

# Introduction to Coalescent Theory

## What is Coalescent Theory?

---

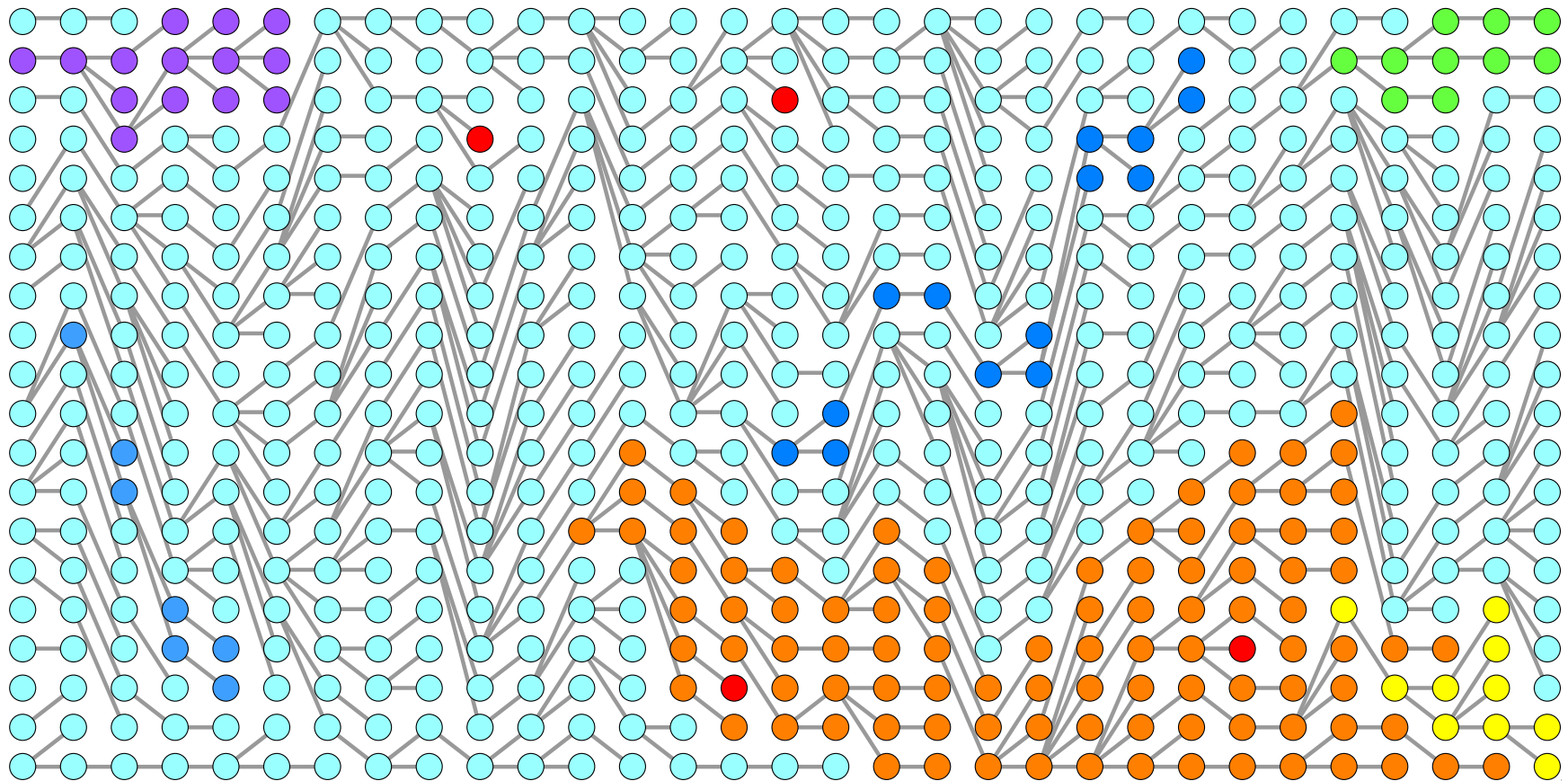
Coalescent theory is a retrospective model of population genetics based on the genealogy of gene copies. It uses mathematics for describing the characteristics of the joining of lineages back in time to a common ancestor. This lineage joining is referred to as coalescence. The theory provides the basis for estimation of the expected time to coalescence and for establishing the relationships of coalescence times to population size, age of the most recent common ancestor, and other population genetic parameters.

# Characteristics of the Wright-Fisher Population Model

- constant population size ( $N$ )
- synchronized and discrete (non-overlapping) generations
- subsequent generations drawn at random from previous generation (i.e., random mating)
- genetic drift is the only force affecting frequency

