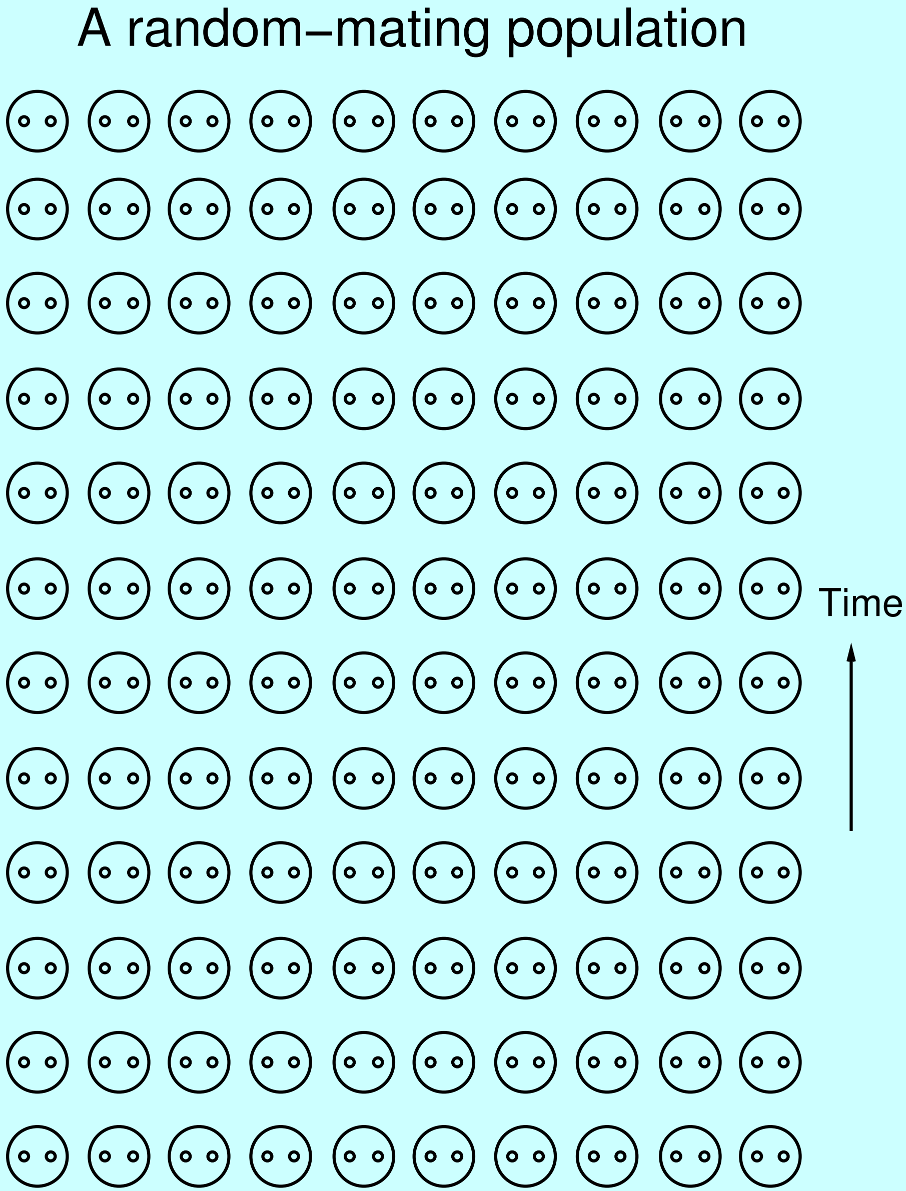
Coalescents

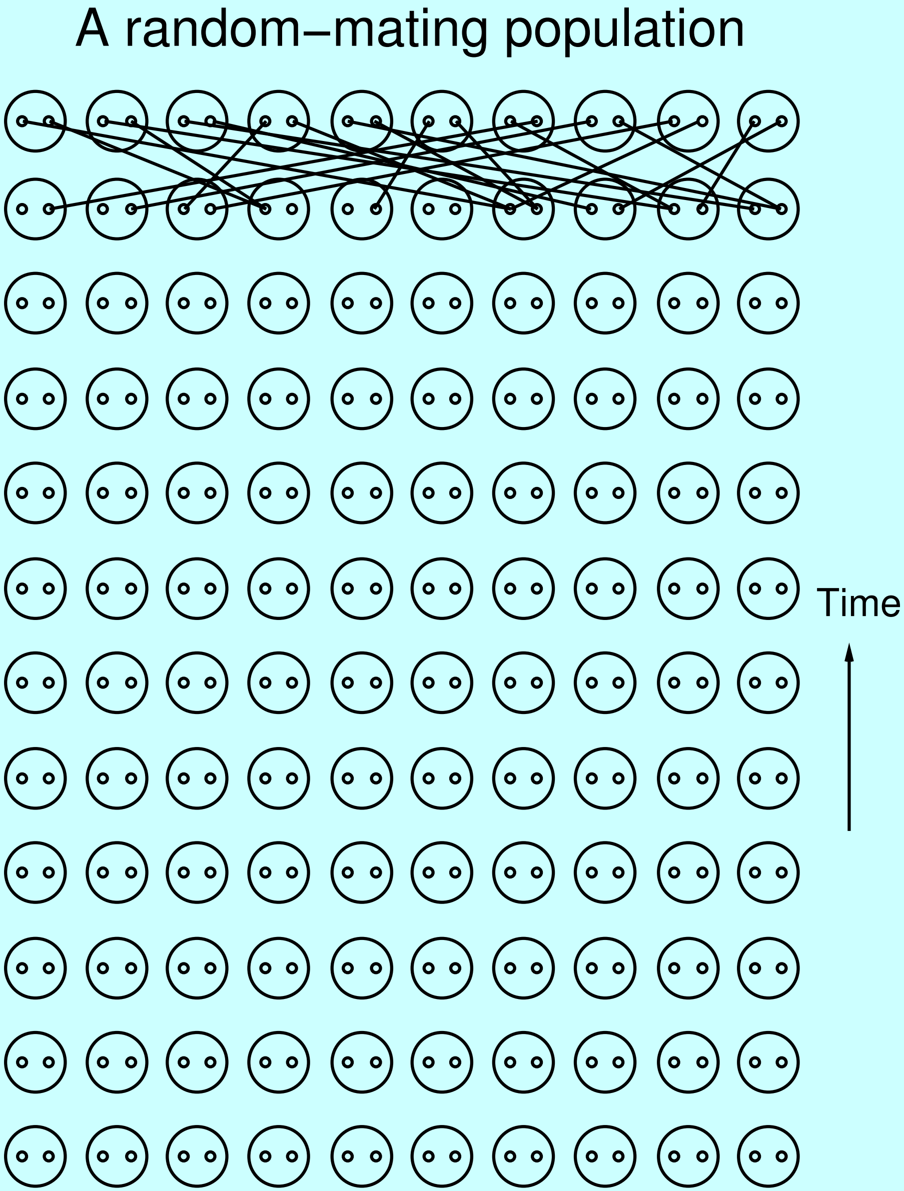
Genome 562

February 18, 2011

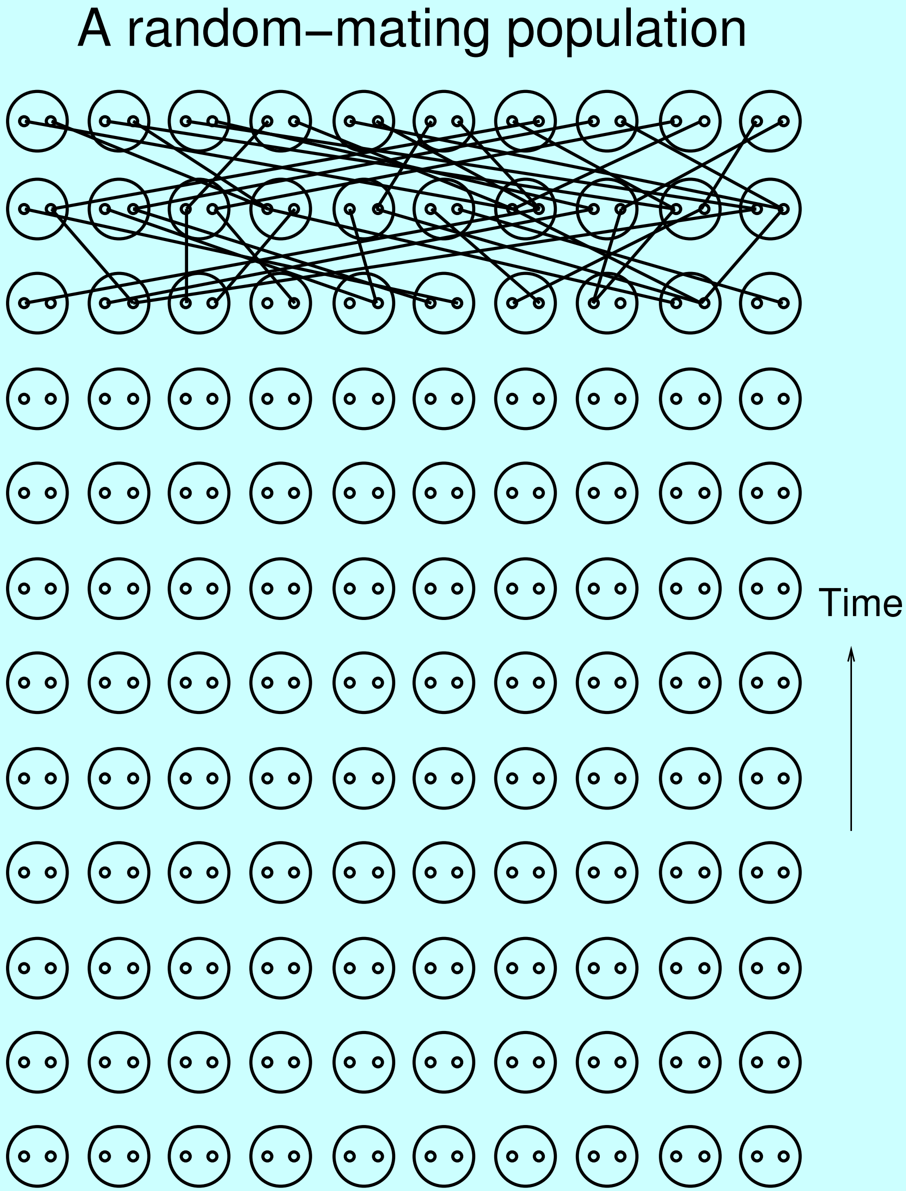
Gene copies in a population of 10 individuals



