Coalescent and multispecies coalescent

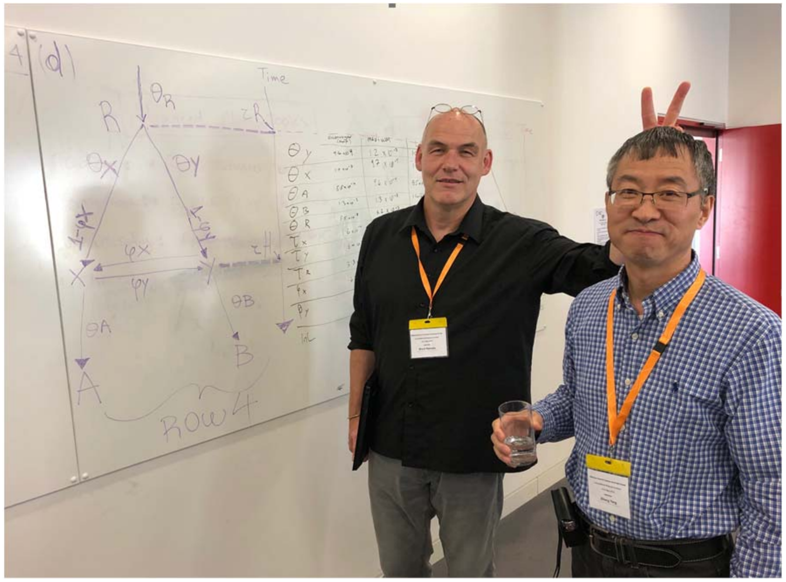
Ziheng Yang

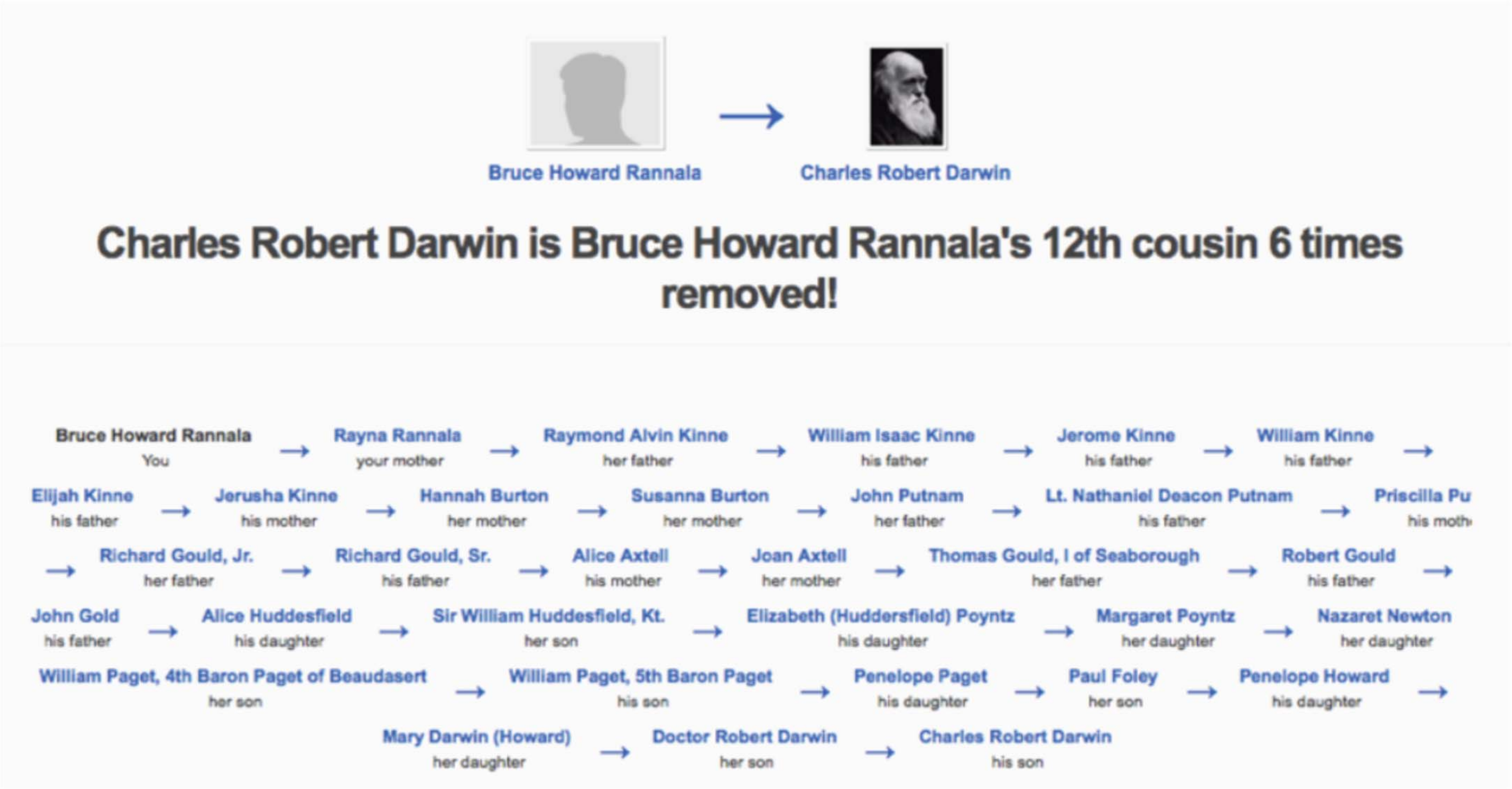
Department of Genetics, Evolution, and Environment

University College London

Outline

- Pedigree, genealogy and coalescent
- Multispecies coalescent (MSC)
- Estimation of parameters under the MSC





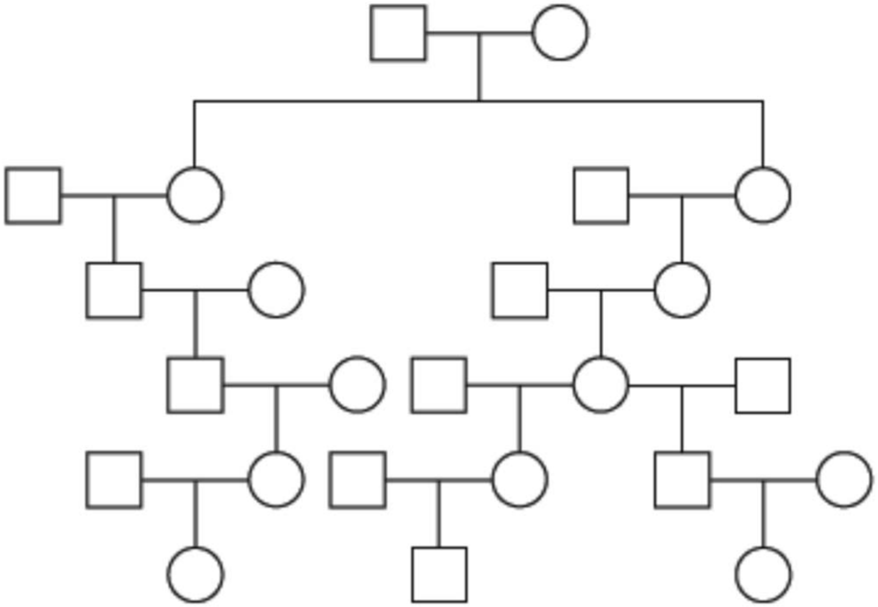
Bruce Howard Rannala → Charles Robert Darwin