# Example Wright-Fisher Population Through Time

---



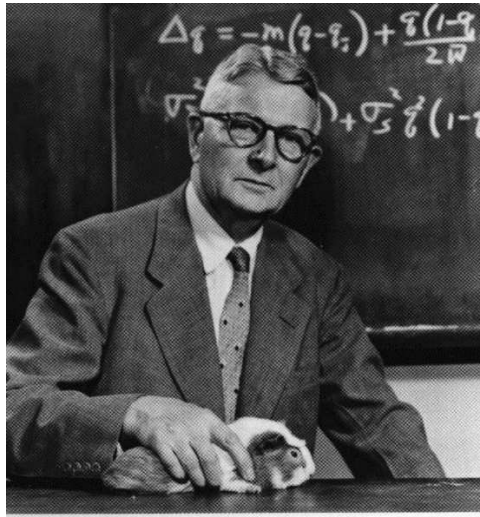
Time →

## Other population models

---

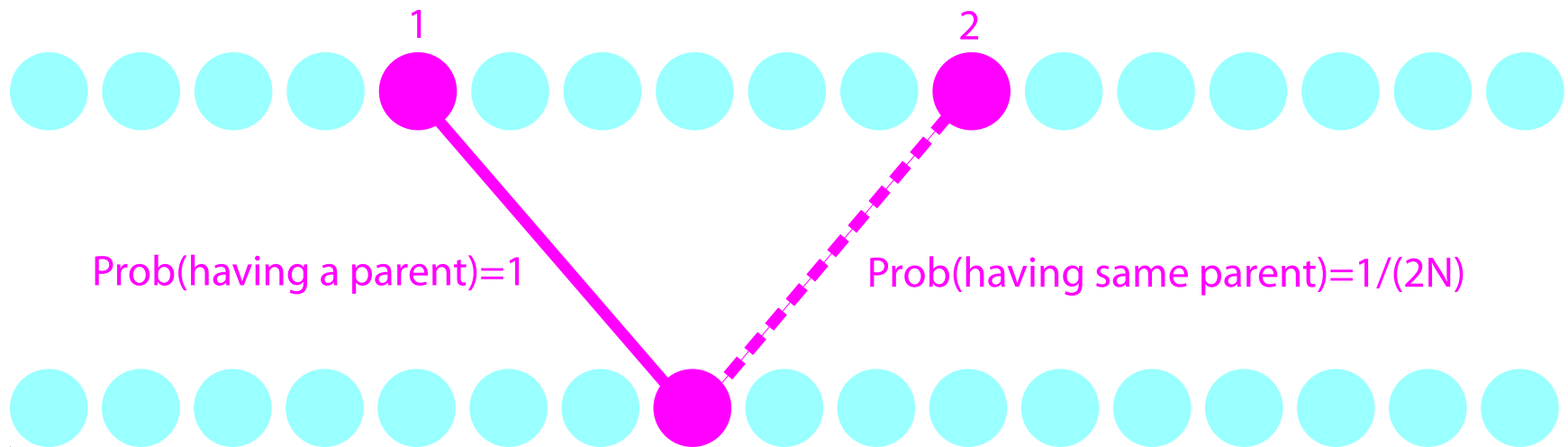
- other population models can often be put in terms of modifications to the Wright-Fisher model
- the  $N$  parameter becomes the effective population size  $N_e$
- for example, cyclic populations have an  $N_e$  that is the harmonic mean of the various sizes

# The Coalescent



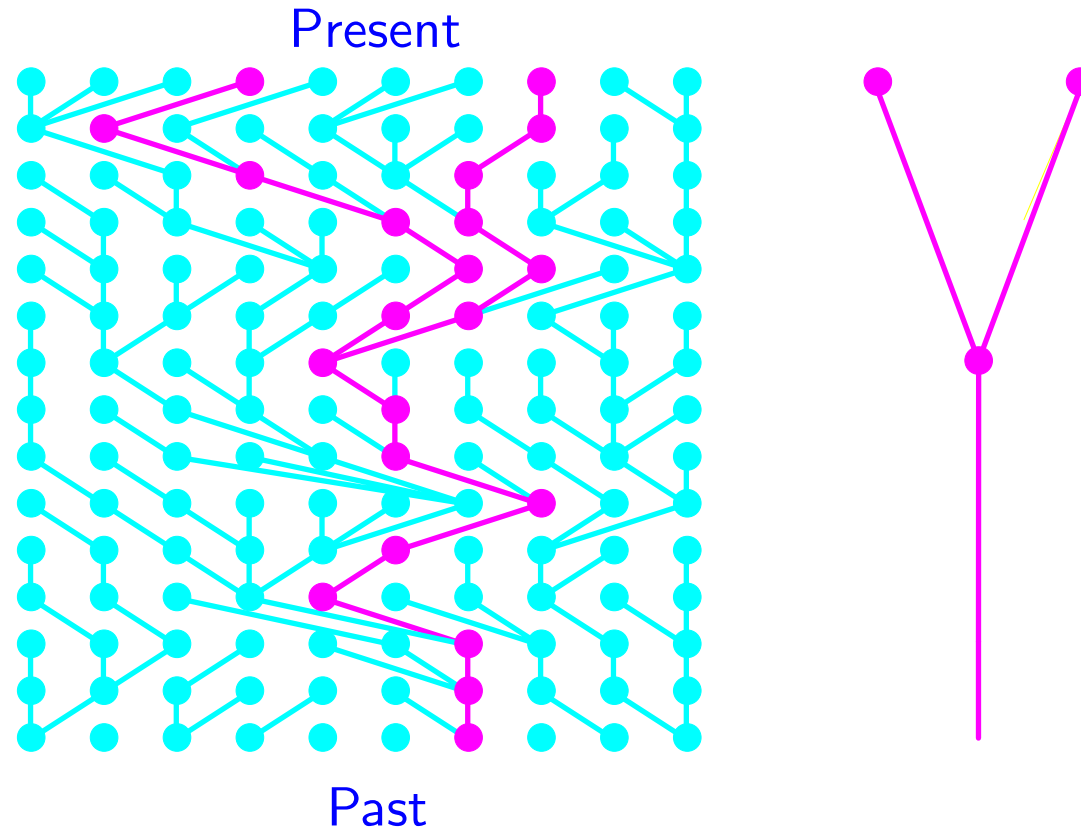
Sewall Wright showed that the probability that 2 gene copies come from the same gene copy in the preceding generation can be described as follows —

$$\text{Prob (two genes copies share a parent)} = \frac{1}{2N}$$



# The Coalescent

---



In every generation there is a chance of  $1/2N$  to coalesce. The coalescence time of the sampled lineages through previous generations follows a geometric distribution with  $\mathbb{E}(u) = 2N$  [the expectation of the time of coalescence  $u$  of two gene copies is  $2N$ ]

