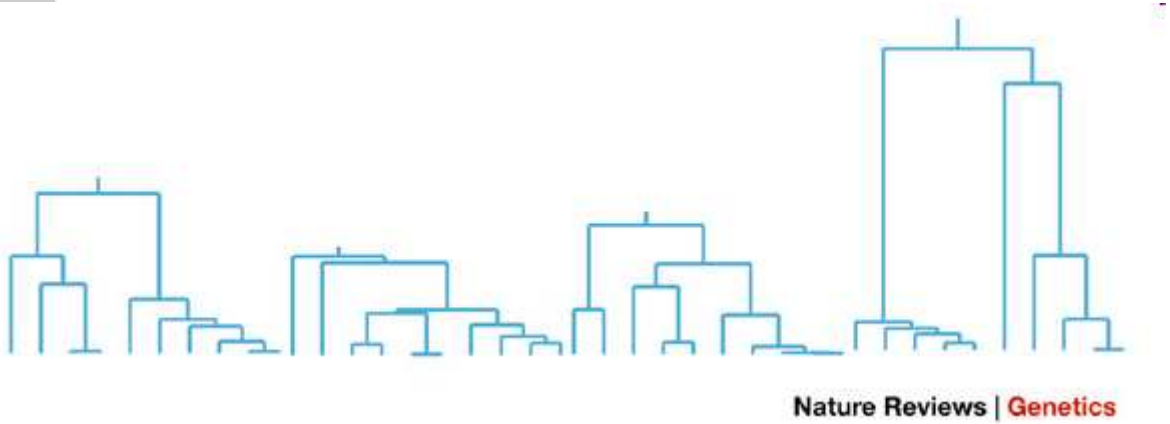
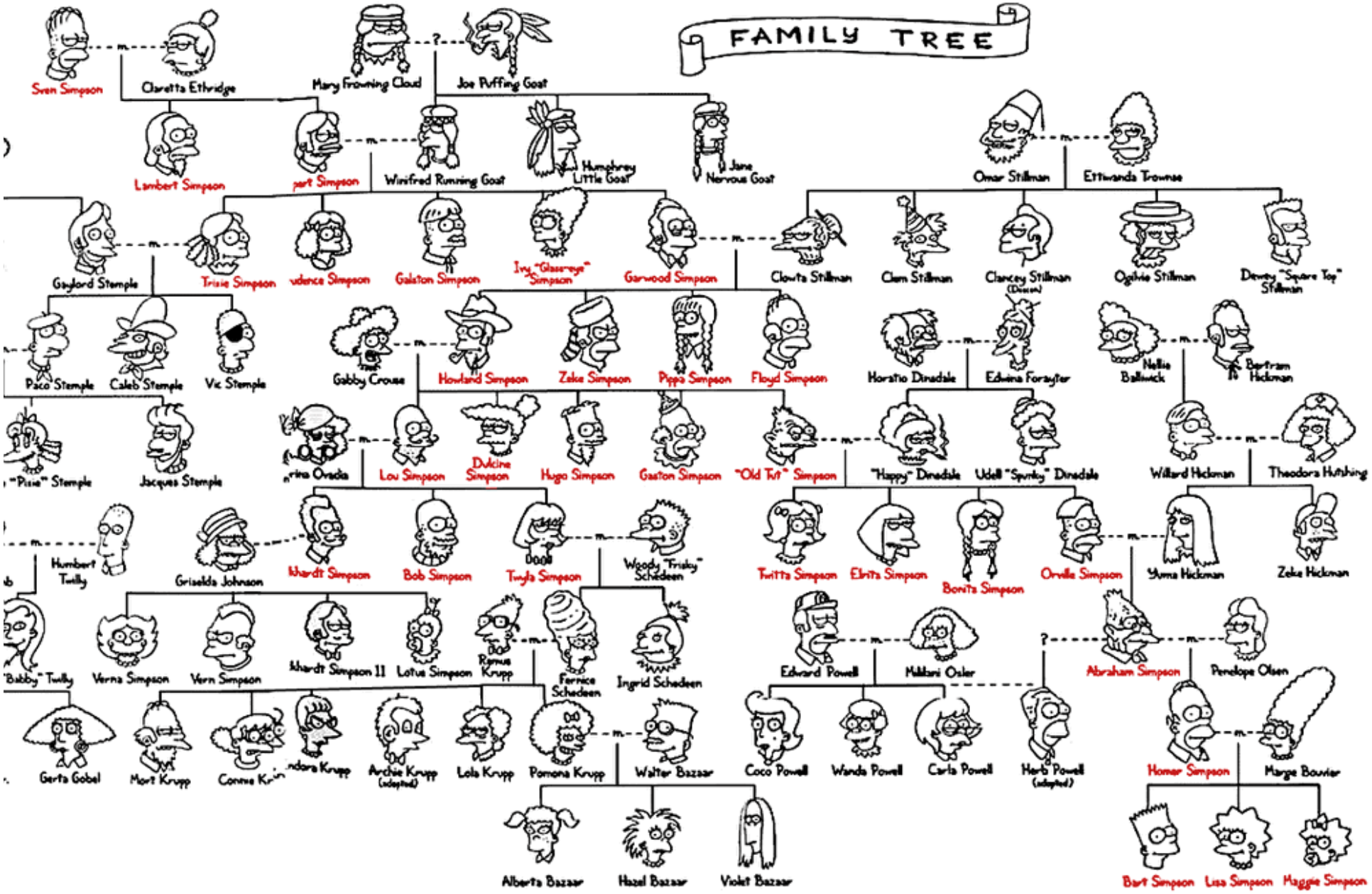
Gene Genealogies Coalescence Theory



Annabelle Haudry

Glasgow, July 2009


FAMILY TREE
Sven Simpson
Claretta Ethridge
Mary Frowning Cloud
Joe Puffing Goat
Lambert Simpson
Winifred Running Goat
Humphrey Little Goat
Jane Nervous Goat
Omar Stillman
Ettwanda Trownse
Gaylord Stemple
Trixie Simpson
Galston Simpson
Ivy "Glass-eye" Simpson
Garwood Simpson
Clowta Stillman
Clem Stillman
Clancey Stillman (Divorced)
Ogilvie Stillman
Dewey "Square Toy" Stillman
Paco Stemple
Caleb Stemple
Vic Stemple
Gabby Crouse
Howland Simpson
Zeke Simpson
Pippa Simpson
Floyd Simpson
Horatio Dinedale
Edwina Forsyter
Nellie Ballmick
Bertram Hickman
"Pixie" Stemple
Jacques Stemple
Lou Simpson
Dulcine Simpson
Hugo Simpson
Gaston Simpson
"Old Tut" Simpson
"Happy" Dinedale
Udell "Spunky" Dinedale
Willard Hickman
Theodora Huthsing
Humbert Twilly
Griselda Johnson
Khardt Simpson
Bob Simpson
Twyla Simpson
Woody "Frisky" Schedeen
Twitta Simpson
Elvita Simpson
Bonita Simpson
Orville Simpson
Yuma Hickman
Zeke Hickman
"Bobby" Twilly
Verna Simpson
Vern Simpson
Khardt Simpson II
Lotus Simpson
Romus Krupp
Fernice Schedeen
Ingrid Schedeen
Edward Powell
Mildoni Osler
Abraham Simpson
Penelope Olsen
Gerta Gobel
Mort Krupp
Archie Krupp (adopted)
Lola Krupp
Pomona Krupp
Walter Bazaar
Coco Powell
Wanda Powell
Carla Powell
Herb Powell (adopted)
Homer Simpson
Marge Bouvier
Alberta Bazaar
Hazel Bazaar
Violet Bazaar
Bart Simpson
Lisa Simpson
Maggie Simpson

What could tell a gene genealogy?

- How much diversity in the population?
- Has the demographic size of the population changed? How? When?
- Is that gene under selection?
- Is/was there gene flow between sub-populations?

Could I describe the evolutionary model under which the species has been evolving?