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

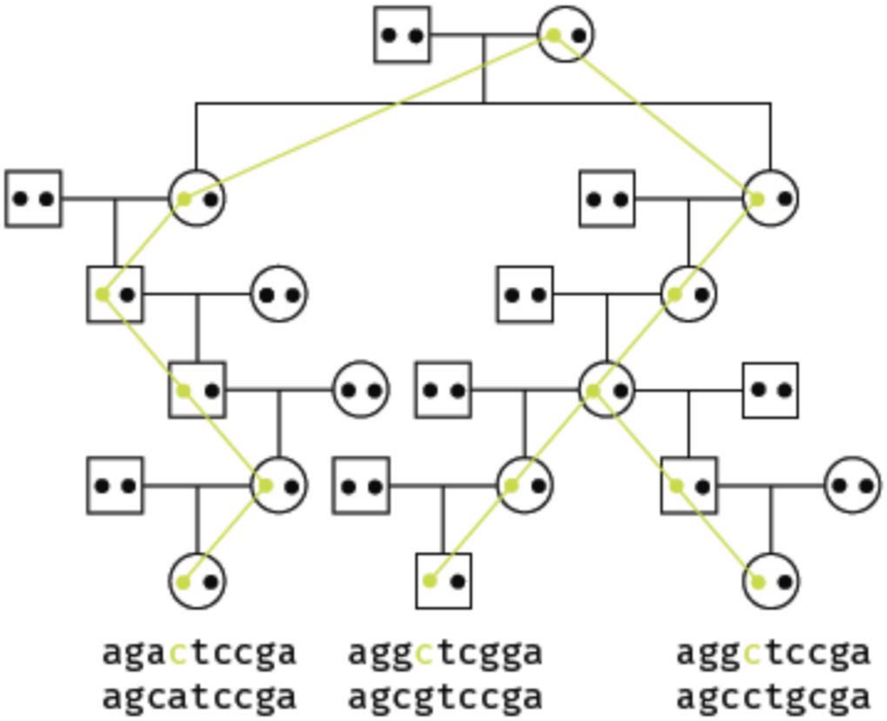
Bruce Howard Rannala
You
→
Rayna Rannala
your mother
→
Raymond Alvin Kinne
her father
→
William Isaac Kinne
his father
→
Jerome Kinne
his father
→
William Kinne
his father
→
Elijah Kinne
his father
→
Jerusha Kinne
his mother
→
Hannah Burton
her mother
→
Susanna Burton
her mother
→
John Putnam
her father
→
Lt. Nathaniel Deacon Putnam
his father
→
Priscilla Pu
his moth
→
Richard Gould, Jr.
her father
→
Richard Gould, Sr.
his father
→
Alice Axtell
his mother
→
Joan Axtell
her mother
→
Thomas Gould, I of Seaborough
her father
→
Robert Gould
his father
→
John Gold
his father
→
Alice Huddesfield
his daughter
→
Sir William Huddesfield, Kt.
her son
→
Elizabeth (Huddersfield) Poyntz
his daughter
→
Margaret Poyntz
her daughter
→
Nazaret Newton
her daughter
William Paget, 4th Baron Paget of Beaudasert
her son
→
William Paget, 5th Baron Paget
his son
→
Penelope Paget
his daughter
→
Paul Foley
her son
→
Penelope Howard
his daughter
→
Mary Darwin (Howard)
her daughter
→
Doctor Robert Darwin
her son
→
Charles Robert Darwin
his son

John Gould

Pedigree

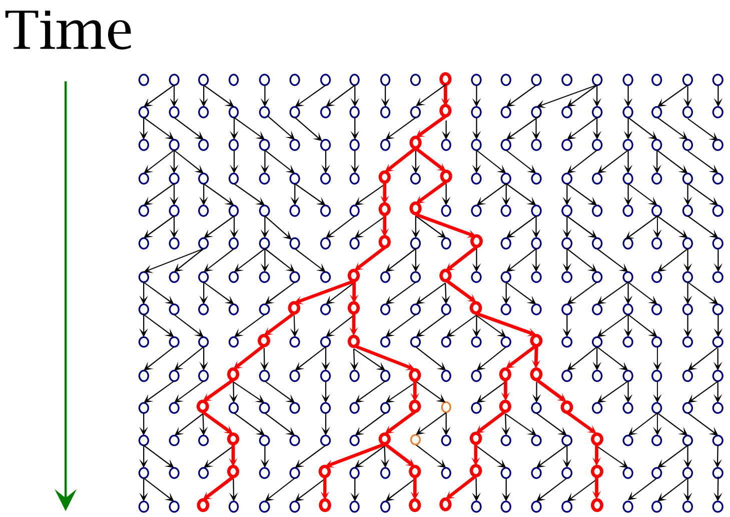


Gene tree within pedigree



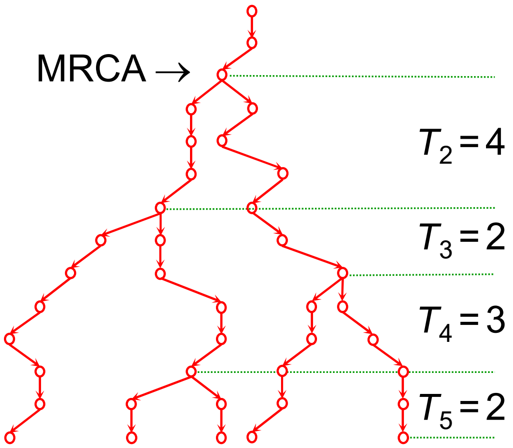
Fisher-Wright model & coalescent

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



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

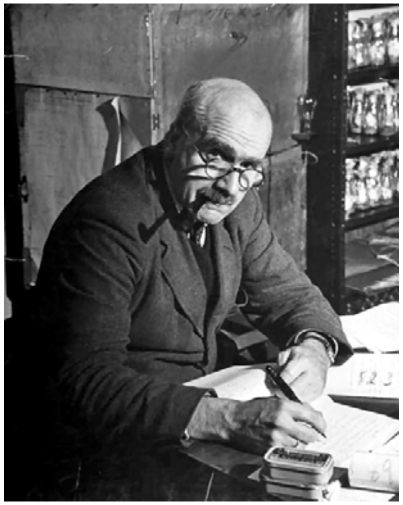
(b) Coalescent process.
The process of lineage joining when one traces the genealogical history of the sample backwards in time.