# The Coalescent

---

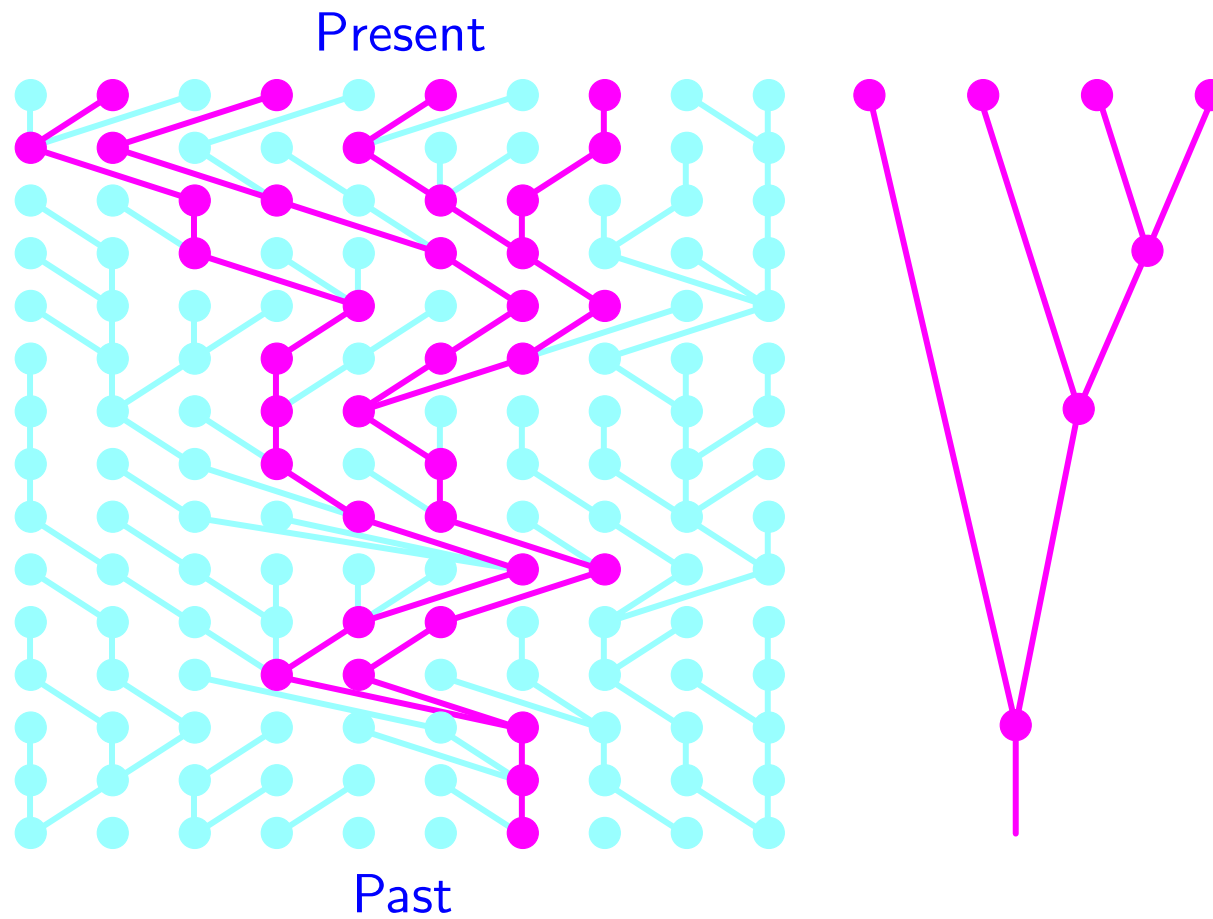


J.F.C. Kingman generalized this for  $k$  gene copies.

$$\text{Prob } (k \text{ copies are reduced to } k - 1 \text{ copies}) = \frac{k(k - 1)}{4N}$$

# Kingman's $n$ -coalescent

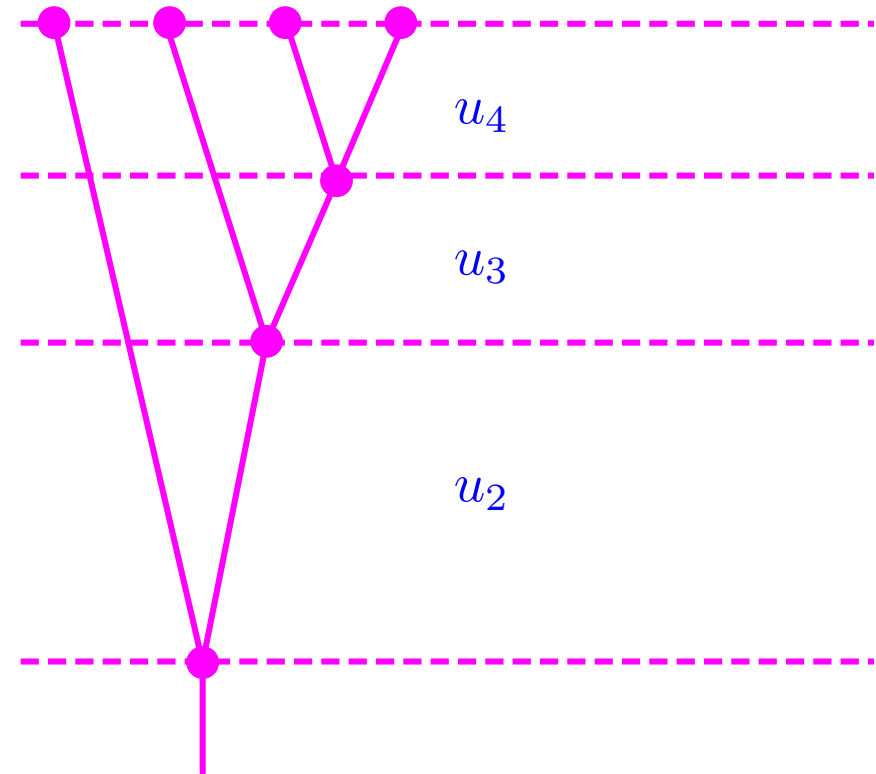
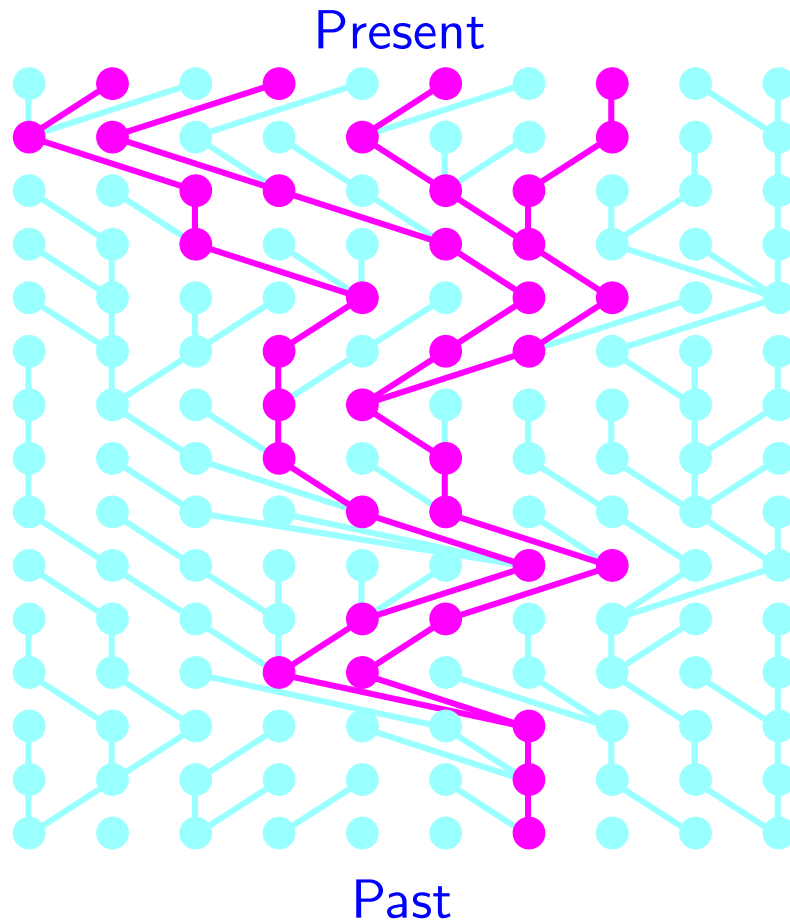
---