Wright-Fisher: a simple model

Finite and constant population size

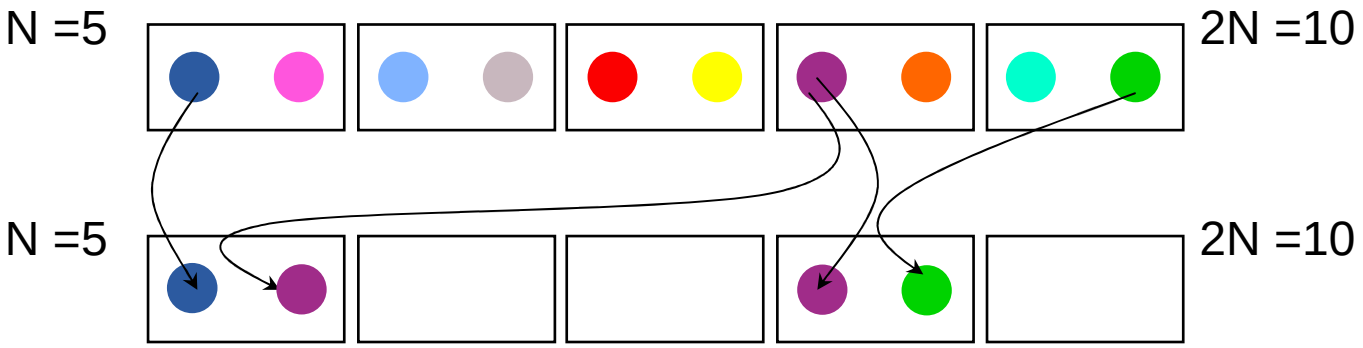
N diploid individuals, 2N genes

Random mating

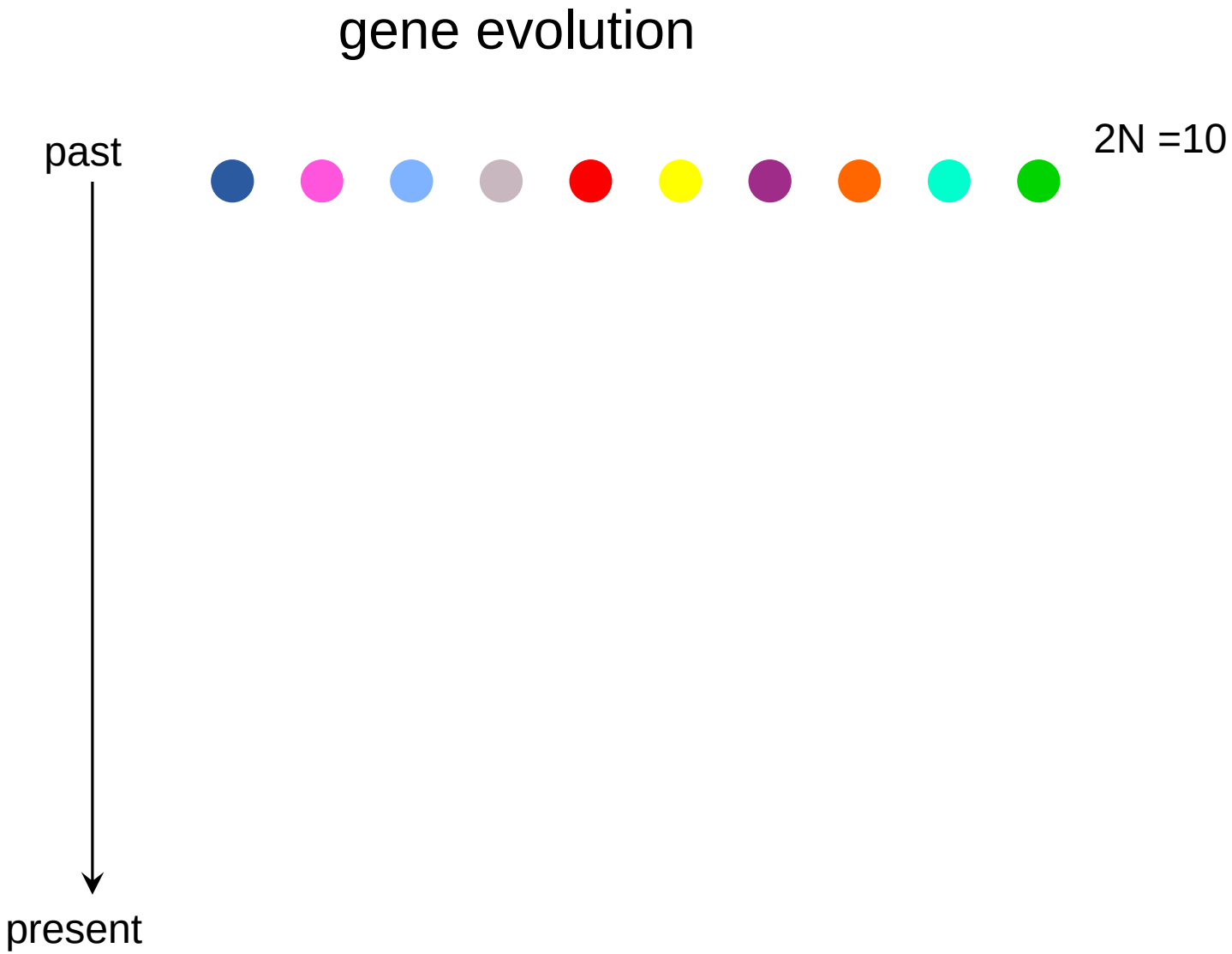
Non overlapping generations

No selection, structure…

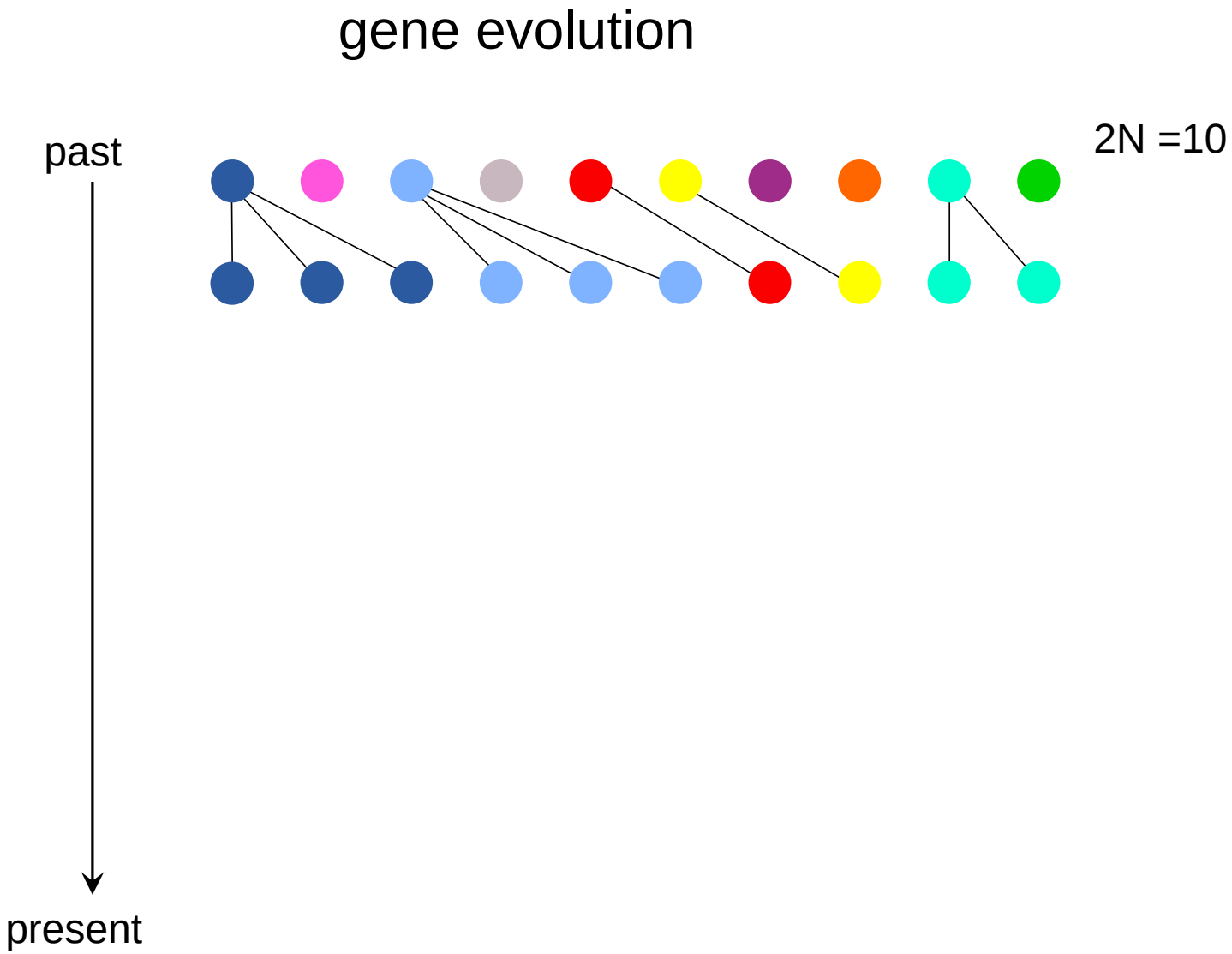
Only Drift



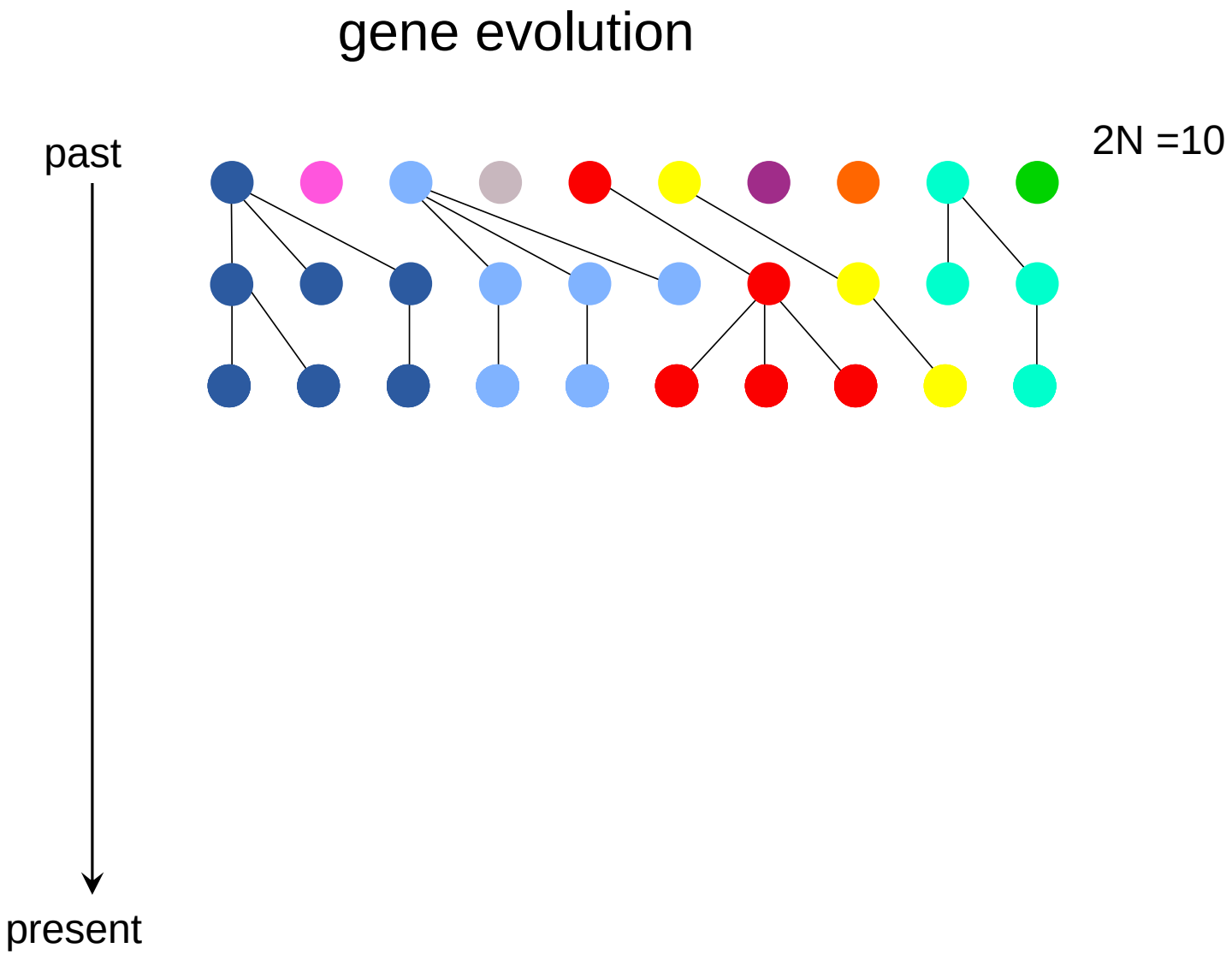
Genetic Drift



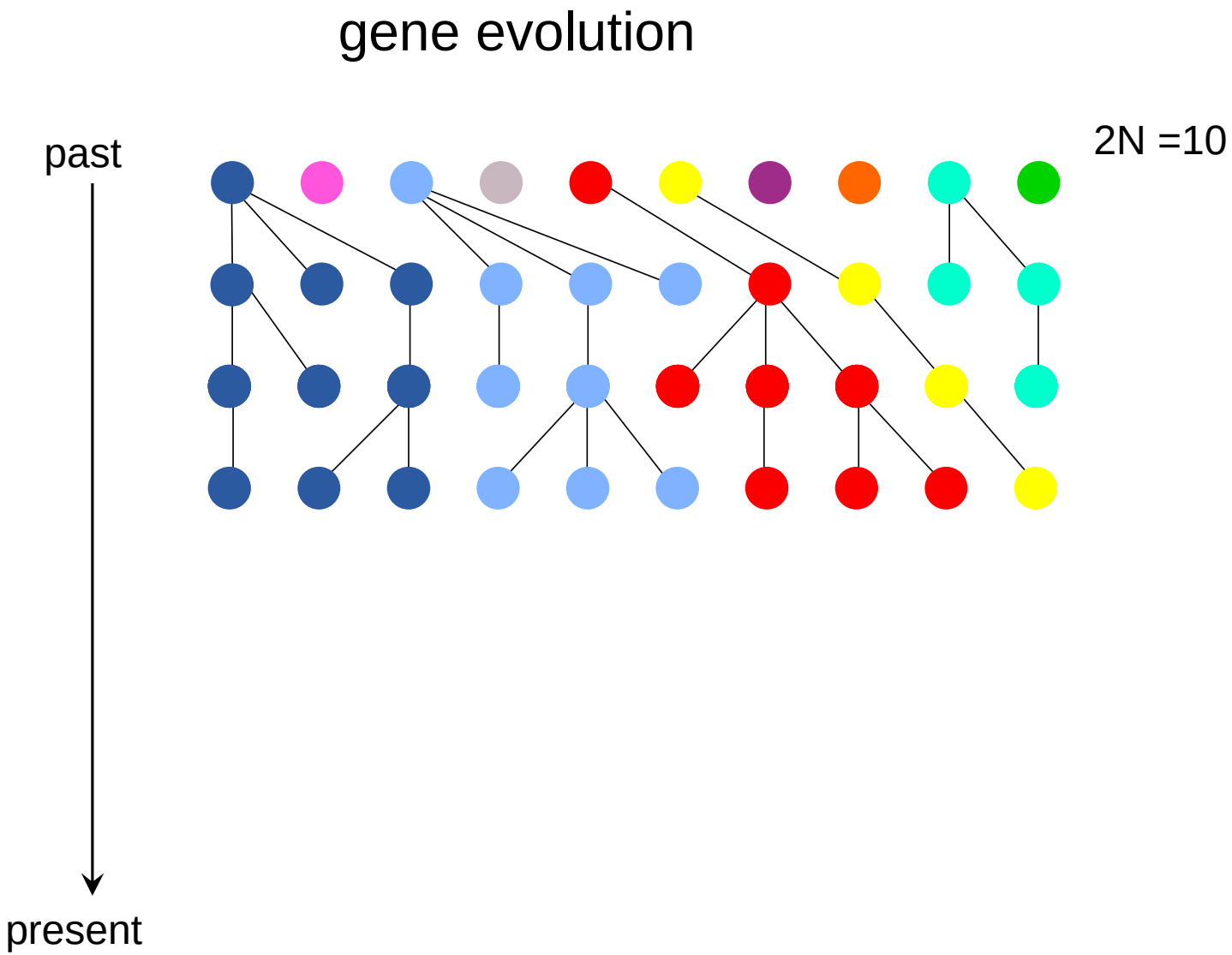
Genetic Drift



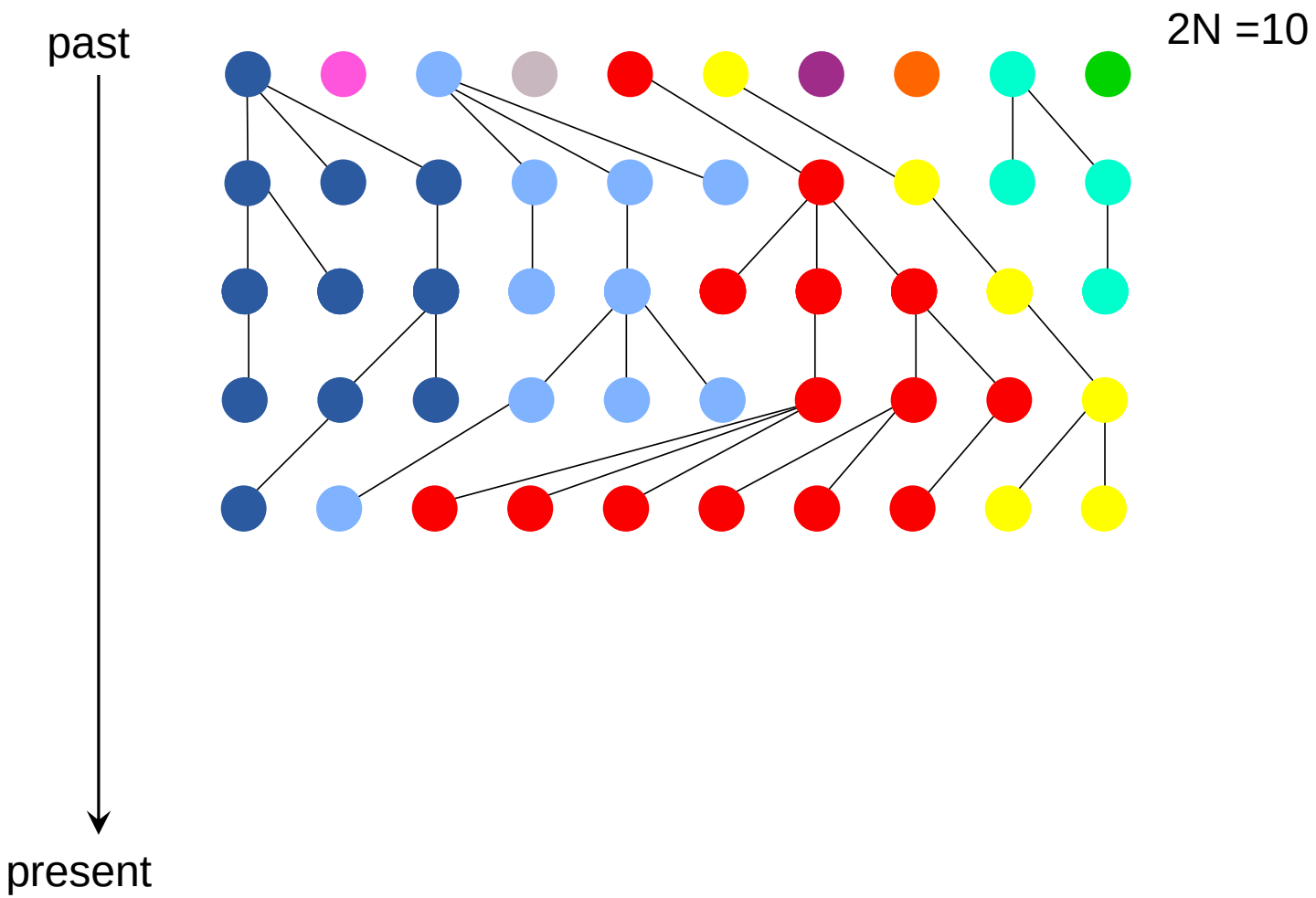
