomplete lineage sorting, gene tree-species discordance

Takahata, N. 1989.
Genetics 122:957-966

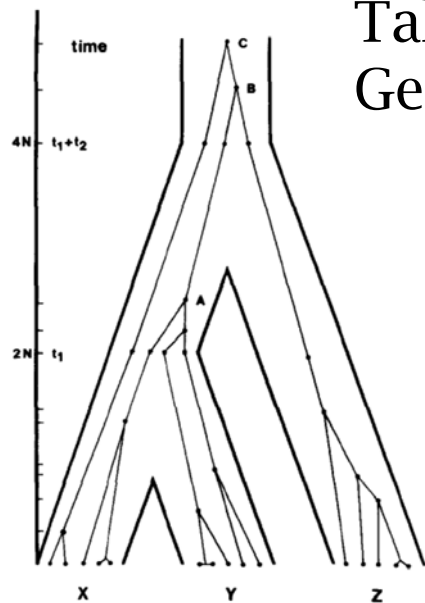


FIGURE 1.—Model of a population tree and a gene tree generated on a computer. X, Y and Z represent three different populations which diverged t_1 and $t_1 + t_2$ generations ago. Five genes were sampled from each population and $t_1 = t_2 = 2N$ were assumed. Dots and lines represent genes and ancestral lineages. Each node corresponds to a coalescence of genes. A, B and C stand for interspecific coalescences and all other nodes for intraspecific coalescences. In this simulation, there remained four ancestral genes from X and Y at t_1 . Note that the probabilities that the first and the first two coalescences are intraspecific are $1/3$ and $1/9$, respectively.

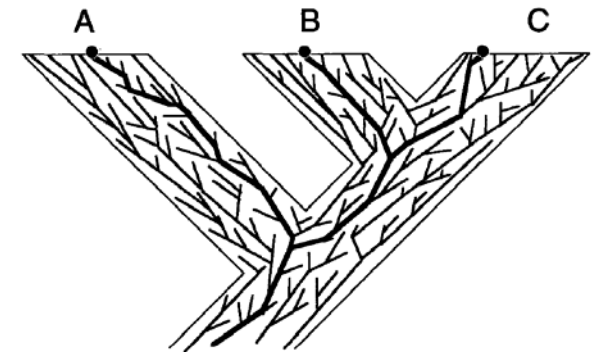


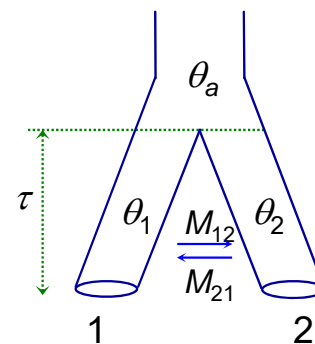
FIGURE 1. A gene tree contained within a species tree leading to three extant species: A, B, and C. Bold branches of gene tree show relationships among the sampled copies of the gene (●). Sampled copies from sister species B and C are sister copies.

Maddison, W.P. 1997.
Syst. Biol. 46:523-536

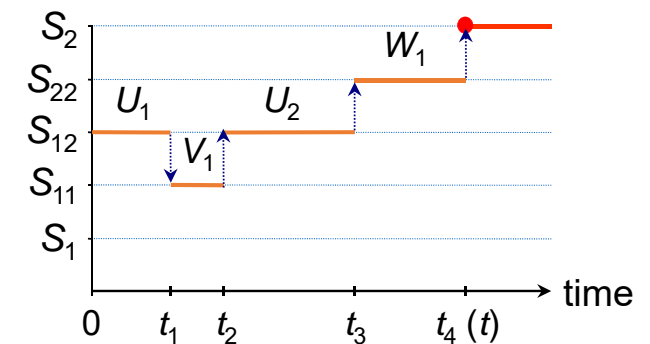
MSC has many applications & extensions

- Inference of species divergences and population sizes
- Estimation of migration patterns and rates (IMa, etc.)
- Introgression & hybridization
- Species tree estimation (STEM, BEST, *BEAST, BPP etc.)
- Species delimitation (BPP)
- ...