# Kingman's $n$ -coalescent

The expectation for the time interval  $u_k$  is

$$\mathbb{E}(u_k) = \frac{4N}{k(k-1)}$$



$$p(\text{Genealogy}|N) = \prod_j^T e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{1}{2N}$$

## Some Properties of the Coalescent

---

- the larger the sample size the greater the rate of coalescence (i.e., the more lineages there are the greater the probability that two will coalesce)

as  $k \uparrow, k(k - 1)/4N \uparrow$

- the larger the population size the slower the rate of coalescence

as  $N \uparrow, 1/2N \downarrow$

- time to coalescence gets longer as the process moves toward the most recent common ancestor

as  $k \downarrow, \mathbb{E}(u_k) = 4N/k(k - 1) \uparrow$

- small samples sizes have a high probability of including the most recent common ancestor of the population

$$Prob(MRCA) = (k - 1)/(k + 1)$$

## Expected Time to the Most Recent Common Ancestor

- diploids:  $\mathbb{E}(u) = 4N$
- haploids:  $\mathbb{E}(u) = 2N$
- uniparentally inherited genes (e.g., cpDNA, mtDNA, Y chromosome):  
 $\mathbb{E}(u) = N$

## The $\Theta$ parameter

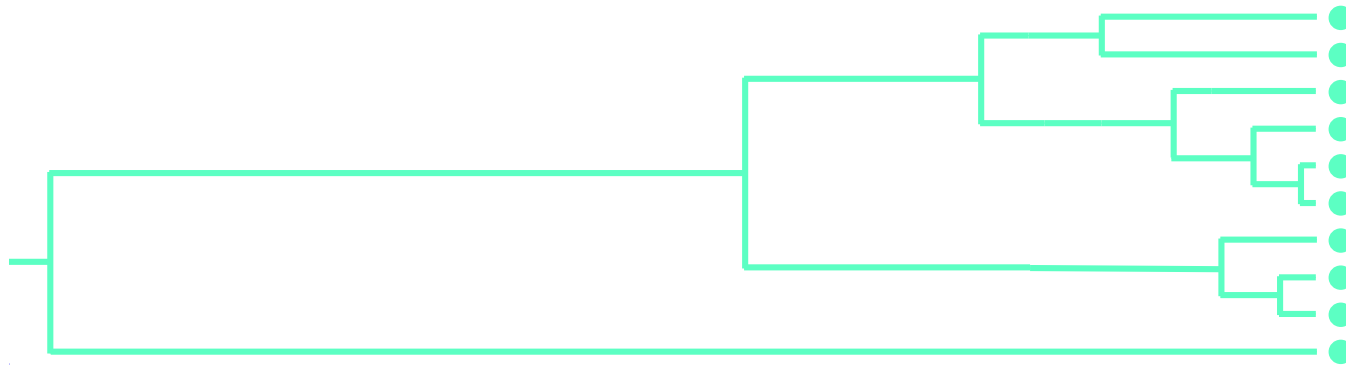
---

- the coalescent is defined in terms of  $N$  ( $N_e$ ) and time
- time cannot be measured directly with genetic data, though we can measure genetic divergence
- the equations are rescaled in terms of  $N_e$ , time, and the mutation rate  $\mu$
- $N_e$  cannot be estimated directly, but rather the composite parameter  $\Theta$  is estimated
- $\Theta = 4N_e\mu$  in diploids

# Idealized Population Size Estimator

---

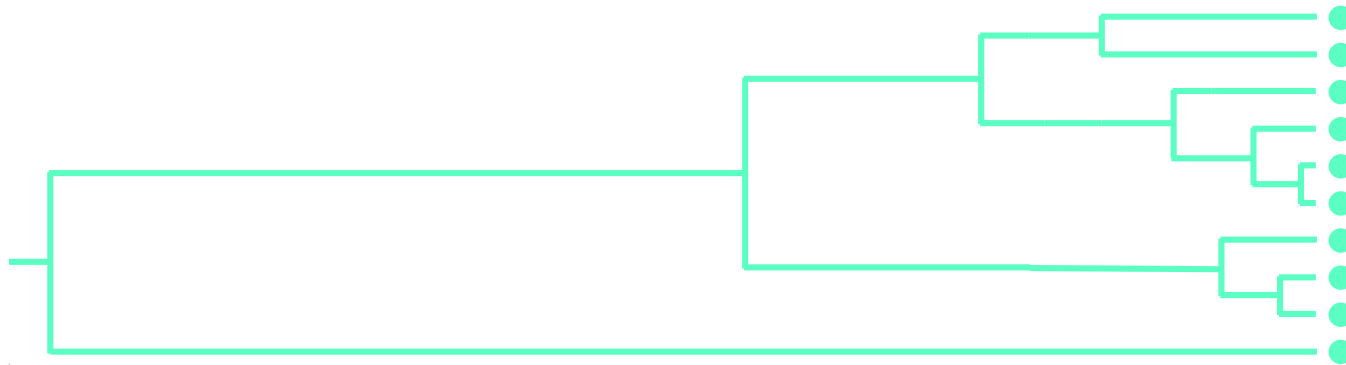
1. have complete and accurate knowledge of the genealogy
2. calculate  $p(\text{Genealogy}|\text{N})$



# Idealized Population Size Estimator

---

1. have complete and accurate knowledge of the genealogy
2. calculate  $p(\text{Genealogy}|\text{N})$