Genetic Drift



Genetic Drift

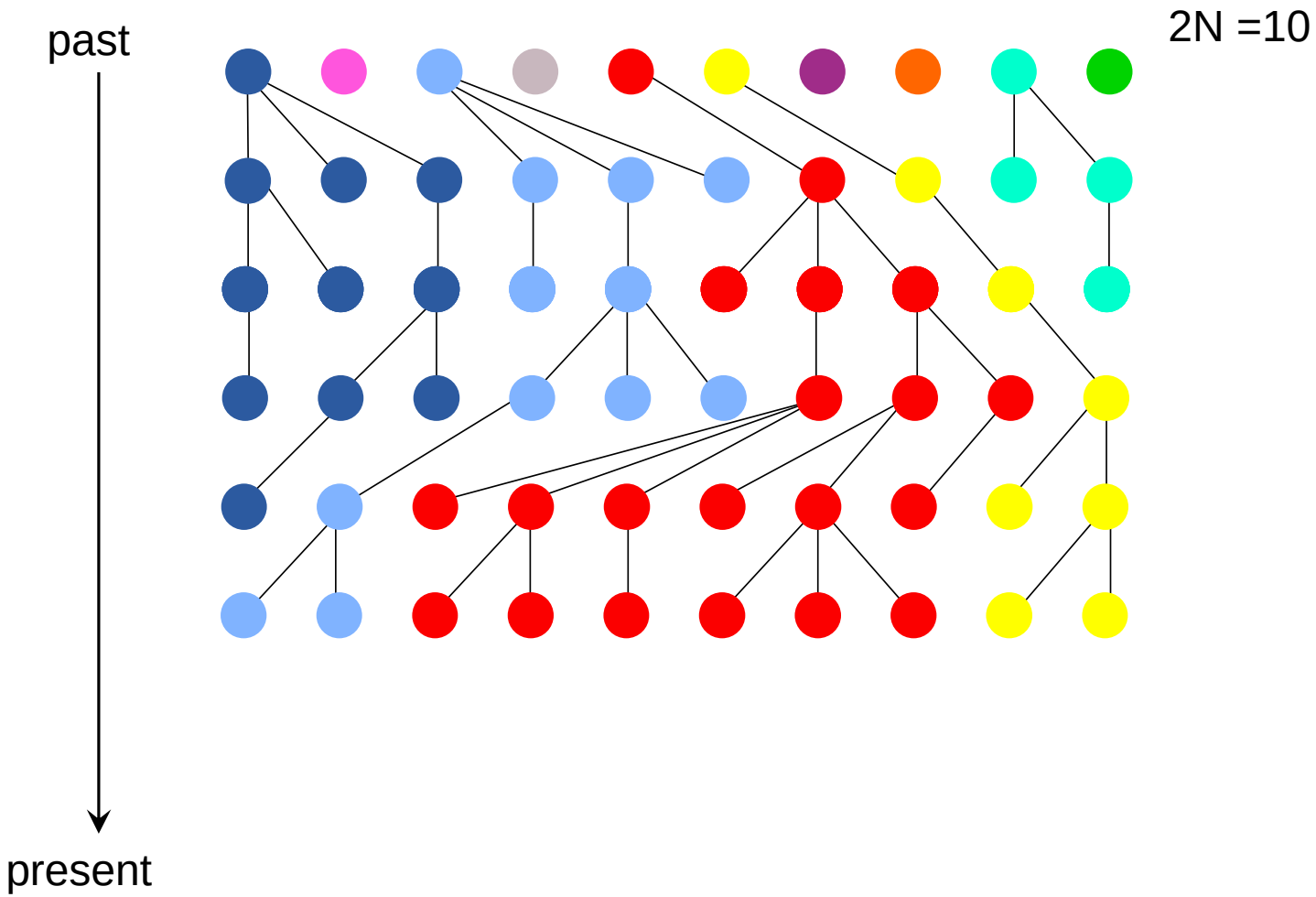


Genetic Drift



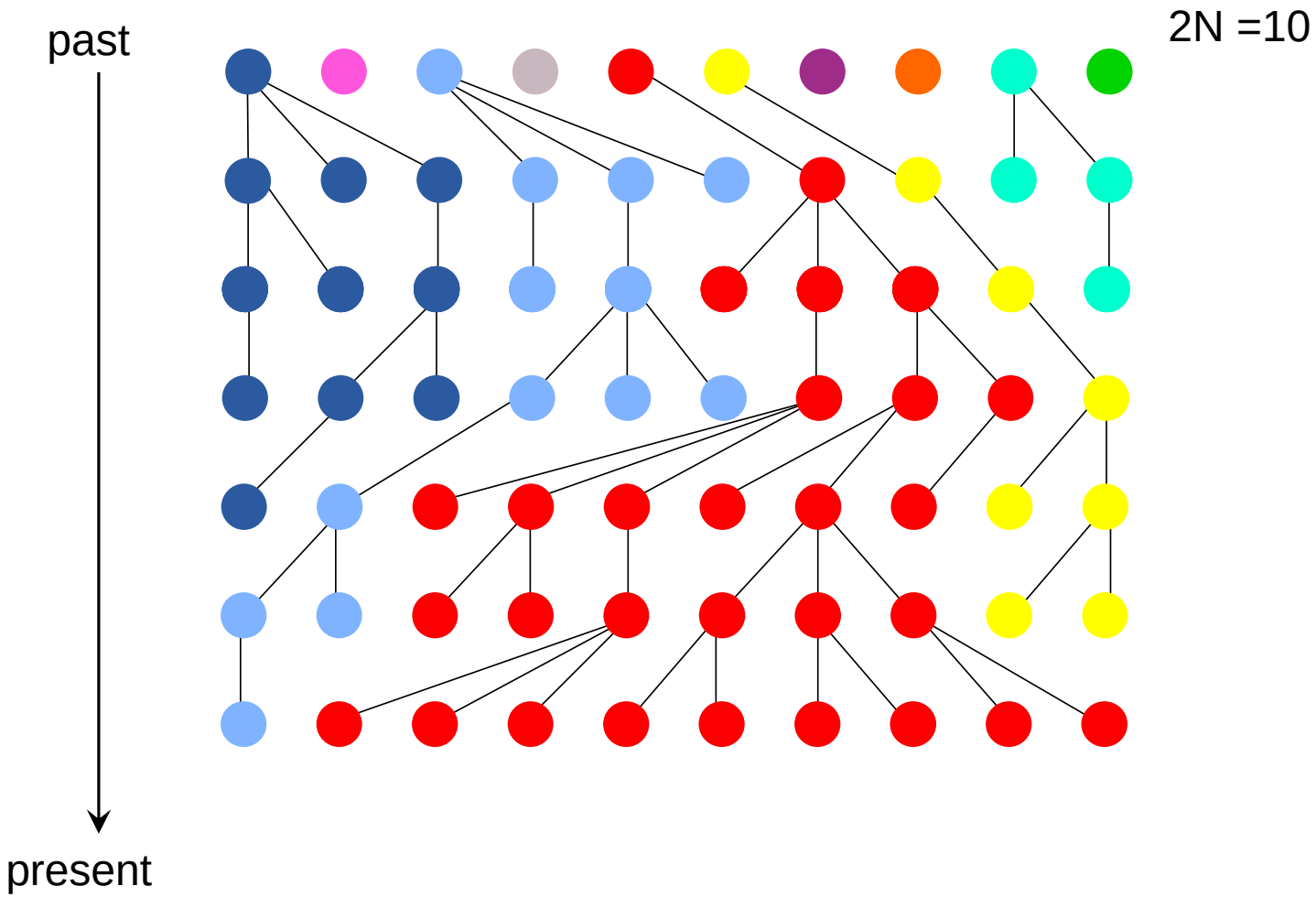
Genetic Drift

gene evolution



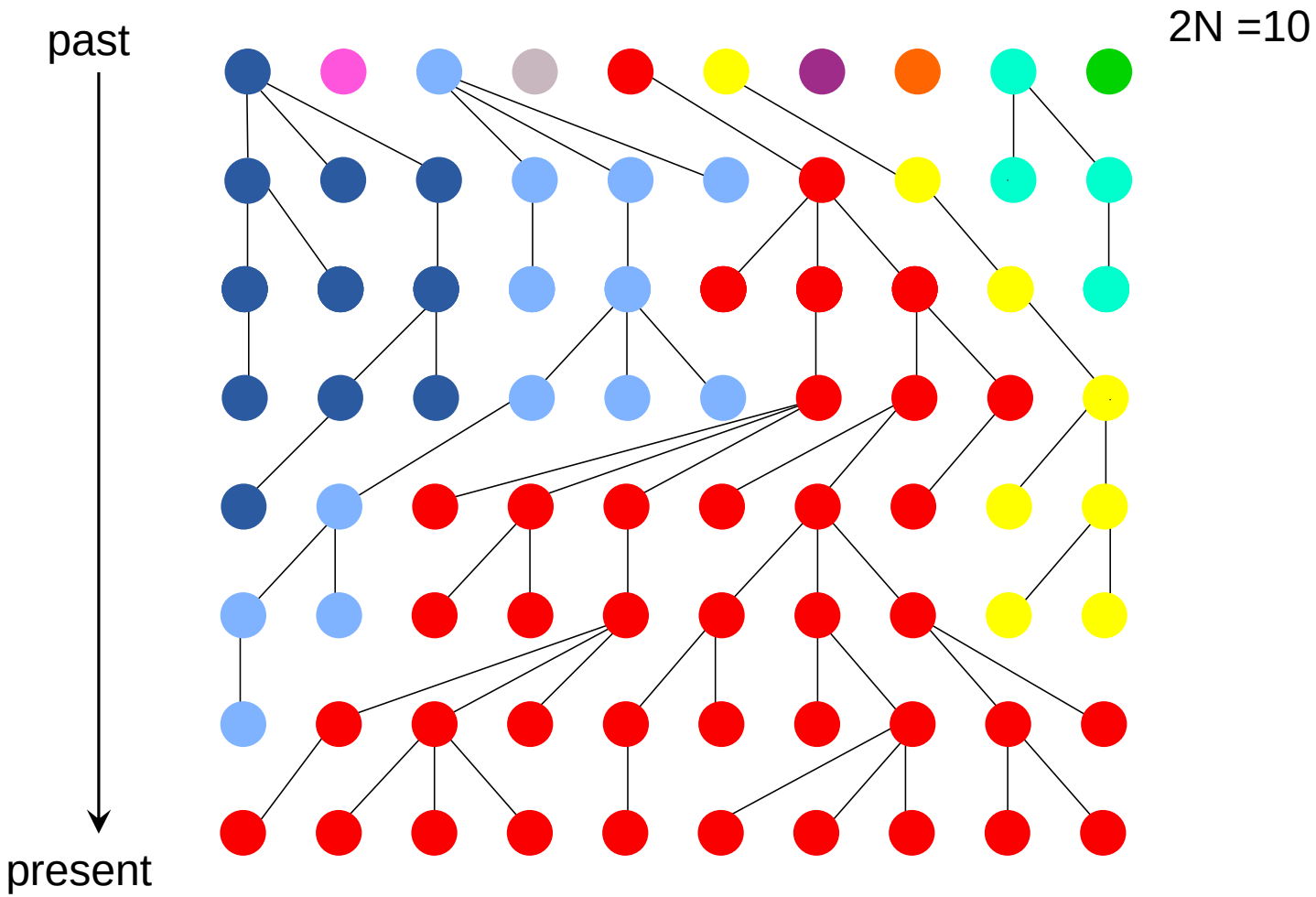
Genetic Drift

gene evolution

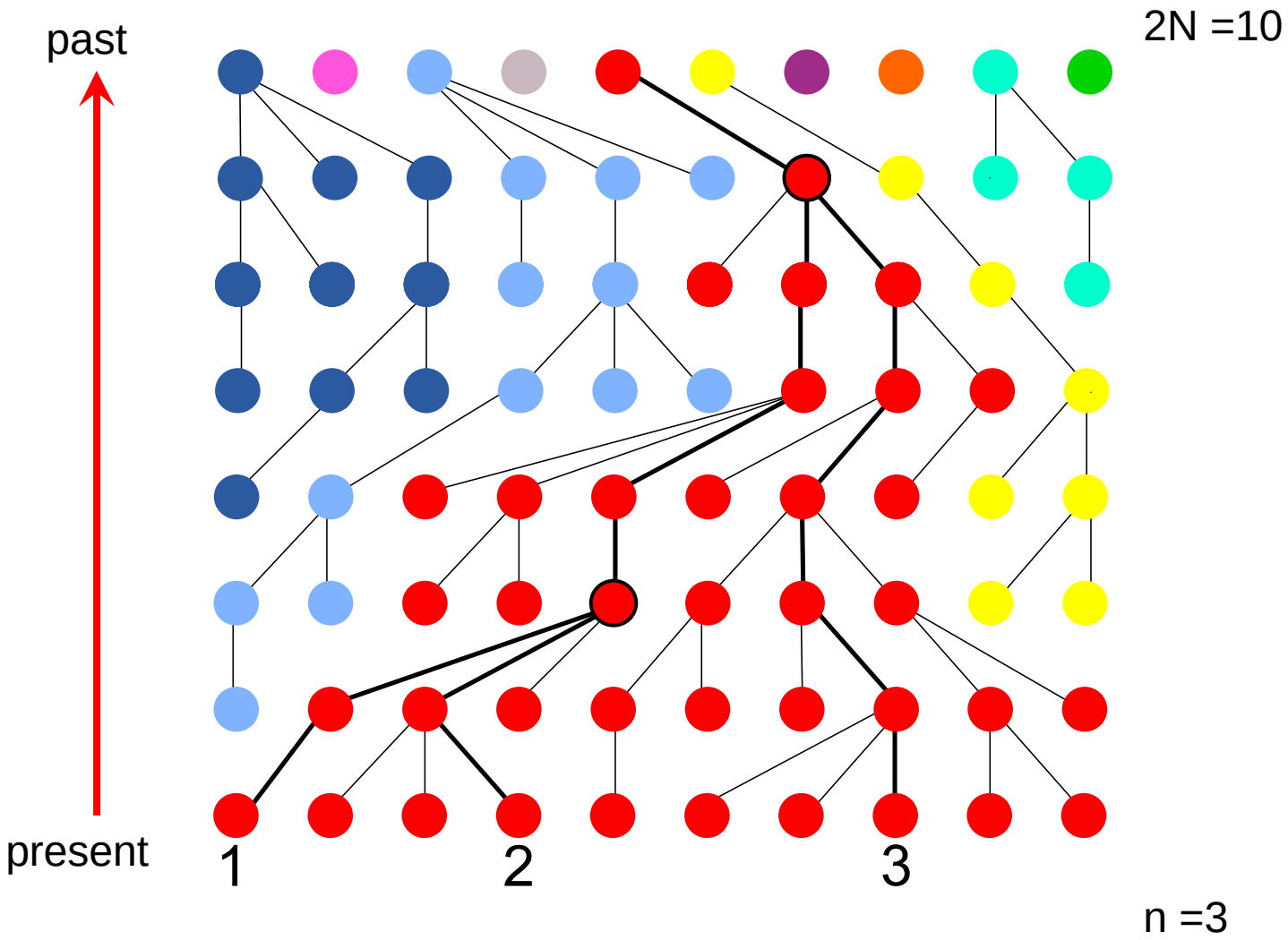


Genetic Drift

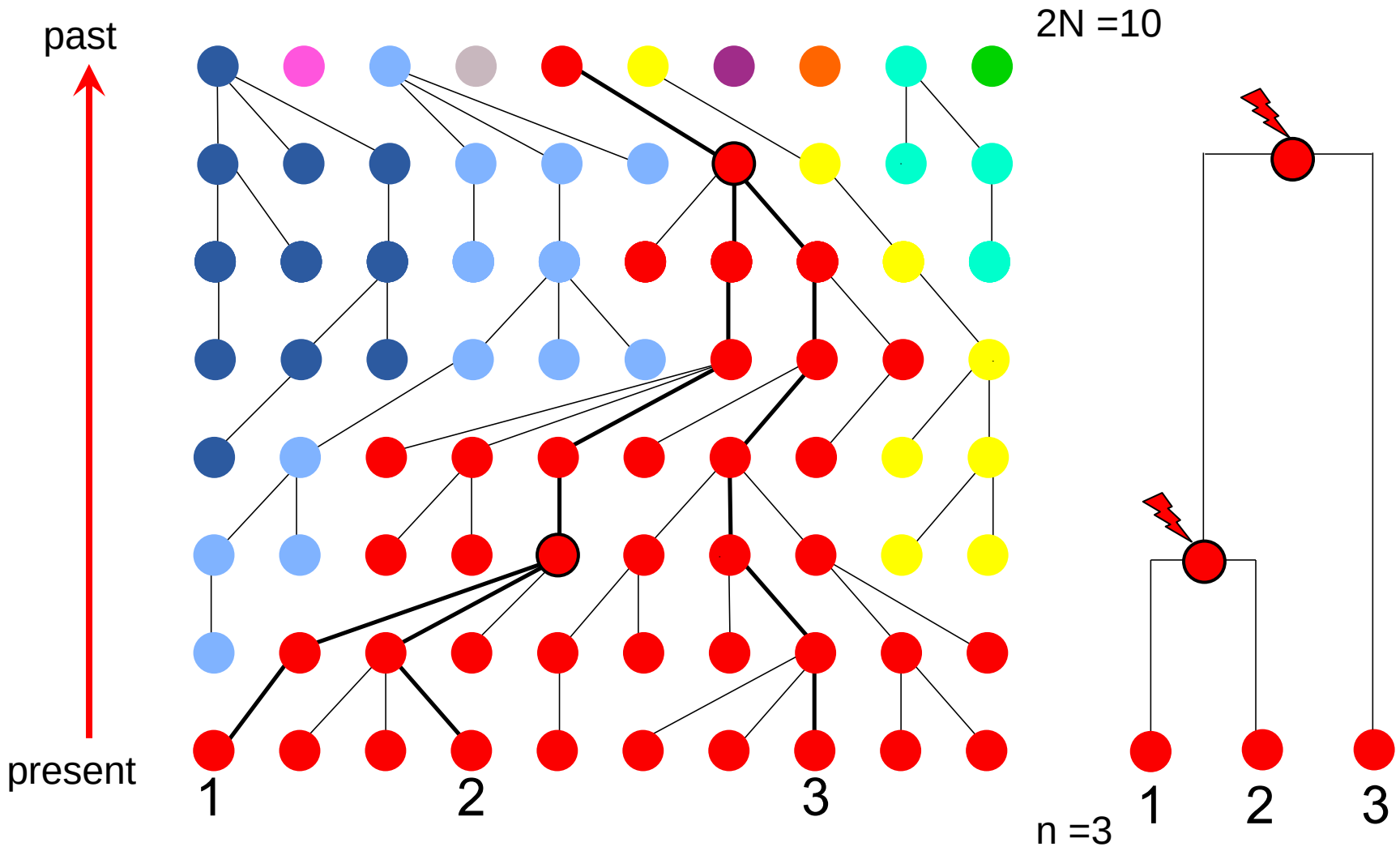
gene evolution



Backward in Time

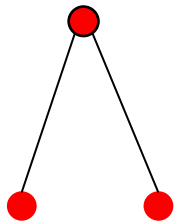


Coalescent Events



Mathematical Theory: Kingman