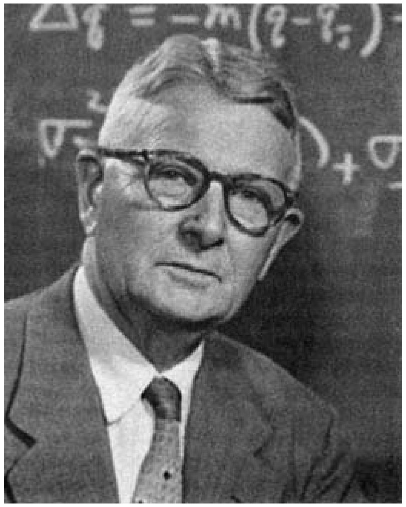
$$T_j \sim \mathrm{Exp}\left(\frac{1}{2N} \times \frac{j(j-1)}{2}\right)$$

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)

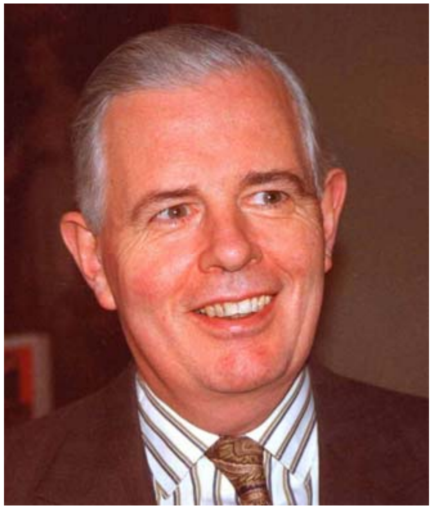


Sewall Wright (1889-1988)

“Three giants in population genetics, two in UCL”

Coalescent runs the time machine backwards

John Kingman (1939-)

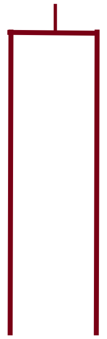


The coalescent: 2 genes

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

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

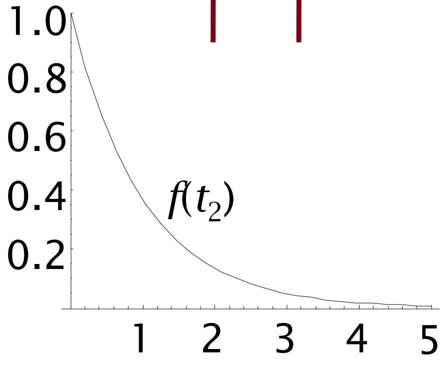
$t_2 = T_2/(2N)$



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:

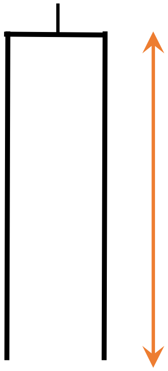
$$f(t_2) = \mathrm{e}^{-t_2}$$



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), 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 $g = 15$y, $\mu = 10^{-9}$/site/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

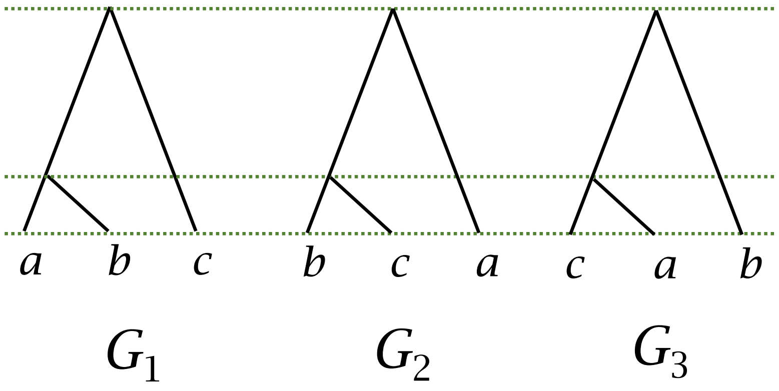
$$\Pr(X > t) = e^{-\lambda t}.$$

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).



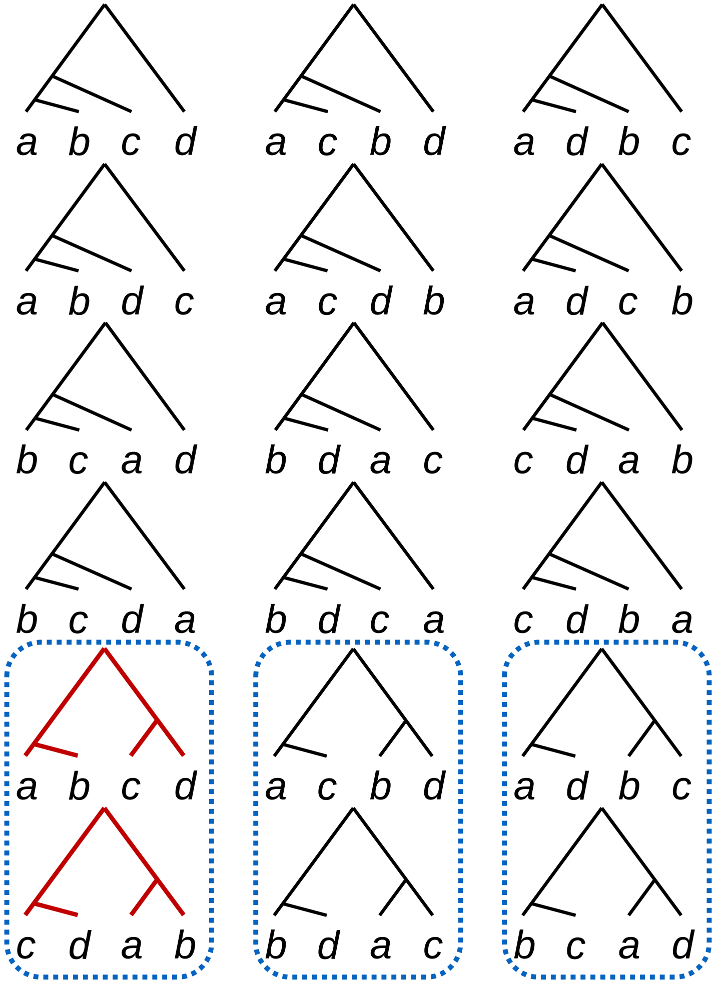
1 coalescent time unit

$\frac{1}{3}$ coalescent time units

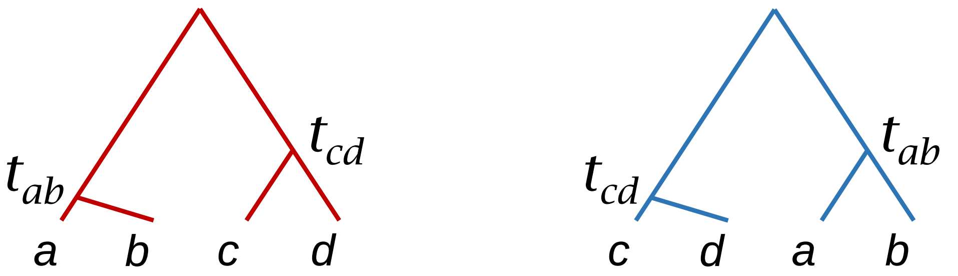
The coalescent: $n = 4$ sequences

- 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

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



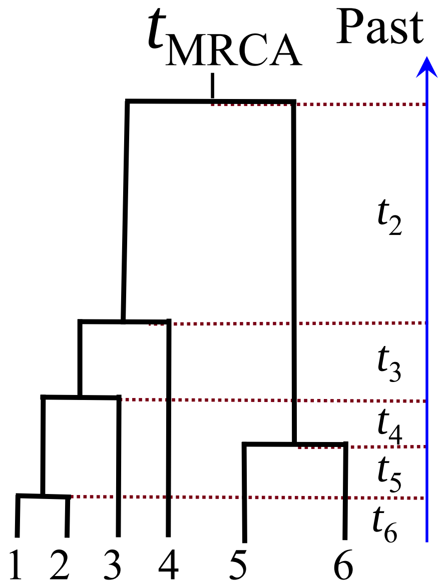
A **labelled history** (or **ranked gene tree**) is a (rooted) gene tree with interior nodes ranked by age