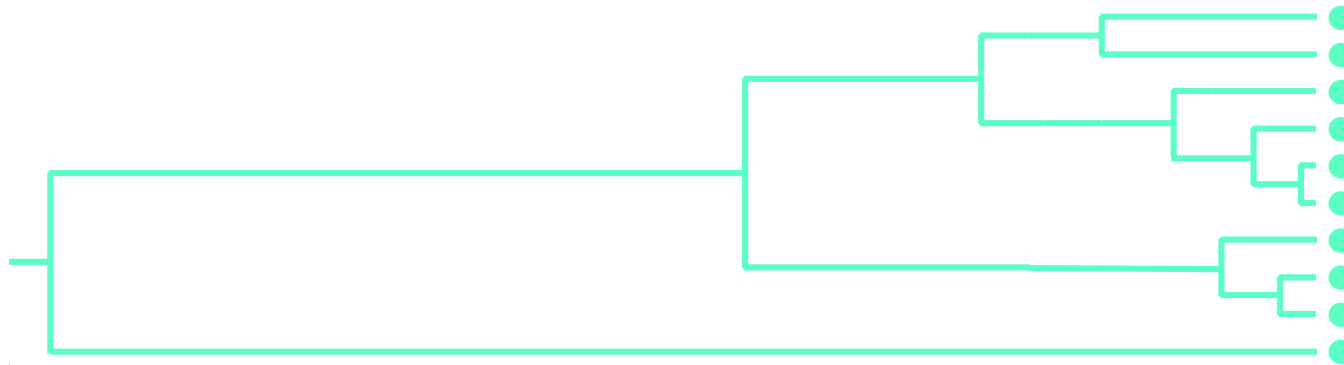


$$p(\text{Genealogy}|\text{N}) = p(u_1|\text{N}, k) \frac{1}{2\text{N}} \times p(u_2|\text{N}, k - 1) \frac{1}{2\text{N}} \times \dots$$

# Idealized Population Size Estimator

---

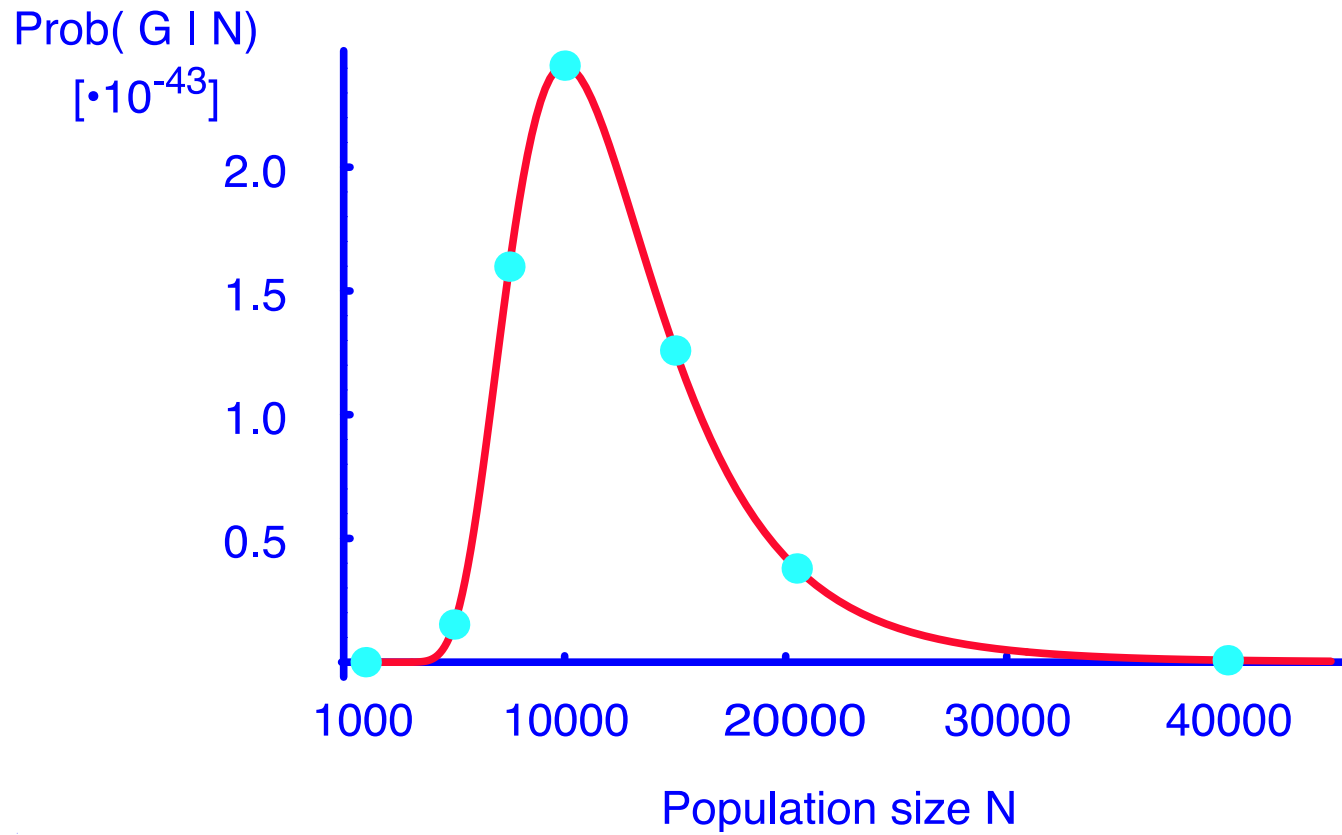
1. have complete and accurate knowledge of the genealogy
2. calculate  $p(\text{Genealogy}|\text{N})$



$$p(\text{Genealogy}|\text{N}) = \prod_j^T e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{1}{2N}$$