(a) Species tree



(b) Migration trajectory



MSC or coalescent is the biological process of reproduction viewed backwards in time

H_0 : MSC (null model)

H_1 : MSC + population structure

H_2 : MSC + hybridization

H_3 : MSC + recombination

H_4 : MSC + population structure + hybridization

etc.

(Degnan JH. 2018.
Syst. Biol. 67:786-799)

Some terminologies are confusing:

“to distinguish hybridization from lineage sorting”

“investigate whether the conditions of applicability of coalescence-based methods are met ...”

Multispecies coalescent (MSC)

(i) $f(G)$

- Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. *Evolution* 59:24-37.
- Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68.
- Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24:332-340.
- Rosenberg, N. A., and M. Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3:380-390.
- Rosenberg, N. A., and R. Tao. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. *Systematic Biology* 57:131-140.

...

(ii) $f(G, t)$

- Rannala, B., and Z. Yang. 2003. *Genetics* 164:1645-1656.

$f(G)$ is useful for two-step summary methods.
 $f(G, t)$ is useful for full-likelihood methods (ML & Bayesian).

Multispecies coalescent (MSC) density

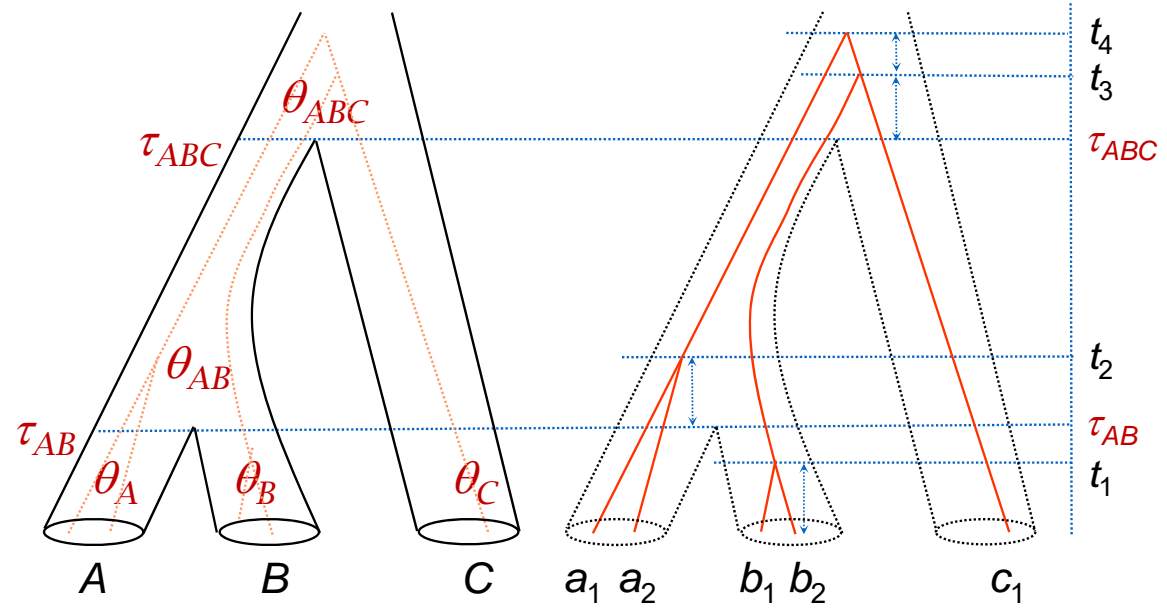
$$f(G, t_1, t_2, t_3, t_4 \mid S, \Theta) =$$

$$e^{-\frac{2}{\theta_A} \tau_{AB}}$$

$$\times \frac{2}{\theta_B} \times e^{-\frac{2}{\theta_B} t_1}$$

$$\times \frac{2}{\theta_{AB}} \times e^{-\left(3 \times \frac{2}{\theta_{AB}}\right)(t_2 - \tau_{AB})} \times e^{-\frac{2}{\theta_{AB}}(\tau_{ABC} - t_2)}$$

$$\times \frac{2}{\theta_{ABC}} \times e^{-\left(3 \times \frac{2}{\theta_{ABC}}\right)(t_3 - \tau_{ABC})} \times \frac{2}{\theta_{ABC}} e^{-\frac{2}{\theta_{ABC}}(t_4 - t_3)}$$



