Lecture 7. Coalescents

Joe Felsenstein

Department of Genome Sciences and Department of Biology

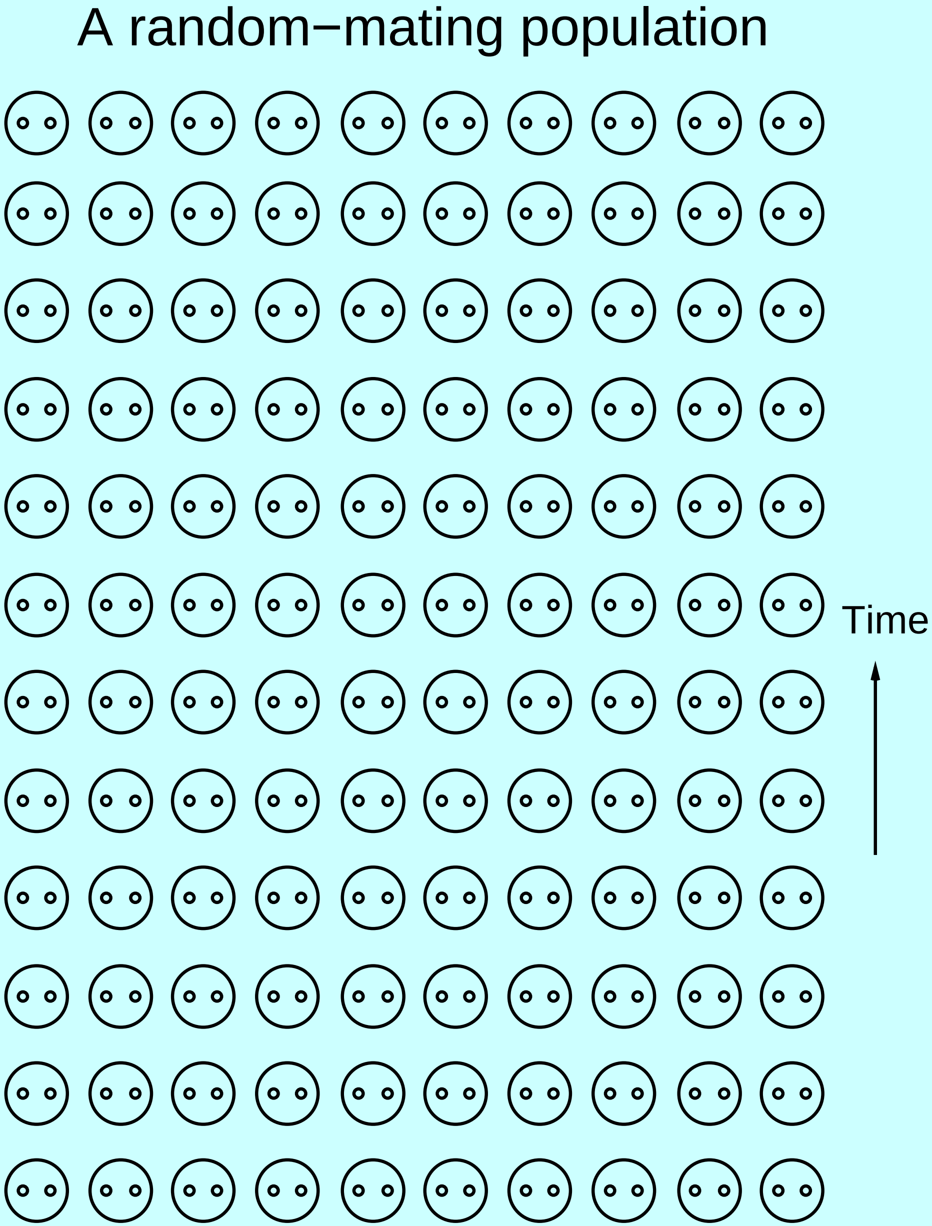
The Wright-Fisher model

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

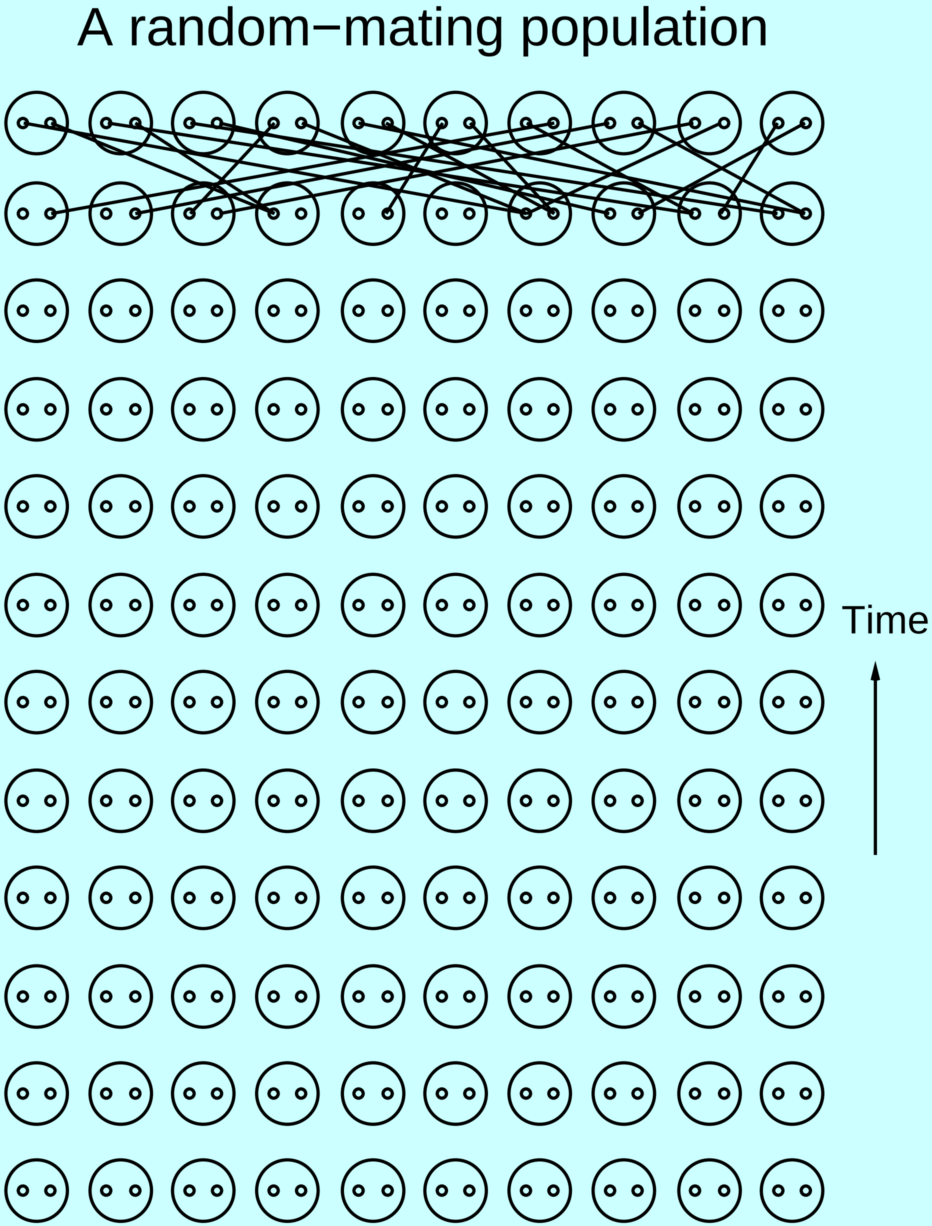
- Choose two individuals *with replacement* (including the possibility that they are the same individual) to be parents,
- Each produces one gamete, these become a diploid individual,
- Repeat these steps until N diploid individuals have been produced.

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.

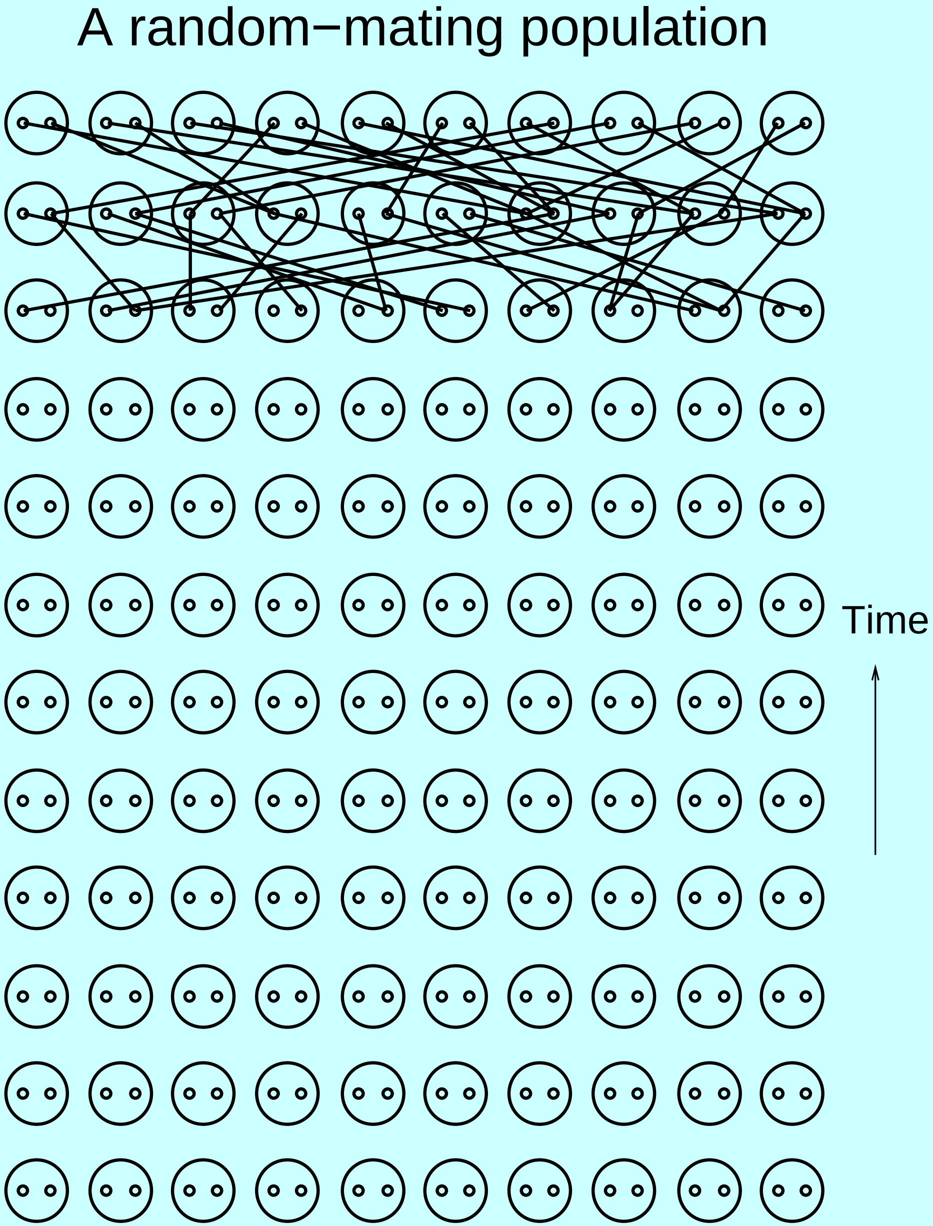
The ancestry of gene copies in a Wright-Fisher model



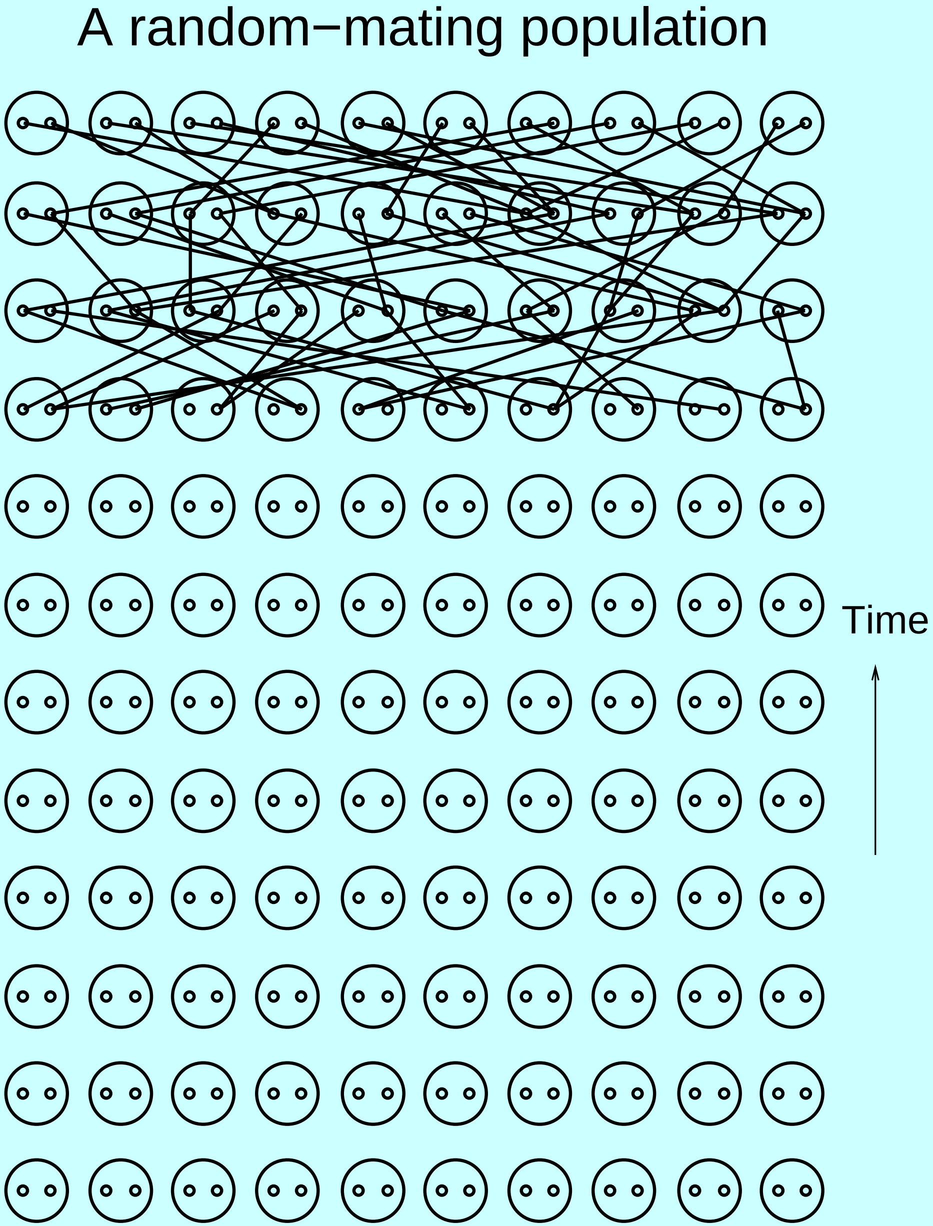
Each copy drawn from a random one in previous generatio