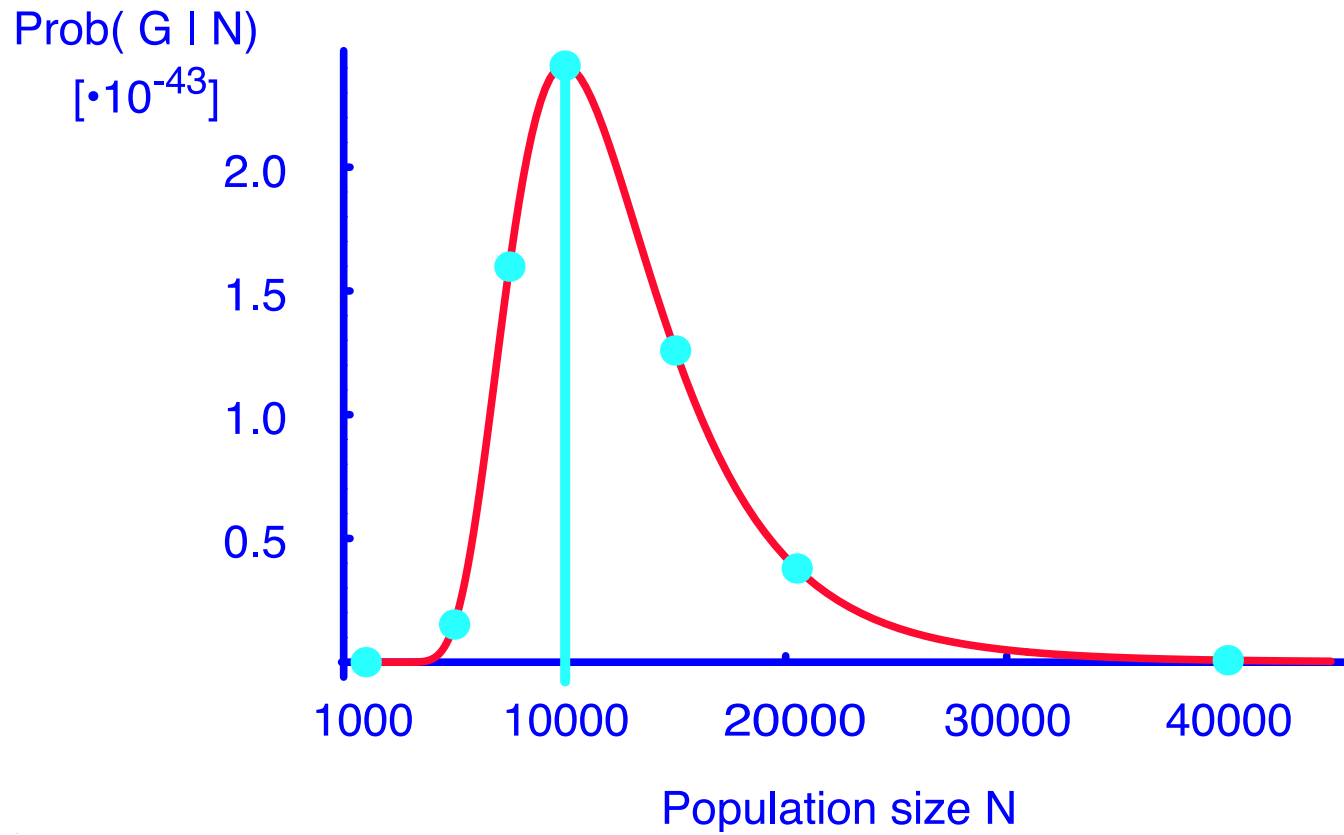
# Idealized Population Size Estimator

---



# Idealized Population Size Estimator

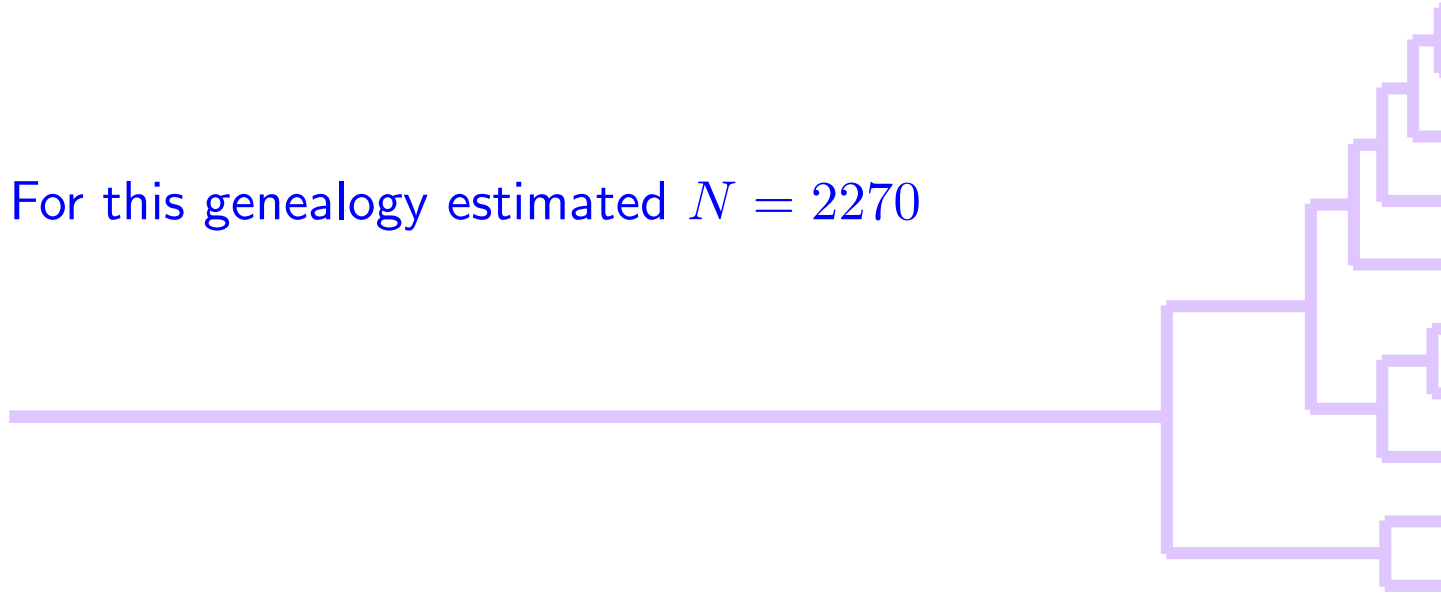
---



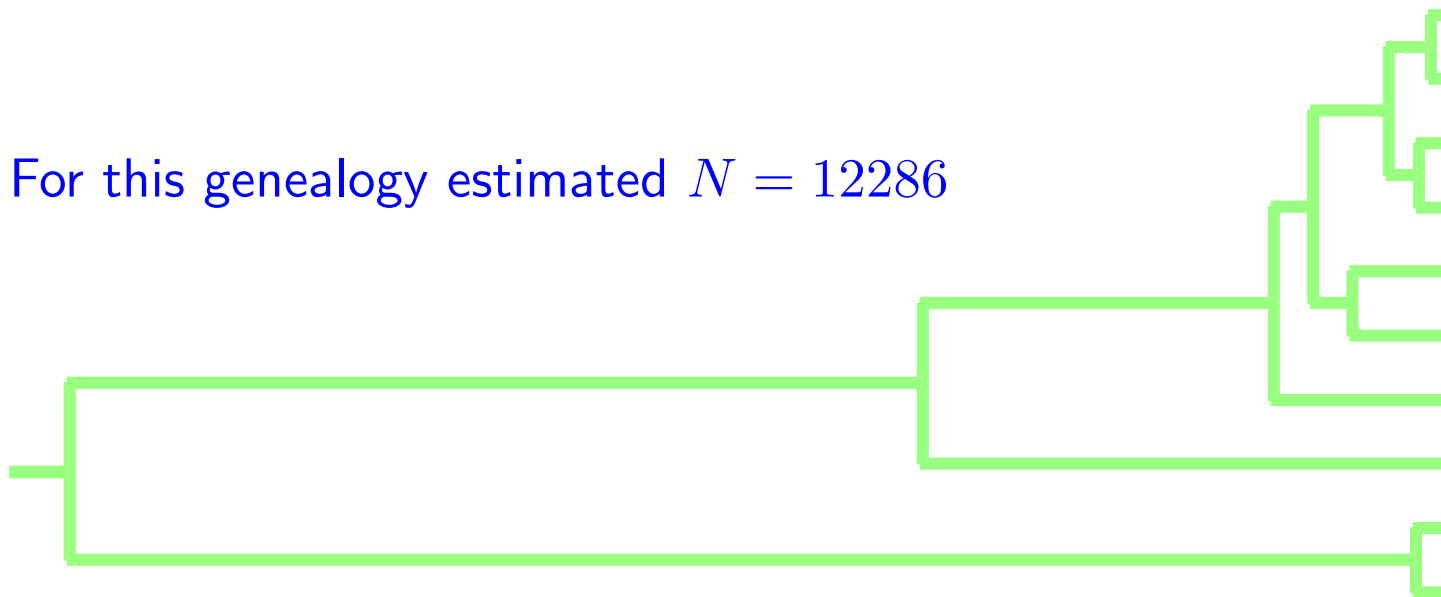
# Idealized Population Size Estimator

---

For this genealogy estimated  $N = 2270$



For this genealogy estimated  $N = 12286$



## Back to Reality . . .

---

- complete and accurate knowledge of the genealogy is unavailable (except in simulations)
- accurate inference of genealogies is a hard problem in general
- population data usually does not have enough information for accurate inference of a genealogy

# Non-likelihood use of the Coalescent

---

- Summary statistics
  - Watterson's estimator of  $\theta$