Wright-Fisher Model

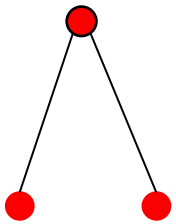


Coalescence probability of 2 genes at a given generation

$$p(2, 2N) = 1/2N$$

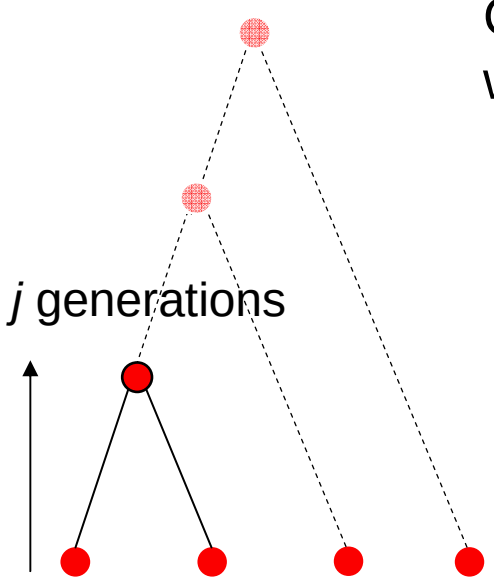
Mathematical Theory: Kingman

Wright-Fisher Model



Coalescence probability of 2 genes at a given generation

$$p(2, 2N) = 1/2N$$

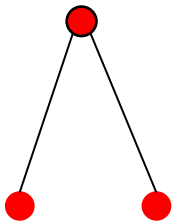


Coalescence probability of 2 genes *j* generations back, within a sample of *n* genes