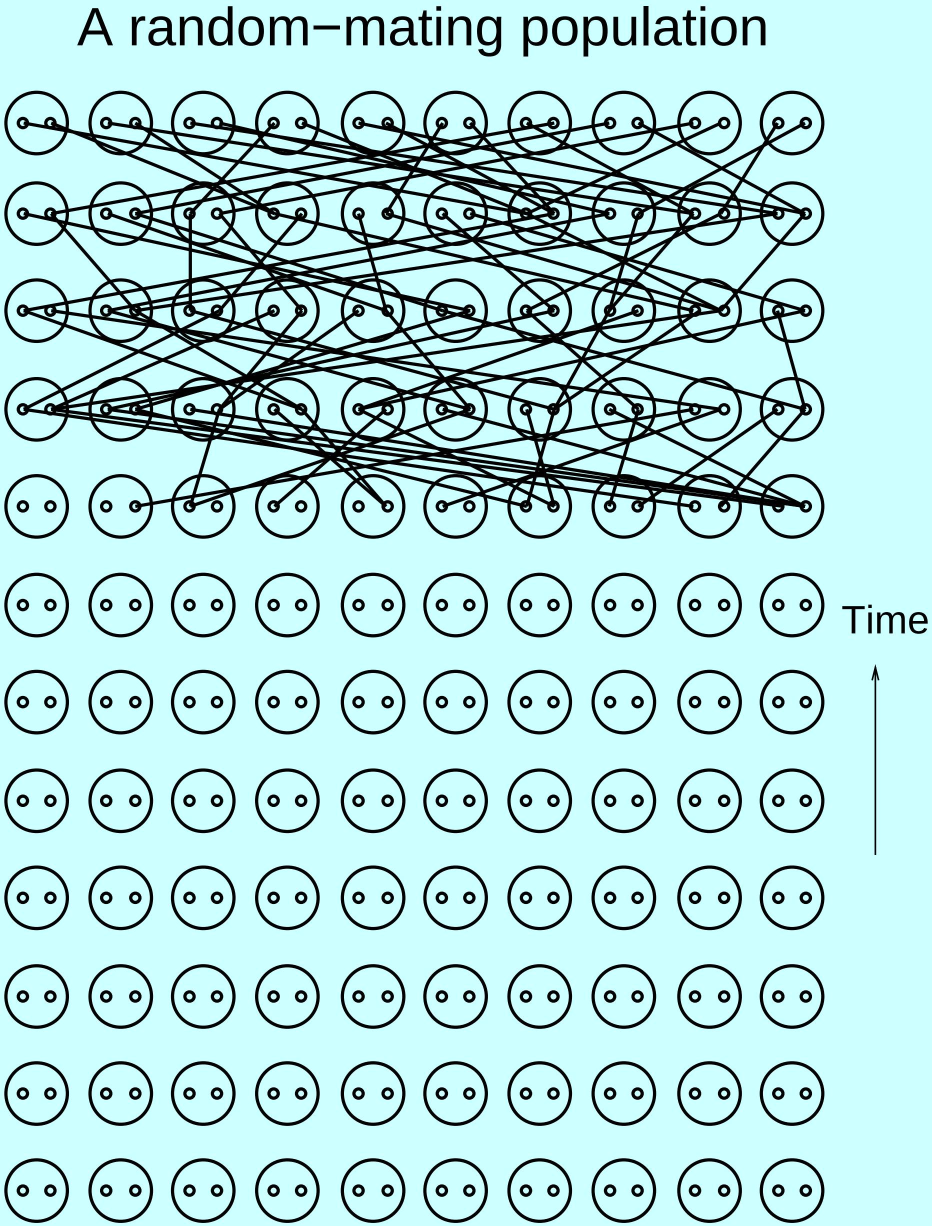
and going further back ...