  - $F_{ST}$  (estimates  $\theta$  and/or migration rate)
  - Hudson's and Wakeley's estimators of recombination rate
- Known-tree methods
  - UPBLUE (Yang)
  - Skyline plots (Strimmer, Pybus, Rambaut)

These methods are conceptually easy, but not always powerful, and they are difficult to extend to complex cases.

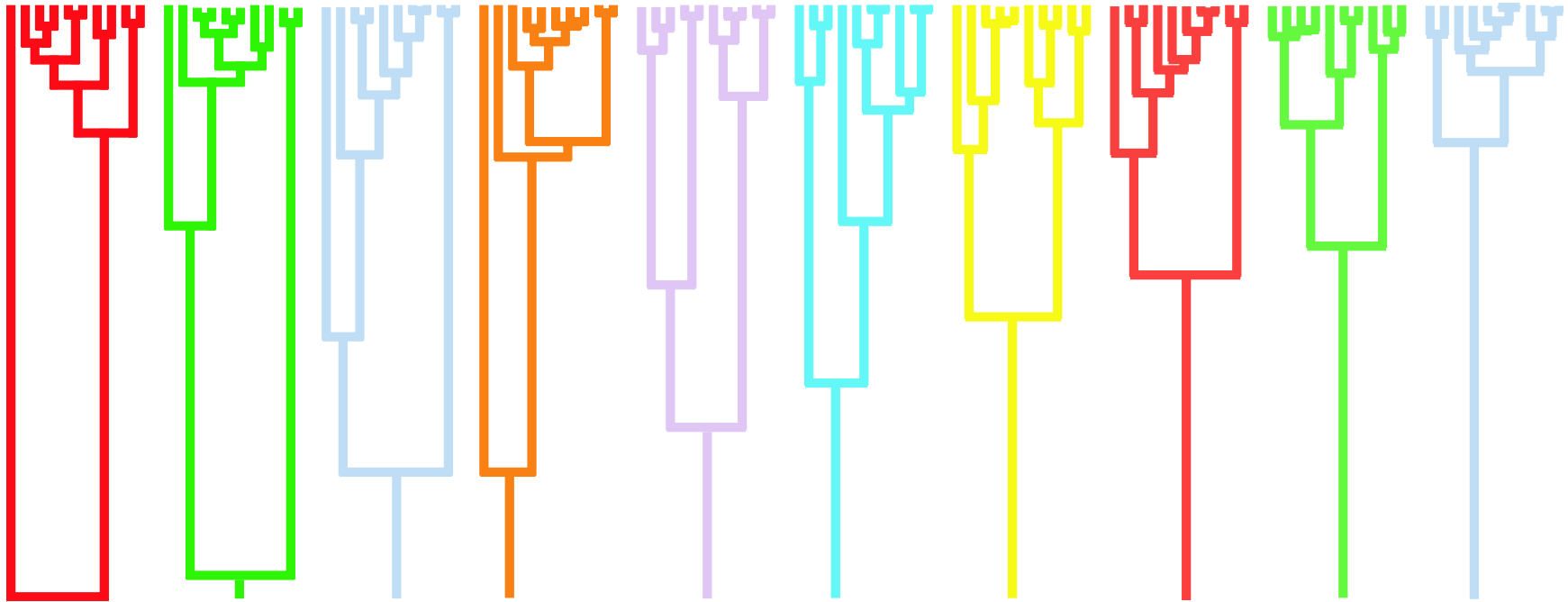
# Variance of the Coalescent

---

- Stochastic variance: variance of the evolutionary genetic processes
  - variance caused by the variance in times to coalescence
  - variance caused by the distribution of mutations on the genealogy
- Sampling variance: variance of experimentation
  - variance caused by sampling of individuals, loci, and gene copies