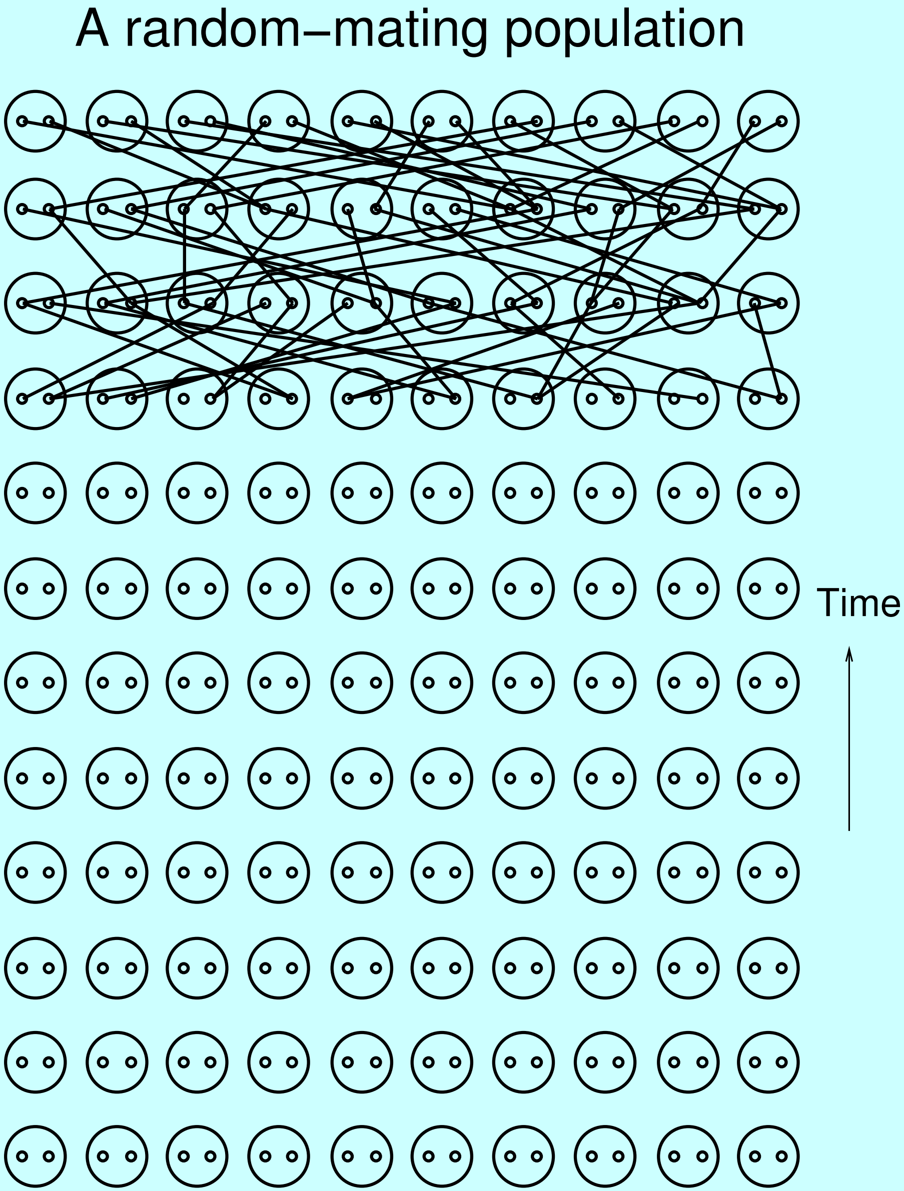
Going back one generation



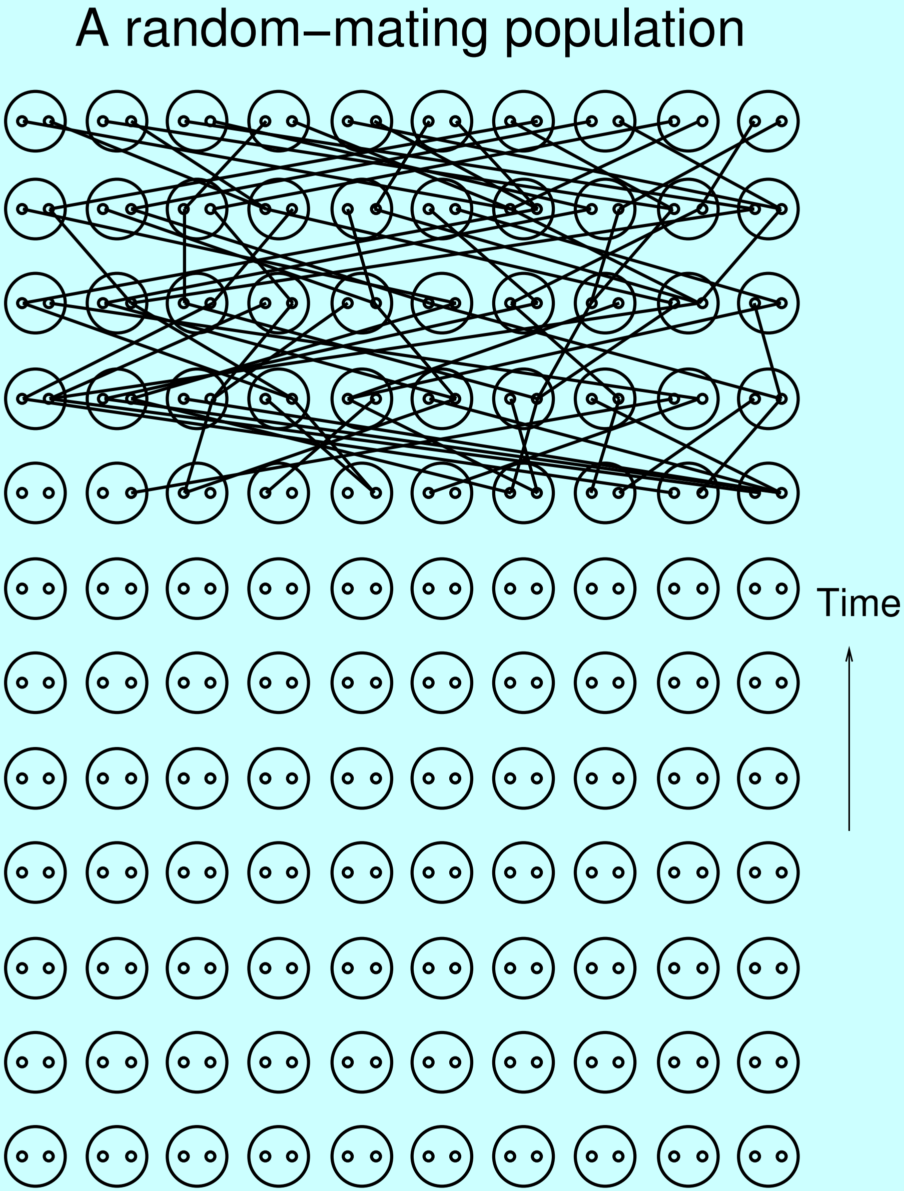
... and one more



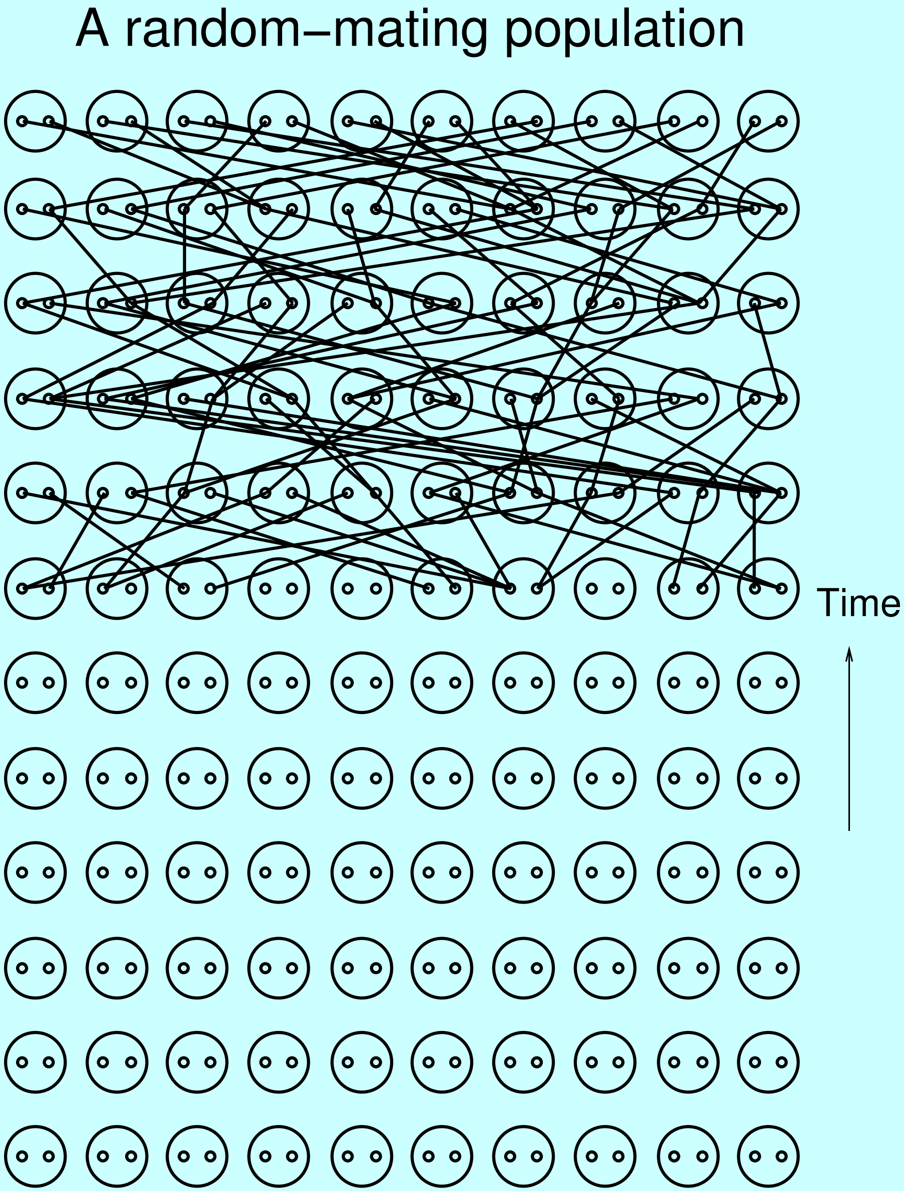
... and one more



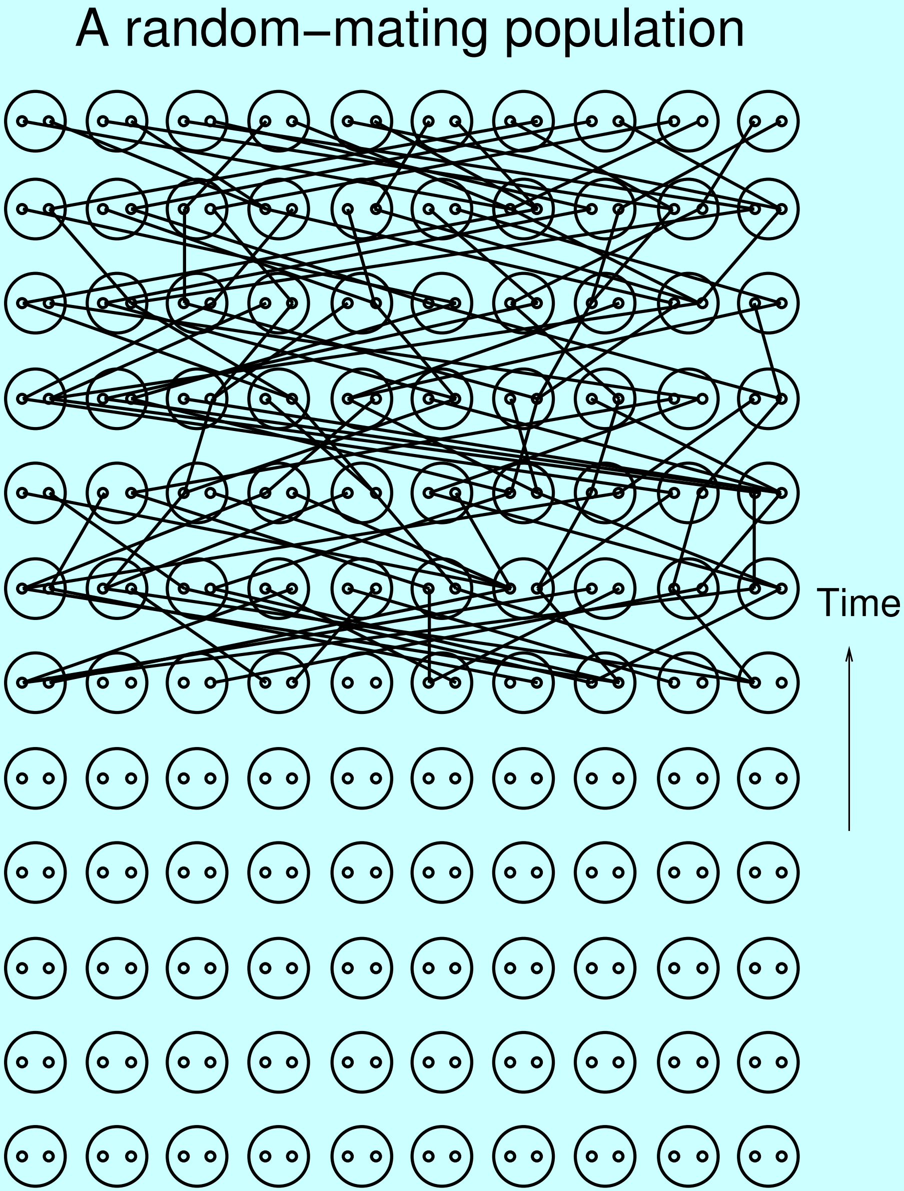
... and one more



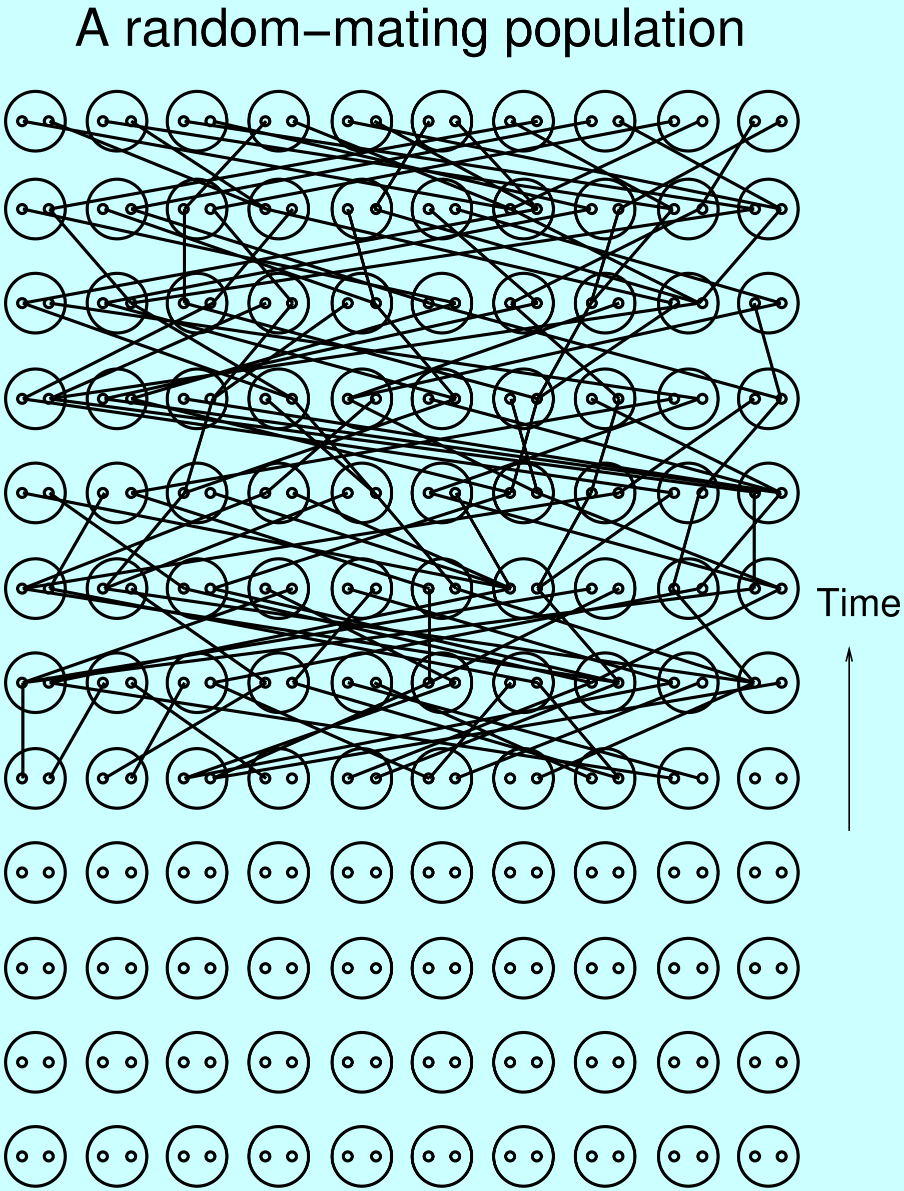
... and one more



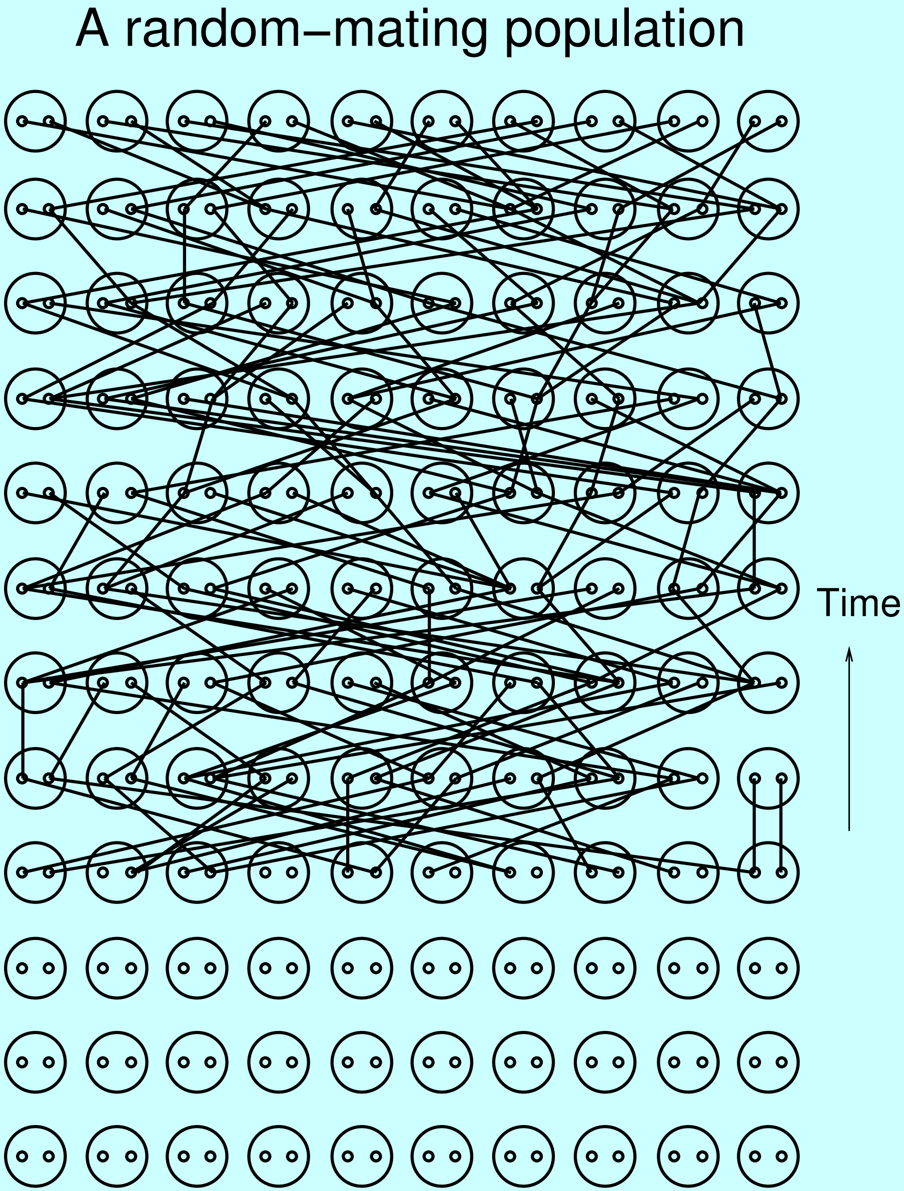
... and one more



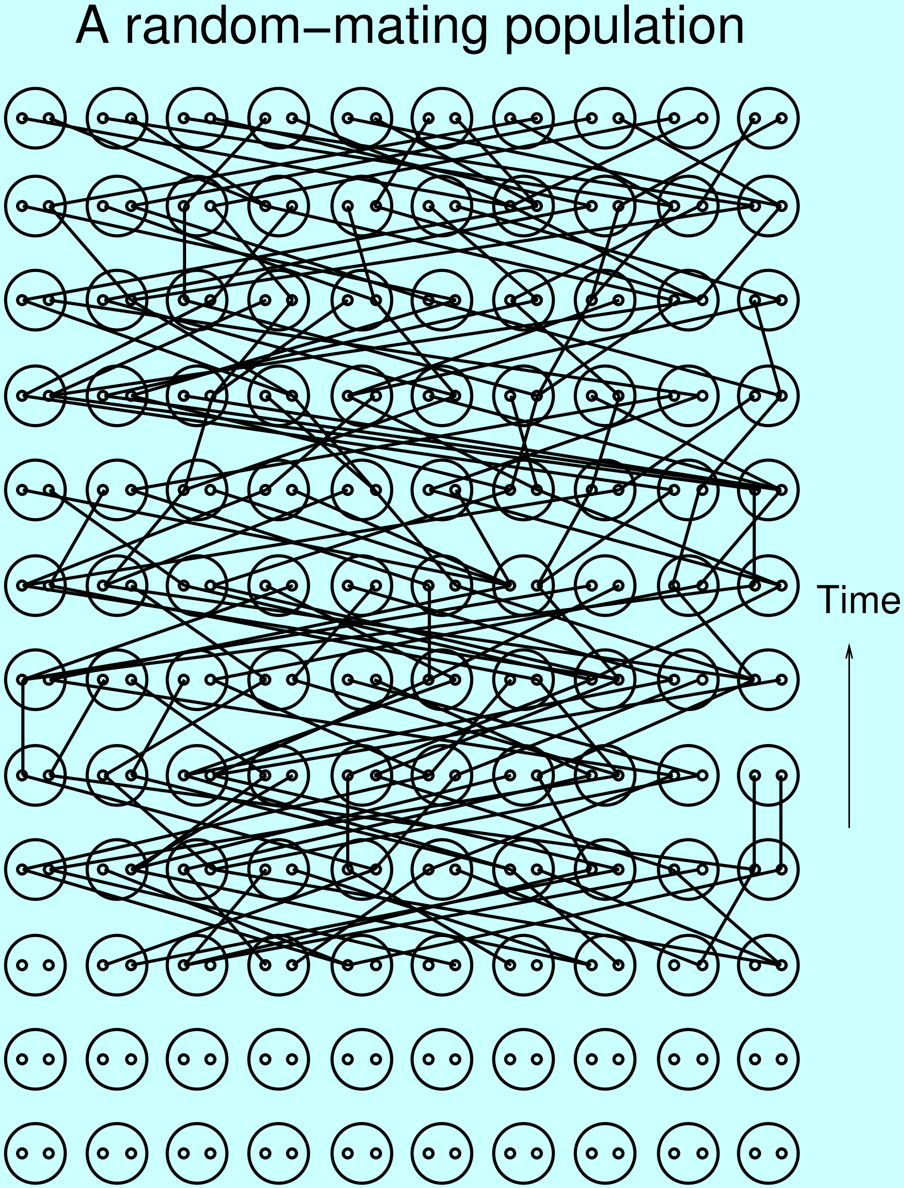
... and one more



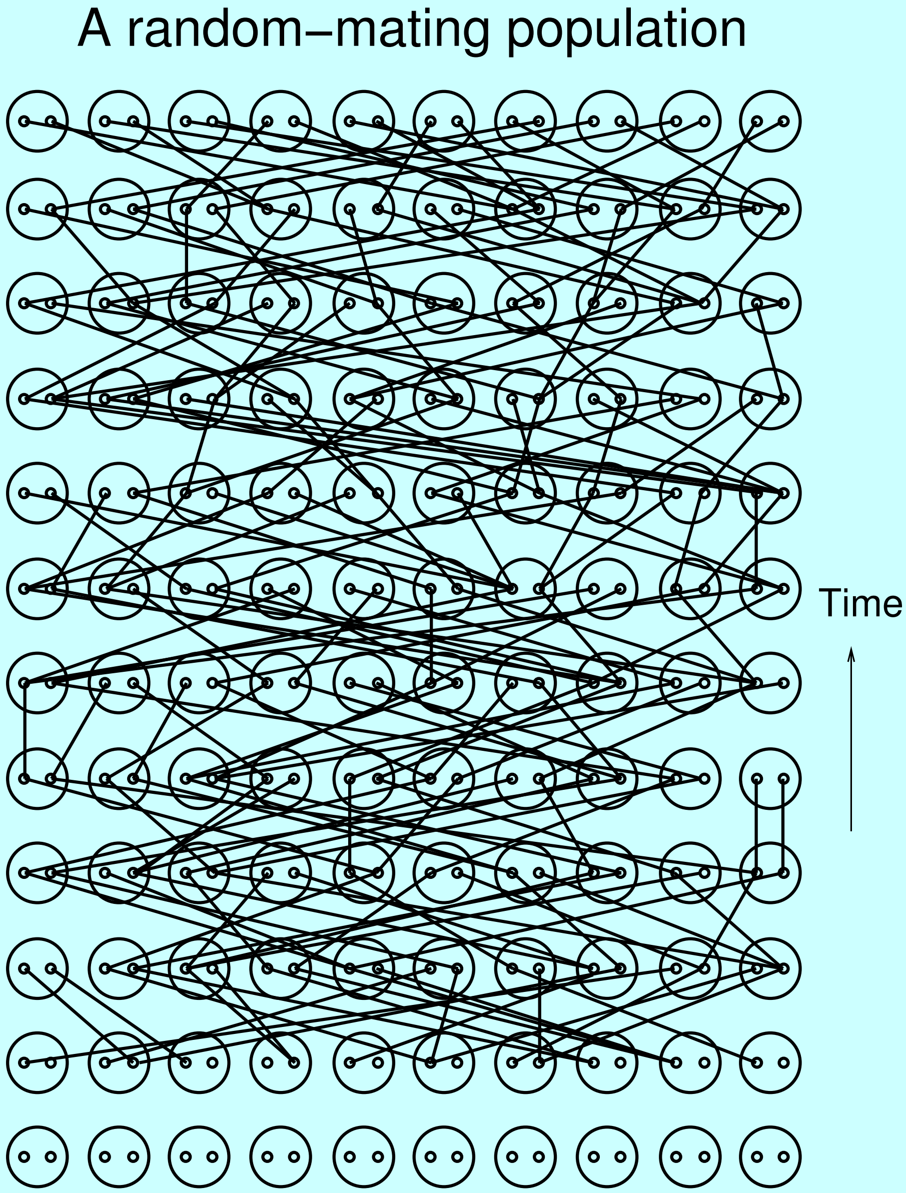
... and one more



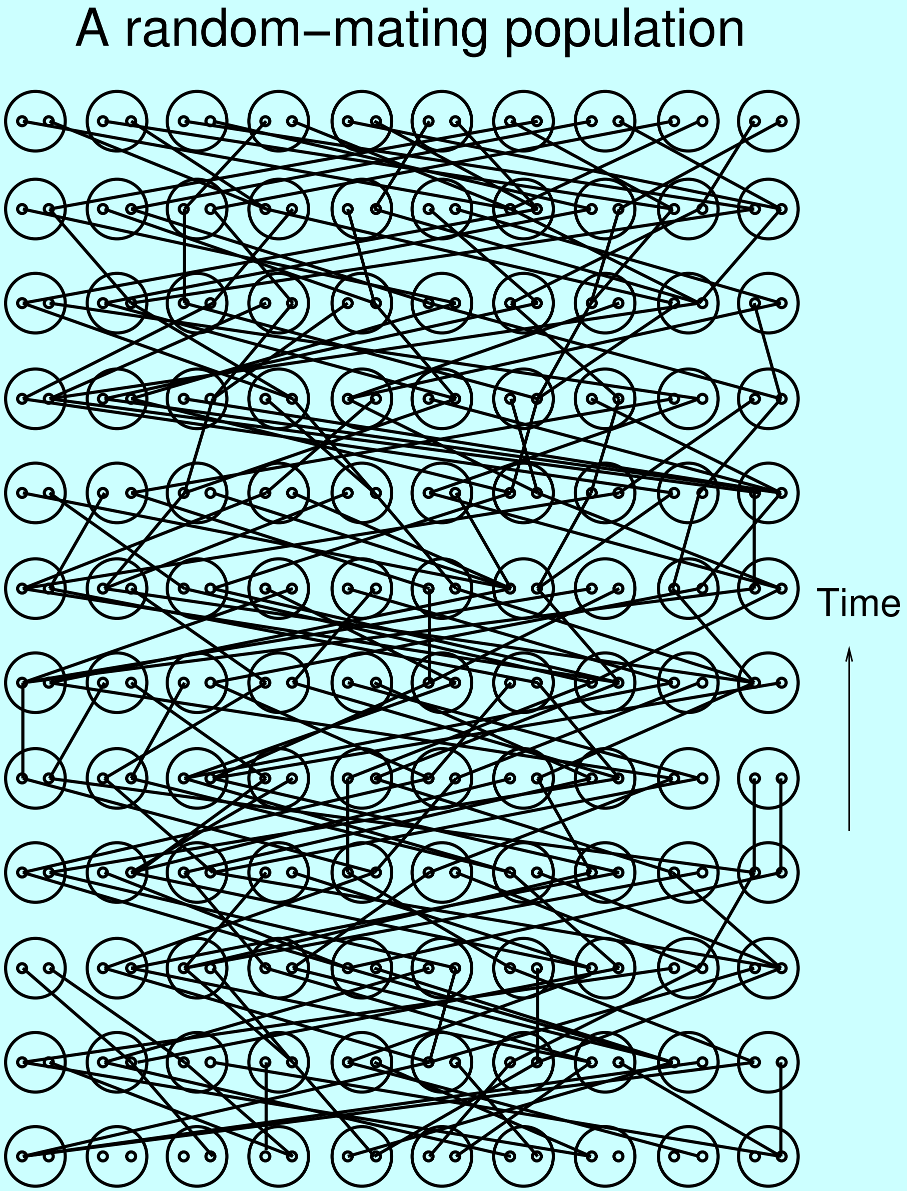
... and one more



... and one more

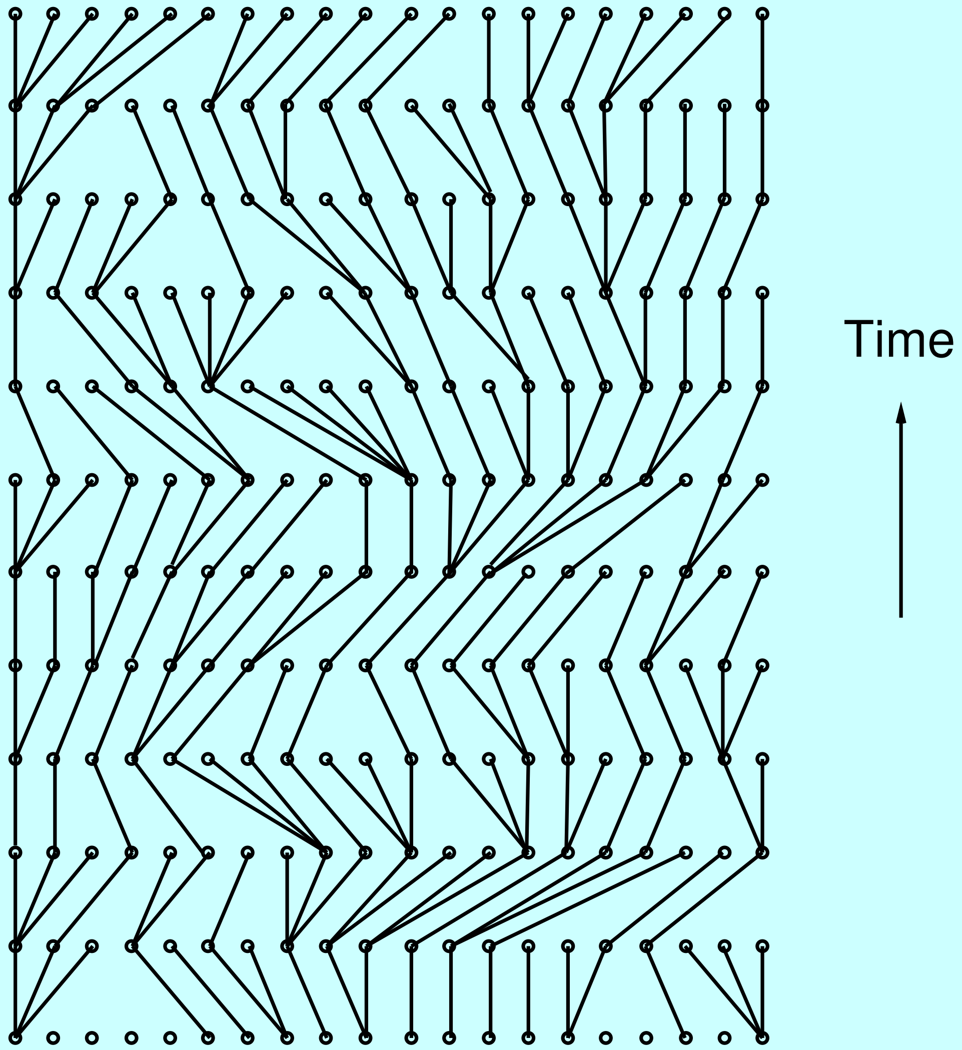


... and one more



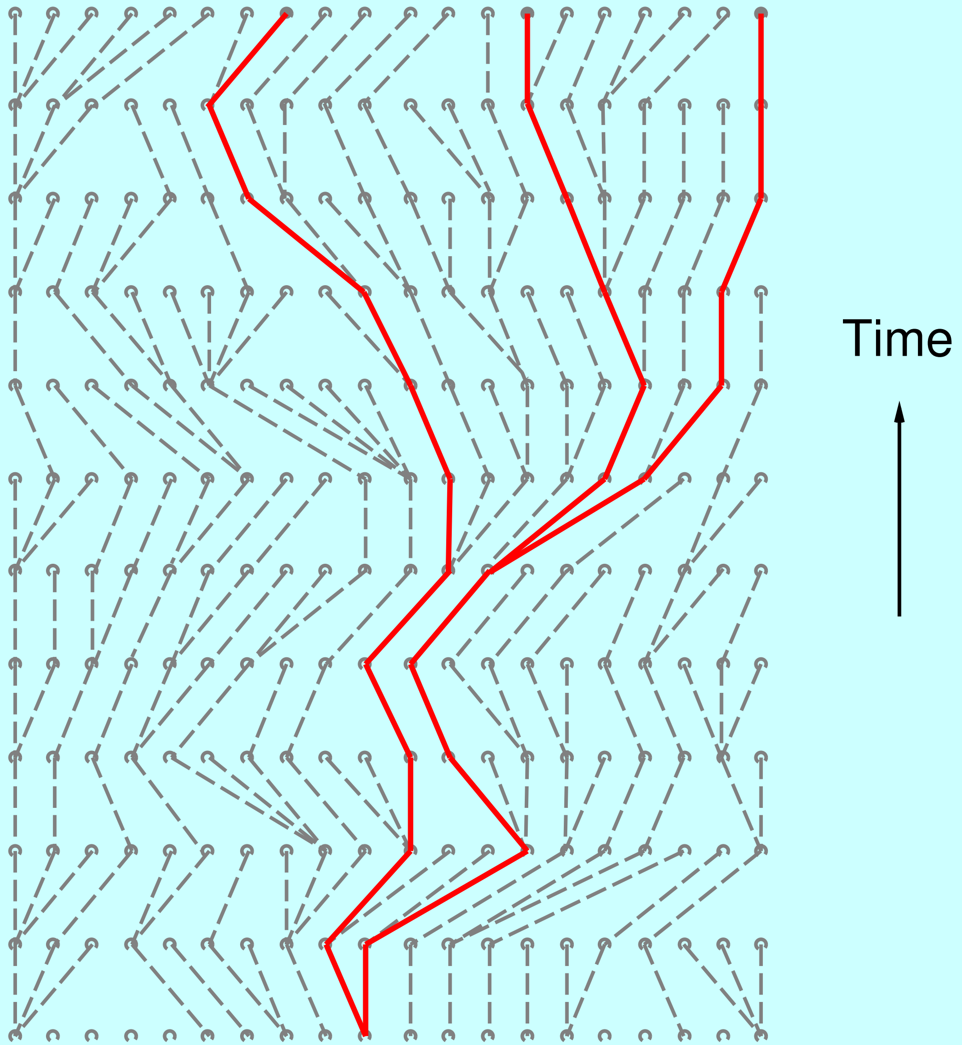
The genealogy of gene copies is a tree