and even further

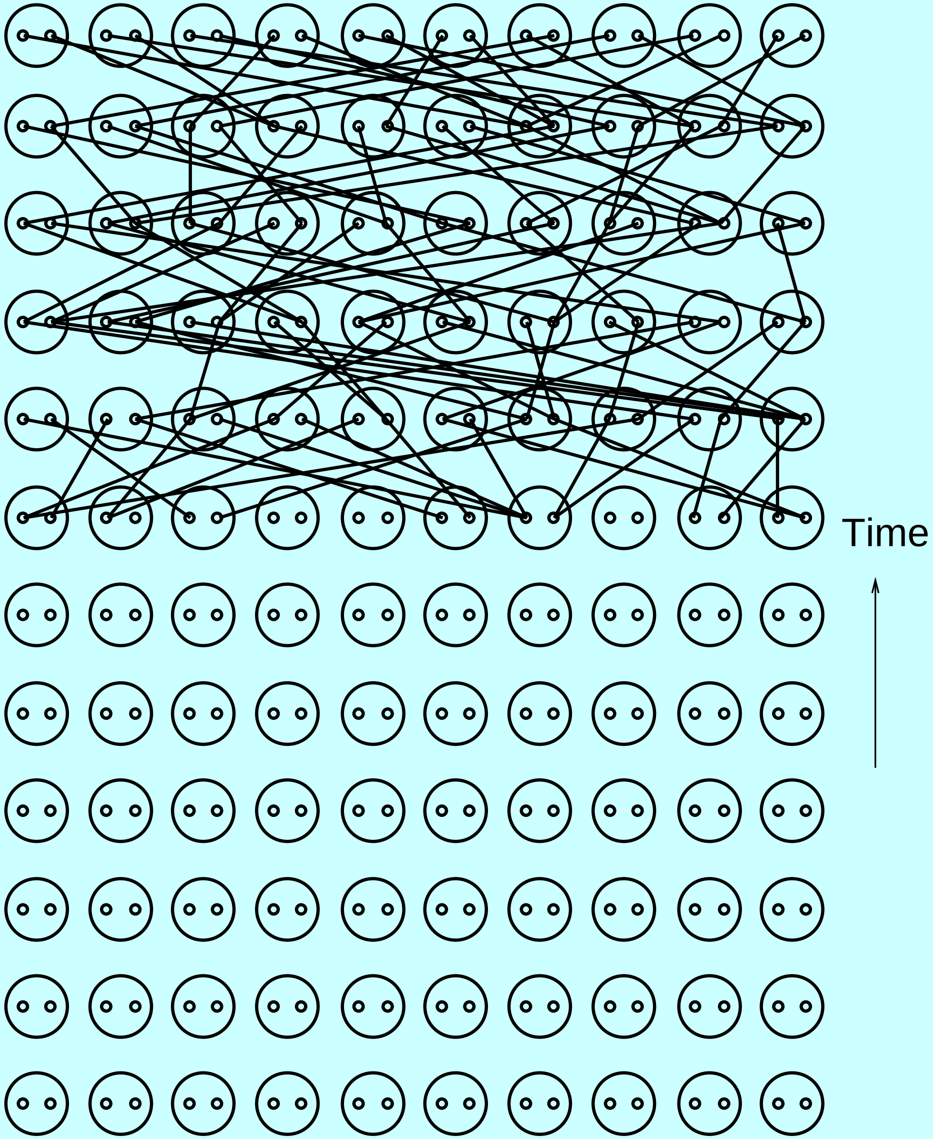


and further

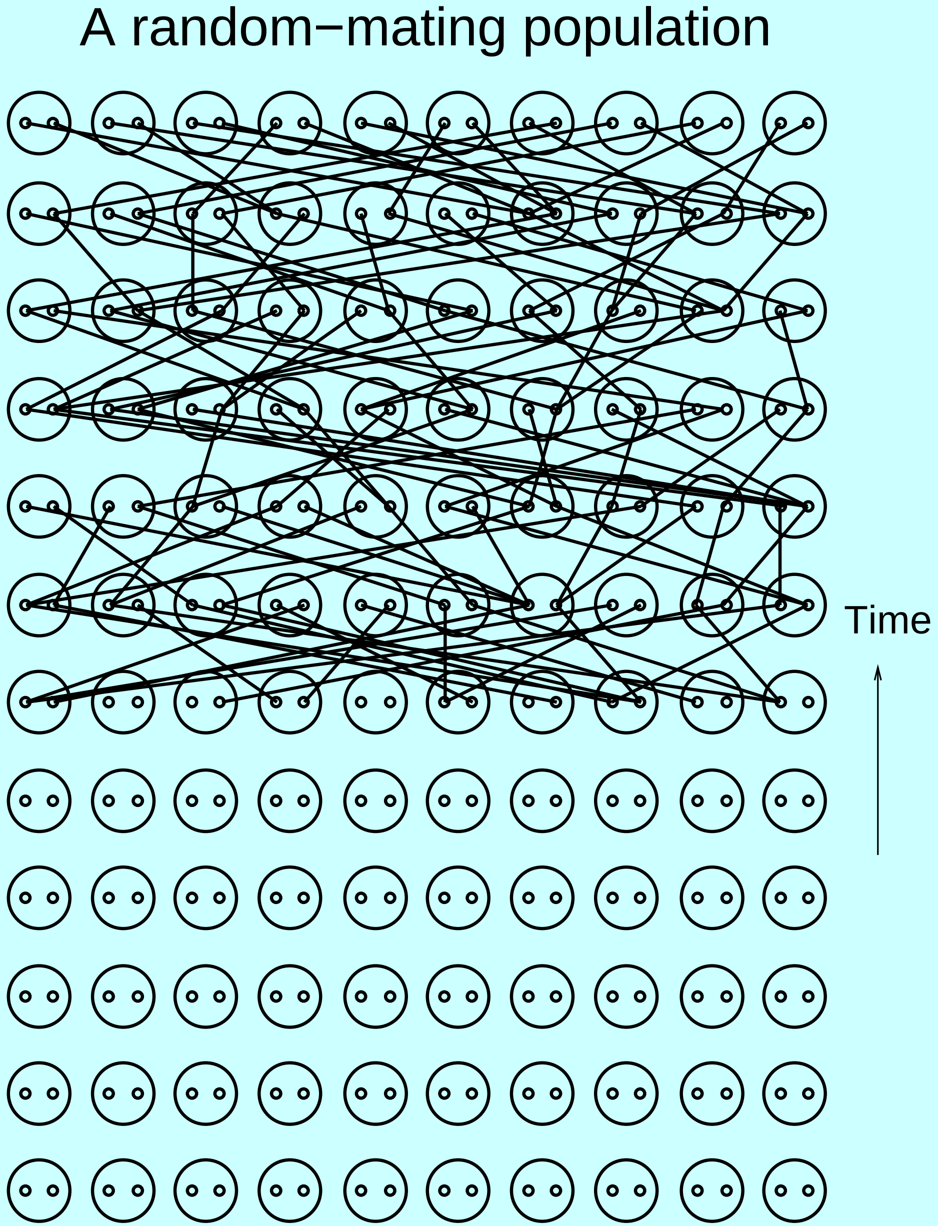


and so on

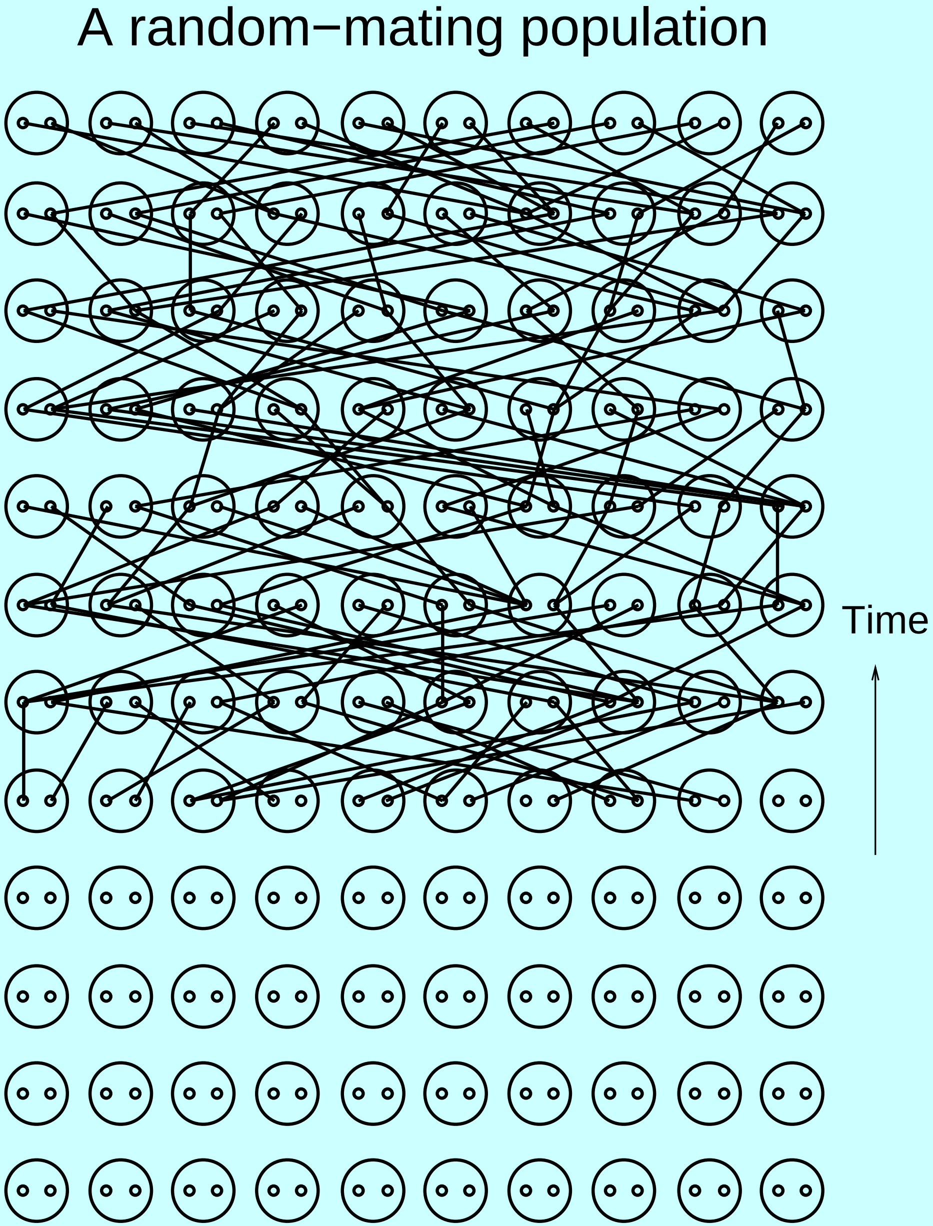
A random−mating population



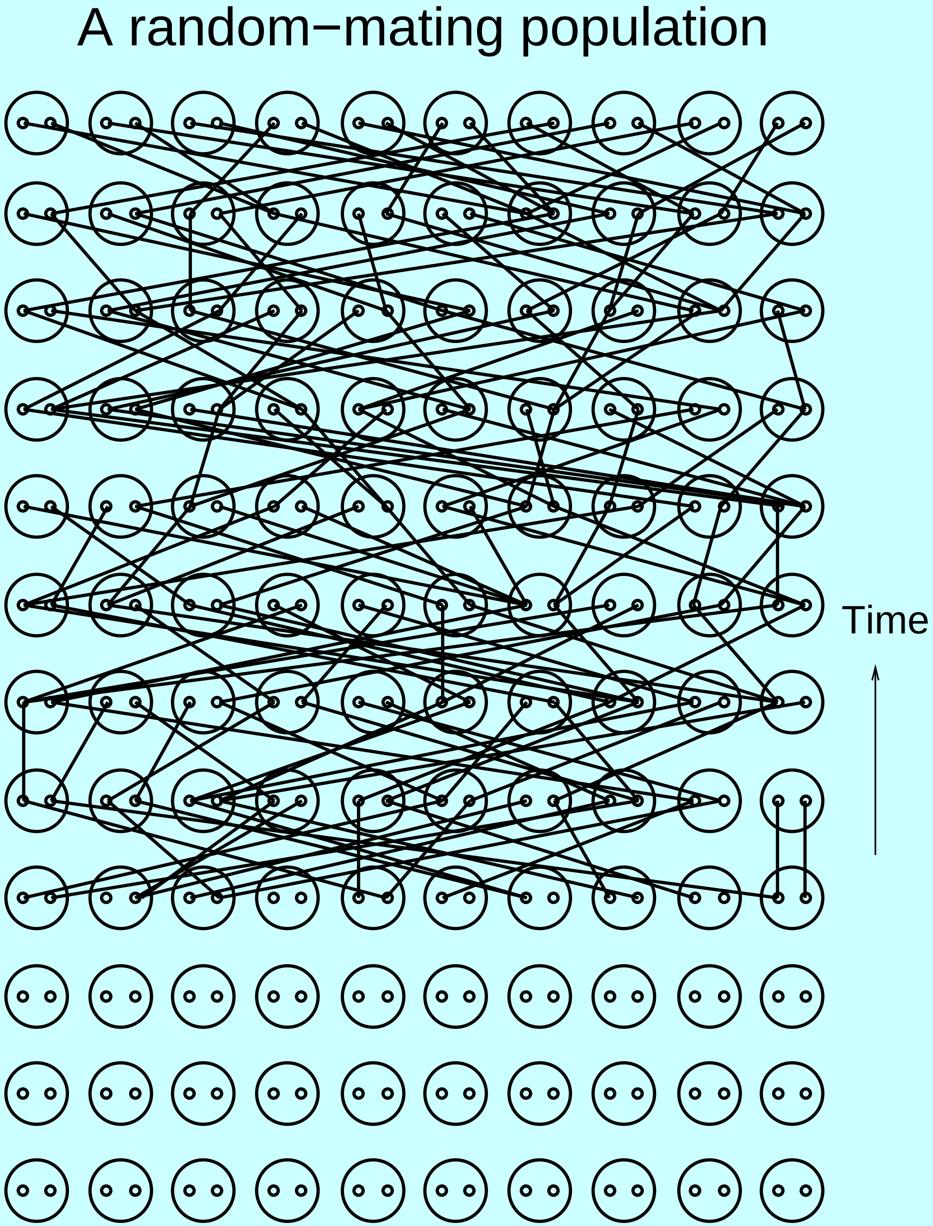
and on



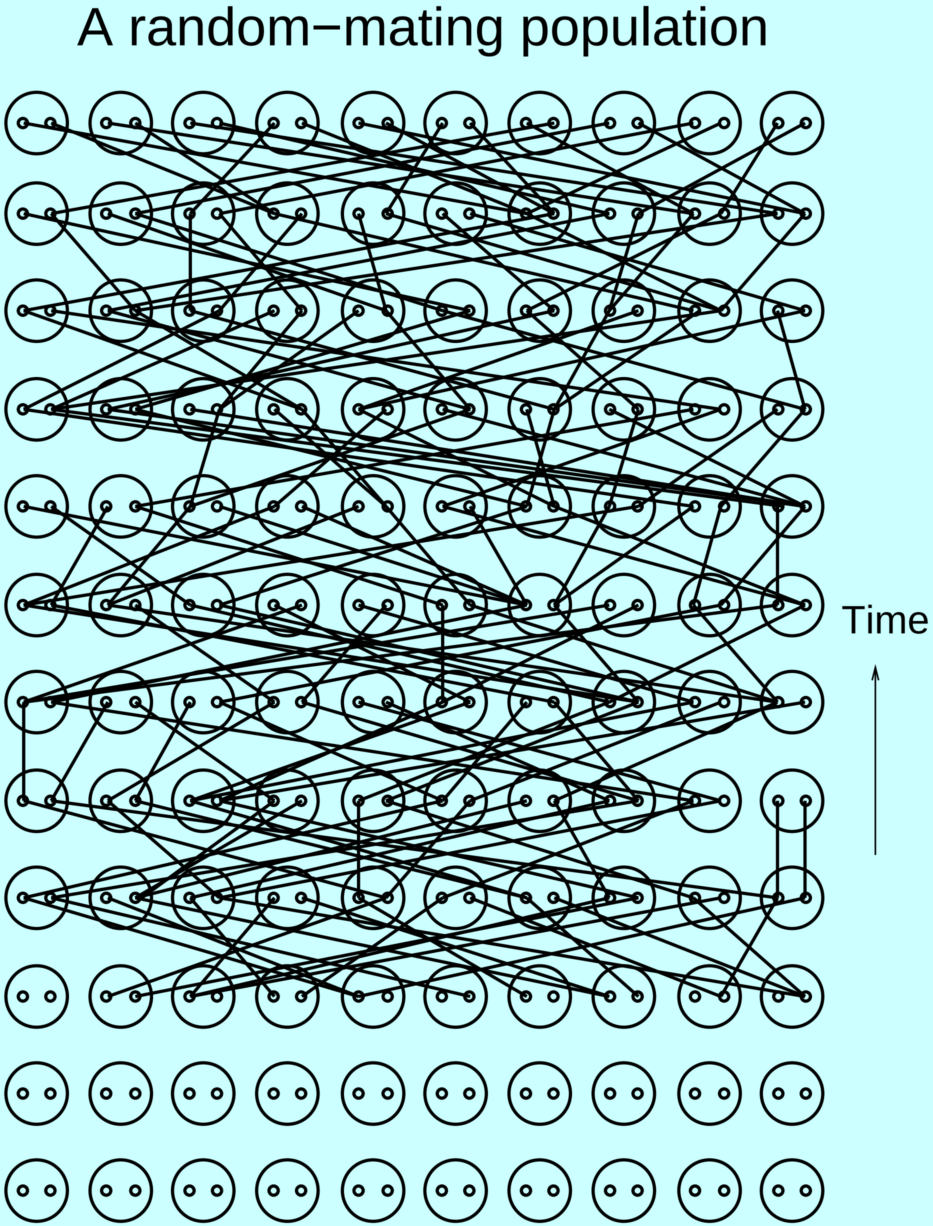
(yawn)



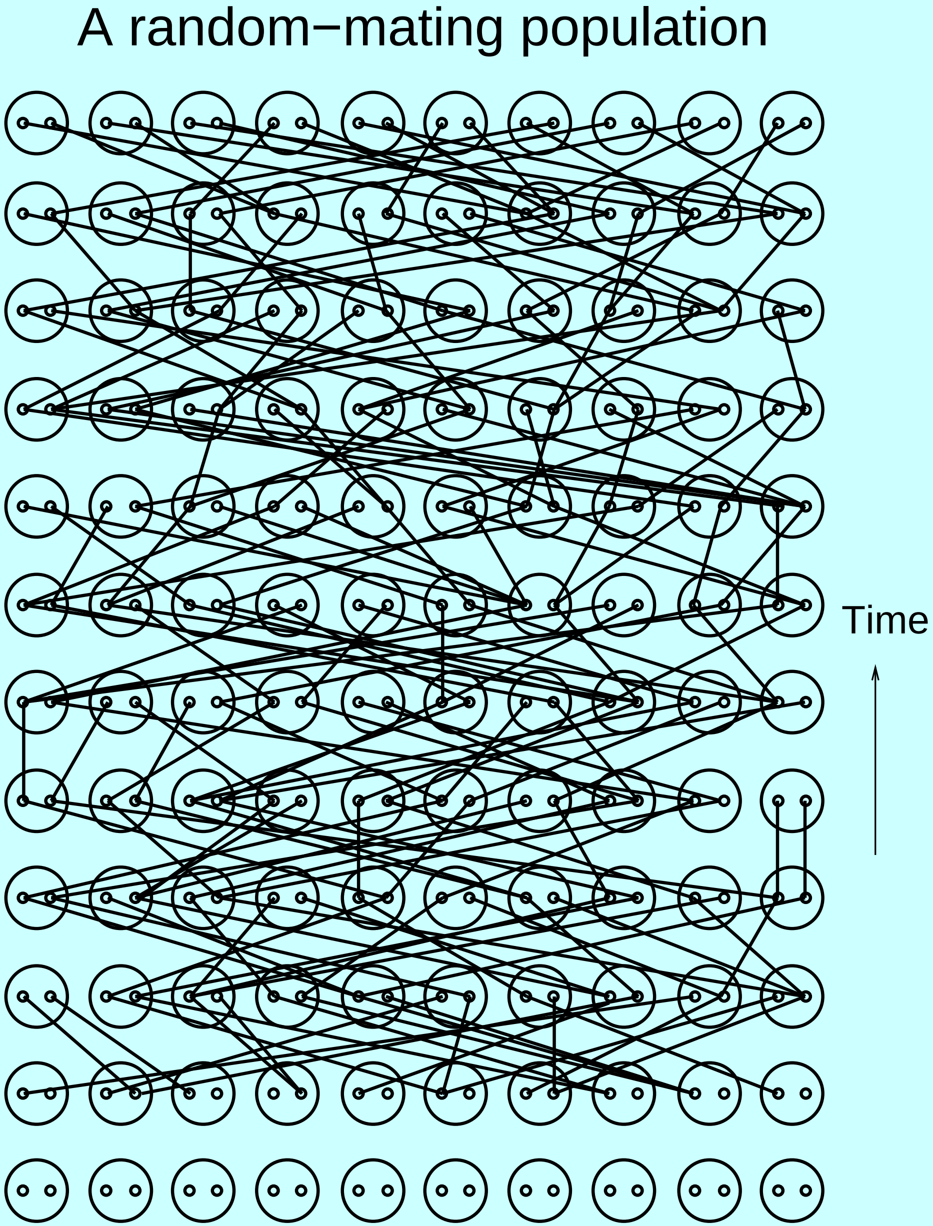
nearly there



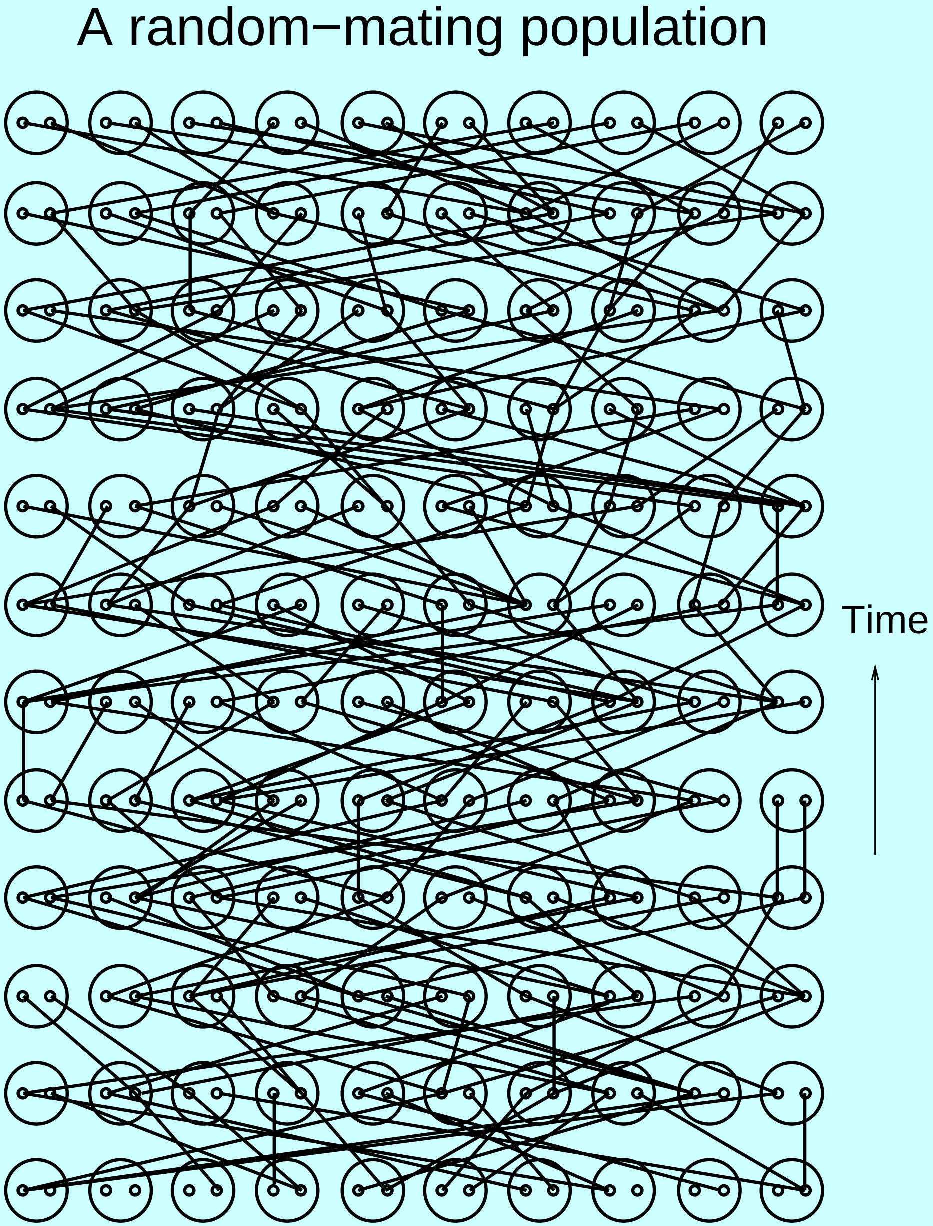
almost!