Full likelihood methods of species tree estimation integrate over the unknown gene trees

Maximum likelihood is feasible for 3 species (3 sequences) (3S)

$$L(S, \Theta) = \prod_i f(X_i | S, \Theta) = \prod_i \left[\sum_{G_i} \int f(G_i, t_i | S, \Theta) f(X_i | G_i, t_i) dt_i \right]$$

Bayesian method averages over the gene trees through MCMC marginalisation (*BEAST, BP&P, ...)

$$f(S, \Theta, \{G_i, t_i\} | X) \propto f(S) f(\Theta | S) \prod_i f(G_i, t_i | S, \Theta) f(X_i | G_i, t_i)$$

S : species tree.

Θ : θ and τ parameters on the species tree.

G_i & t_i : gene tree topology and branch lengths at locus i .

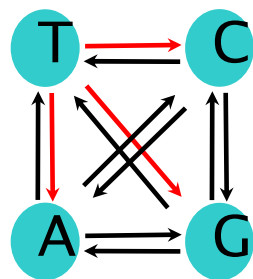
MSC density

Phylogenetic likelihood

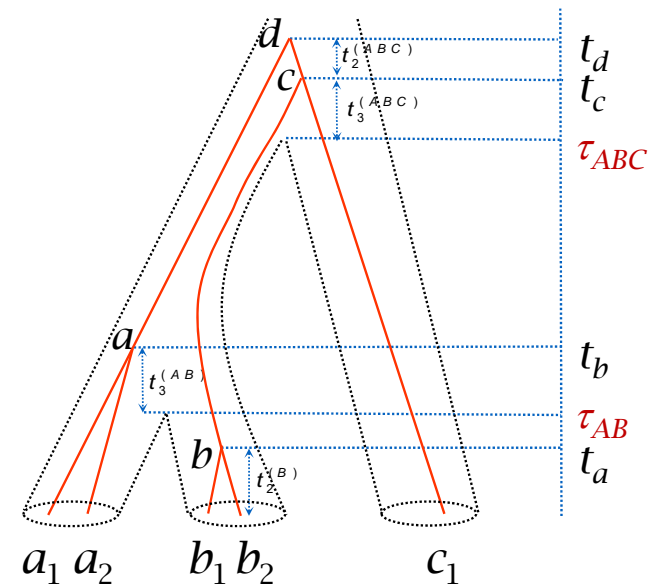
Model & data

- The **MSC** gives the distribution of the gene tree and branch lengths (Rannala & Yang 2003 *Genetics* 164:1645-1656).
- The **phylogenetic likelihood** is the probability of the sequence data at each locus (Felsenstein 1981 *J. Mol. Evol.* 17:368-376).

A1	TCCATTCAAG	AGTCTATTAT	CAGTTAATTC	...
A2	TCCATTCAAG	AGTCTATTAT	CAGTTAG G TTC	...
B1	TCCATTCAAG	AGT T TATTAT	CAGTTAATTC	...
B2	TCCATTCAAG	AGT T TATTAT	CAGTTAATTC	...
C1	TCCATTCAAG	G GTCTATTAT	CAGTTAATTC	...



JC69 model



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 (θ_S, τ_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.

People who did the work



Tomas Flouri



Paschalia Kapli



Xiyun Jiao



Jun Huang



Yuttapong Thawornwattana

Funding
BBSRC



Bruce Rannala (@BruceRannala) | T...
twitter.com