Genealogy of gene copies, after reordering the copies

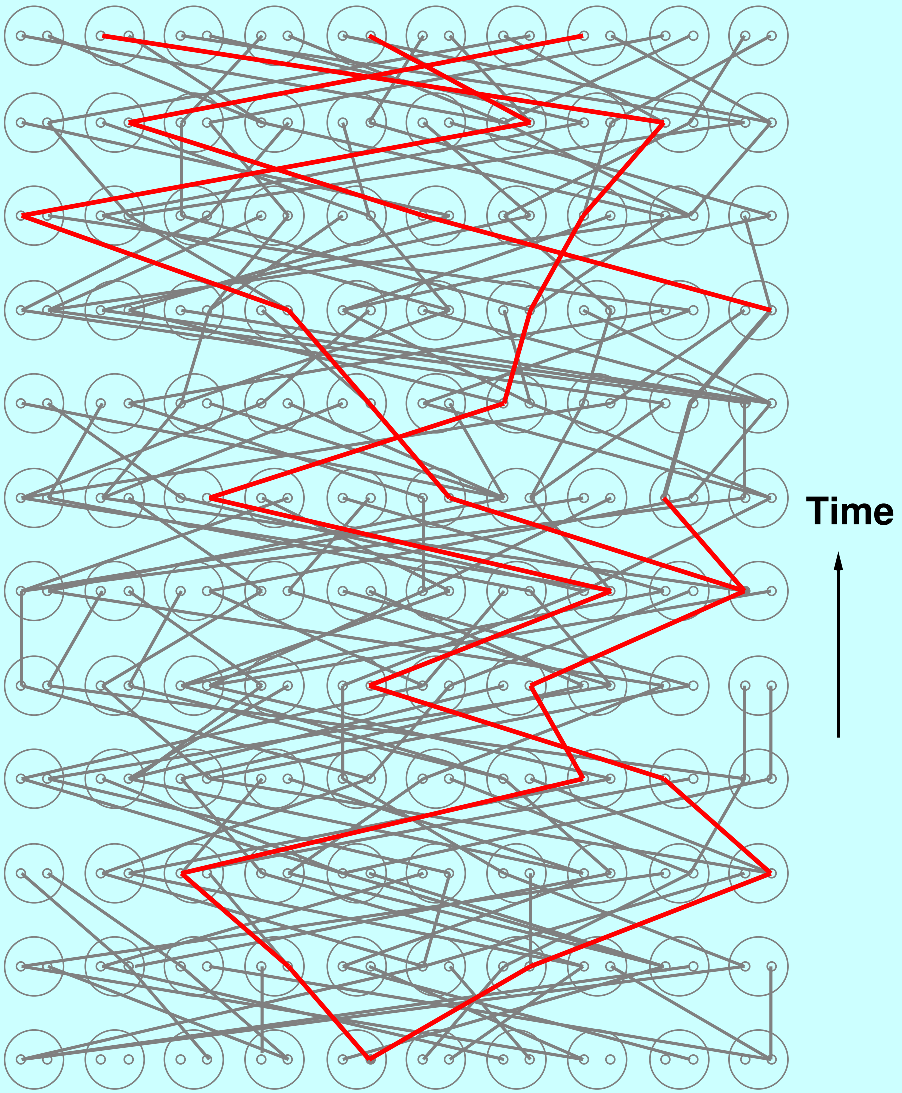


Ancestry of a sample of 3 copies

Genealogy of a small sample of genes from the population



Here is that tree of 3 copies in the pedigree



Sir John Kingman

J. F. C. Kingman in about 1983

Currently Emeritus Professor of Mathematics at Cambridge University, U.K., and former head of the Isaac Newton Institute of Mathematical Sciences.

Kingman's coalescent

Random collision of lineages as go back in time (sans recombination)

Collision is faster the smaller the effective population size

Average time for

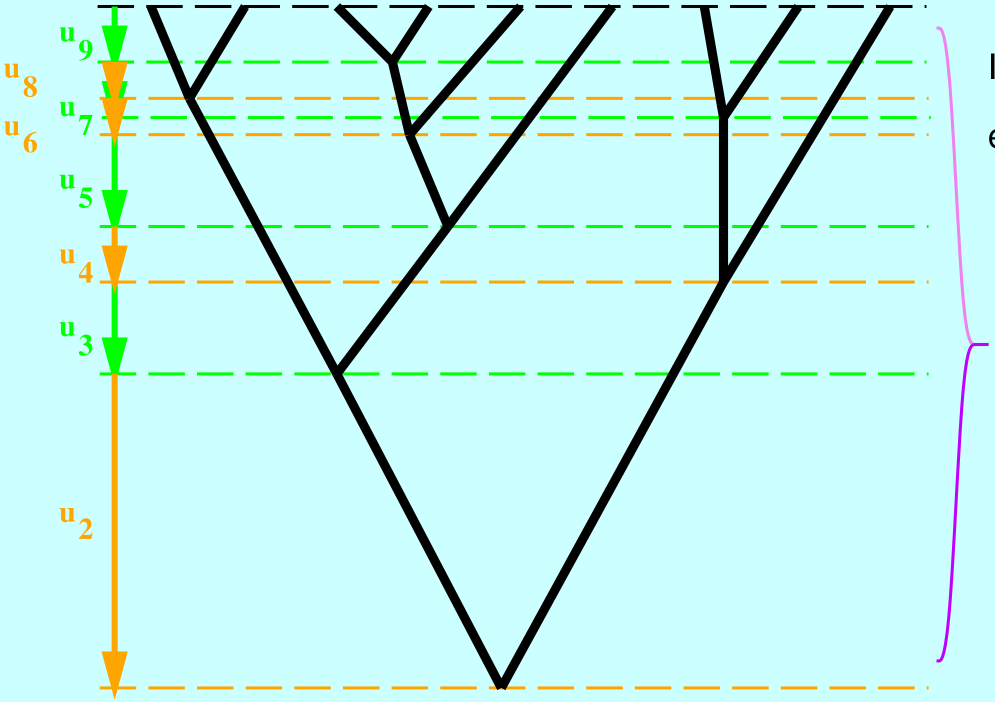
k copies to coalesce to

$$k-1 \quad = \quad \frac{4N}{k(k-1)}$$

Average time for

two copies to coalesce

$= \; 2N$ generations



In a diploid population of

effective population size N,

Average time for n

copies to coalesce

$$= \; 4N\left(1 - \frac{1}{n}\right) \text{ generations}$$

The Wright-Fisher model

This is the canonical model of genetic drift in populations. It was invented in 1930 and 1932 by Sewall Wright and R. A. Fisher.
In this model the next generation is produced by doing this:

- Choose two individuals *with replacement* (including the possibility that they are the same individual) to be parents,
- Each produces one gamete, these become a diploid individual,
- Repeat these steps until N diploid individuals have been produced.