one more after this

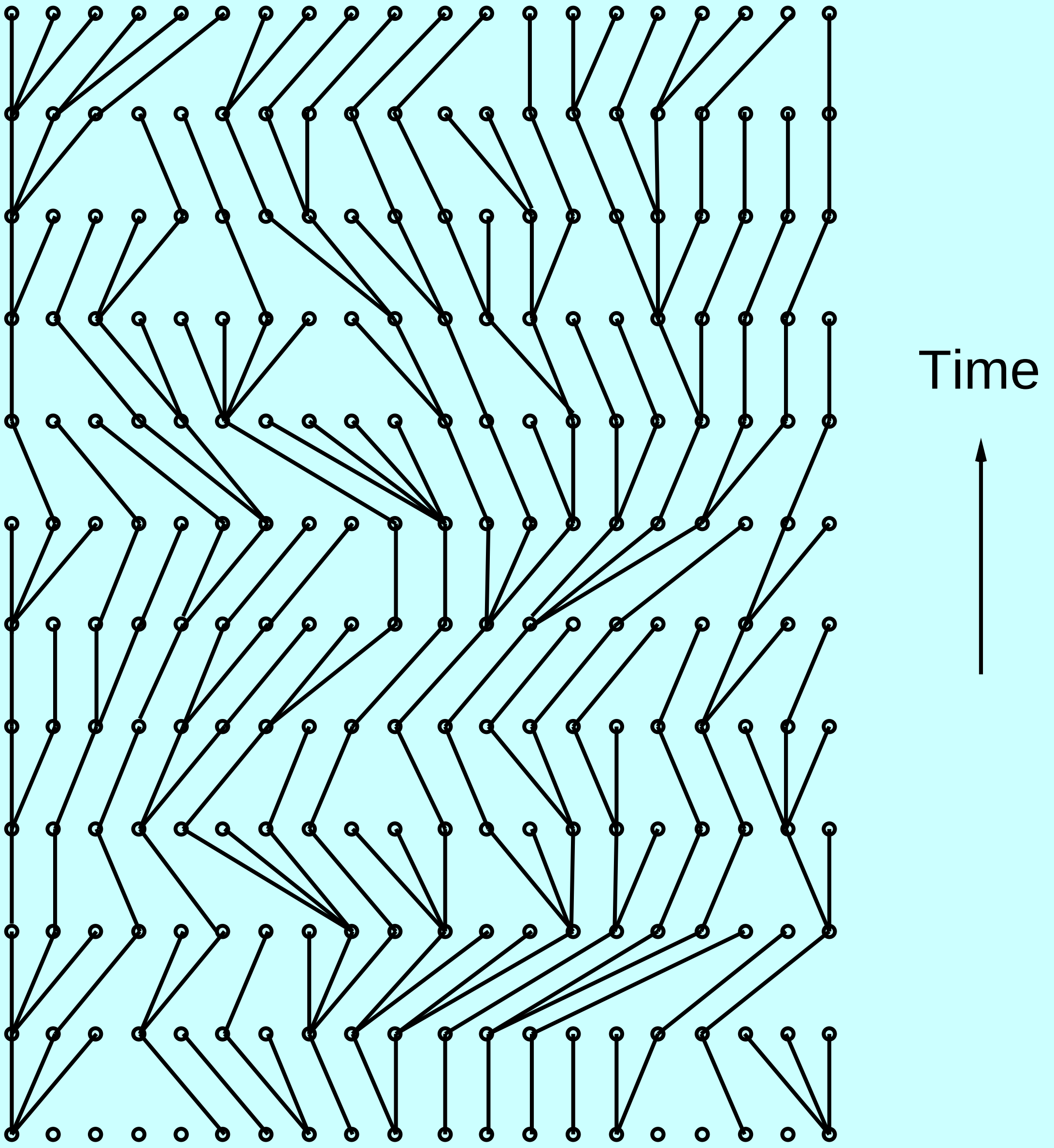


OK, so this is the pedigree of genes



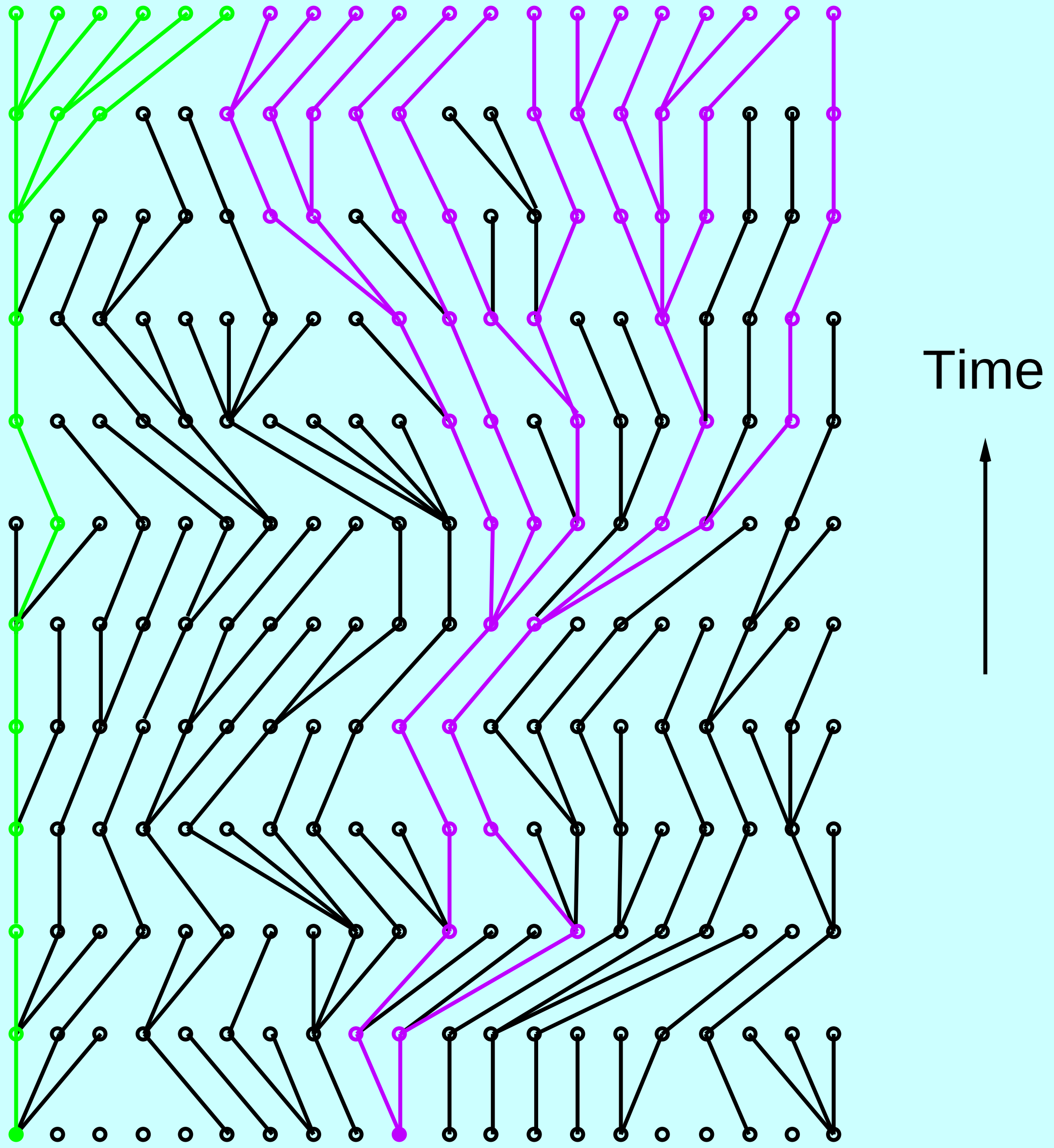
The ancestry of gene copies, untangled

Genealogy of gene copies, after reordering the copies



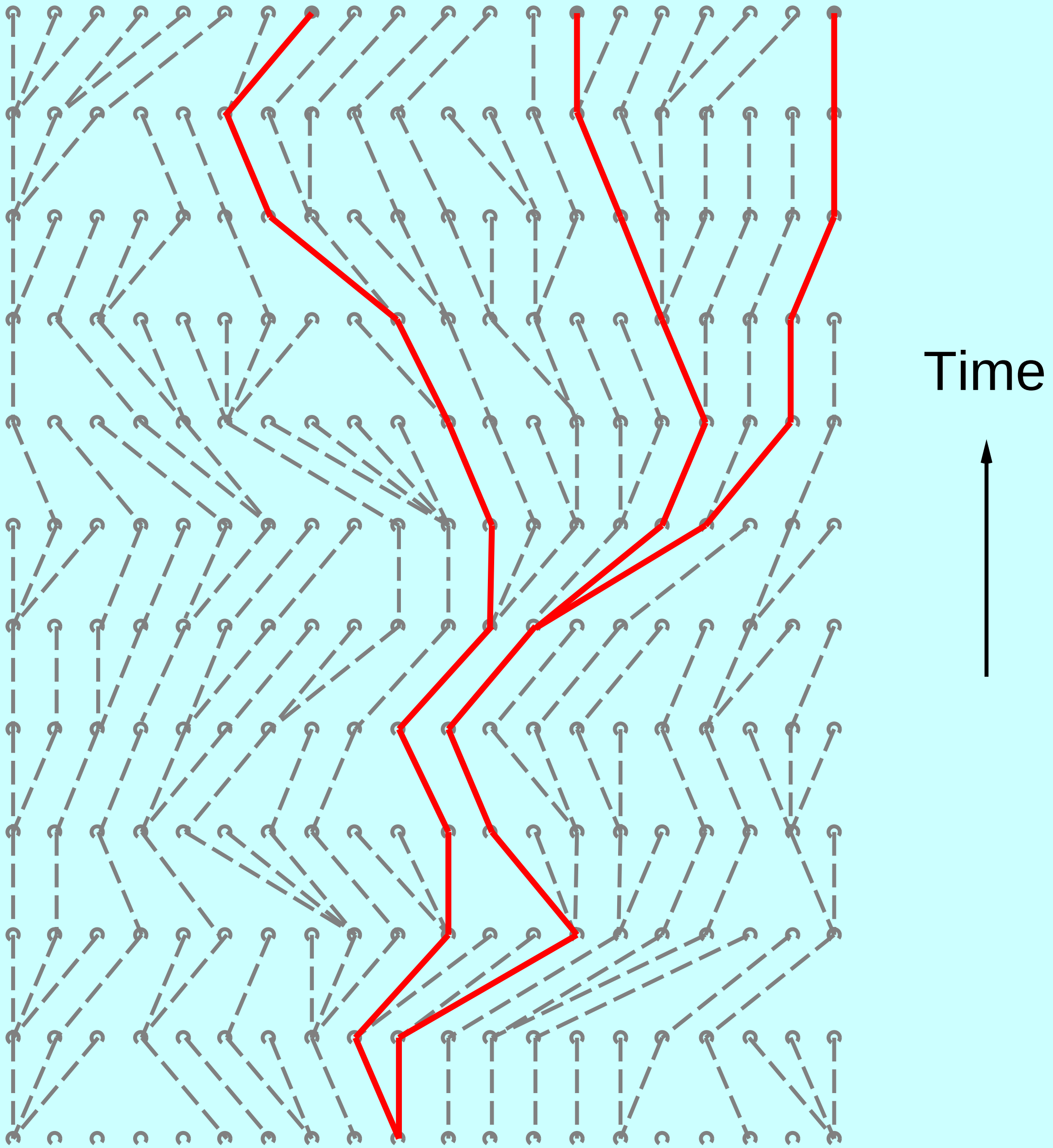
The ancestry of present-day gene copies (untangled)