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

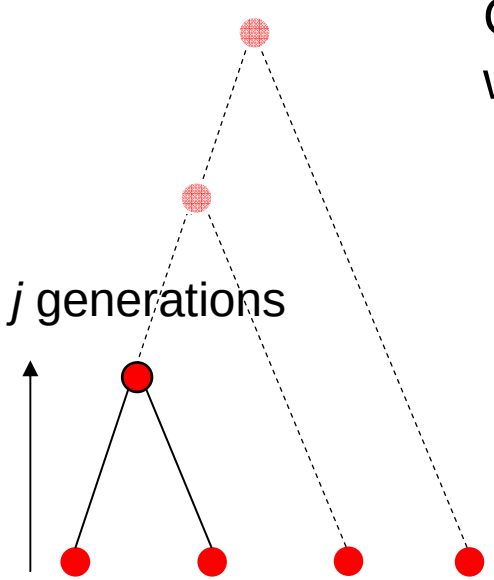
Mathematical Theory: Kingman

Wright-Fisher Model



Coalescence probability of 2 genes at a given generation

$$p(2, 2N) = 1/2N$$



Coalescence probability of 2 genes *j* generations back, within a sample of *n* genes

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

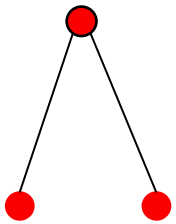
$p(2, 2N)$

probability 2 genes haven't coalesced during the first *j*-1 generations

number of possibles pairs of 2 genes within *n* genes

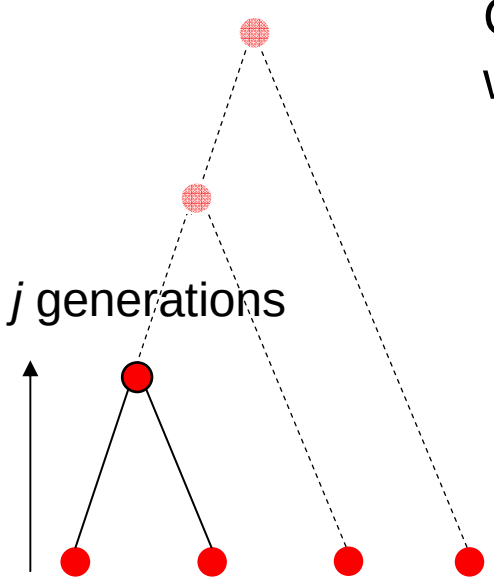
Mathematical Theory: Kingman

Wright-Fisher Model



Coalescence probability of 2 genes at a given generation

$$p(2, 2N) = 1/2N$$



Coalescence probability of 2 genes *j* generations back, within a sample of *n* genes

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

$T_2 \sim$ geometric distribution

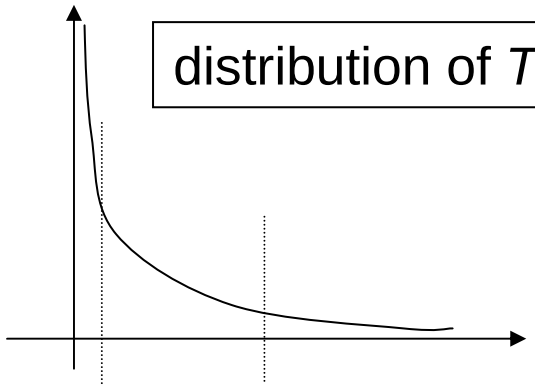
For $n << N$, approximation by an exponential distribution

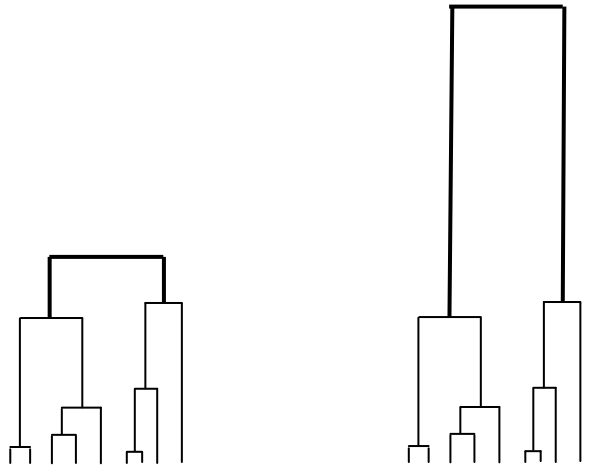
Tn coalescence time between 2 genes: ***Tn* ~ exponential distribution**:

E (*Tn*)= 4*N* / (*n*(*n*-1)) so **E(T_2) = 2*N***

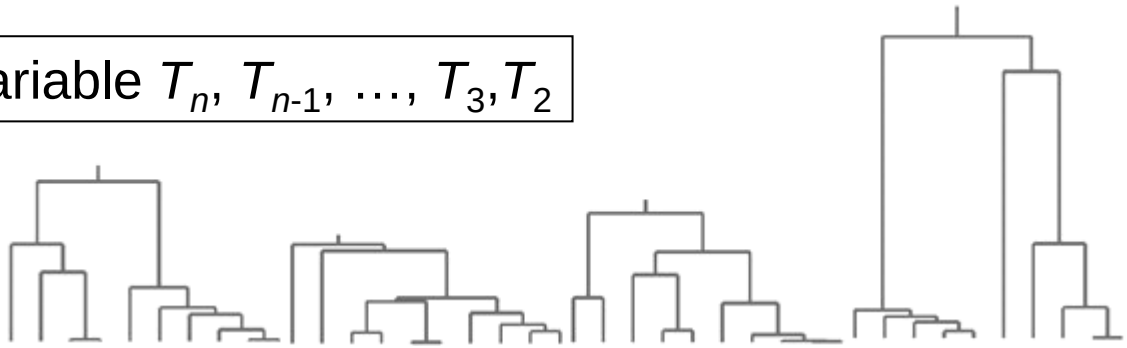
Highly Variable Process

Same T_n, T_{n-1}, …, T_3, but different T_2



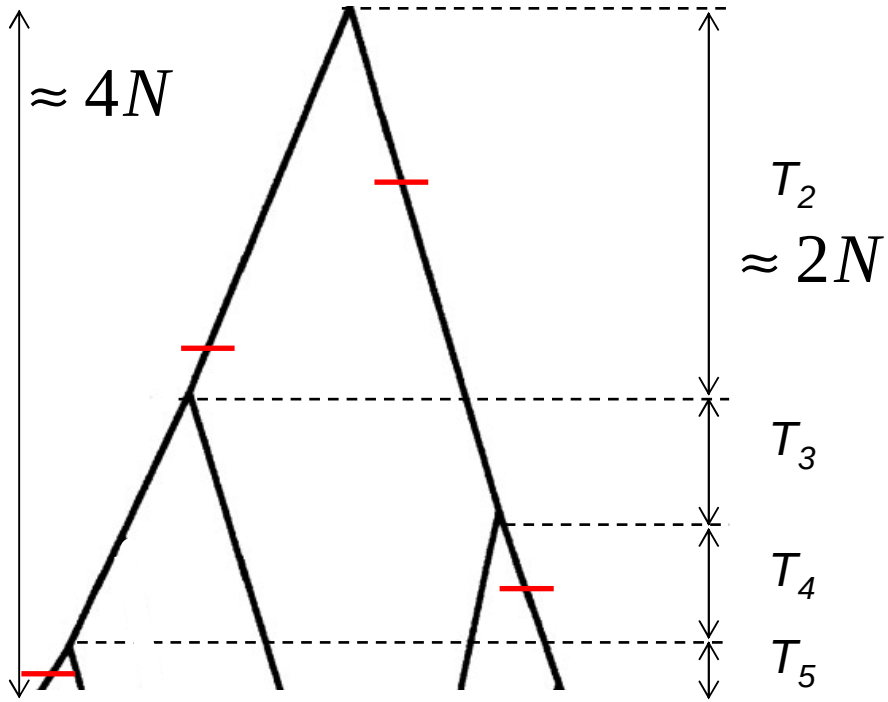


Variable T_n, T_{n-1}, …, T_3,T_2



All trees have same probability

Coalescent with Mutations



—

mutations added on each branch of the tree following a *Poisson* process *mi* :

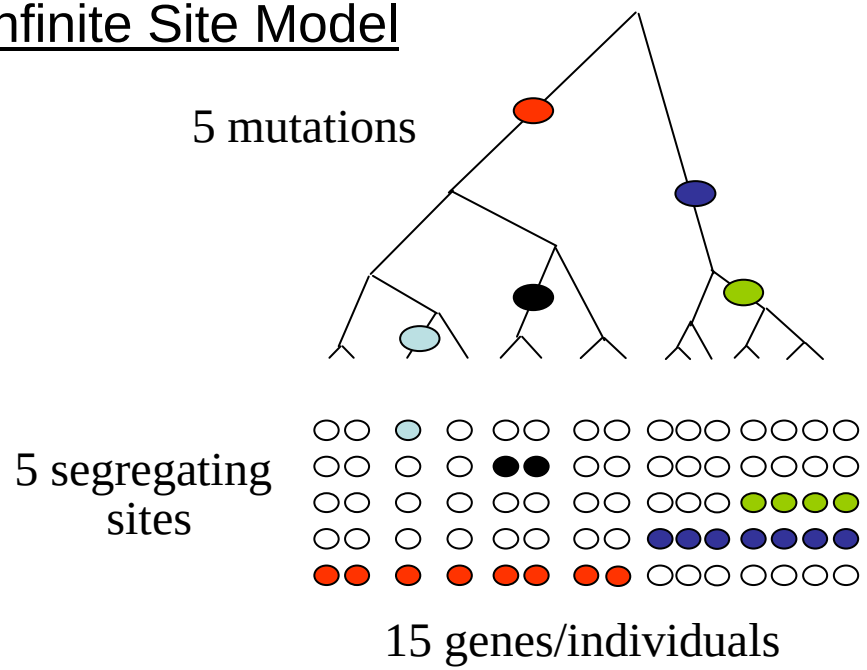
E(*mi*)= ***μt***

μ : mutation rate/gene/generation

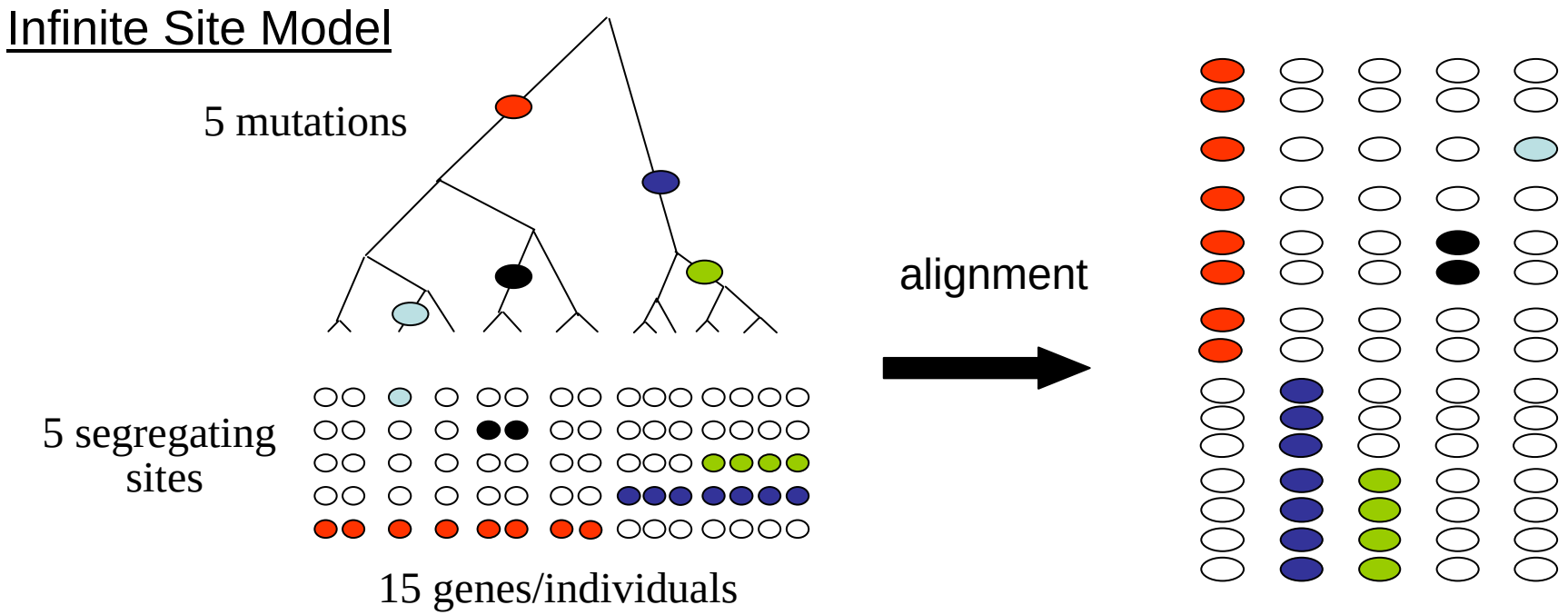
t : length of branch

How to summarize a genealogy?