## Stochastic Variance of the Coalescent

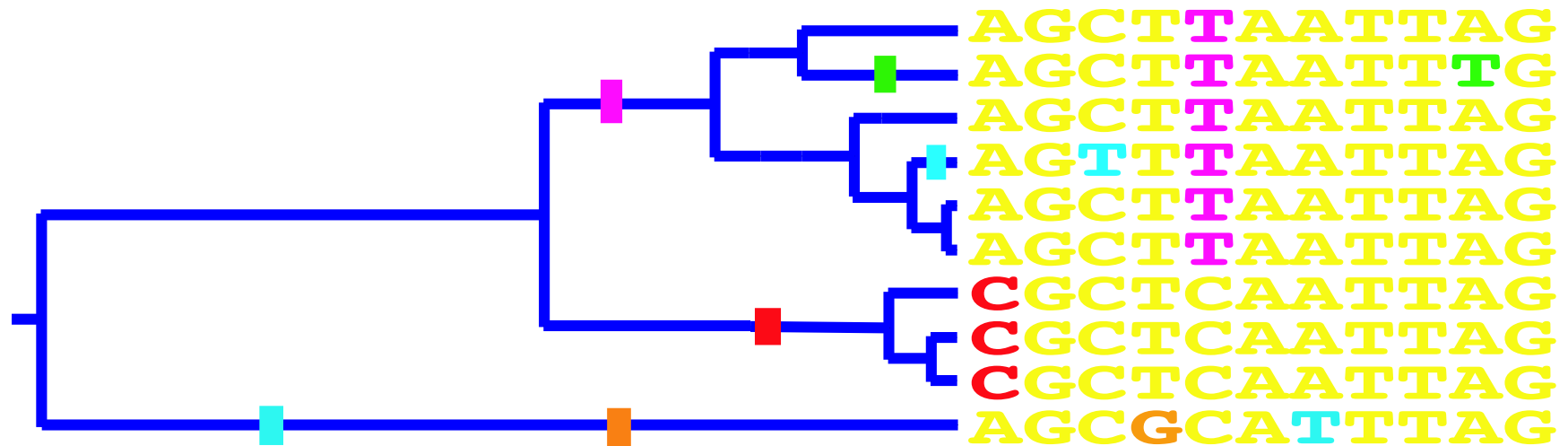
---



10 coalescent genealogies of the same sample size,  $n = 10$ , with the same population size,  $N = 10,000$ , sorted by age to the most recent common ancestor (decreasing order)

## Variability of mutations

---

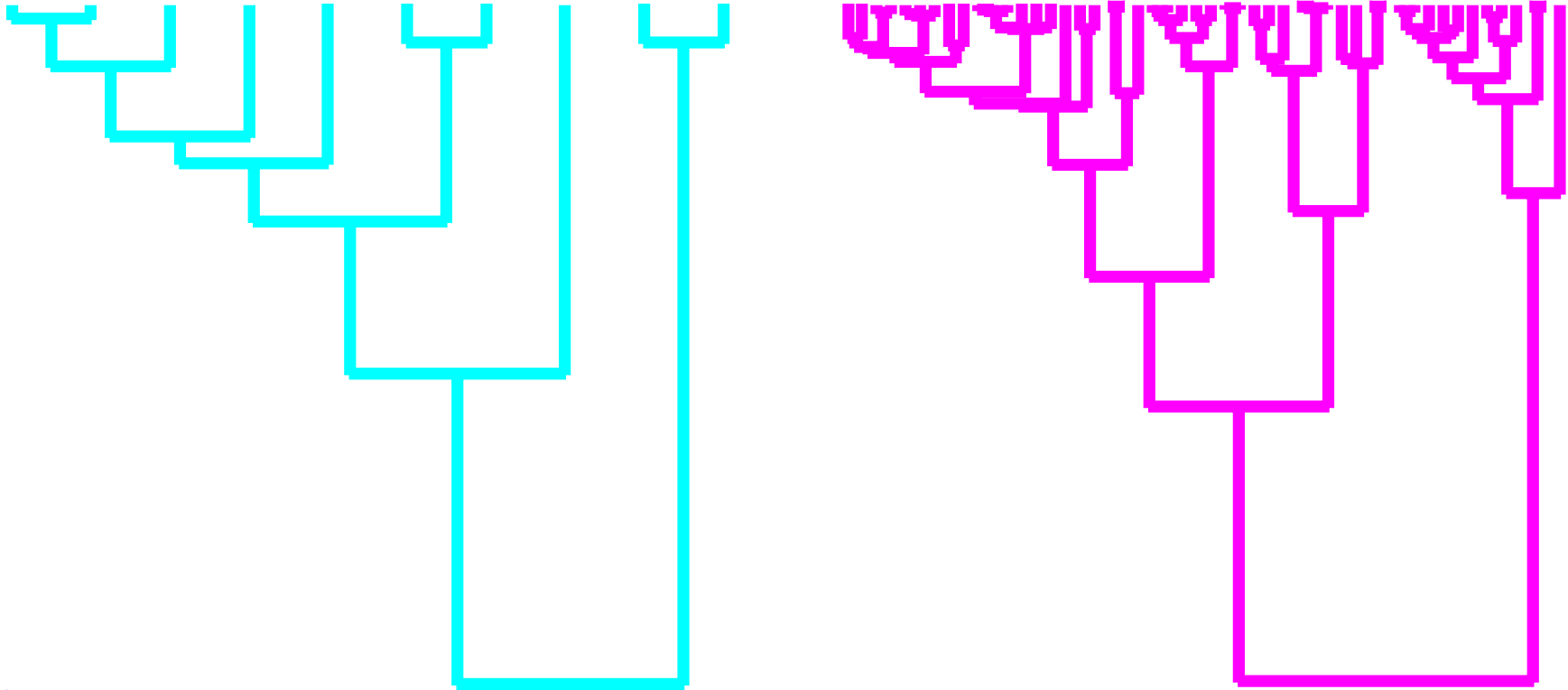


the genealogical and mutational processes are independent — it is sometimes useful to think of the genealogy having been created first, and the mutations then occurring along branches of the genealogy (this is the way that coalescent simulations are often done)

note: analysis of the sequences would not result in the genealogy shown — although there is no inconsistency between the genealogy and the sequences, the information in the sequences is insufficient for accurate inference of the genealogy

## Does adding more gene copies help?

---



added gene copies to the genealogy on the left join the genealogy in the more recent generations (evident as increased short branches at top of the genealogy on the right), and thus do not provide any additional information regarding the time to the most recent common ancestor

## **The bottom line**

---

1. The coalescent history of a population contains usable information.
2. The information content of a single locus is limited.
3. Additional sequence length or gene copies are only mildly helpful.
4. Multiple loci allow the best estimates.