Genealogy of gene copies, after reordering the copies

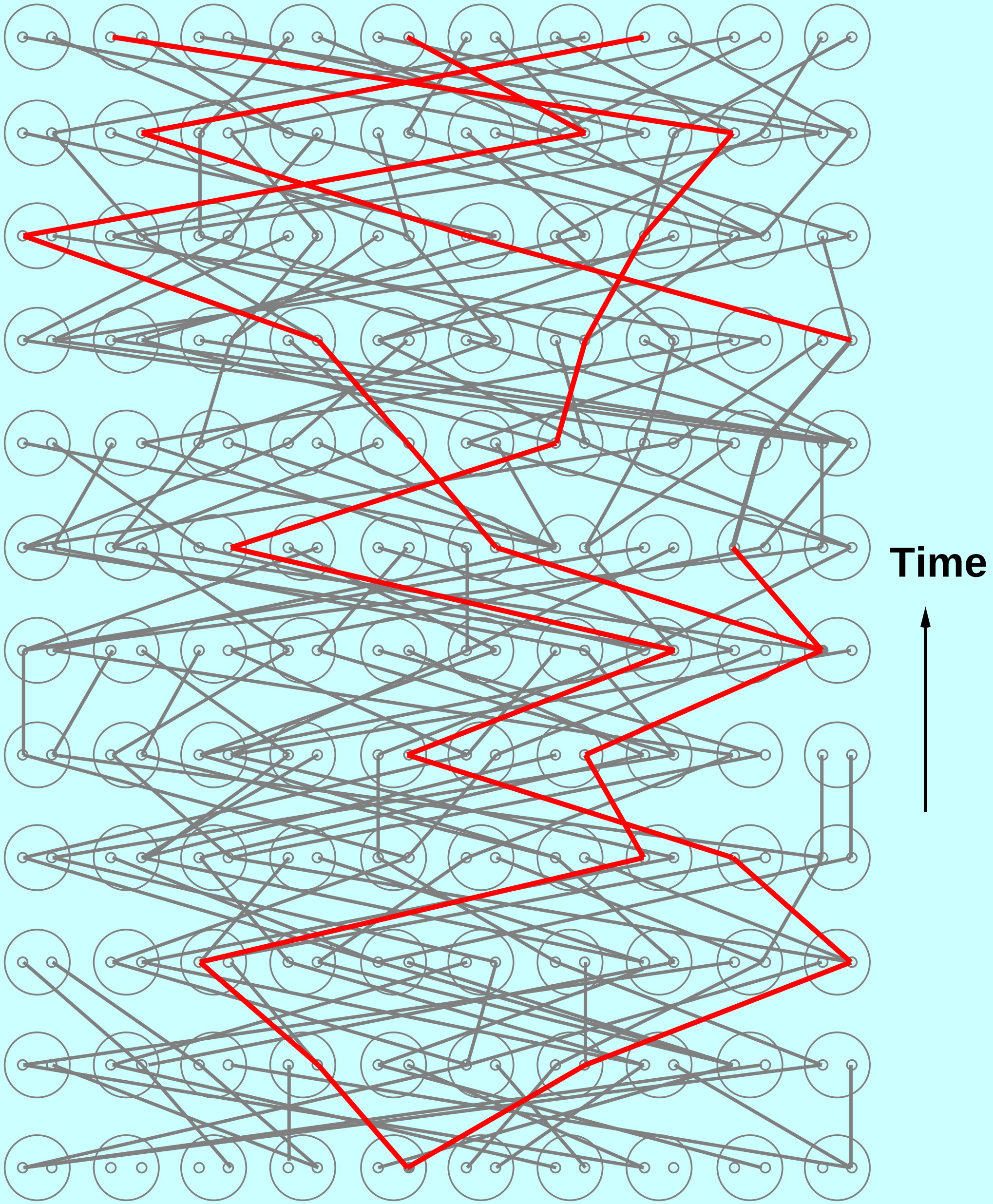


The ancestry of a sample of 3 genes

Genealogy of a small sample of genes from the population



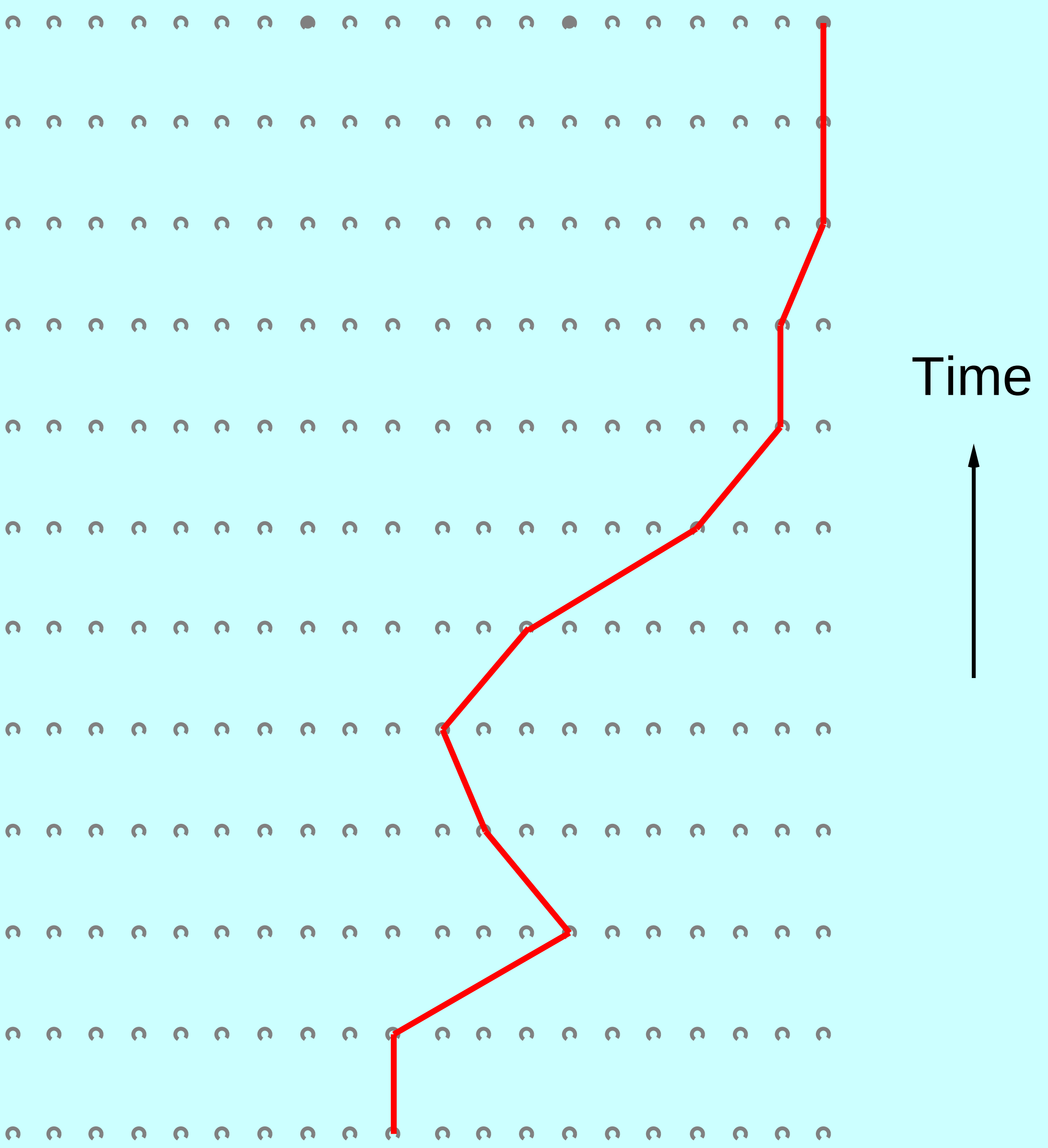
Where the tree of 3 copies is in the genealogy



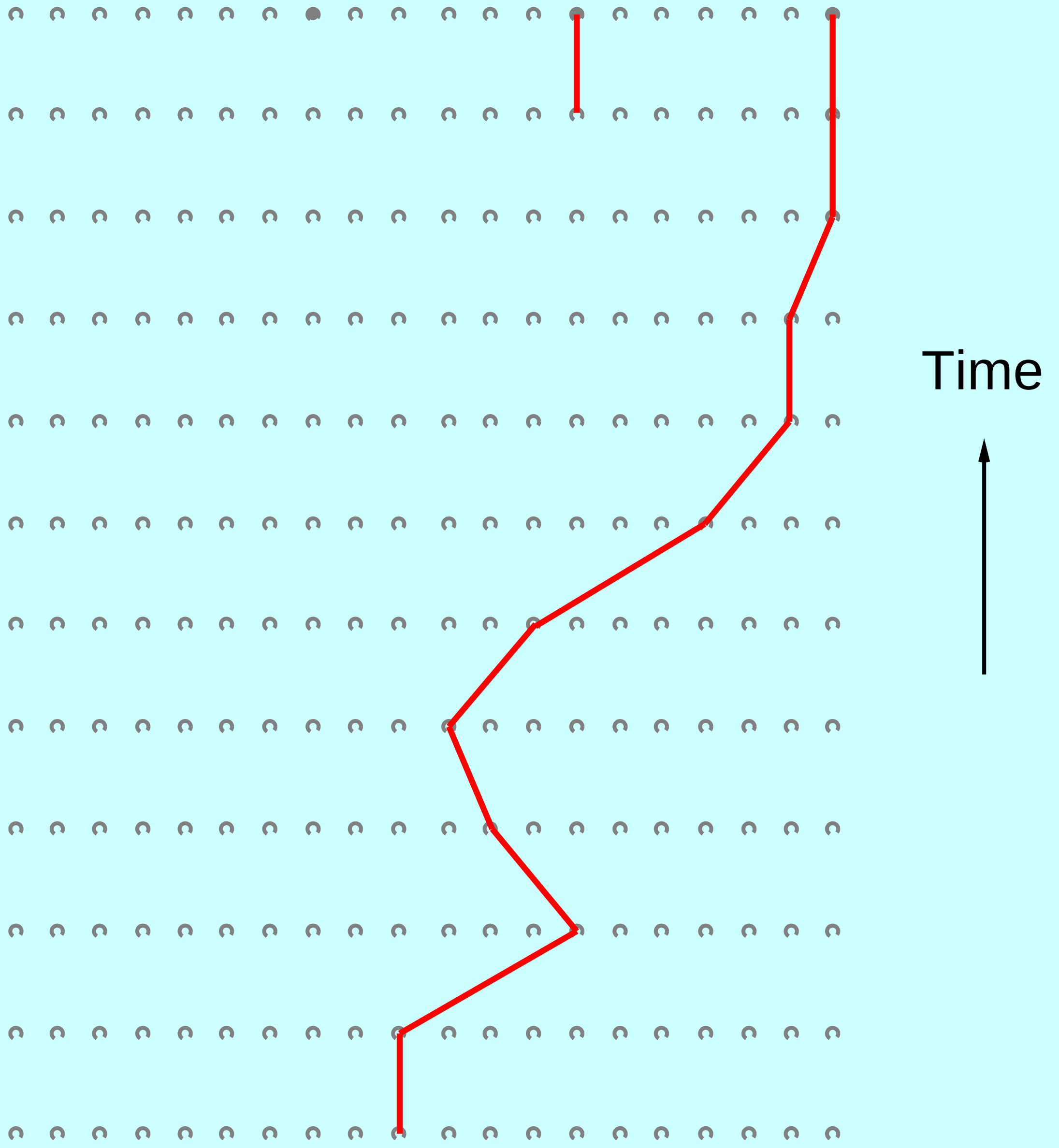
J. F. C. Kingman, about 1980

Invented the coalescent process, making the study of genealogies of samples from populations possible