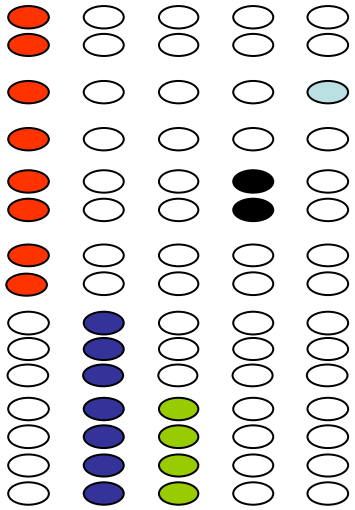
Infinite Site Model



How to summarize a genealogy?



How to summarize a genealogy?



π = 0.37

S = 5

θ_S = 0.31

Estimations de $4N\mu$

- **nucleotide diversity π** (Tajima 1983)
  π = average # of pairwise differences

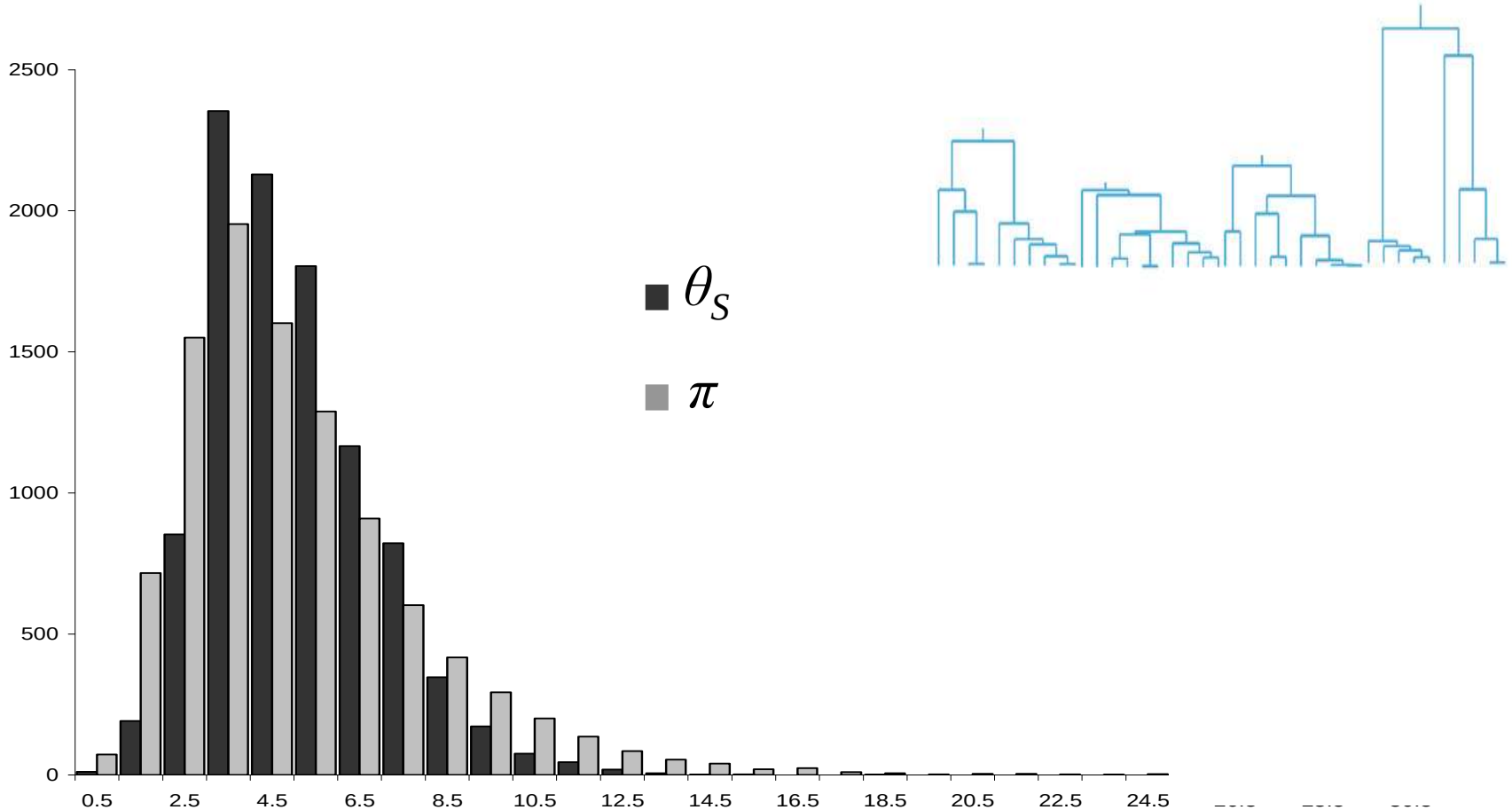
$$E(\pi) = \mu.t = \mu.2.2N = 4N\mu$$

- **theta θ_S** (Watterson 1975)
  S = # segregating sites

$$\theta_S = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

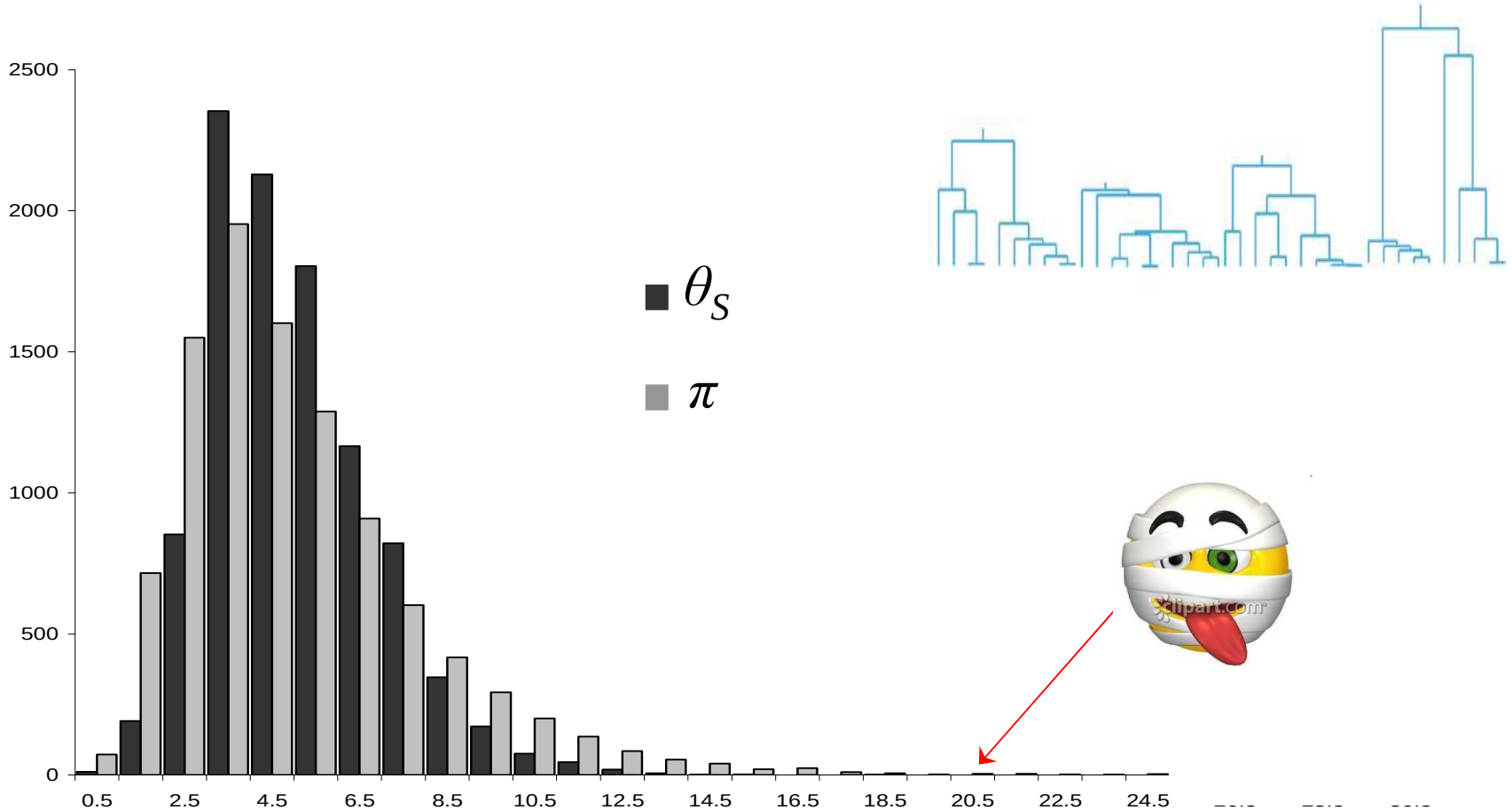
And a set of genealogies?

Expected distributions of summary statistics in a population under a standard model, 10,000 simulations with $\theta = 5$



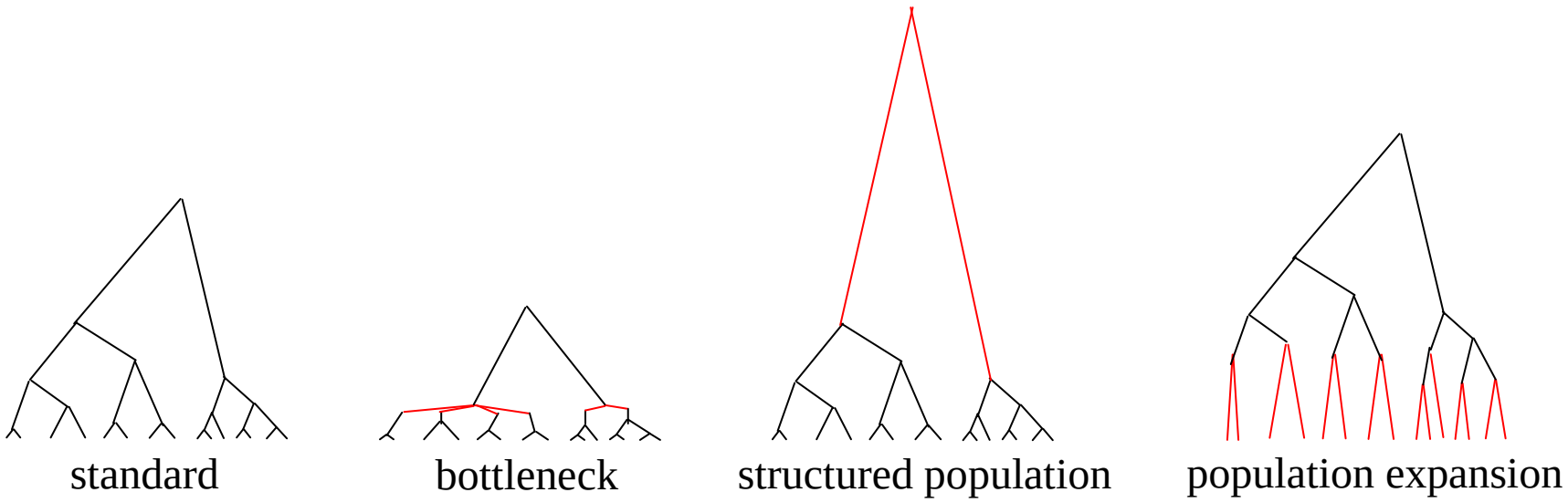
And a set of genealogies?