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

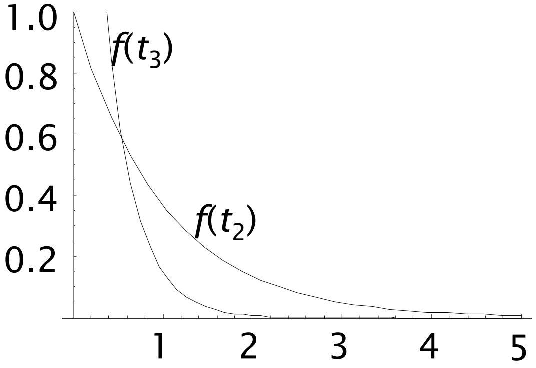


(i) Each of the $H_n = \binom{n}{2}\binom{n-1}{2} \ldots \binom{2}{2}$ labelled histories (G_i) has equal probability, $P(G_i) = \frac{1}{H_n}$.

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

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

It takes on average ~2 × 2*N* (±2.15*N*) generations for the whole sample to coalesce, and 2*N* generations for the last two lineages to coalesce.



Poisson process & exponential waiting time

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

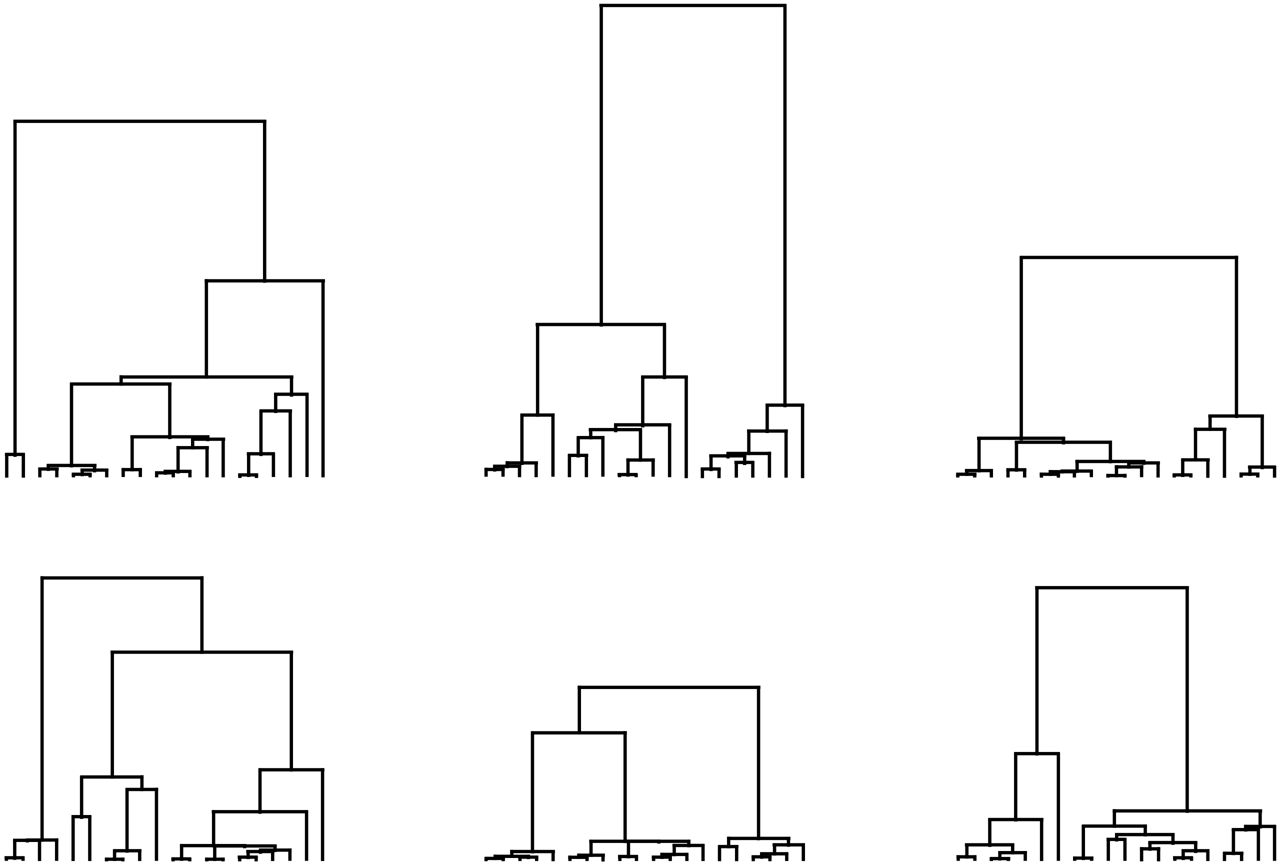
$$f(t) = \lambda e^{-\lambda t}.$$

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

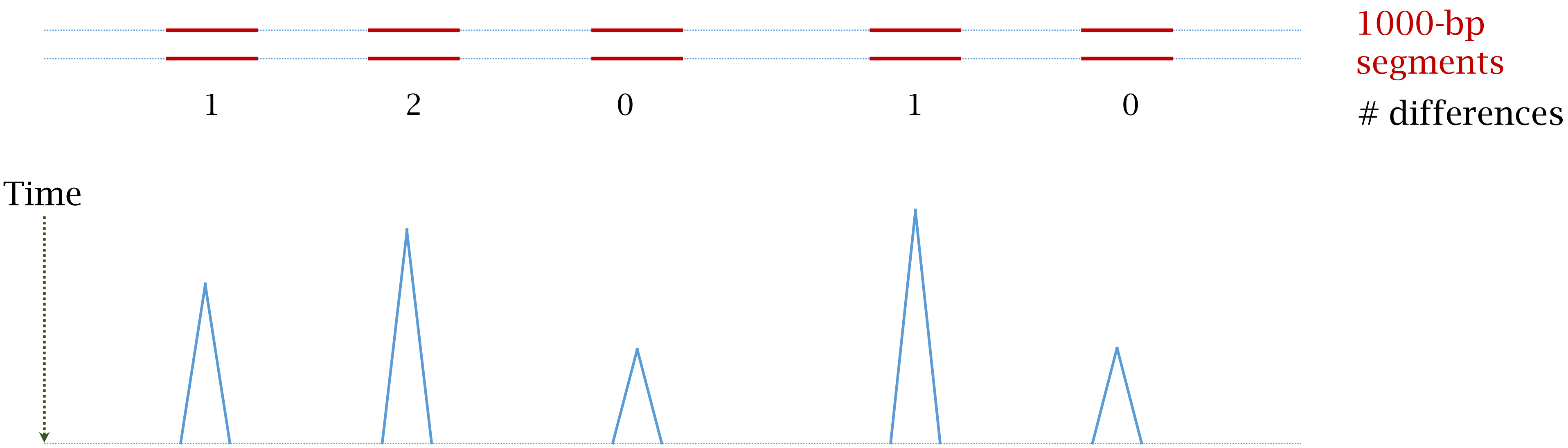
$$\Pr(X > t) = e^{-\lambda t}.$$

Coalescent. Two sequences coalesce according to a Poisson process, at the rate of $\frac{1}{2N}$ per generation. The average waiting time until the coalescent is $2N$ generations.

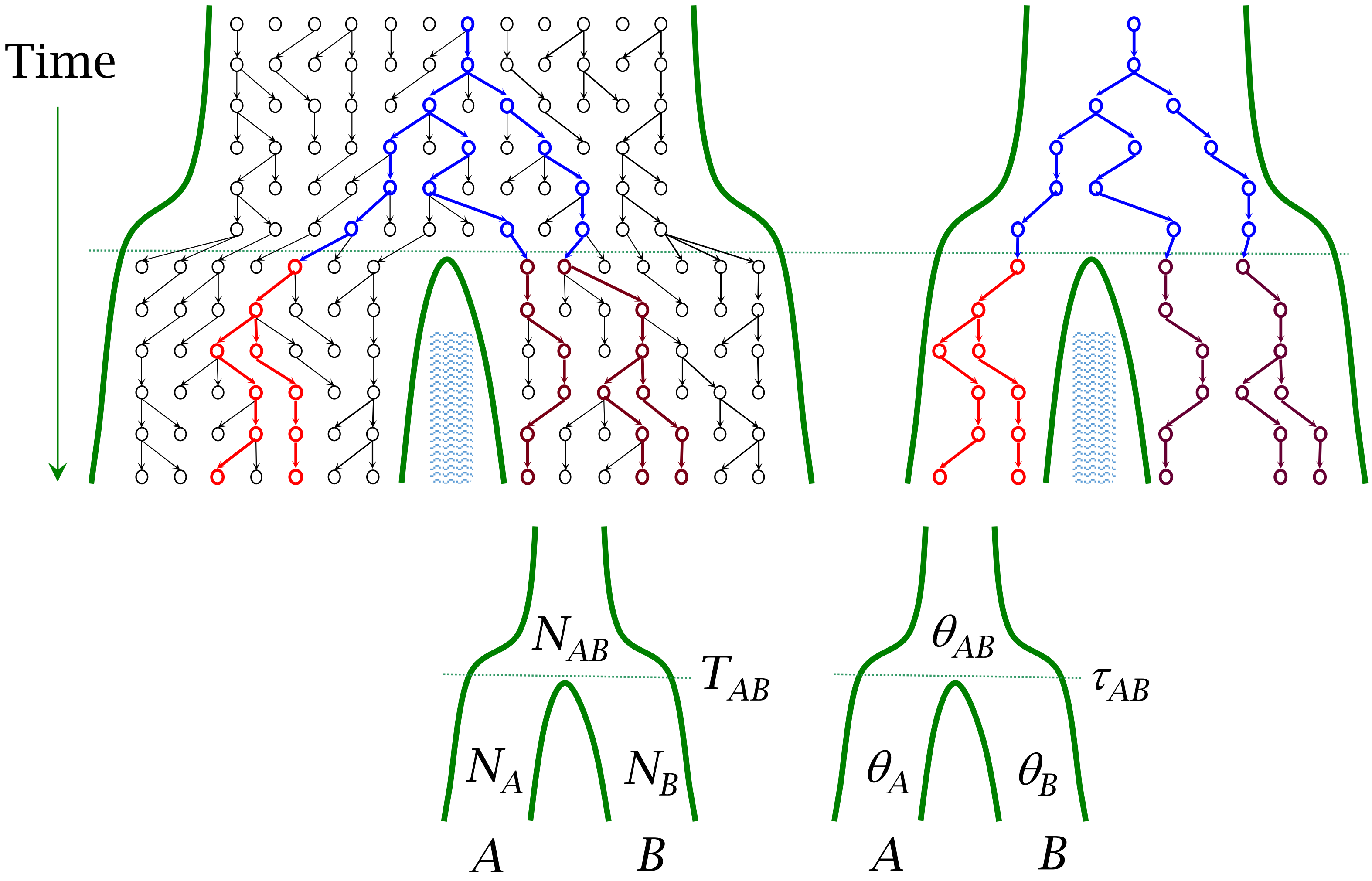
The coalescent ($n = 20$)



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



Multispecies coalescent (MSC)