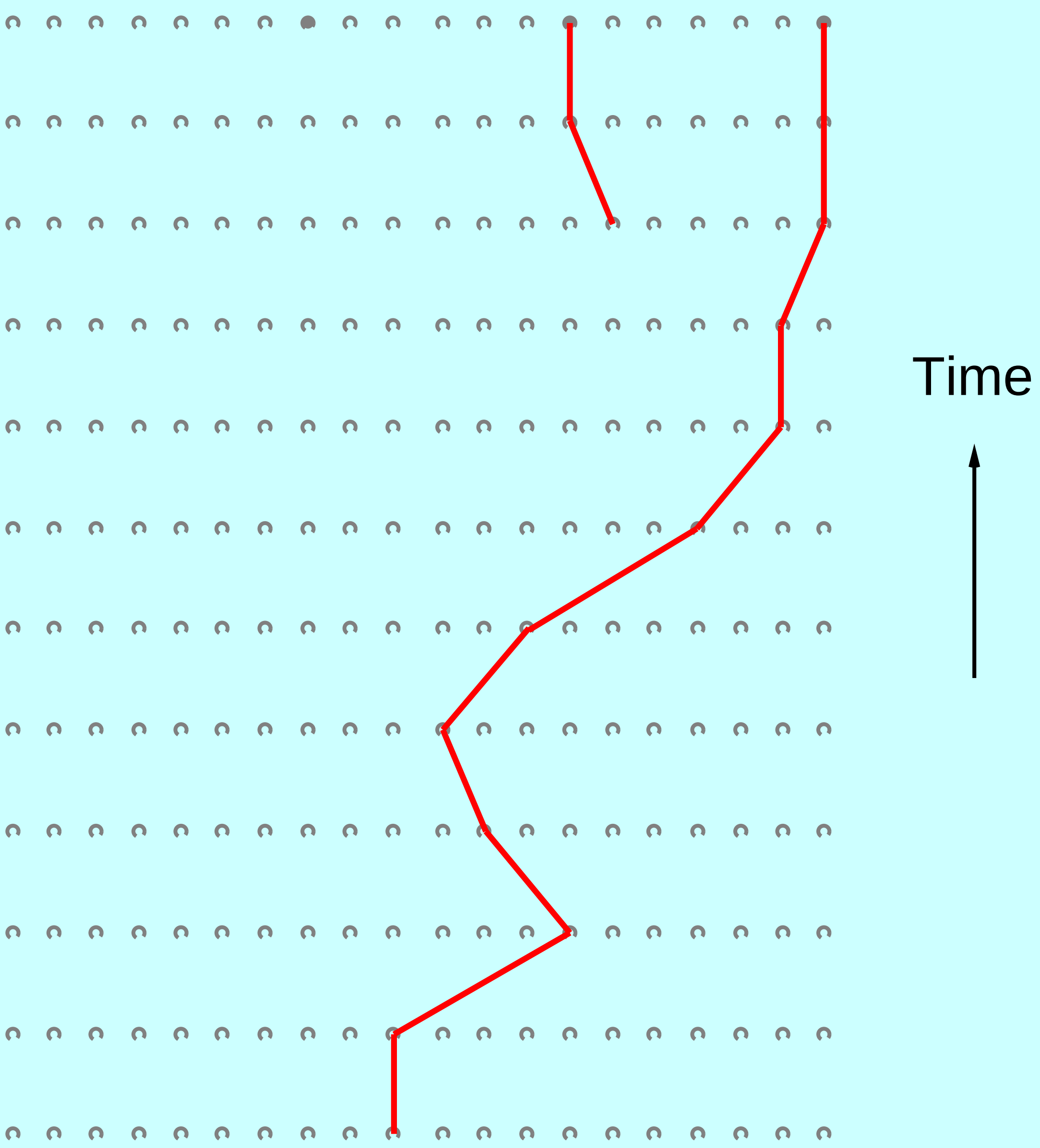
A pair of lineages going back in time



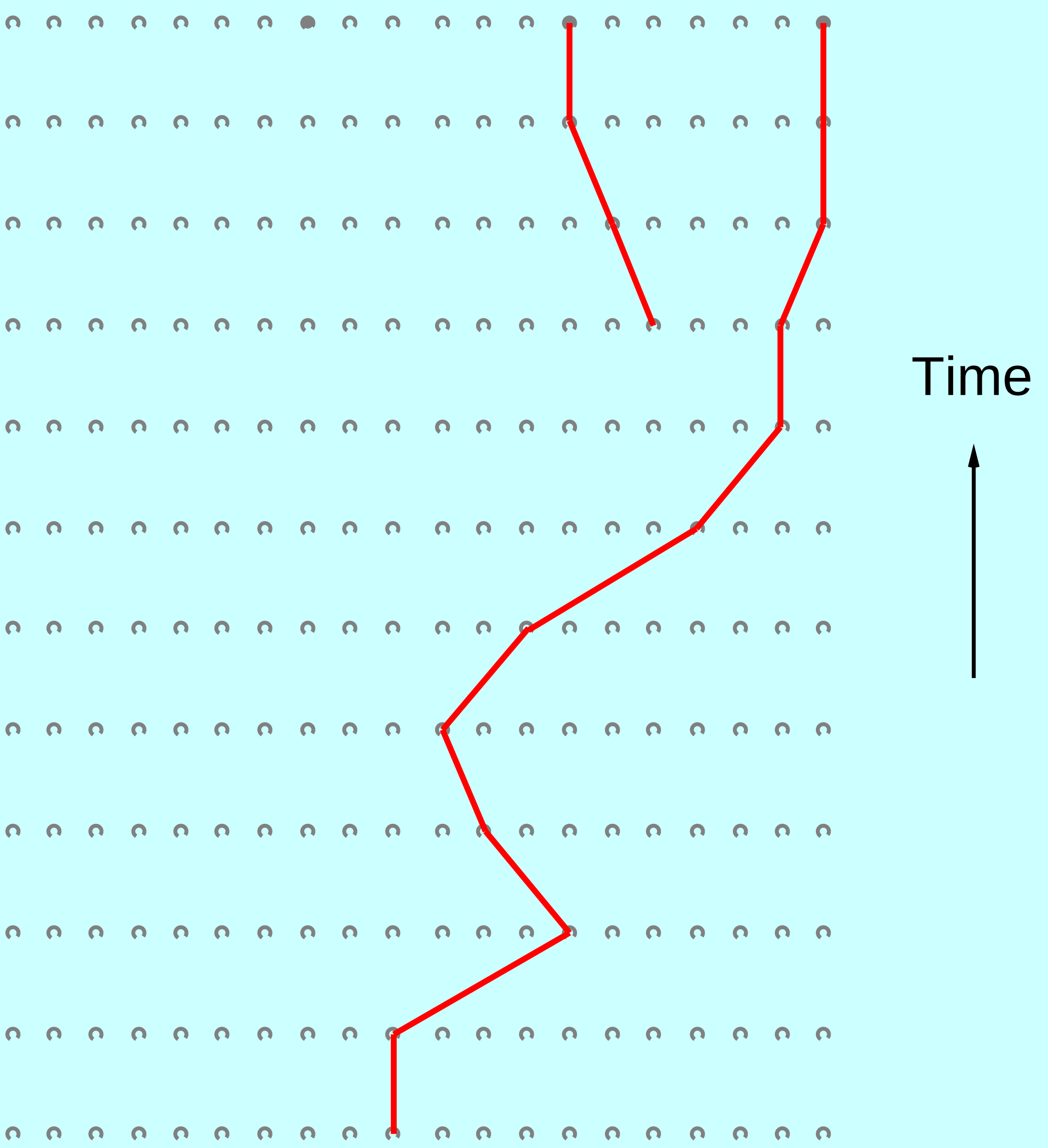
A pair of lineages going back in time



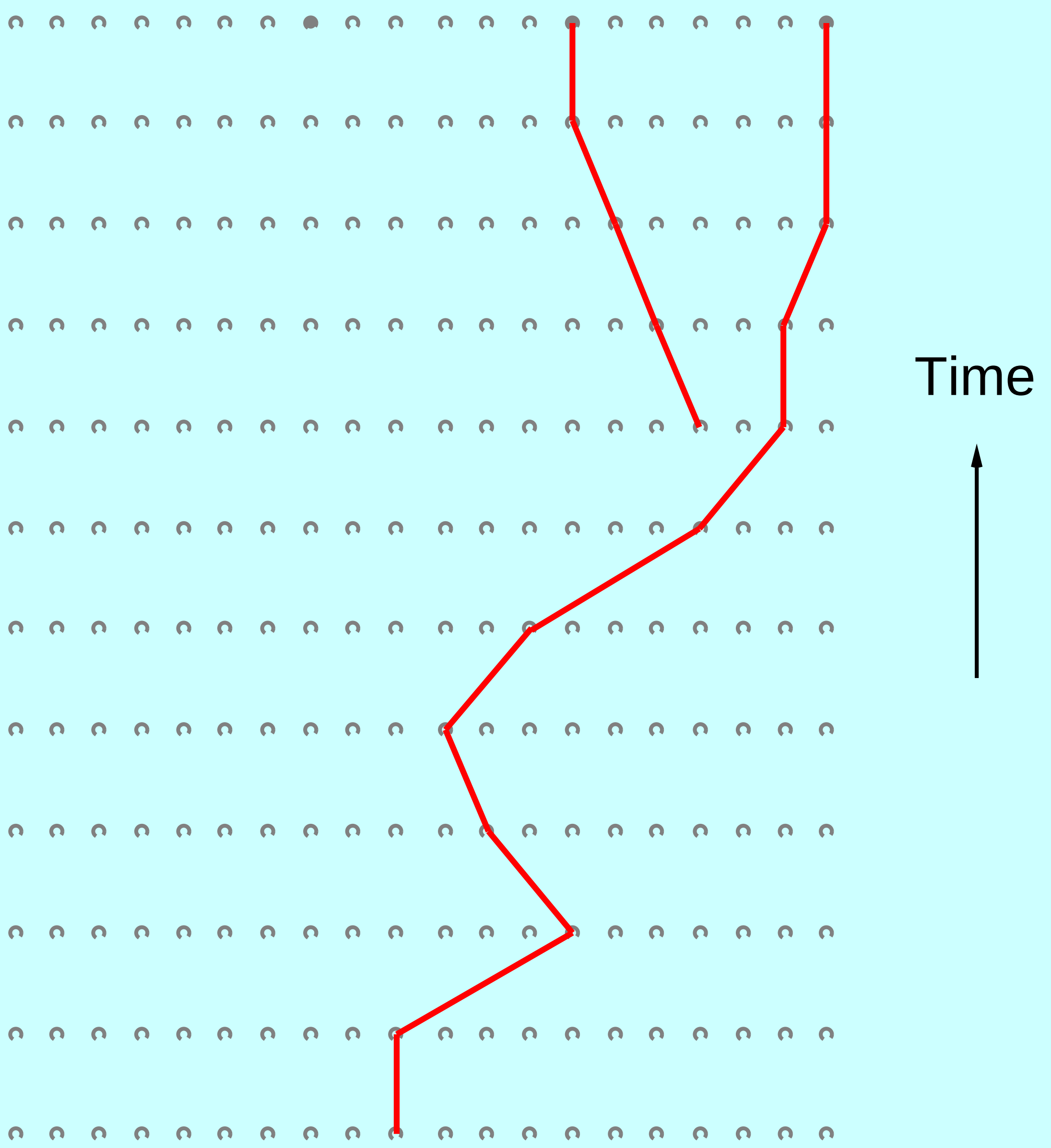
A pair of lineages going back in time



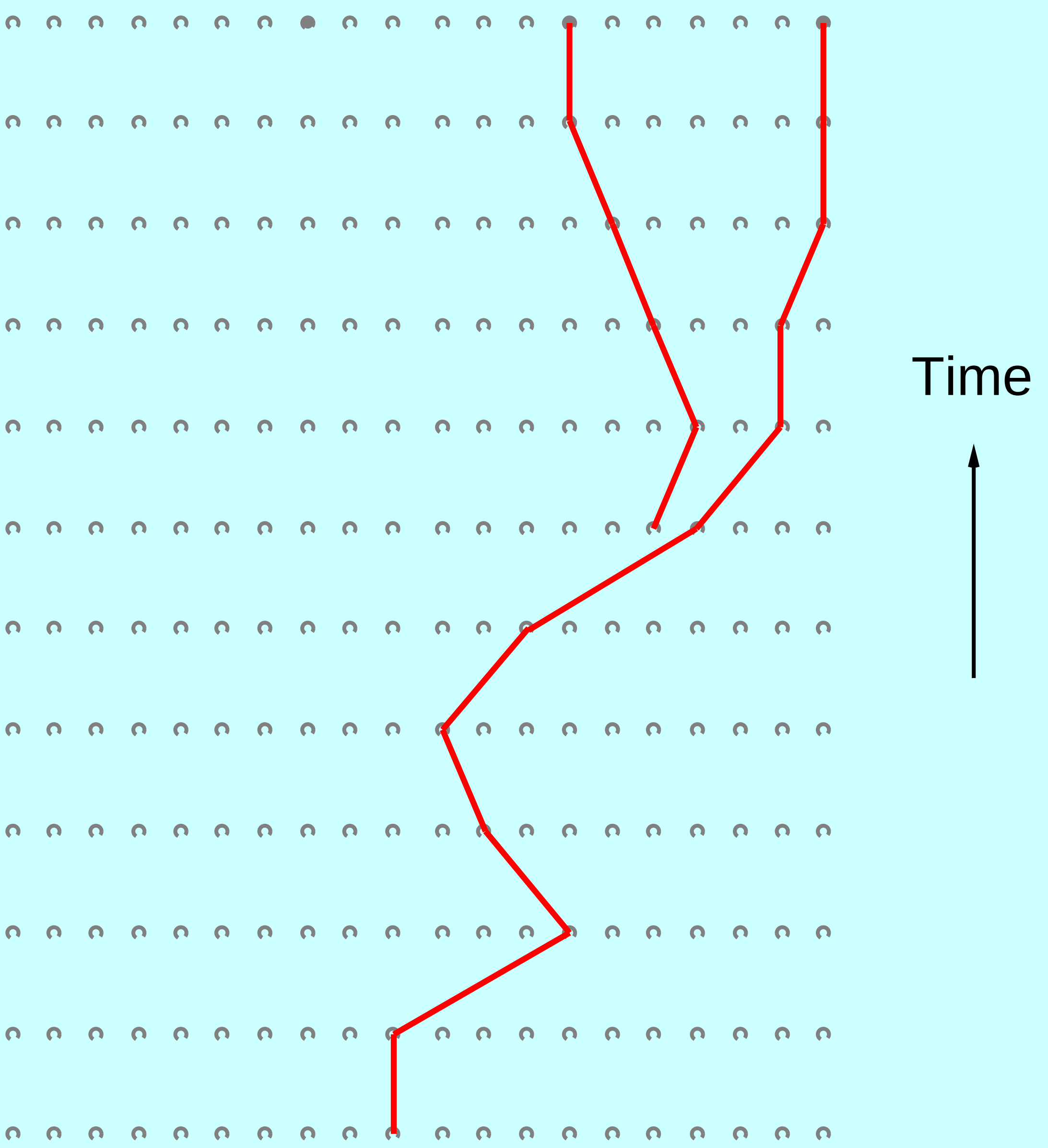
Each generation there is a probability



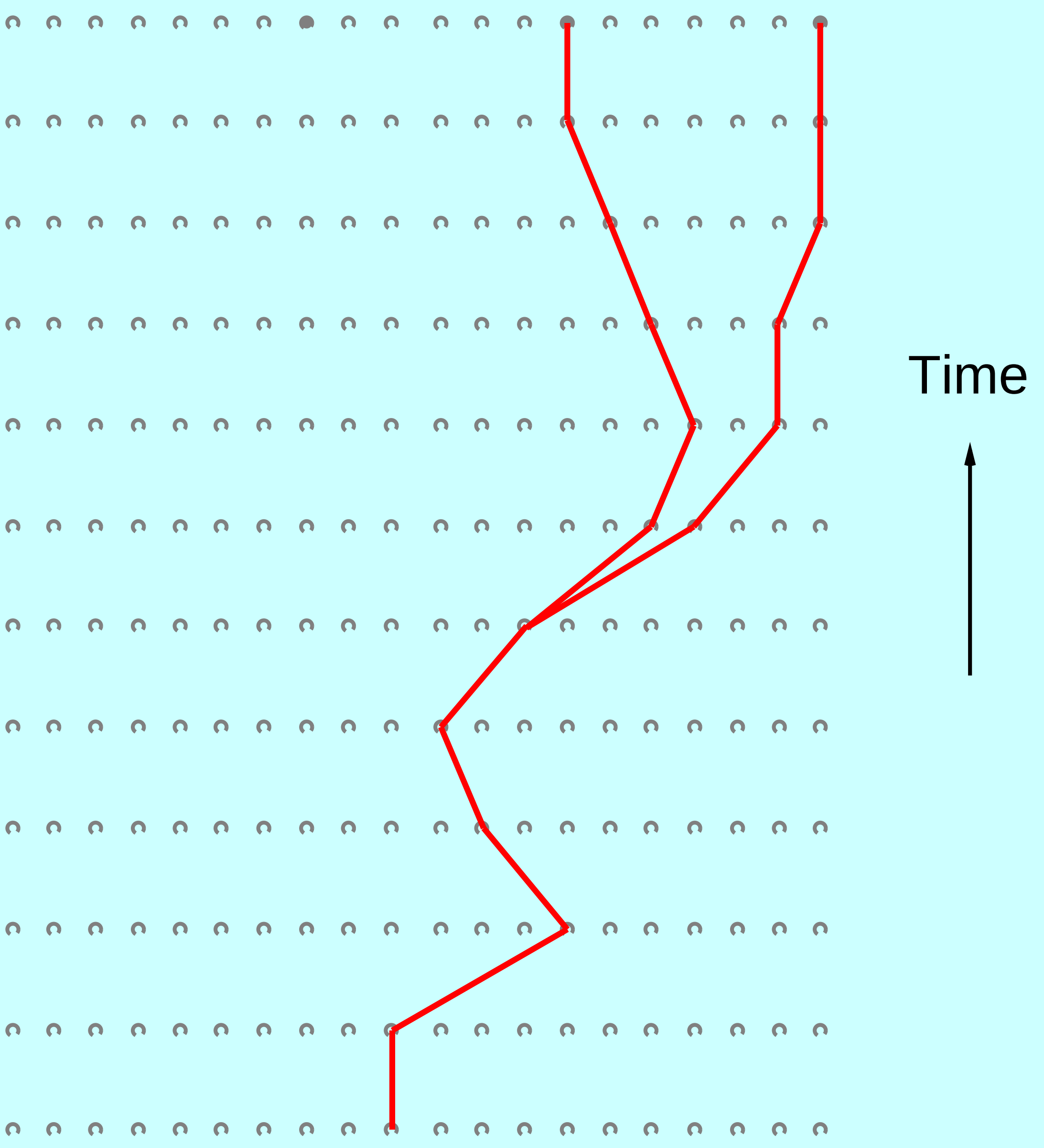
... of 1 in 20 that they will collide



... and if we toss enough times ...



... they will finally collide



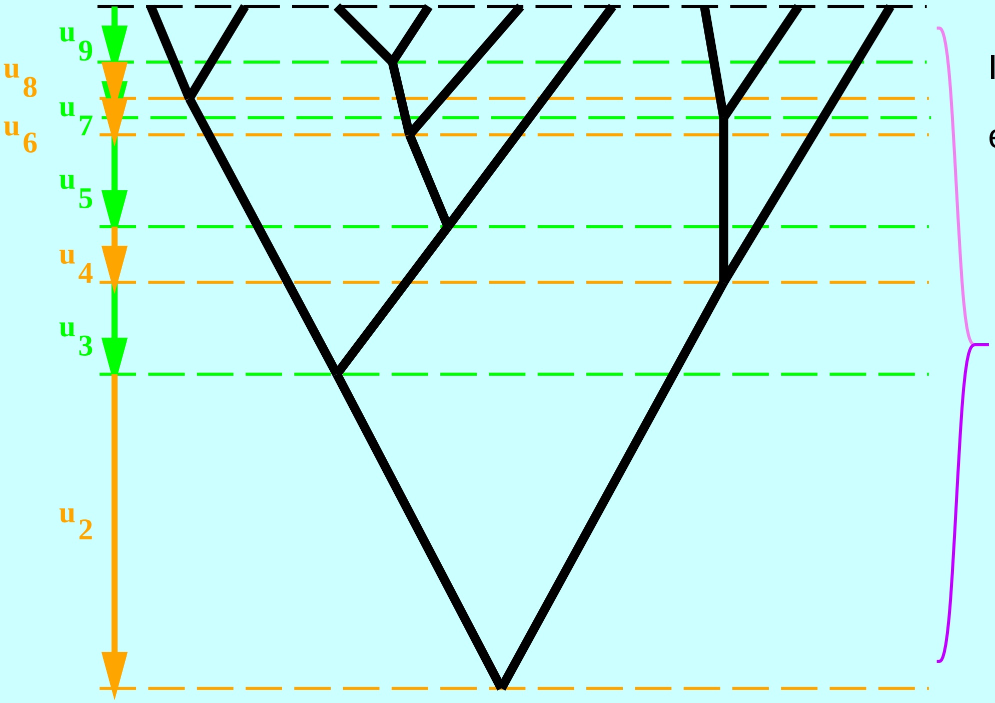
Kingman’s coalescent process

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

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



In a diploid population of effective population size N,

Average time for n copies to coalesce

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

Average time for two copies to coalesce

$= 2N$ generations

The coalescent – a derivation

The probability that k lineages becomes $k-1$ one generation earlier is (as each lineage “chooses" its ancestor independently):

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

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

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

so that the total probability that a pair coalesces is

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

probability that someone coalesces

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

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

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

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

and since

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

(continued)

the quantity

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

showing that the events involving 3 or more lineages simultaneously coalescing are in the terms of order $1/N^2$ and thus become unimportant if N is large. For example, when $k = 10$ and $N = 100$, there is a 0.7956 chance that there is no coalescence, 0.1874 that one pair coalesces, and only 0.01695 that more than one coalesces.