The effect of this is to have each locus in an individual in the next generation consist of two genes sampled from the parents' generation at random, with replacement.

The coalescent – a derivation

The probability that k lineages becomes $\mathsf{k}-1$ one generation earlier turns out to be (as each lineage “chooses” its ancestor independently):

$$\mathsf{k}(\mathsf{k}-1)/2 \times \mathrm{Prob}\ (\mathrm{First\ two\ have\ same\ parent,\ rest\ are\ different})$$

(since there are $\binom{\mathsf{k}}{2} = \mathsf{k}(\mathsf{k}-1)/2$ different pairs of copies)
We add up terms, all the same, for the $\mathsf{k}(\mathsf{k}-1)/2$ pairs that could coalesce; the sum is:

$$\mathsf{k}(\mathsf{k}-1)/2 \ \times\ 1\ \times\ \frac{1}{2\mathsf{N}}\ \times\ \left(1-\frac{1}{2\mathsf{N}}\right)$$

$$\times\ \left(1-\frac{2}{2\mathsf{N}}\right)\ \times\cdots\times\ \left(1-\frac{\mathsf{k}-2}{2\mathsf{N}}\right)$$

so that the total probability that a pair coalesces is

$$=\ \mathsf{k}(\mathsf{k}-1)/4\mathsf{N} + \mathsf{O}(1/\mathsf{N}^2)$$

Probabilities of two or more lineages coalescing

Note that the total probability that some combination of lineages coalesces is

$$1 - \mathrm{Prob}\ (\mathrm{Probability\ all\ genes\ have\ separate\ ancestors})$$

$$= \ 1 - \left[1 \times \left(1 - \frac{1}{2N}\right) \left(1 - \frac{2}{2N}\right) \ldots \left(1 - \frac{k-1}{2N}\right) \right]$$

$$= \ 1 \ - \ \left[1 \ - \ \frac{1 + 2 + 3 + \cdots + (k-1)}{2N} \ + \ O(1/N^2) \right]$$

and since

$$1 \ + \ 2 \ + \ 3 \ + \ \ldots \ + \ (n-1) \ = \ n(n-1)/2$$

the quantity

$$= \ 1 \ - \ \left[\, 1 - k(k-1)/4N + O(1/N^2) \right] \ \simeq \ k(k-1)/4N \ + \ O(1/N^2)$$

Can calculate how many coalescences are of pairs

This shows, since the terms of order $1/\mathrm{N}$ are the same, that the events involving 3 or more lineages simultaneously coalescing are in the terms of order $1/\mathrm{N}^2$ and thus become unimportant if N is large.

Here are the probabilities of 0, 1, or more coalescences with 10 lineages in populations of different sizes:

N	0	1	> 1
100	0.79560747	0.18744678	0.01694575
1000	0.97771632	0.02209806	0.00018562
10000	0.99775217	0.00224595	0.00000187

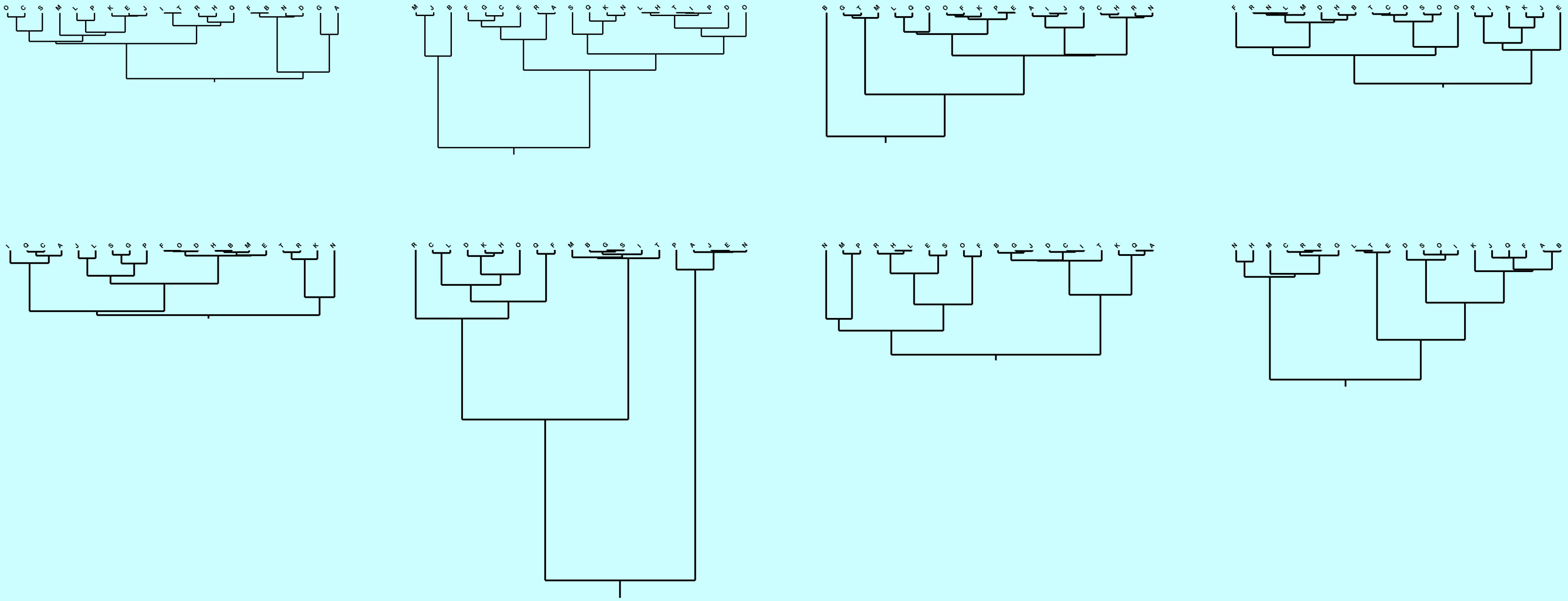
Note that increasing the population size by a factor of 10 reduces the coalescent rate for pairs by about 10-fold, but reduces the rate for triples (or more) by about 100-fold.

The coalescent

To simulate a random genealogy, do the following:

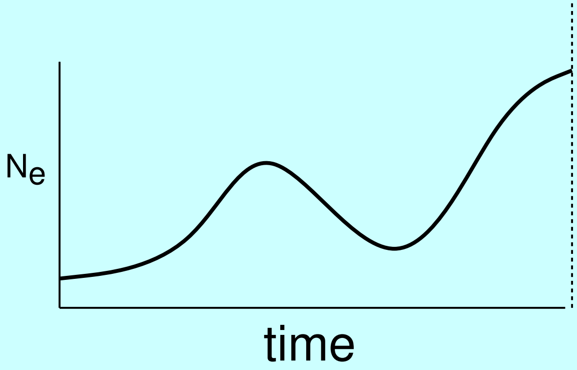
1. Start with k lineages
2. Draw an exponential time interval with mean $4N/(k(k-1))$ generations.
3. Combine two randomly chosen lineages.
4. Decrease k by 1.
5. If $k = 1$, then stop
6. Otherwise go back to step 2.

Random coalescent trees with 16 lineages



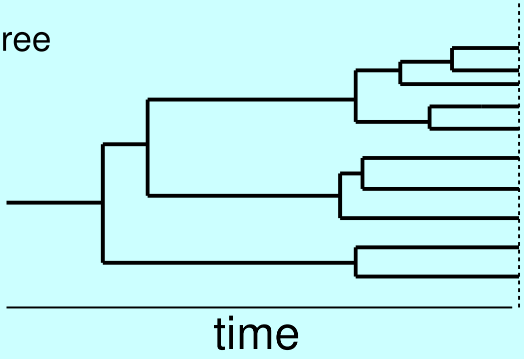
Coalescence is faster in small populations