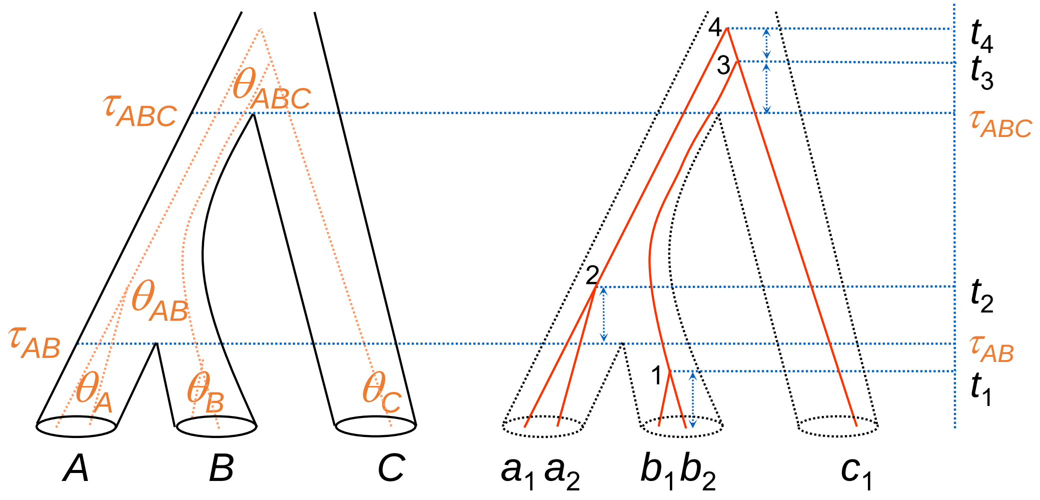


$$\theta = 4N\mu$$

$$\tau = T\mu$$

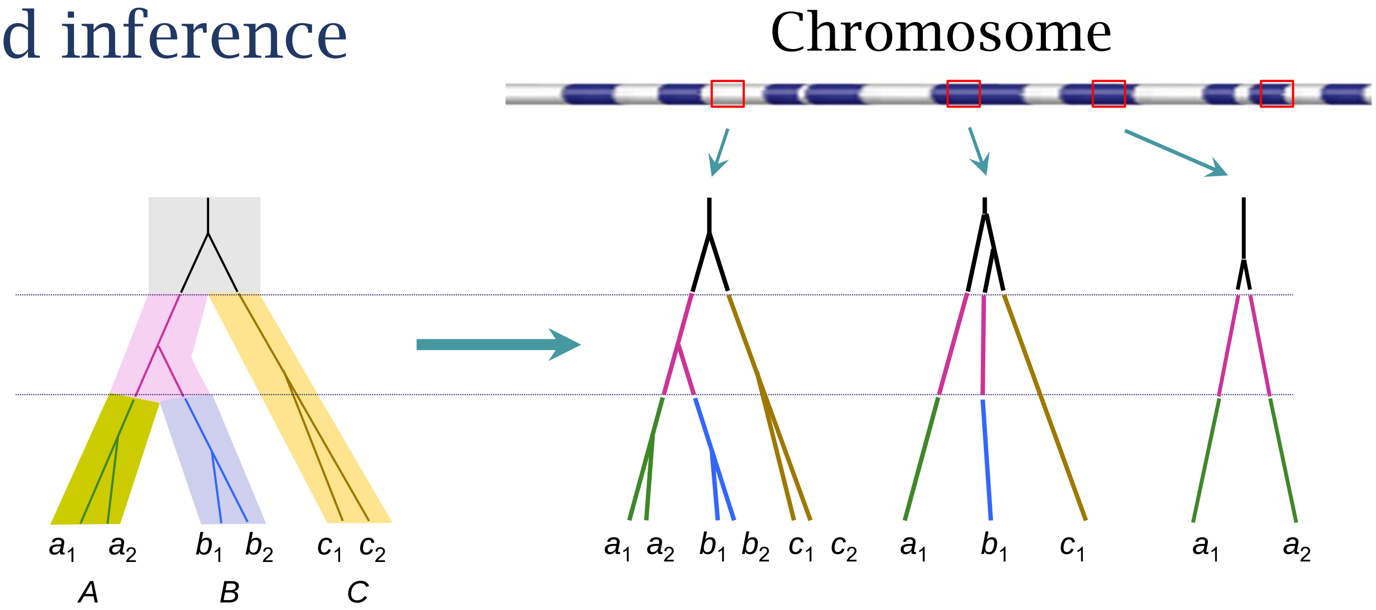
Multispecies coalescent (MSC, Liu *et al.* 2009) or censored coalescent (Rannala and Yang 2003) or inter-specific coalescent (Takahata 1989)

- 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


Chromosome
a_1 a_2 b_1 b_2 c_1 c_2
A B C
a_1 a_2 b_1 b_2 c_1 c_2
a_1 b_1 c_1
a_1 a_2

(**a**) Species tree

(**b**) Gene trees

```
a1  AAGCTTCACC …
a2  AAGCTTCACC …
b1  AAGCTTCACC …
b2  AAGCTTCACC …
c1  AAGCTTCATC …
c2  AAGCTTCATC …
```

```
a1  GCTATCACCA …
b1  GCTATCACCA …
c1  GCTATCACCG …
```

```
a1  TAGCATCACC …
a2  TAGCATCACC …
```

MSC provides a framework for analyzing genomic data from different species

(**c**) Multi-locus sequence data

Comments about compiling multilocus sequence data

Two ways of generating multilocus sequence alignments for analysis under the MSC.

- Reduced representation data: ddRAdseq, transcriptomes, etc.
- Short genomic segments compiled from sequenced genomes. Loci are short genomic segments that are far apart (e.g., <2kb with 10kb gaps). You may write or edit your own pipelines (see Thawornwattana, et al. 2022. *Syst. Biol.* **71**:1159-1177).

Do's

- For species tree estimation, one sequence per species is fine.
- For inference of gene flow, include multiple samples per species, especially from the species receiving migrants.
- A few good-quality samples may be better than many low-coverage samples.

Don'ts

- Avoid haploid consensus sequences, which resolve the phase of heterozygote sites at random, creating chimeric sequences that do not exist in nature. Use ambiguities to represent heterozygotes (Y for T/C, R for A/G, etc.)
- Beware of possible biases due to data filtering (e.g., according to bootstrap support values for gene trees).

Information content in the data may depend on the problem

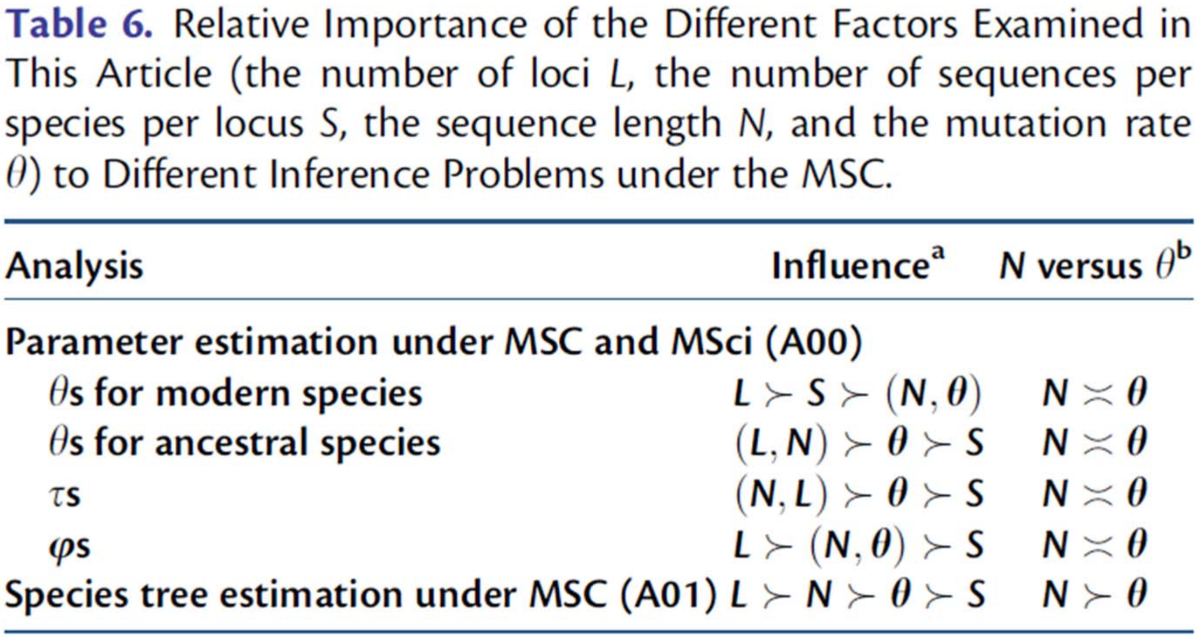


Table 6. Relative Importance of the Different Factors Examined in This Article (the number of loci L, the number of sequences per species per locus S, the sequence length N, and the mutation rate θ) to Different Inference Problems under the MSC.