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.

How far back to the common ancestor?

Adding up the expectations of the $n-1$ coalescent events this is

$$\frac{4N_e}{n(n-1)} + \frac{4N_e}{(n-1)(n-2)} + \ldots + \frac{4N_e}{2(1)}$$

But since

$$\frac{1}{n(n-1)} = \frac{1}{n-1} - \frac{1}{n}$$

This is then

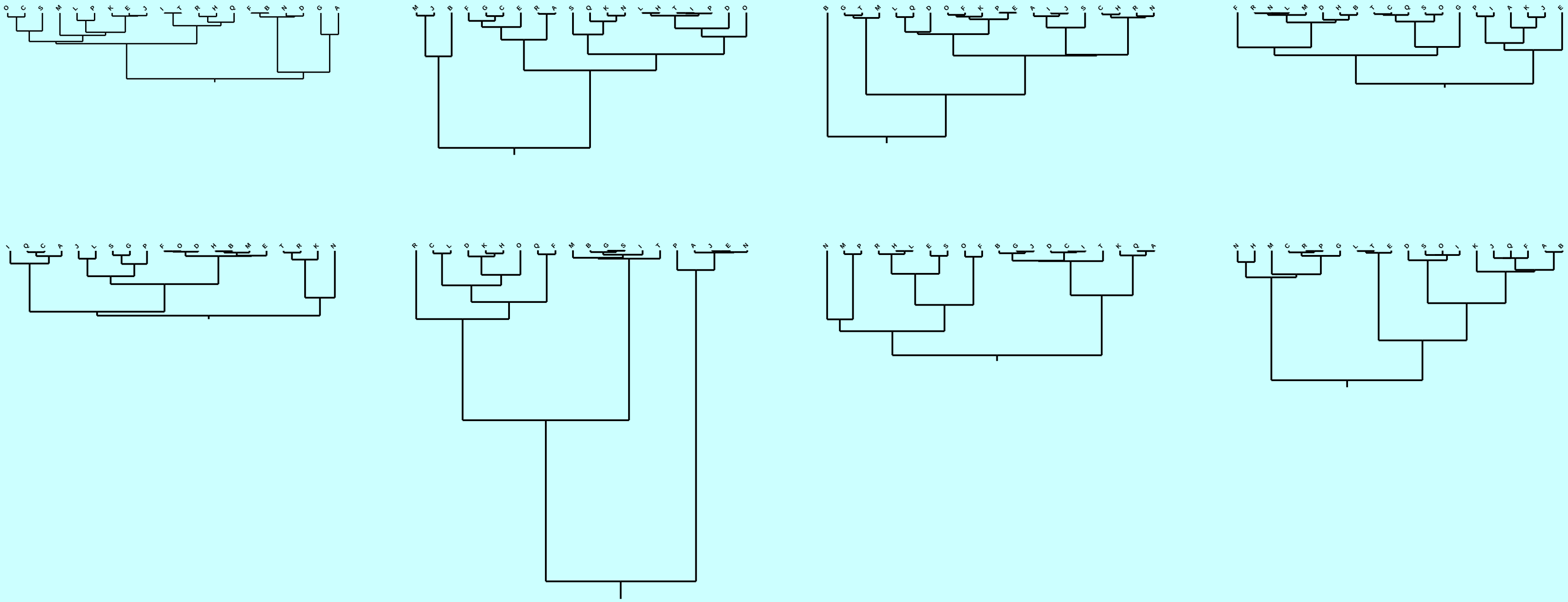
$$4N_e\left[\frac{1}{n-1} - \frac{1}{n} + \frac{1}{n-2} - \frac{1}{n-1} + \frac{1}{n-3} - \frac{1}{n-2} + \ldots + \frac{1}{1} - \frac{1}{2}\right]$$

$$= \qquad 4N_e\left(1 - \frac{1}{n}\right)$$

An important property of coalescents

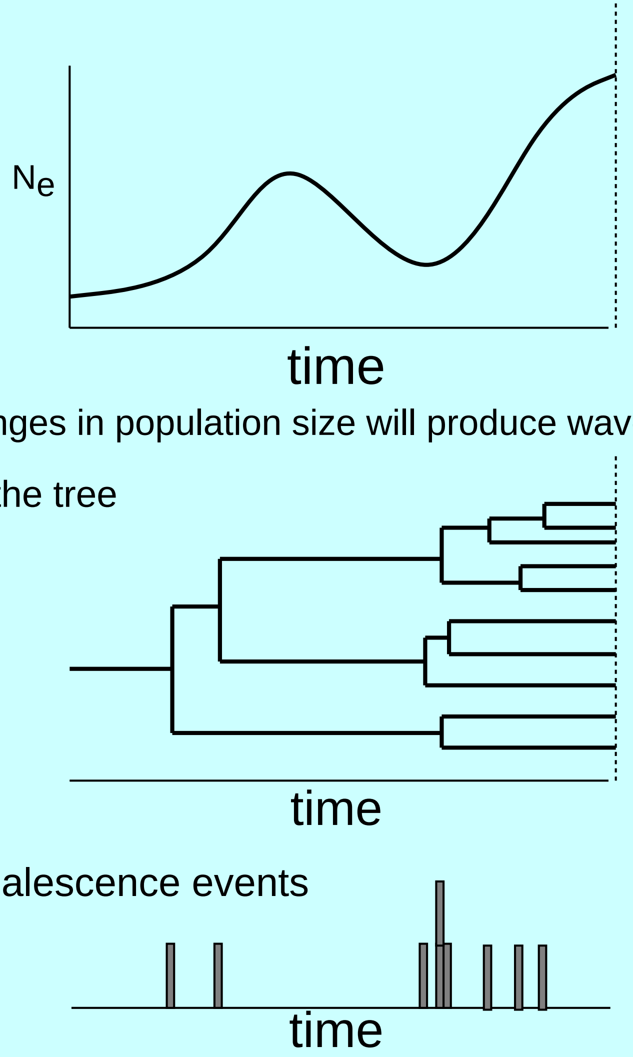
You can sample a genealogy of ancestry of a sample of genes without bothering to reconstruct the ancestry of any other copies.

Random coalescent trees with 16 lineages



Effect of varying population size

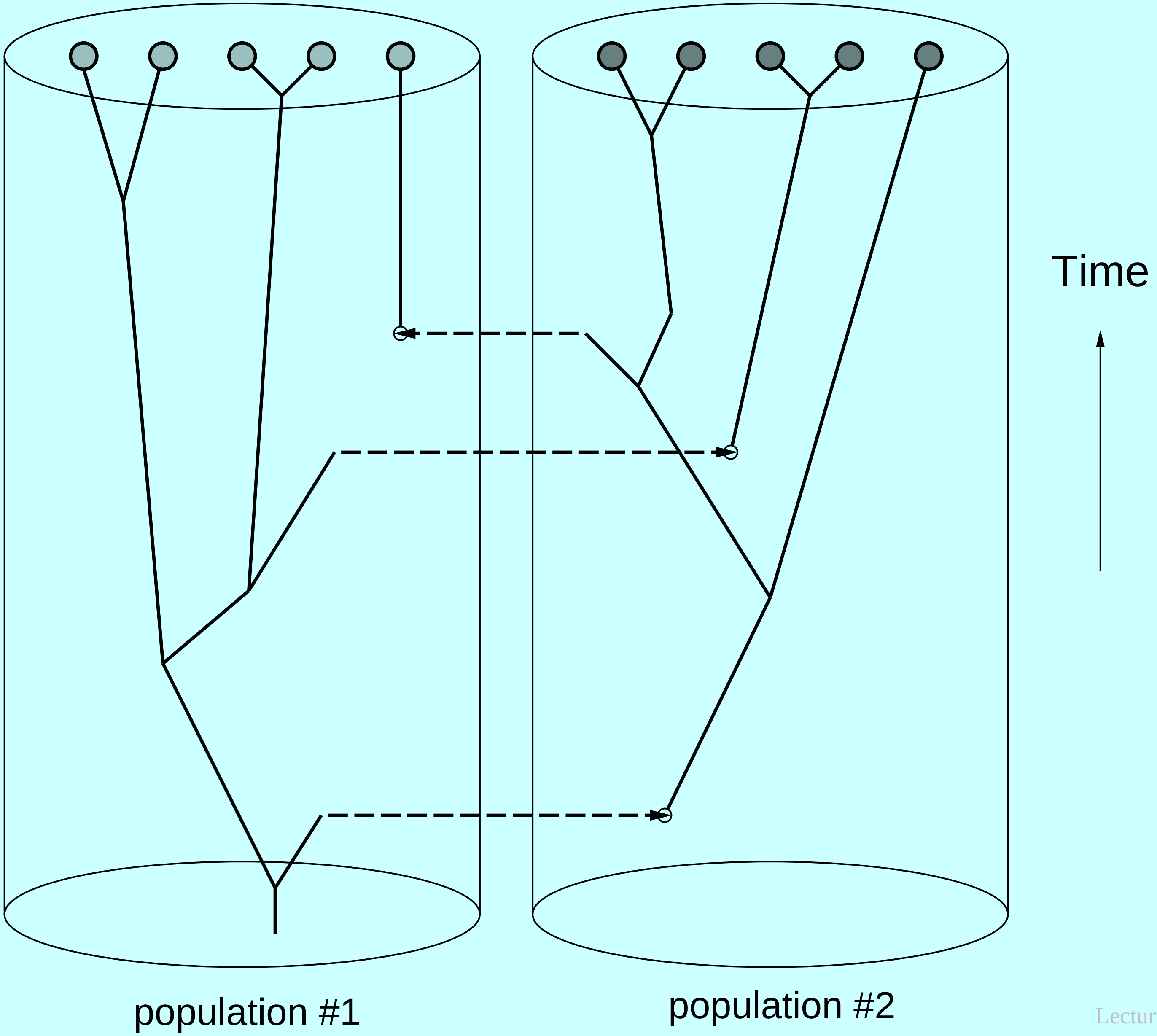
Change of population size and coalescents



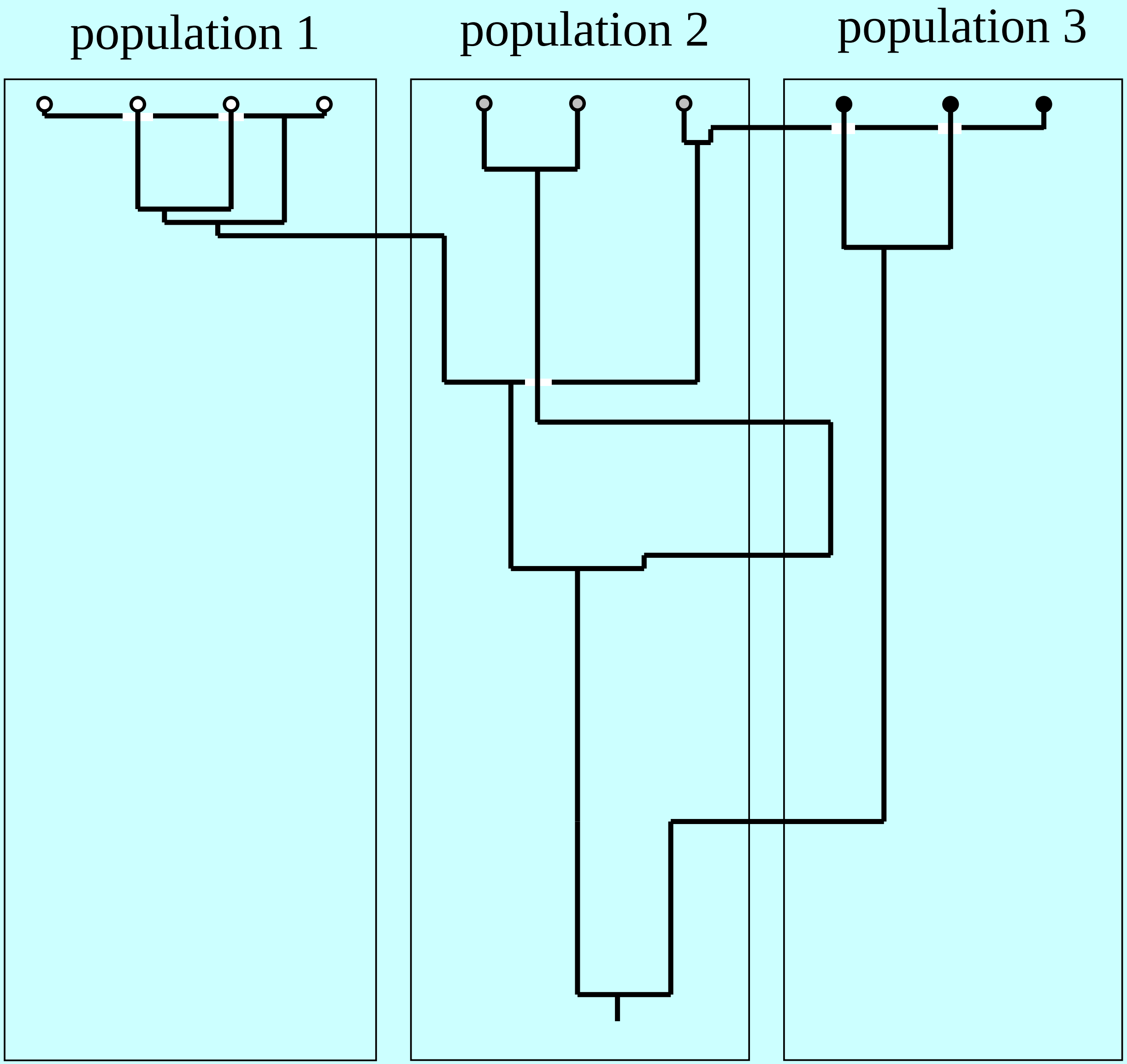
the changes in population size will produce waves of coalescence

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.