Change of population size and coalescents

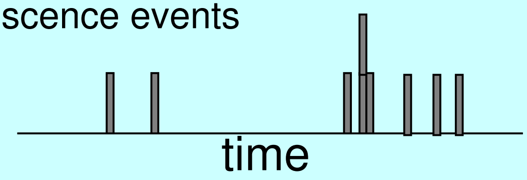


the changes in population size will produce waves of coalescence

the tree

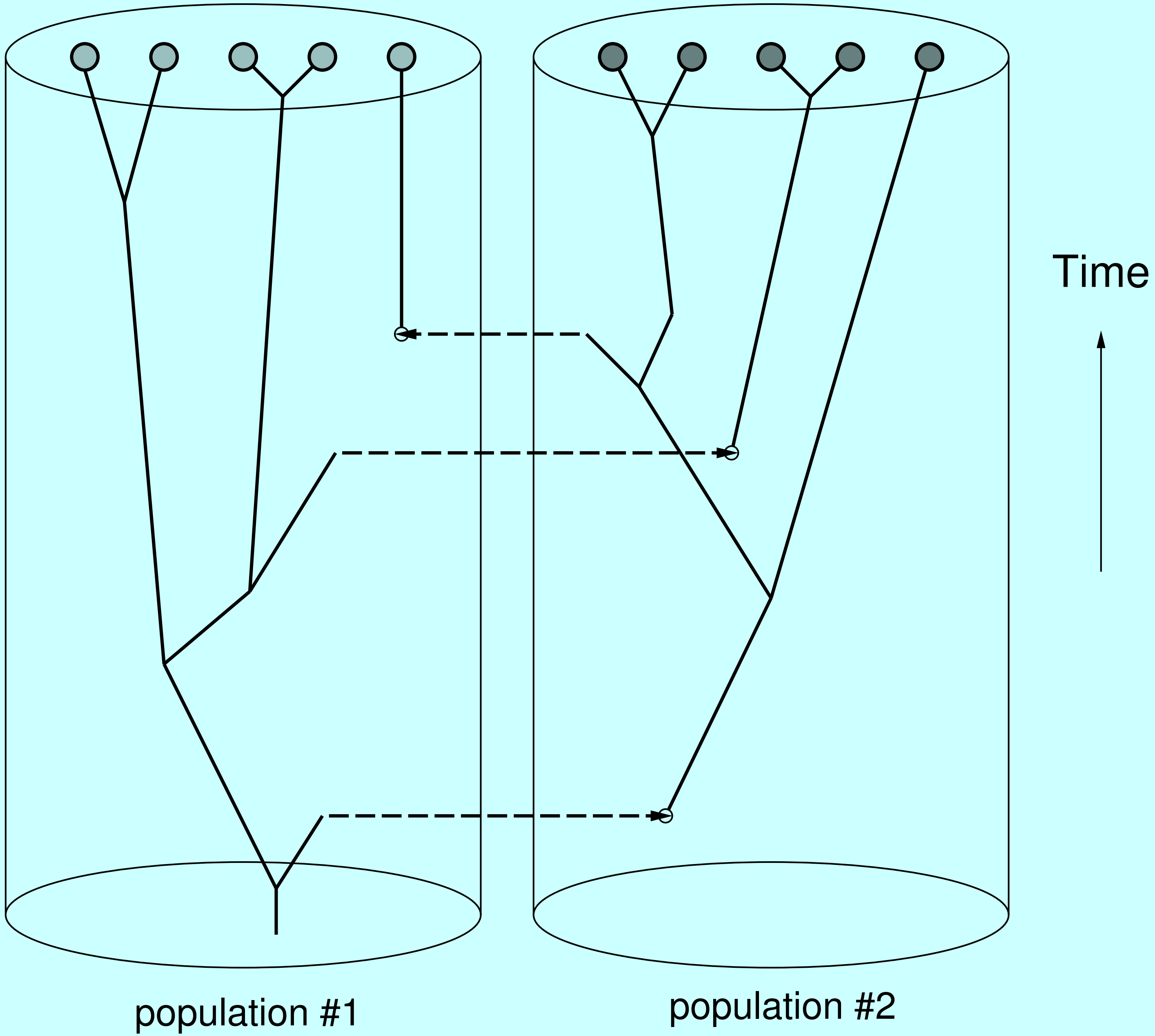


Coalescence events

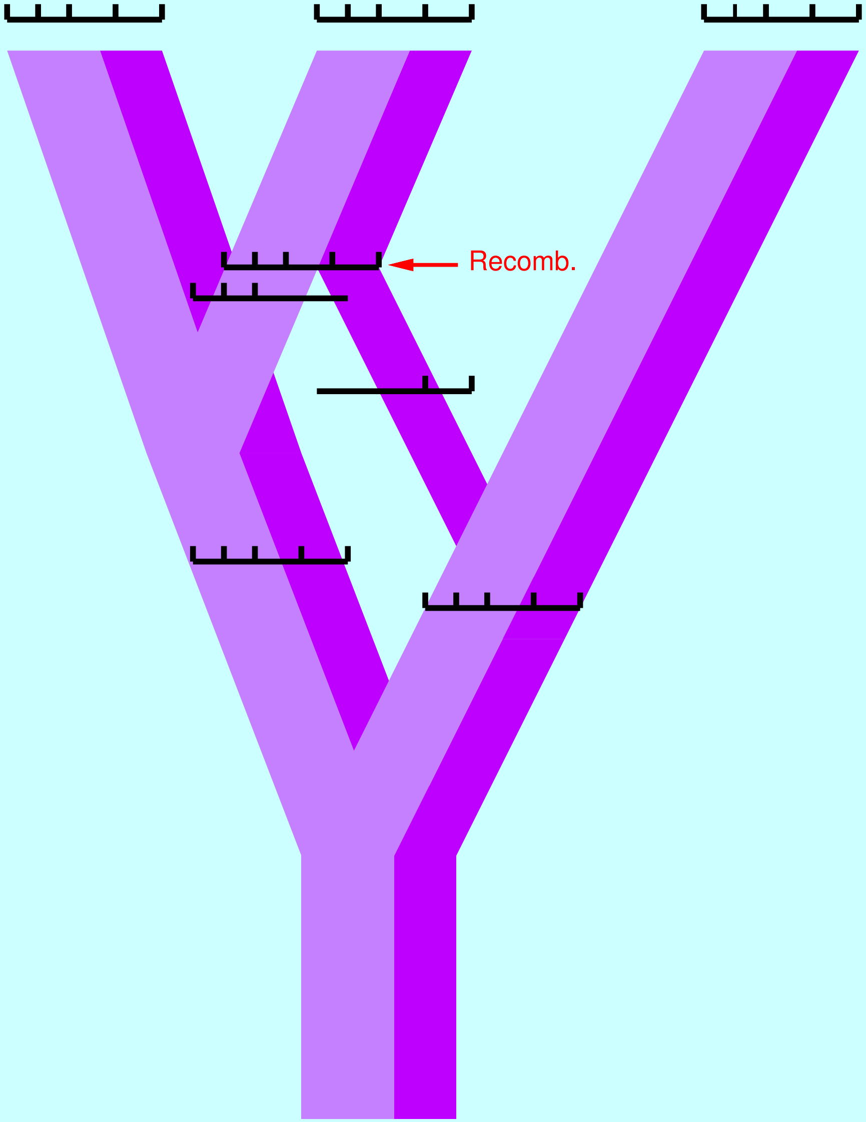


The parameters of the growth curve for N_e can be inferred by likelihood methods as they affect the prior probabilities of those trees that fit the data.

Migration can be taken into account

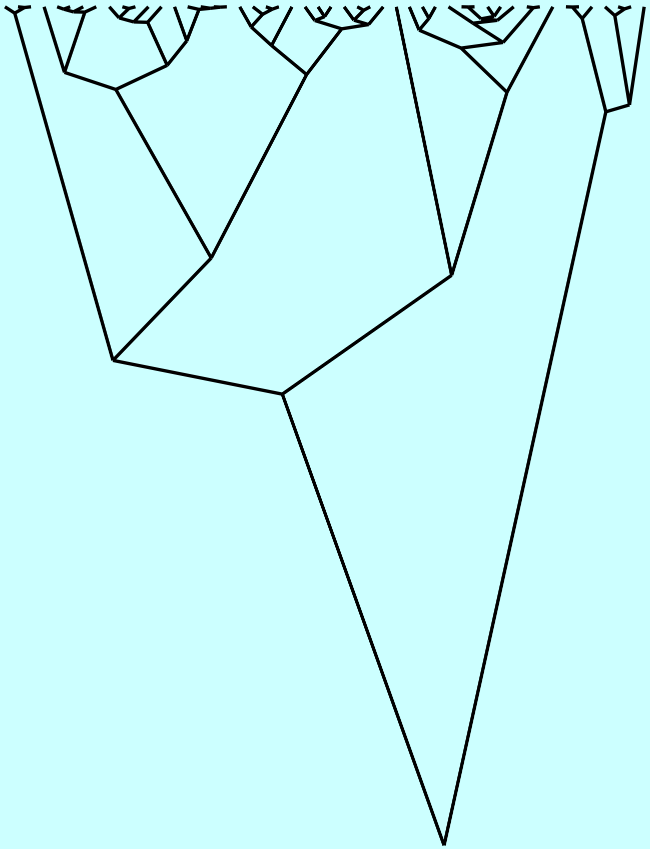


Recombination creates loops



Different markers have slightly different coalescent trees

If we have a sample of 50 copies

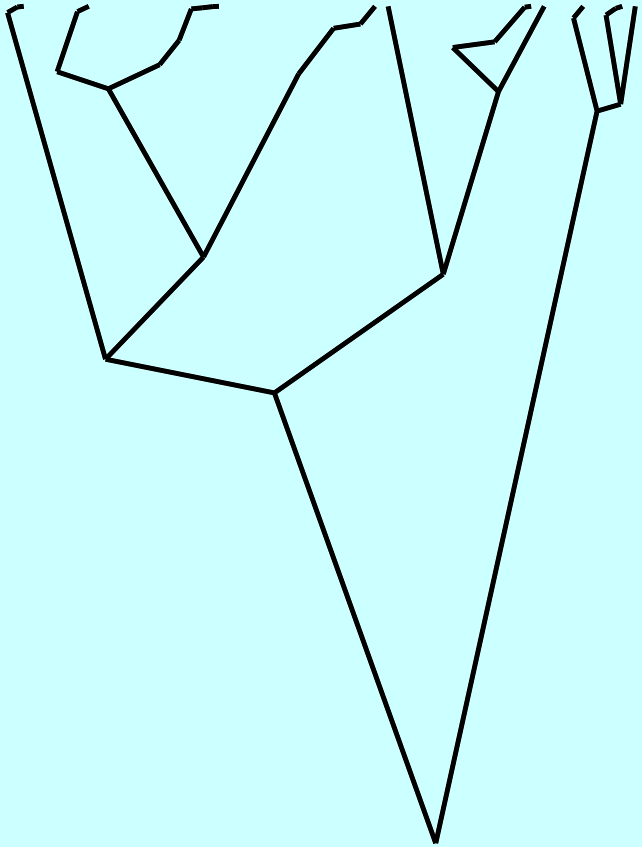


50−gene sample in a coalescent tree

The first 10 account for most of the branch length

10 genes sampled randomly out of a

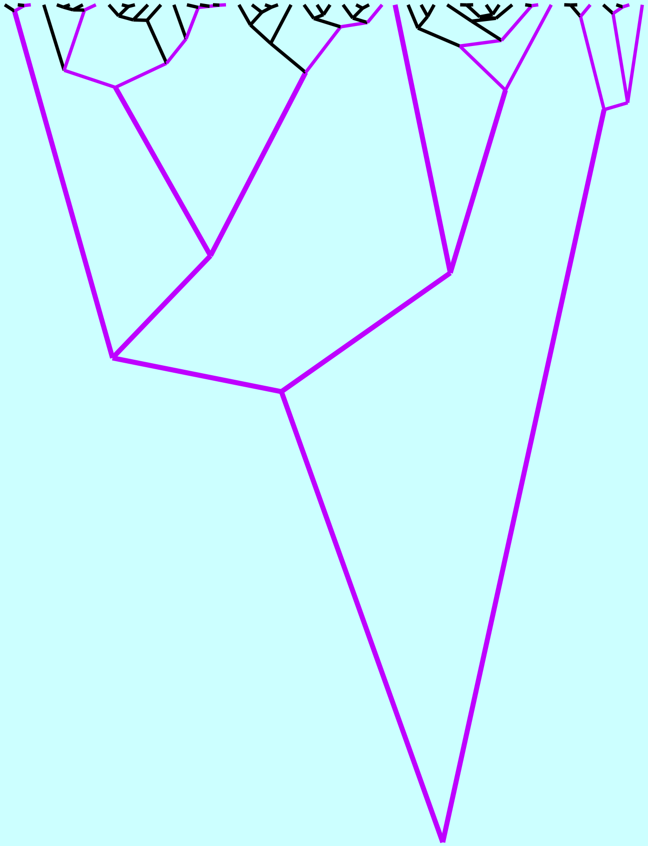
50−gene sample in a coalescent tree



... and when we add the other 40 they add less length

10 genes sampled randomly out of a

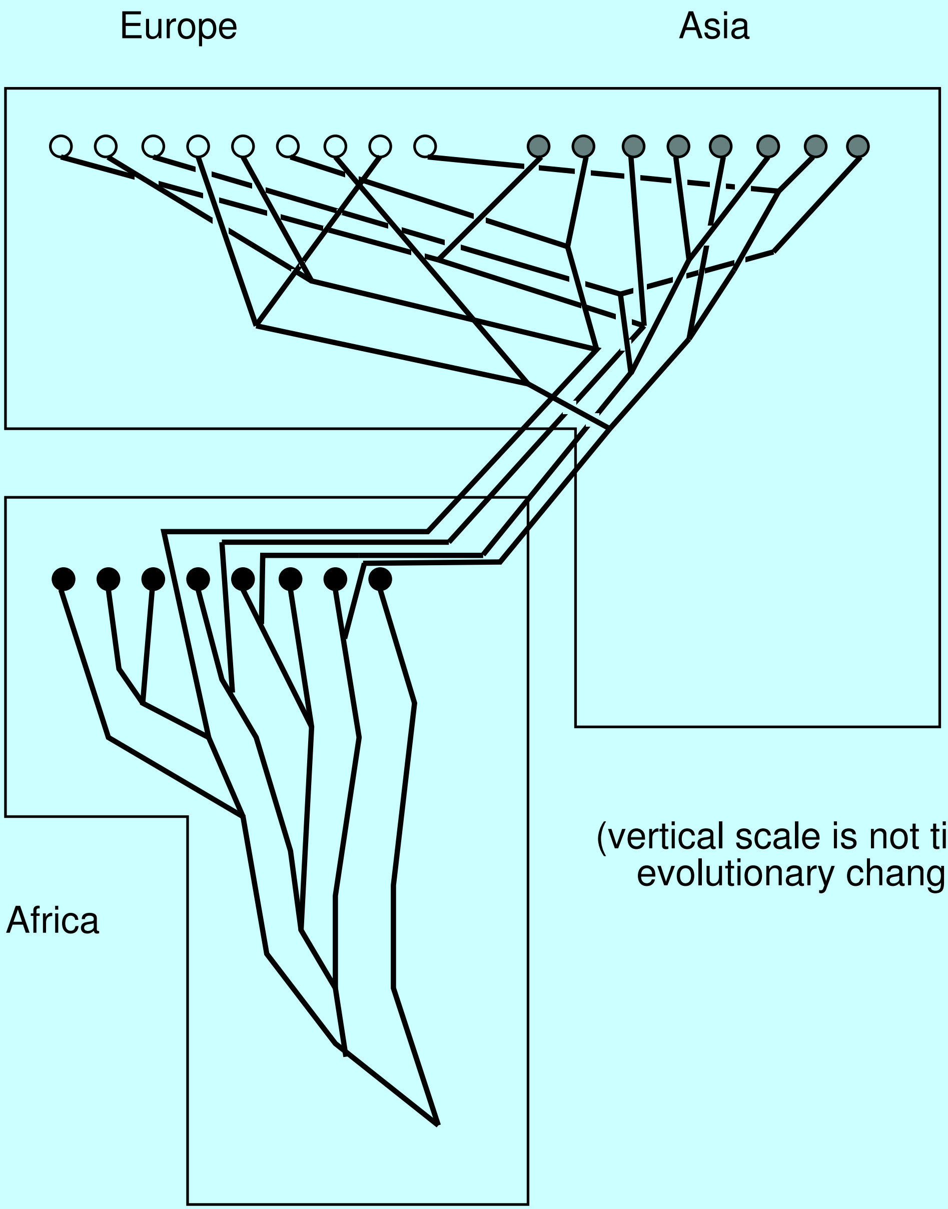
50–gene sample in a coalescent tree



(purple lines are the 10–gene tree)

We want to be able to analyze human evolution

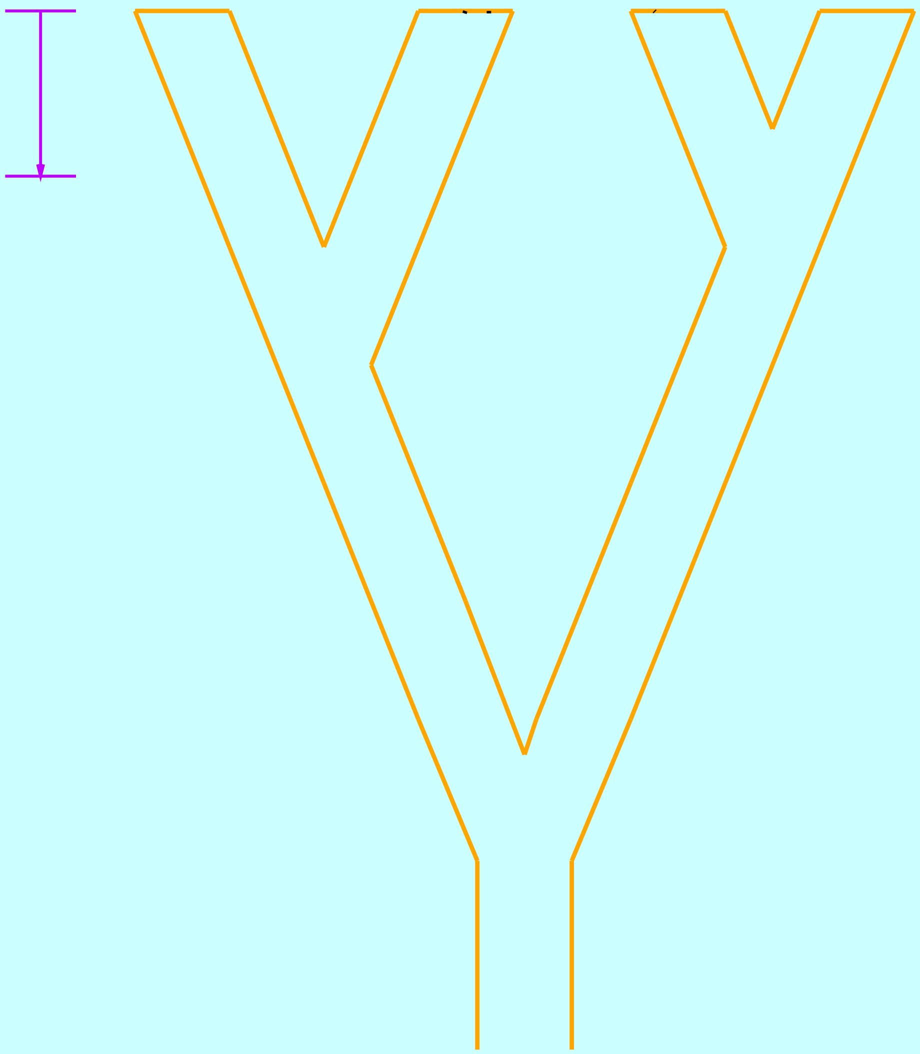
"Out of Africa" hypothesis



(vertical scale is not time or evolutionary change)