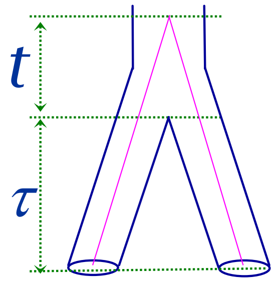
Analysis	Influence[a]	N versus θ[b]
Parameter estimation under MSC and MSci (A00)		
θs for modern species	$L \succ S \succ (N, \theta)$	$N \asymp \theta$
θs for ancestral species	$(L, N) \succ \theta \succ S$	$N \asymp \theta$
τs	$(N, L) \succ \theta \succ S$	$N \asymp \theta$
φs	$L \succ (N, \theta) \succ S$	$N \asymp \theta$
Species tree estimation under MSC (A01)	$L \succ N \succ \theta \succ S$	$N \succ \theta$

L: number of loci
S: number of sequences
N: sequence length (# of sites)
θ: mutation rate (exons vs. introns)

For most problems, the number of loci is the most important factor.
For inference of gene flow, it is important to have multiple samples per species ($S > 1$) as otherwise there may be problems with unidentifiability

Huang J, Flouri T, Yang Z. 2020. *Mol. Biol. Evol.* 37:3211-3224.

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

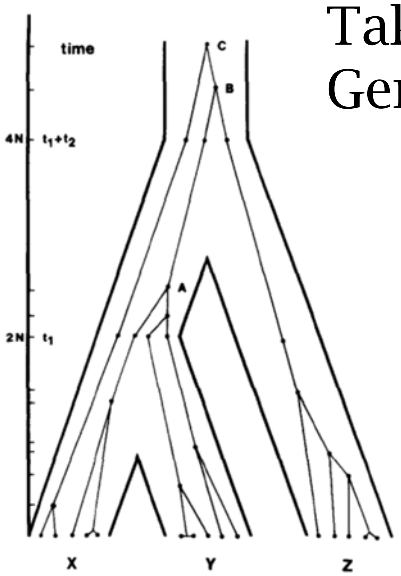


FIGURE 1.—Model of a population tree and a gene tree generated on a computer. *X*, *Y* and *Z* represent three different populations which diverged t_1 and $t_1 + t_2$ generations ago. Five genes were sampled from each population and $t_1 = t_2 = 2N$ were assumed. Dots and lines represent genes and ancestral 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.

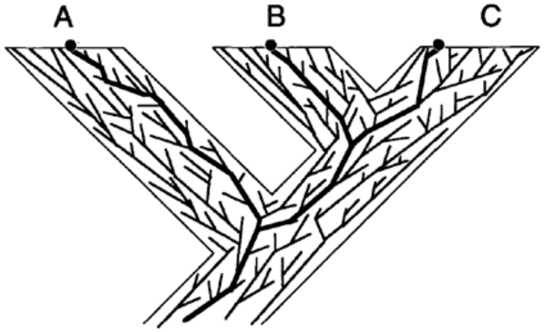
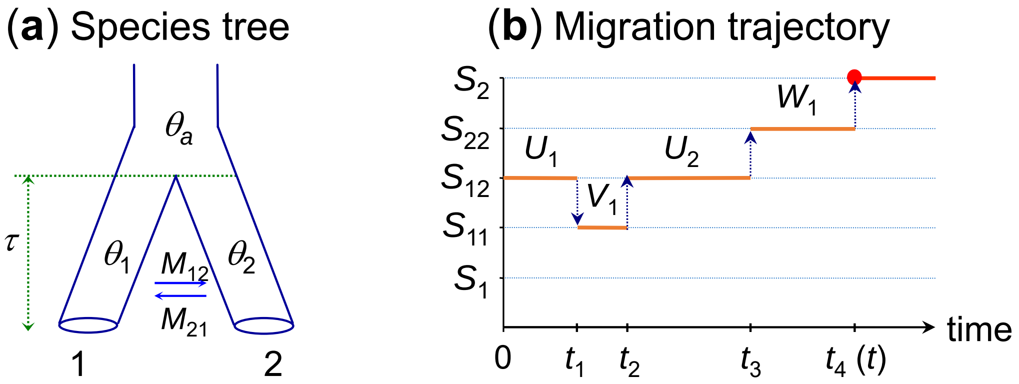


FIGURE 1. A gene tree contained within a species tree leading to three extant species: A, B, and C. Bold branches of gene tree show relationships among the sampled copies of the gene (●). Sampled copies from sister species B and C are sister copies.

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

MSC has many applications & extensions

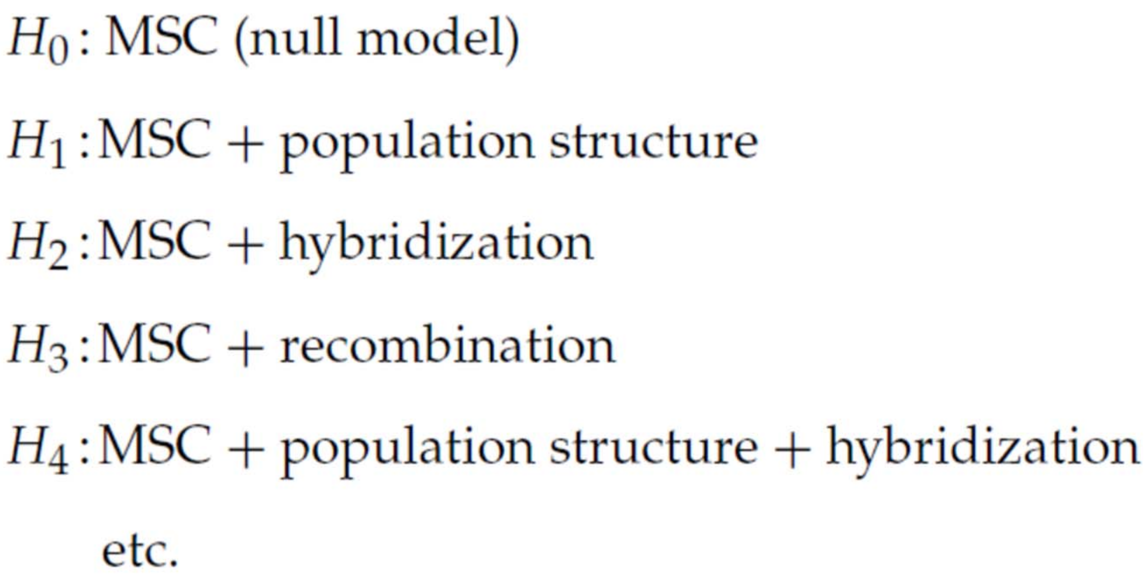
- Inference of species divergences and population sizes
- Estimation of migration patterns and rates (IMa, etc.)
- Introgression & hybridization
- Species tree estimation (STEM, BEST, *BEAST, BPP etc.)
- Species delimitation (BPP)
- …



(**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)

Confusing terminologies in the literature:

"*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.