Expected distributions of summary statistics in a population under a standard model, 10,000 simulations with $\theta = 5$



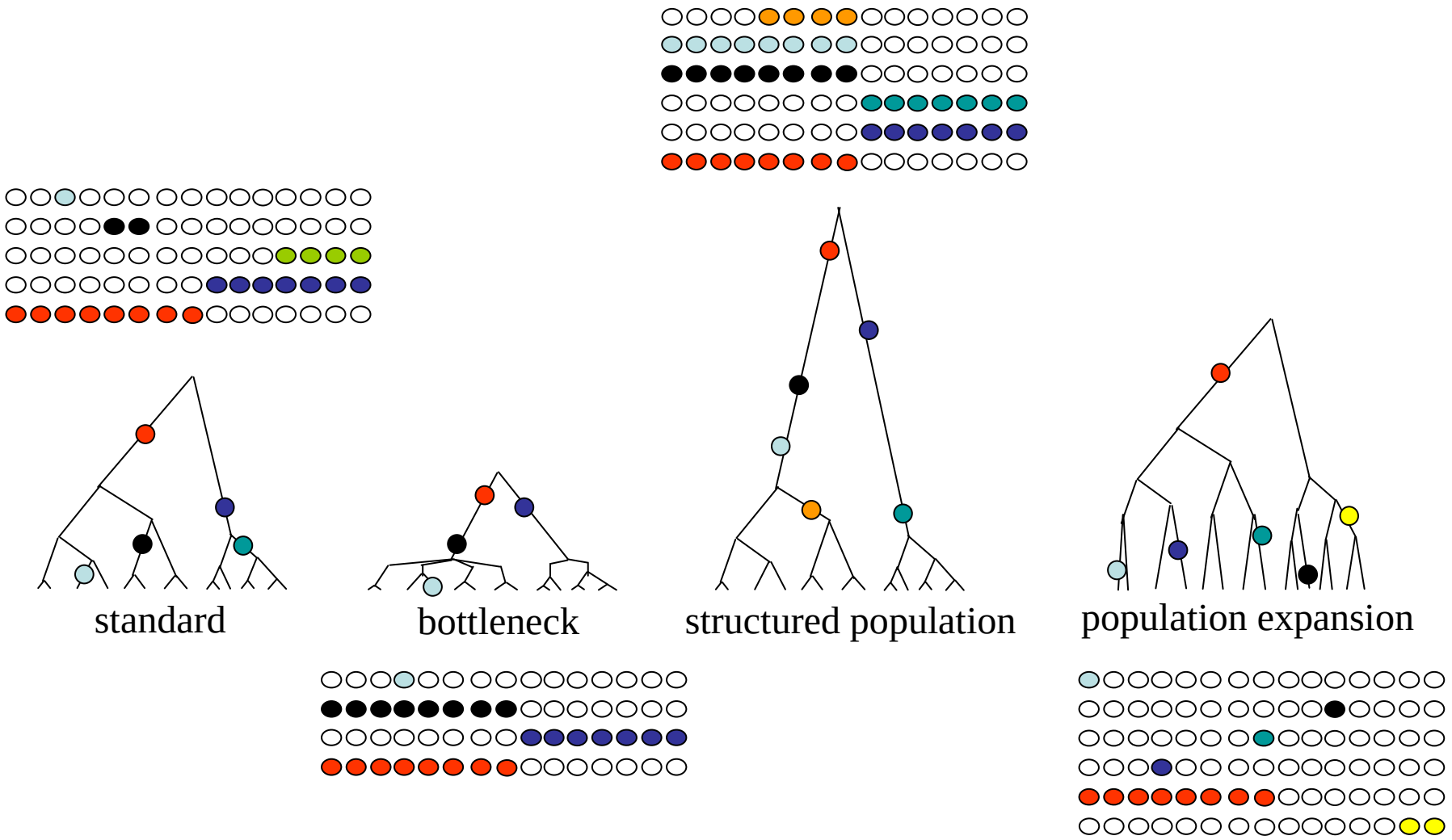
Effect of Demography

Gene genealogy affected by *N* variations



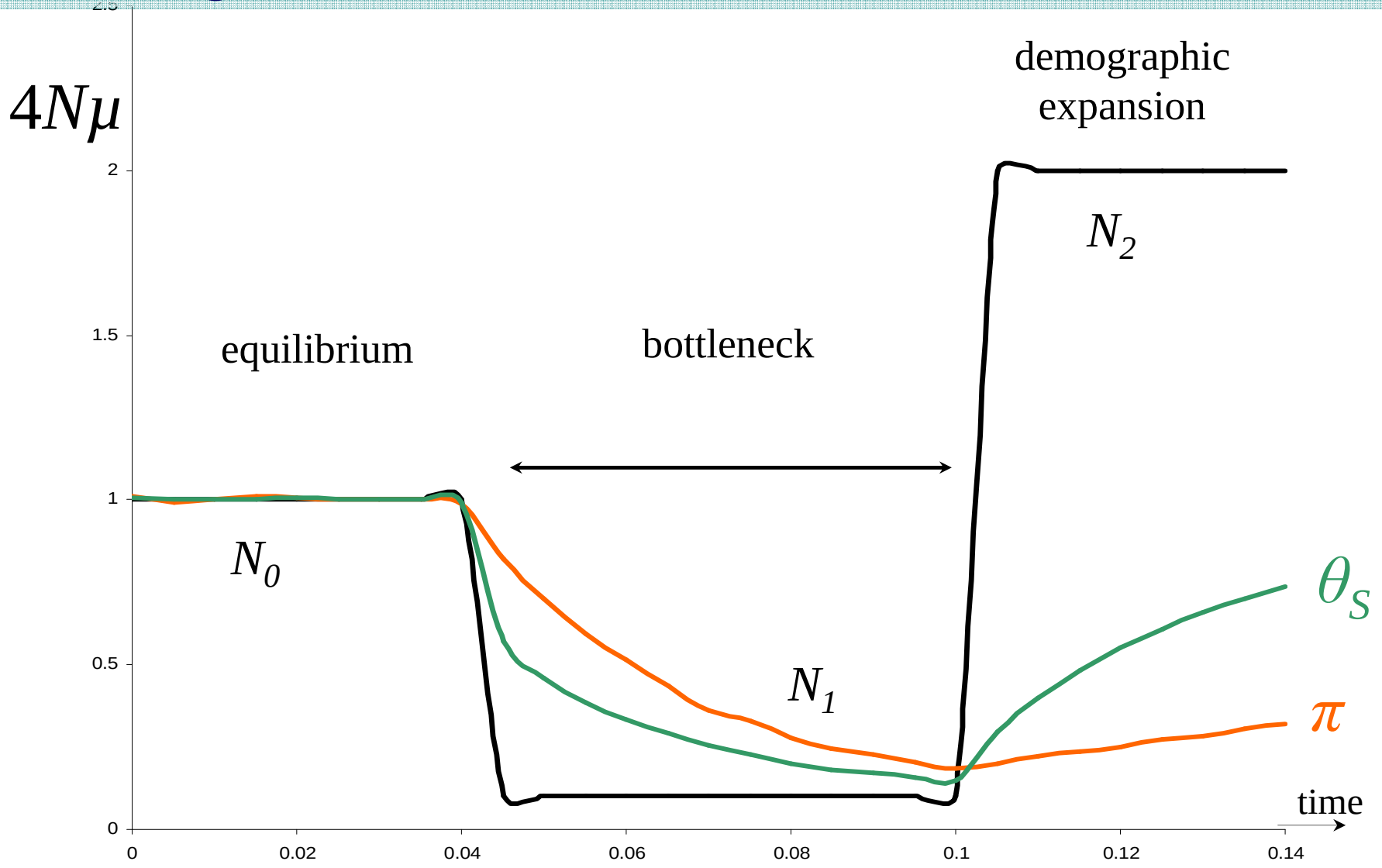
- Formalization of the evolutionary history of populations
- Coalescence theory will describe expected diversity in a sample under different evolution scenarii

Effect of Demography



Affects polymorphism patterns (≠ θ_S and π)

θ_S and π differentially affected



Molecular signatures of variations in population size

Tajima's D

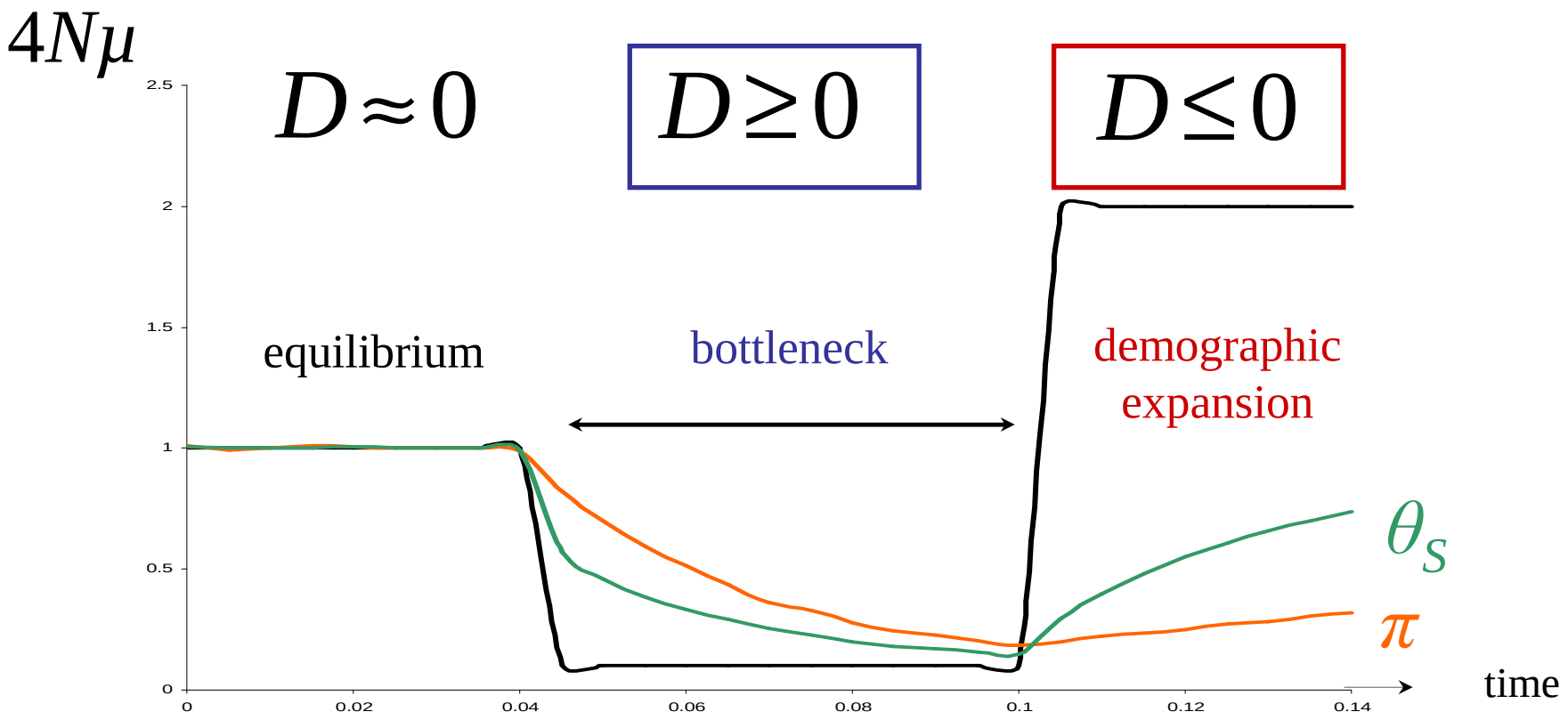
$$D = \frac{\pi - \theta_S}{\sqrt{Var(\pi - \theta_S)}}$$

$D \sim \mathcal{N}(0,1)$ under standard coalescent

Tajima's D

$$D = \frac{\pi - \theta_S}{\sqrt{Var(\pi - \theta_S)}}$$

$D \sim \mathcal{N}(0,1)$ under standard coalescent



Coalescent Simulations

Hudson (2002):

Program *ms* based on coalescence theory to generate simulated gene samples.