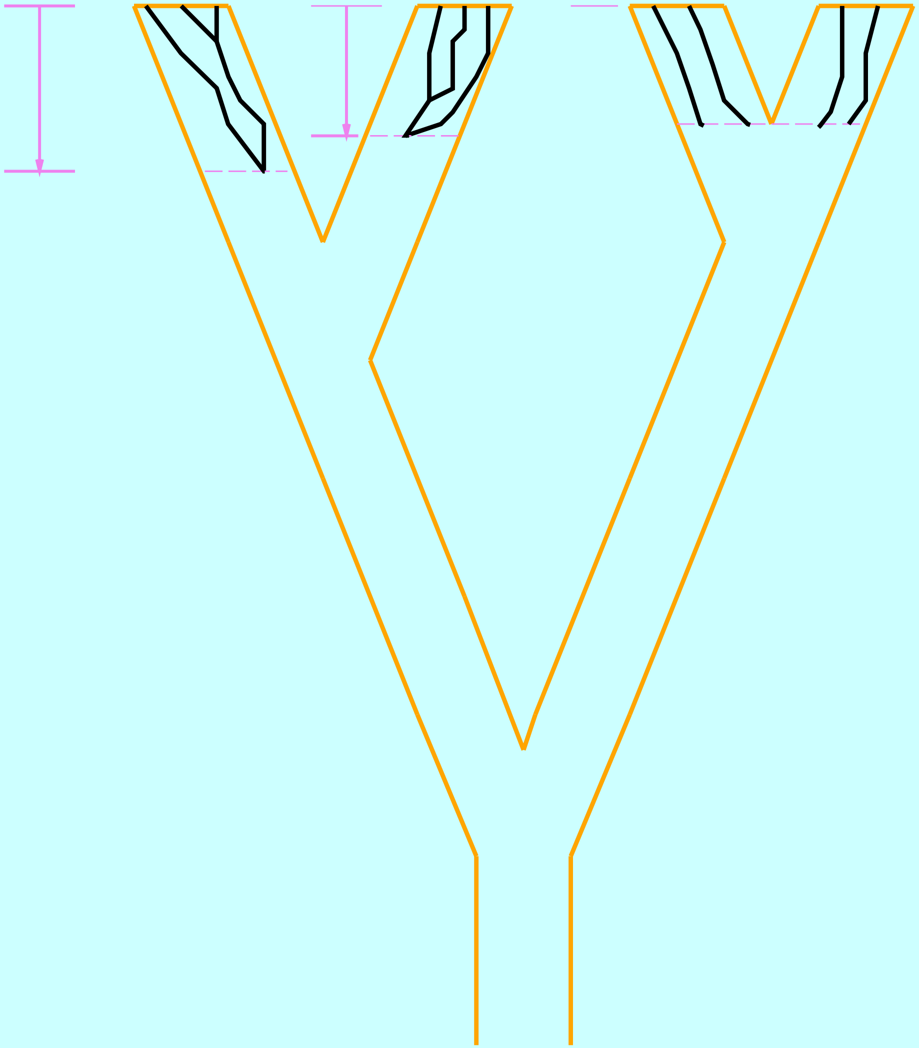
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



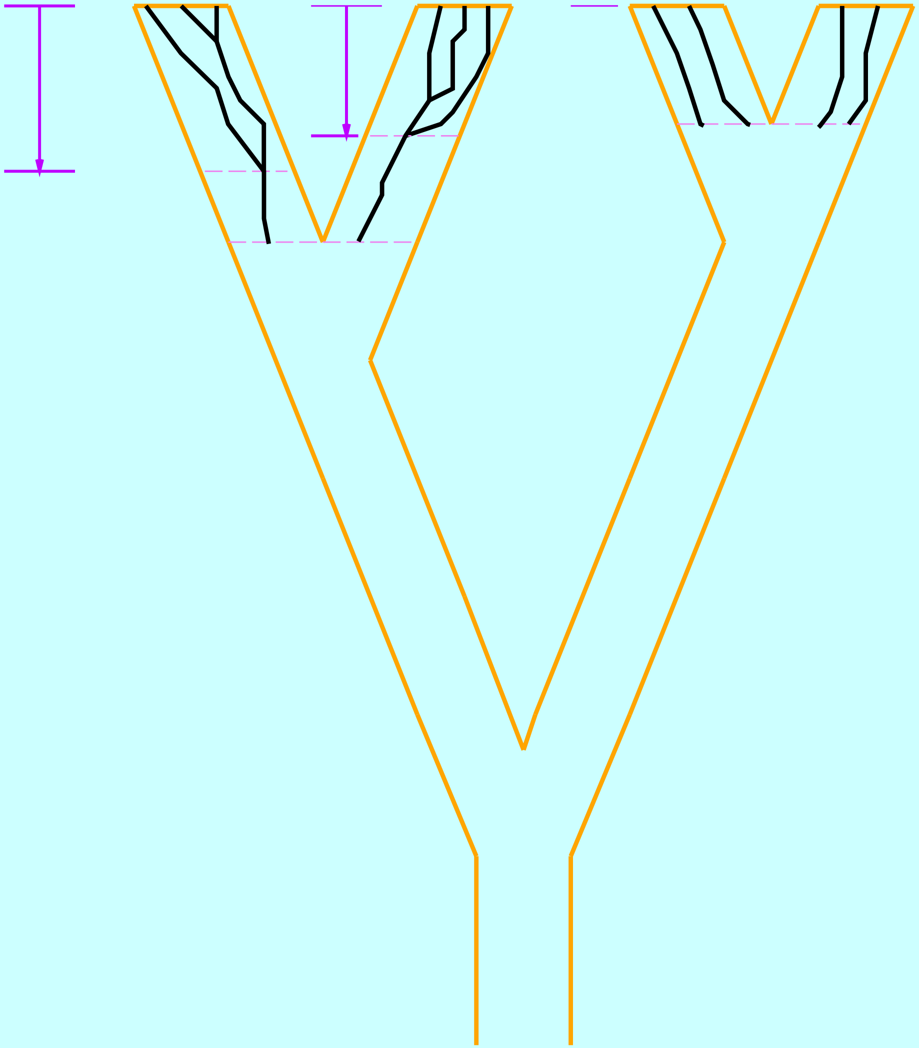
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



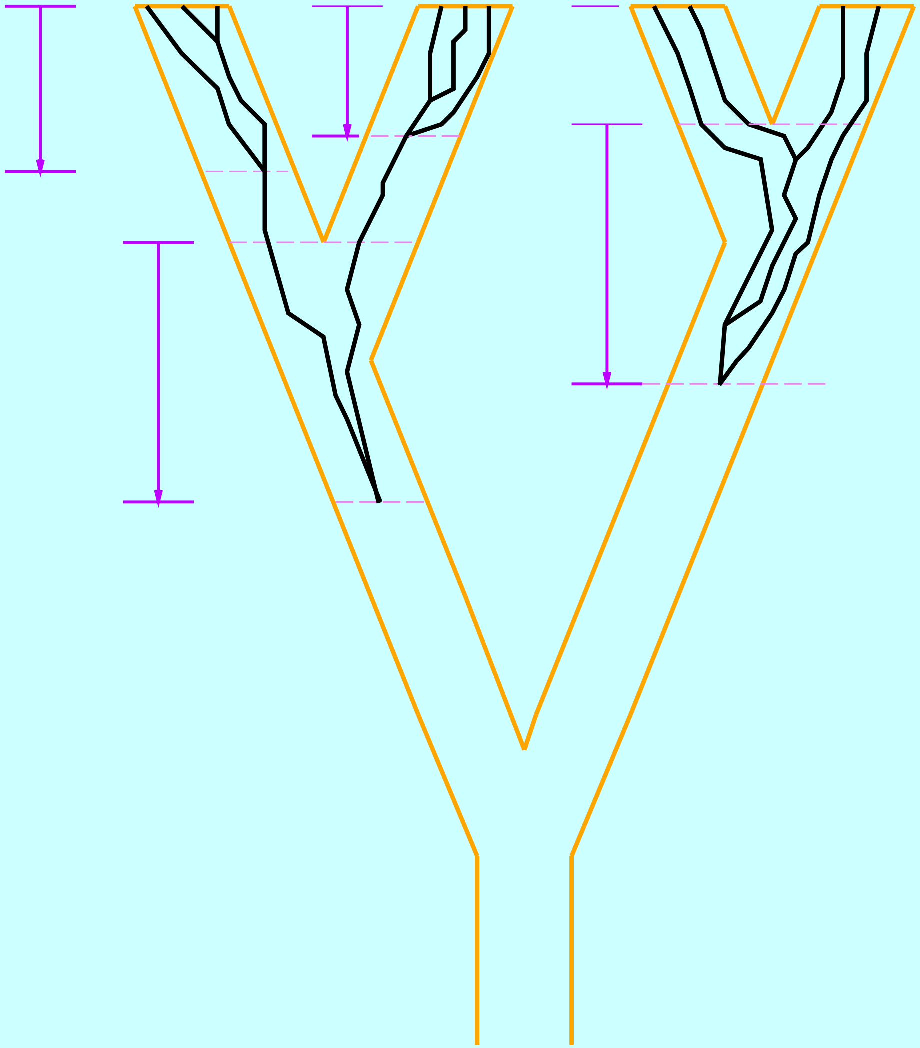
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



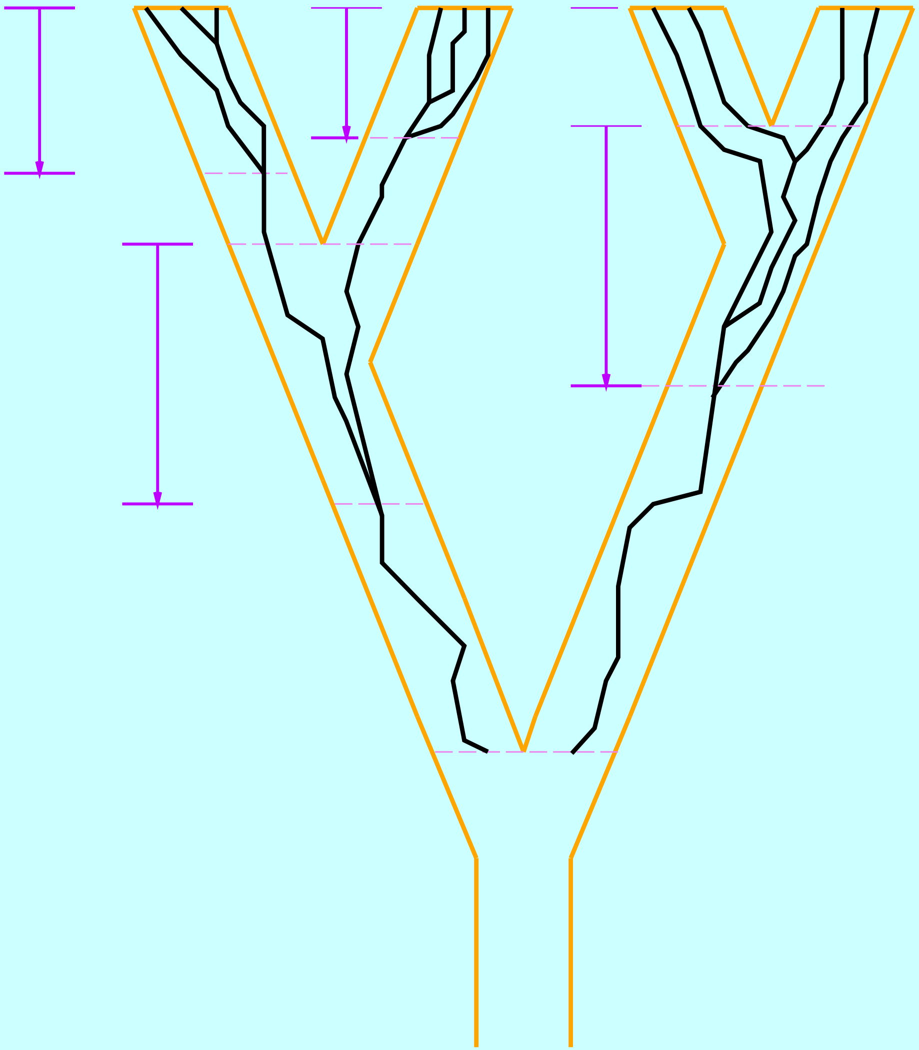
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