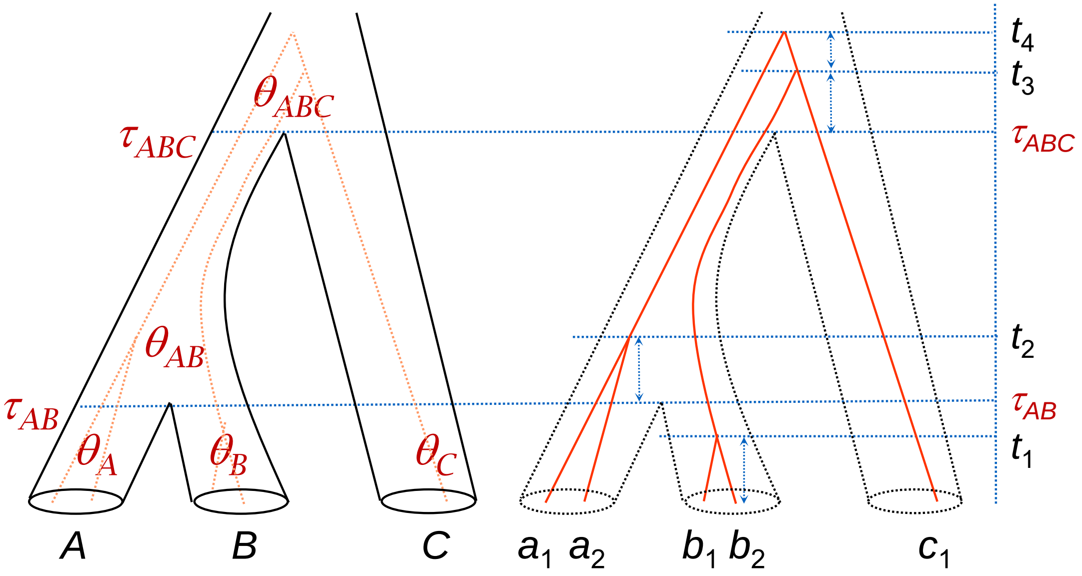
…

(ii) $f(G, \boldsymbol{t})$

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

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

Multispecies coalescent (MSC) density



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

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

pop $B \to$ $$\times \frac{2}{\theta_B} \cdot e^{-\frac{2}{\theta_B}t_1}$$

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

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

Two kinds of terms:

- rate ($\lambda = 2/\theta$)
- Probability of no event: $e^{-\lambda t}$

$$\Theta = (\tau_{AB}, \tau_{ABC}, \theta_A, \theta_B, \theta_C, \theta_{AB}, \theta_{ABC})$$

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

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

$$L(S,\Theta)=\prod_i f(X_i\mid S,\Theta)=\prod_i\left[\sum_{G_i}\int f(G_i,\boldsymbol{t}_i\mid S,\Theta)\,f(X_i\mid G_i,\boldsymbol{t}_i)\,\mathrm{d}\boldsymbol{t}_i\right]$$

MSC density

Phylogenetic likelihood

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

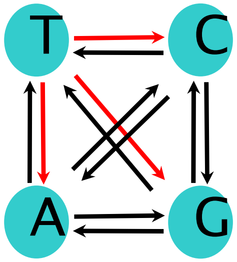
$$f(S,\Theta,\{G_i,\boldsymbol{t}_i\}\mid X)\propto f(S)\,f(\Theta\mid S)\prod_i f(G_i,\boldsymbol{t}_i\mid S,\Theta)\,f(X_i\mid G_i,\boldsymbol{t}_i)$$

S: species tree.
Θ: θ and τ parameters on the species tree.
G_i & $\boldsymbol{t}_i$: gene tree topology and branch lengths at locus i.

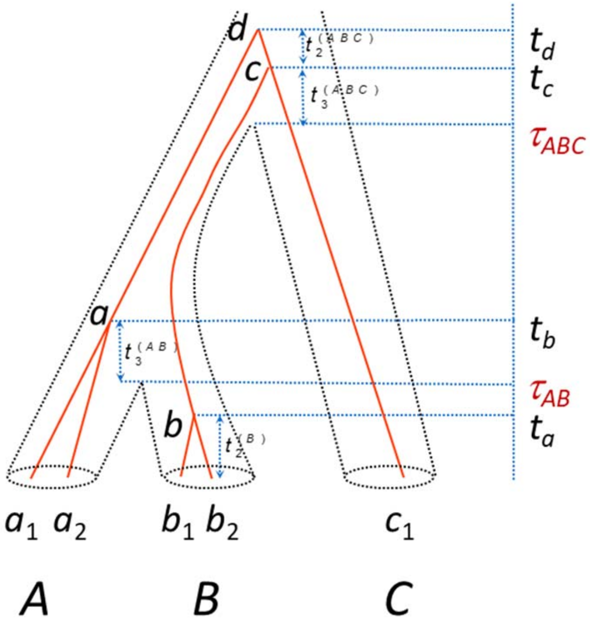
Model & data

- 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 CAGTTAGTTC …
b1    TCCATTCAAG AGTTTATTAT CAGTTAATTC …
b2    TCCATTCAAG AGTTTATTAT CAGTTAATTC …
c1    TCCATTCAAG GGTCTATTAT CAGTTAATTC …
```



JC69 model



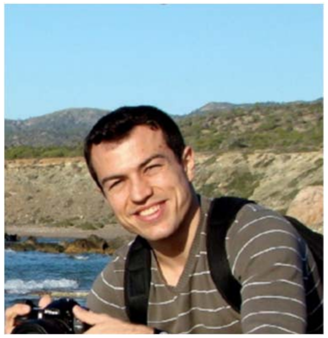
MCMC samples from the posterior: $f(S, \{\tau\text{s}, \theta\text{s}\}, \{G_i, t_i\} \mid \text{Data})$

1. Initialize S, $\{\theta\text{s}, \tau\text{s}\}$, $\{G_i, \boldsymbol{t}_i\}$.
2. Iterate
   - Change parameters (θs, τs in the model).
   - Change gene trees $\{G_i, \boldsymbol{t}_i\}$.
   - Change species tree S (by NNI, SPR, NodeSlider).
   - Save on the disk every k iterations.

S: species tree
$\{\theta\text{s}, \tau\text{s}\}$: parameters in the MSC
$\{G_i, \boldsymbol{t}_i\}$: gene trees and ages

The MCMC algorithm visits the species trees according to their posterior probabilities.

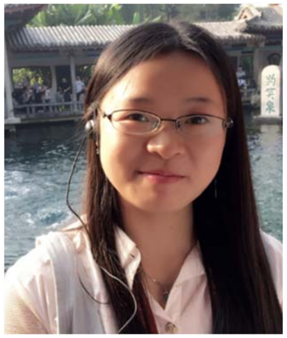
Acknowledgments



Tomas Flouri

Paschalia Kapli

Xiyun Jiao



Jun Huang

Yuttapong Thawornwattana

Tianqi Zhu

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

Funding

BBSRC, NERC