Simplest model = Wright-Fisher population

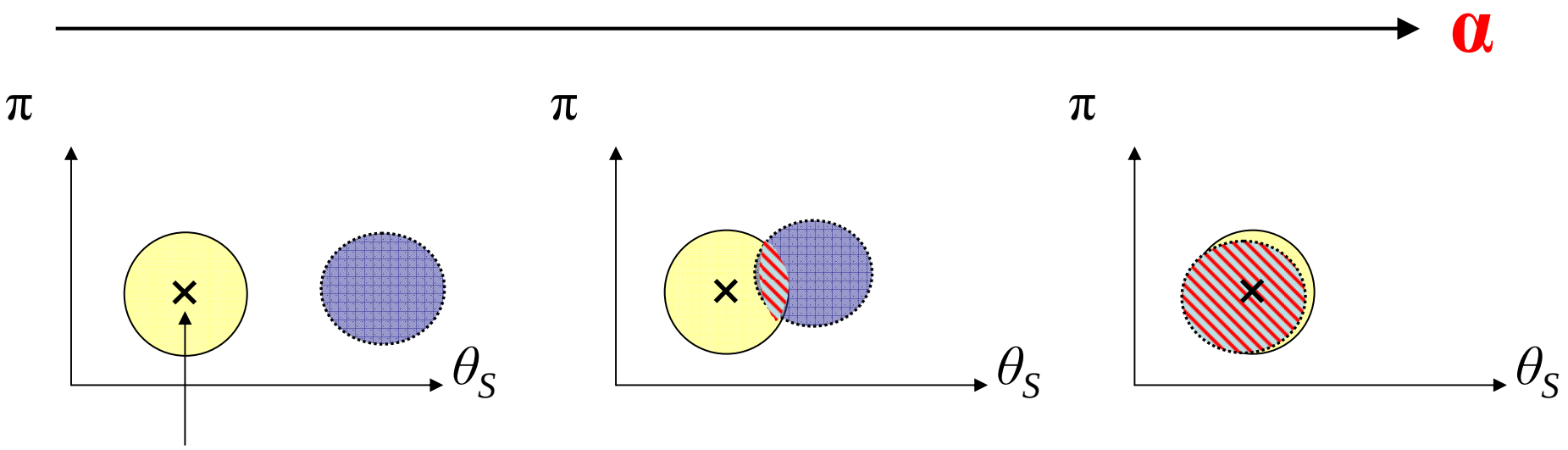
+ complex: - changes in population size
- population structure
- migration
- recombination
- partial selfing
- selection…

Define the best model's parameter

Real samples

α parameter tested in model

Simulated samples



Observed values ± 20%

Estimated Likelihood *i* (α_i)

Example of Maize Domestication

Wright et al. (2005)

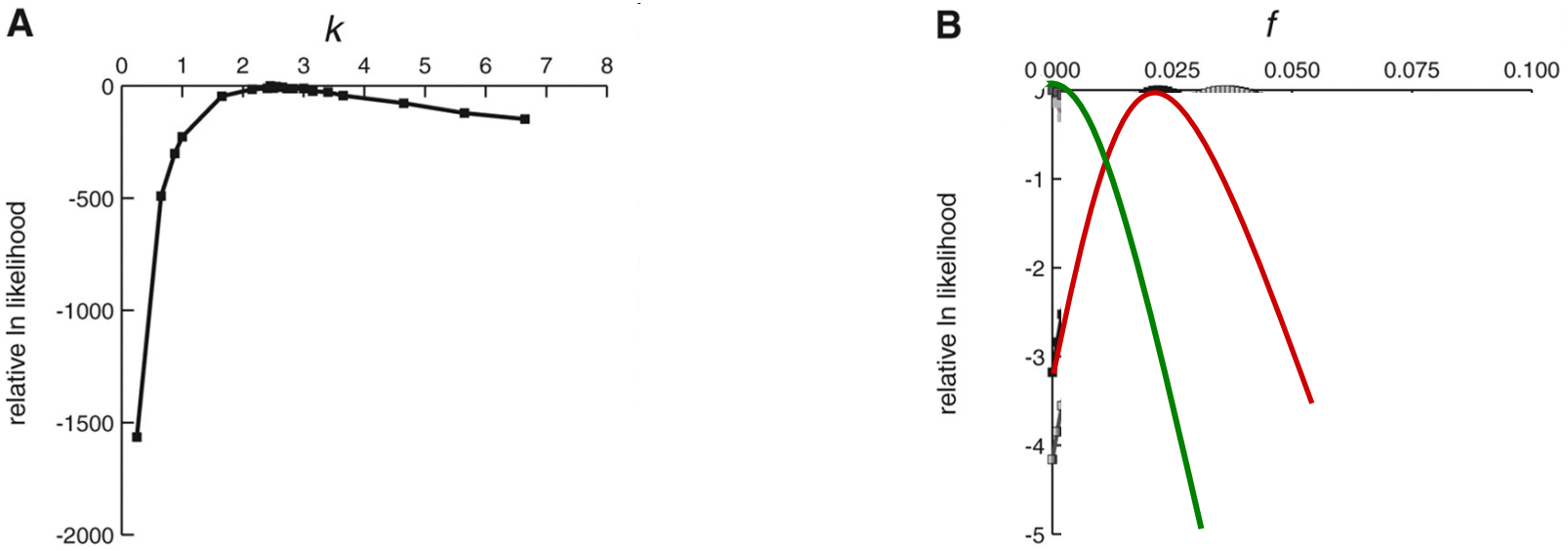
Domestication bottleneck

(k = 2.45)

More severe bottleneck at some genes

candidate selected genes ?

($k1$ = 0.15, $k2$ = 2.45)

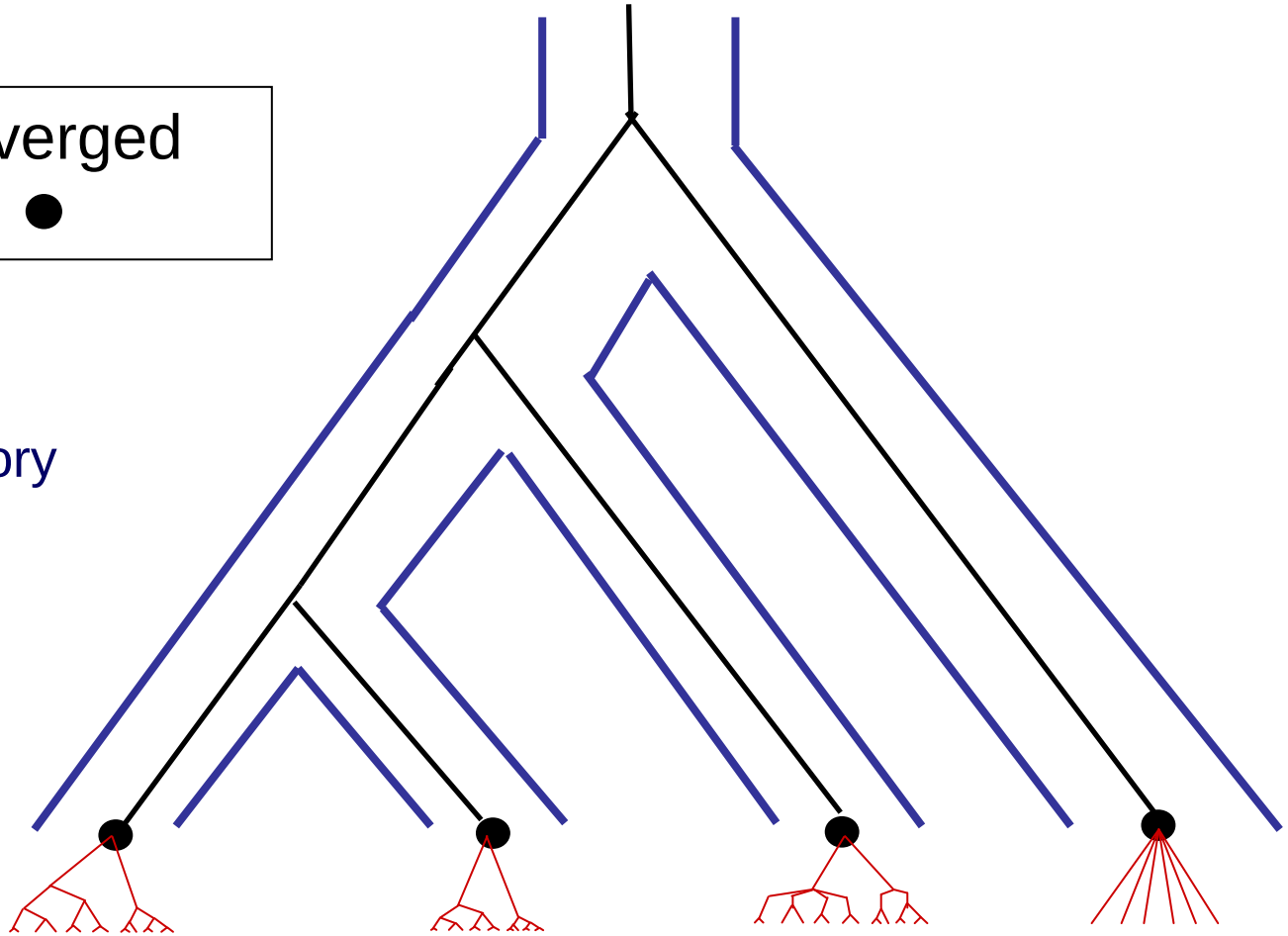


Coalescence and phylogeny

Coalescence: evolutionary history of **genes**

Phylogeny: evolutionary history of **species**

Highly diverged species ●



Speciation history

Population history

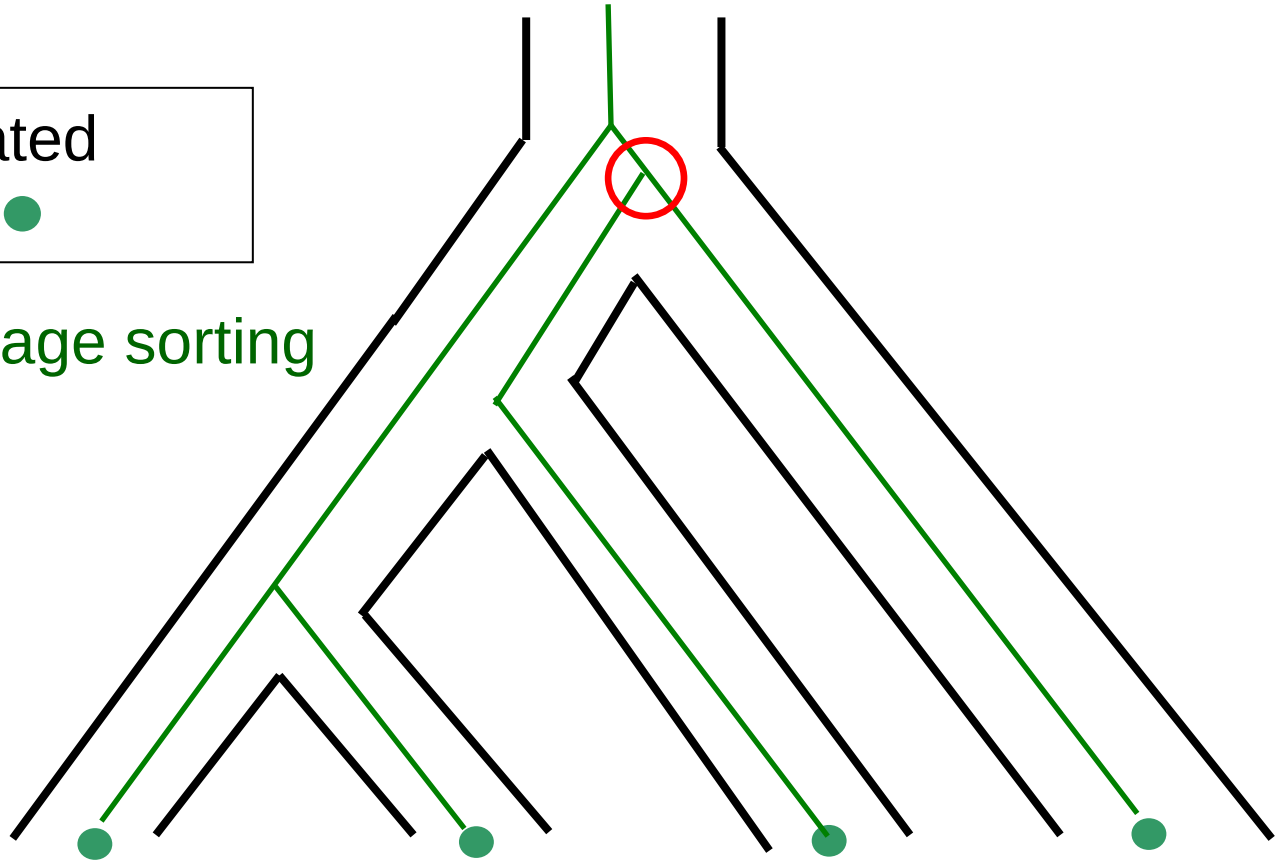
Coalescence and phylogeny

Coalescence: evolutionary history of **genes**

Phylogeny: evolutionary history of **species**

Close related species

Incomplete lineage sorting



Gene history ≠ Species history

More details in:

Hein, J., Schierup, M.H. & Wiuf, C. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. (Oxford University Press, USA: 2005).