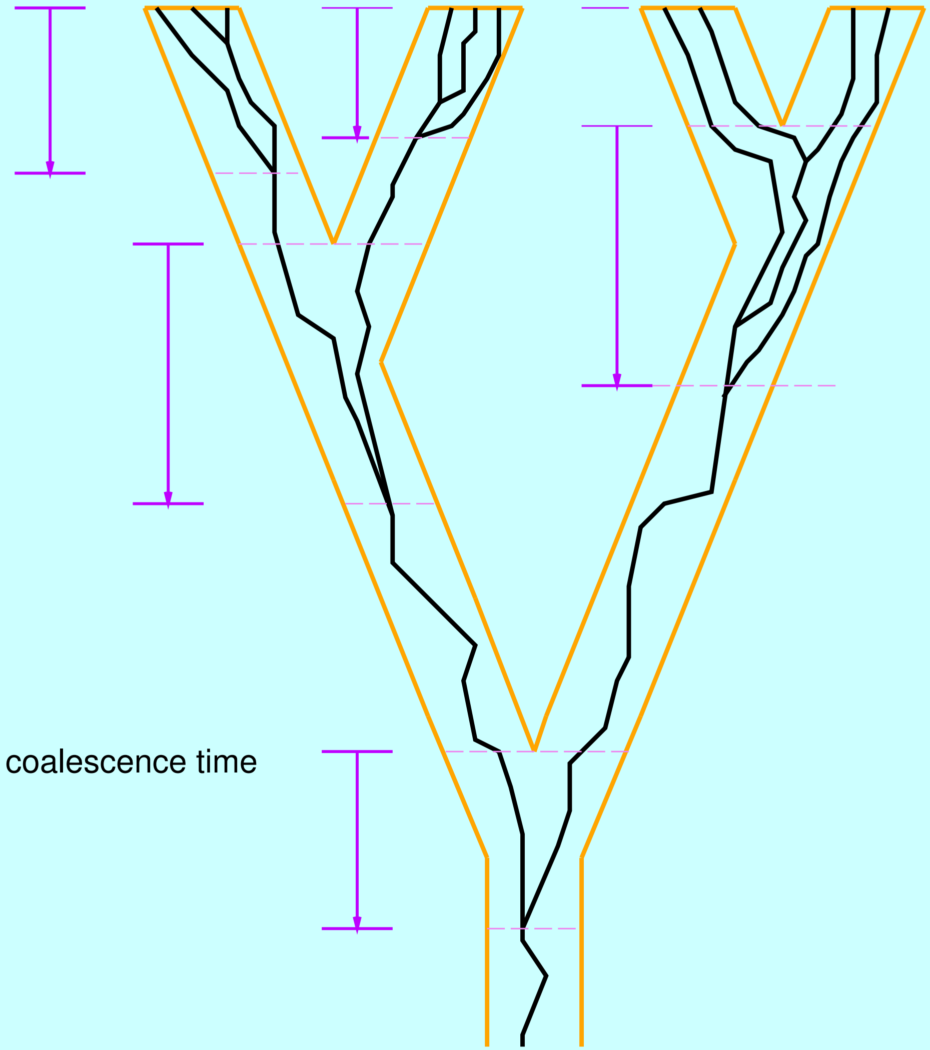
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



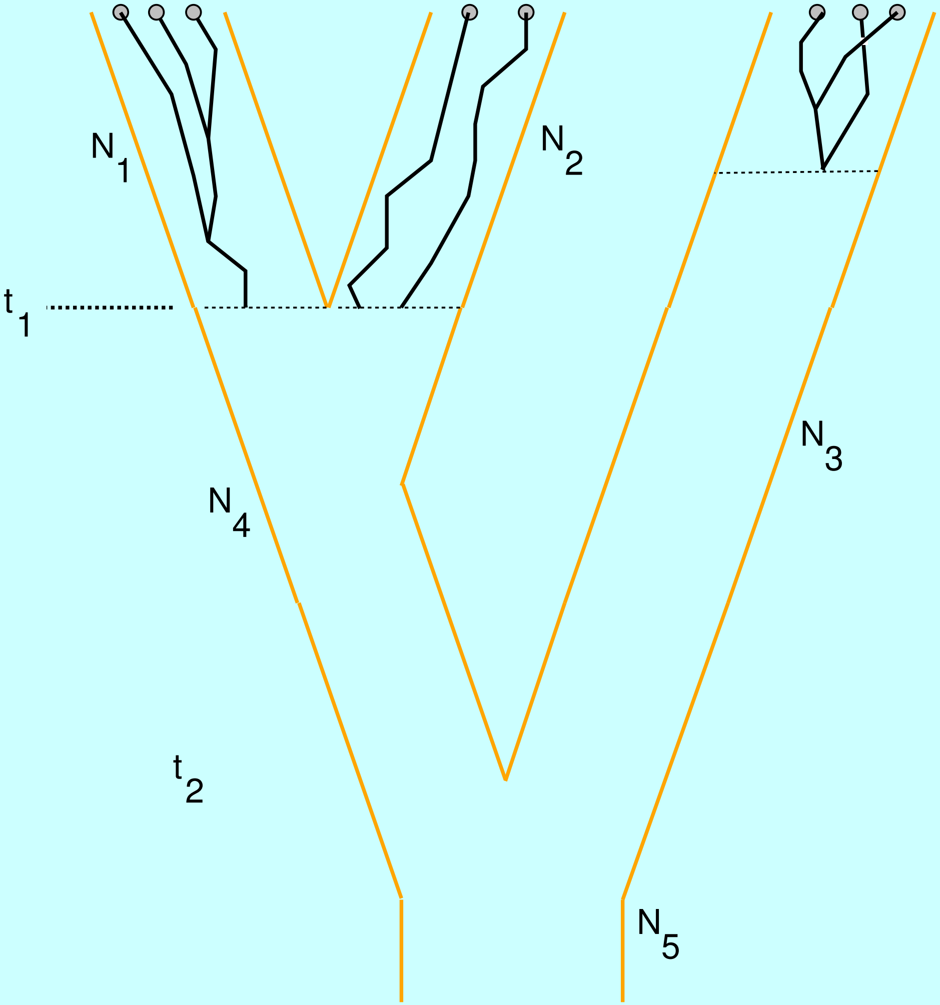
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



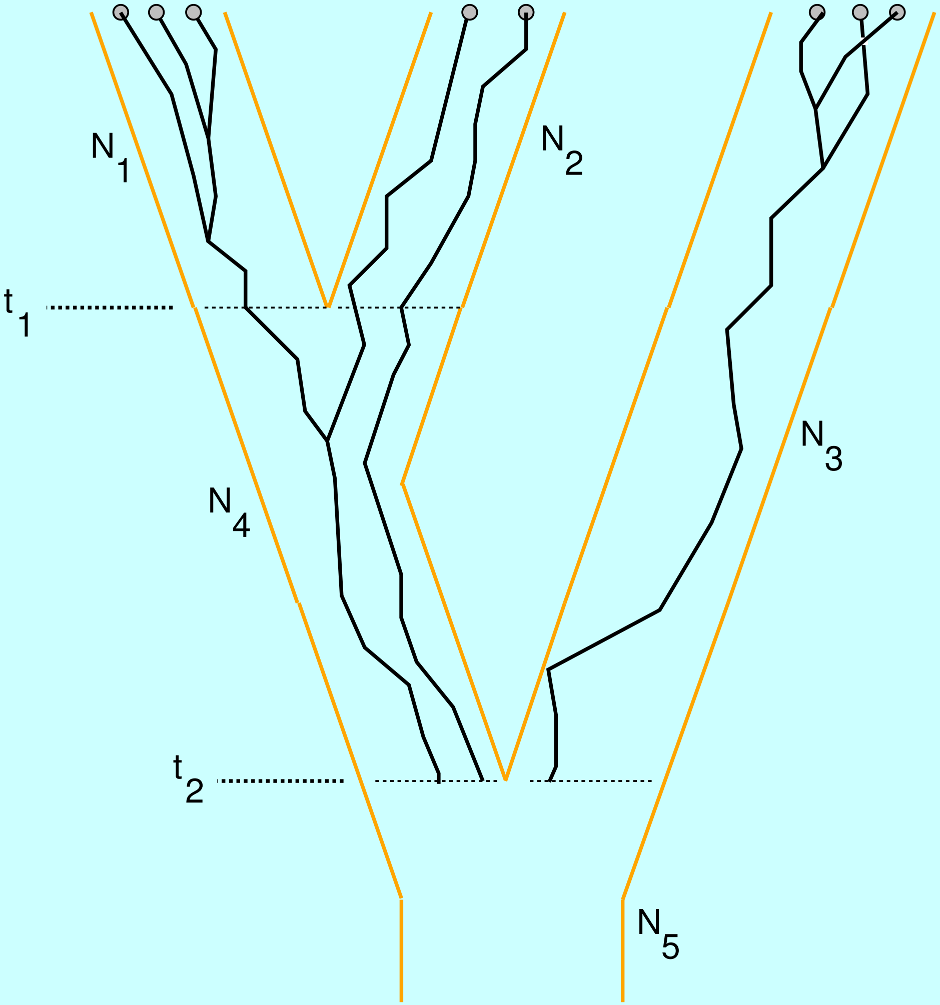
If the branch is more than N_e generations long ...

Gene tree and Species tree



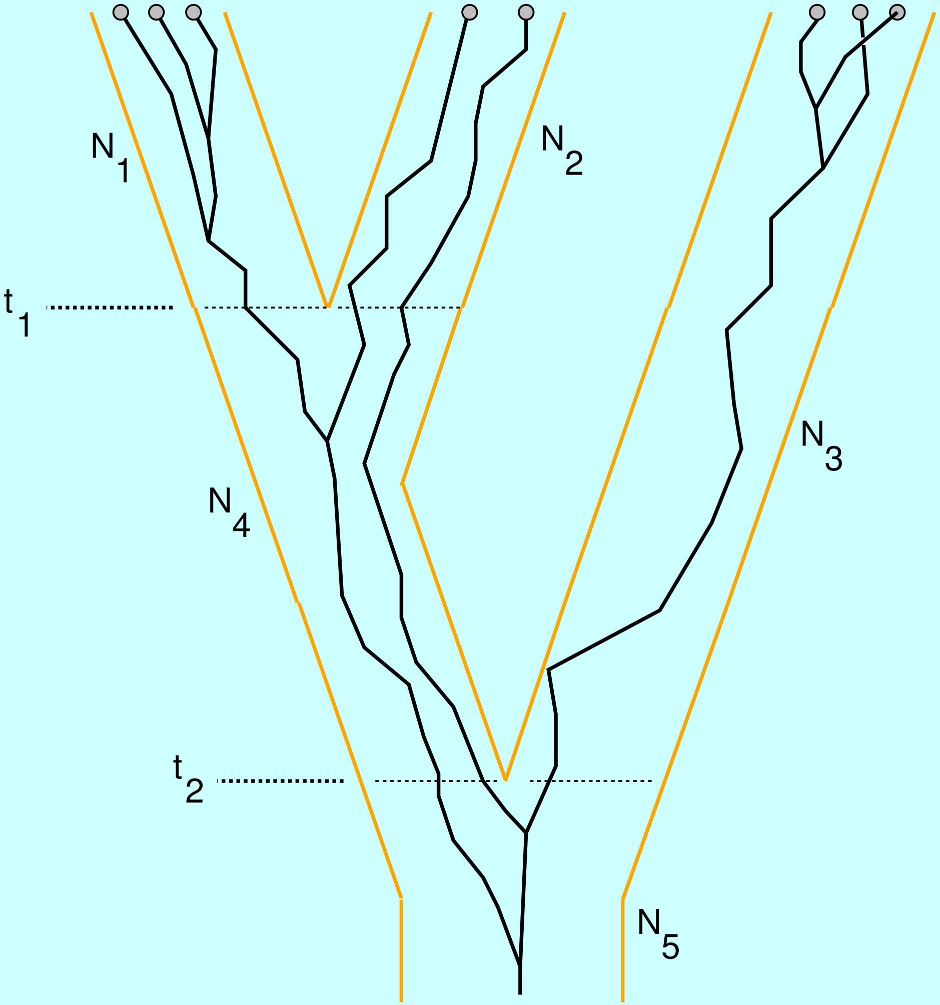
If the branch is more than N_e generations long ...

Gene tree and Species tree



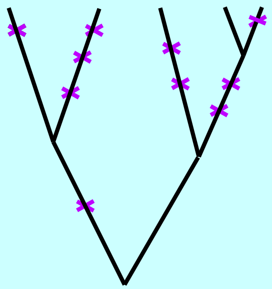
If the branch is more than N_e generations long ...

Gene tree and Species tree



The variability comes from two sources

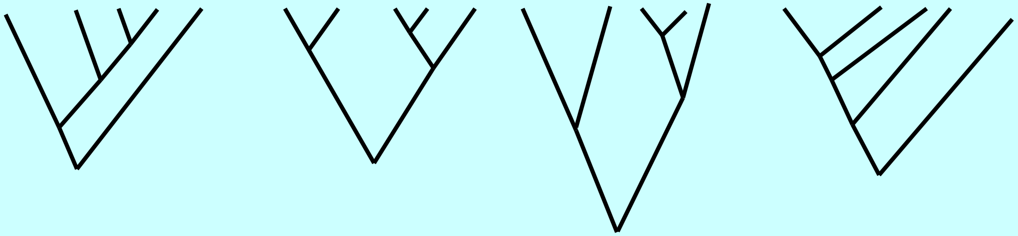
(1) Randomness of mutation



affected by the mutation rate μ

can reduce variance of number of mutations per site per branch by examining more sites

(2) Randomness of coalescence of lineages



affected by effective population size N_e

coalescence times allow estimation of N_e

can reduce variability by looking at

(i) more gene copies, or

(ii) more loci