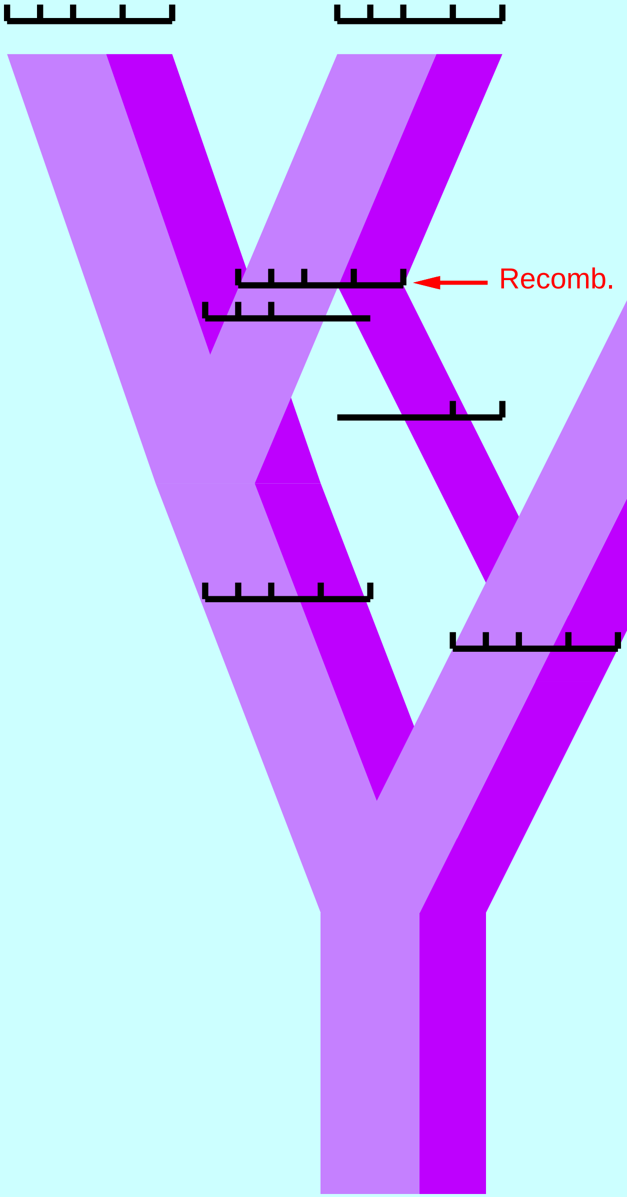
A coalescent with migration (2 populations)



Migration (3 populations) with $4Nm = 1$

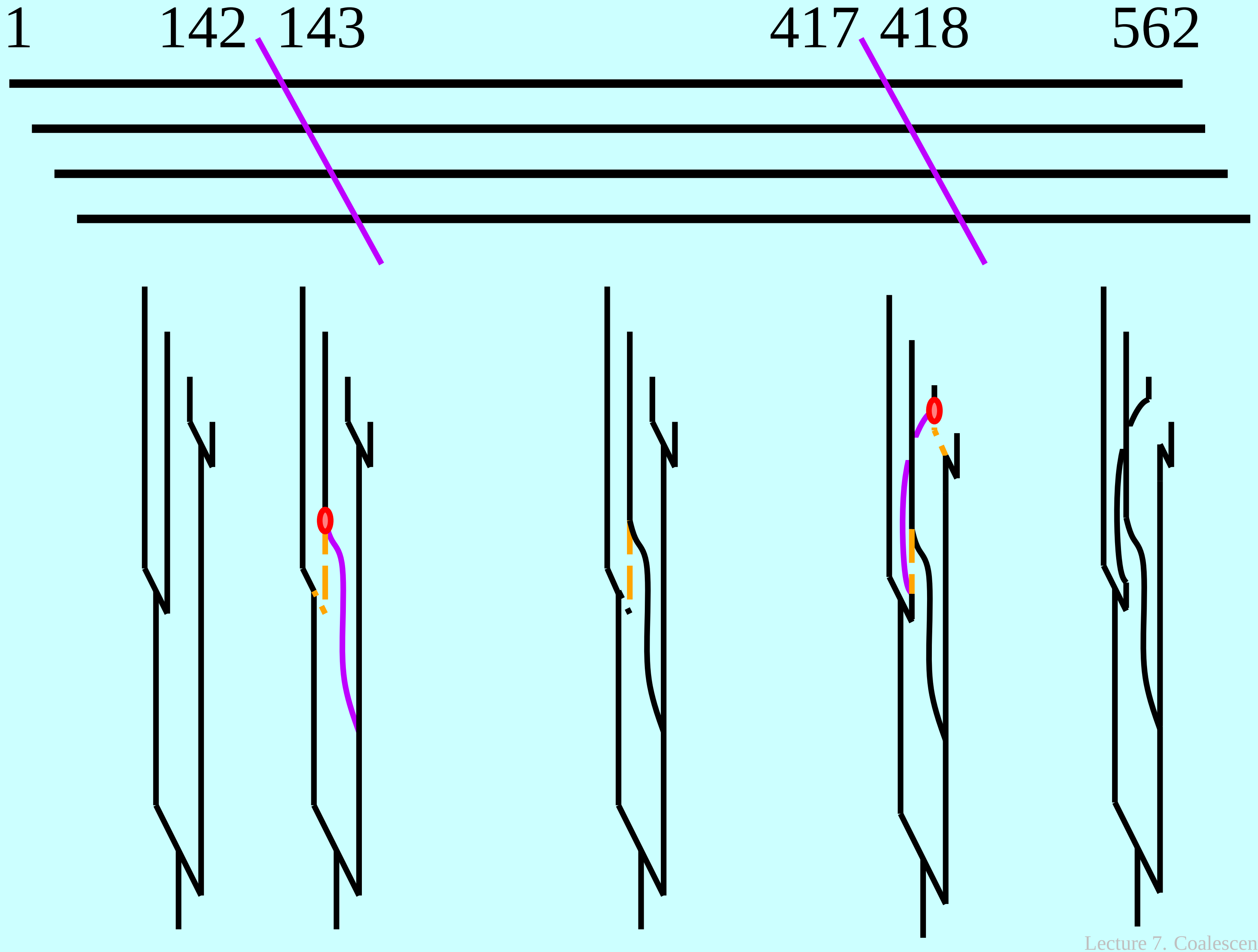


A recombining coalescent



Different markers have slightly different coalescent trees

Trees changing by recombination along a genome



How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.
- We want to find how far along the genome to go so that $4N_e\,r = 1$

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.
- We want to find how far along the genome to go so that $4N_e\,r = 1$
- ... or $40000r = 1$

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.
- We want to find how far along the genome to go so that $4N_e\, r = 1$
- ... or $40000r = 1$
- That is $10^8/(4 \times 10^4) = 2500$ nucleotides.

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.
- We want to find how far along the genome to go so that $4N_e\, r = 1$
- ... or $40000r = 1$
- That is $10^8/(4 \times 10^4) = 2500$ nucleotides.
- If population sizes were bigger, it is less!

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.
- We want to find how far along the genome to go so that $4N_e\,r = 1$
- ... or $40000r = 1$
- That is $10^8/(4 \times 10^4) = 2500$ nucleotides.
- If population sizes were bigger, it is less!
- But this assumes evenly distributed recombinations. With “hot spots” the spacing is greater because you are most of the time in a “cold” region.

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

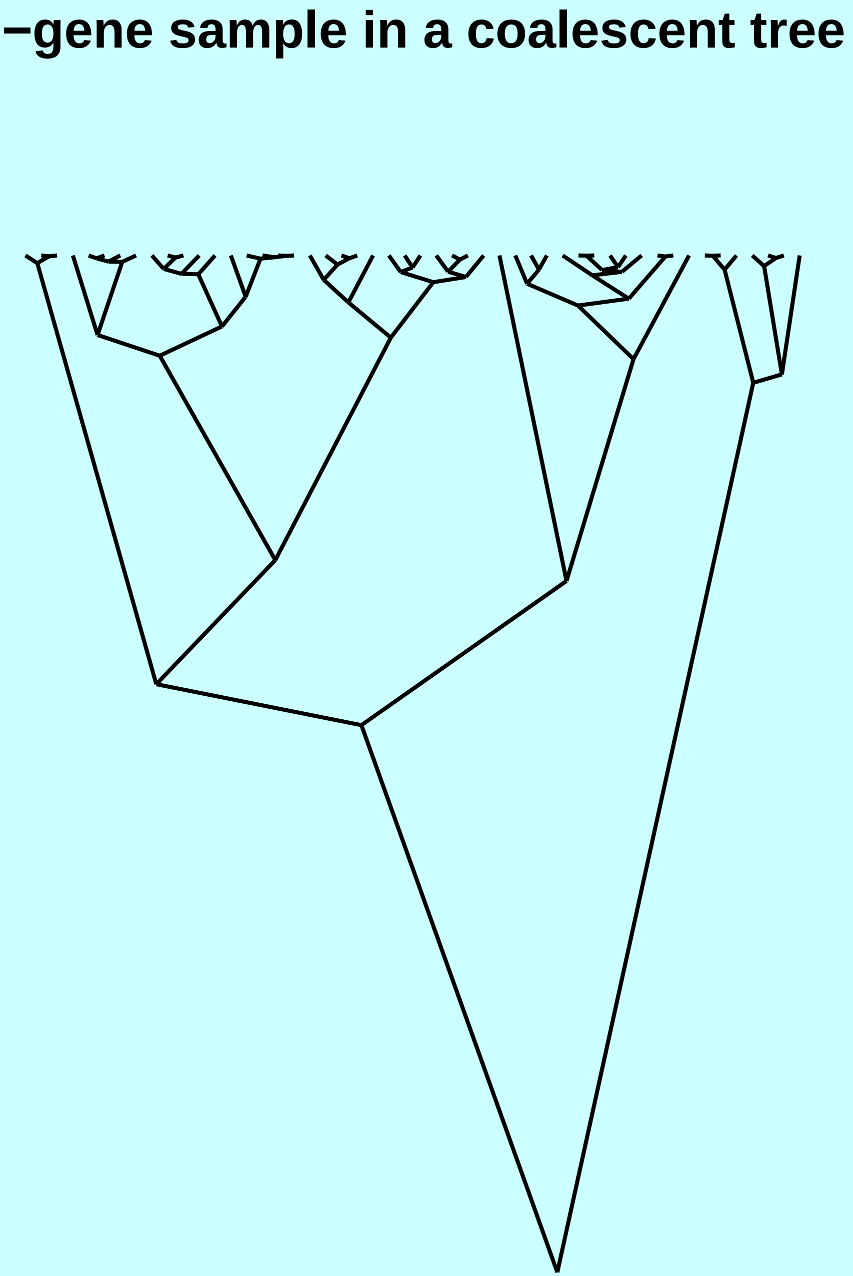
- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.
- We want to find how far along the genome to go so that $4N_e\,r = 1$
- ... or $40000r = 1$
- That is $10^8/(4 \times 10^4) = 2500$ nucleotides.
- If population sizes were bigger, it is less!
- But this assumes evenly distributed recombinations. With “hot spots” the spacing is greater because you are most of the time in a “cold” region.
- It is really the same thing as regions of linkage disequilibrium – maybe 50,000 bases long. This is no accident, they are actually the same phenomenon.

How far until the tree is substantially different?

Roughly, until a branch from the tip down to the root is expected to have one recombination, when markers are this far apart.

- The time back to the root is about $4N_e$ generations.
- In humans, for times long ago the effective population size was as low as $N_e = 10000$
- Humans have about one recombination every 10^8 nucleotides.
- We want to find how far along the genome to go so that $4N_e\, r = 1$
- ... or $40000r = 1$
- That is $10^8/(4 \times 10^4) = 2500$ nucleotides.
- If population sizes were bigger, it is less!
- But this assumes evenly distributed recombinations. With “hot spots” the spacing is greater because you are most of the time in a “cold” region.
- It is really the same thing as regions of linkage disequilibrium – maybe 50,000 bases long. This is no accident, they are actually the same phenomenon.
- Which means a sample from humans will have about $3.3 \times 10^9/50000 = 66000$ different trees!

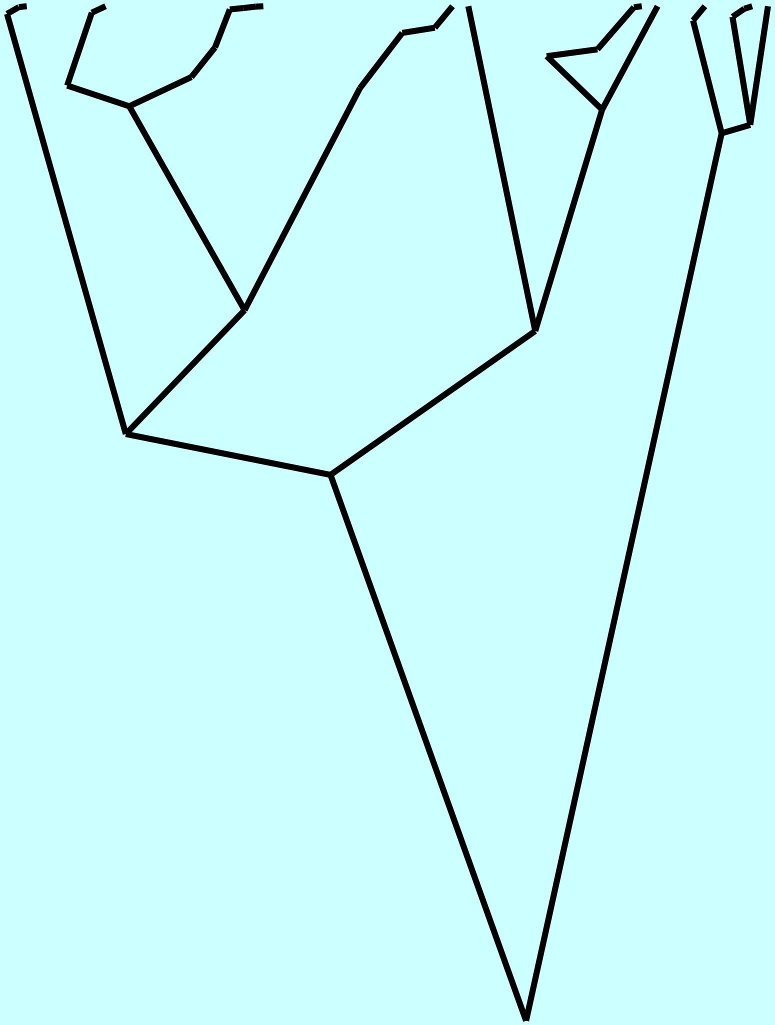
A coalescent of 50 copies



the first 10 copies only

10 genes sampled randomly out of a

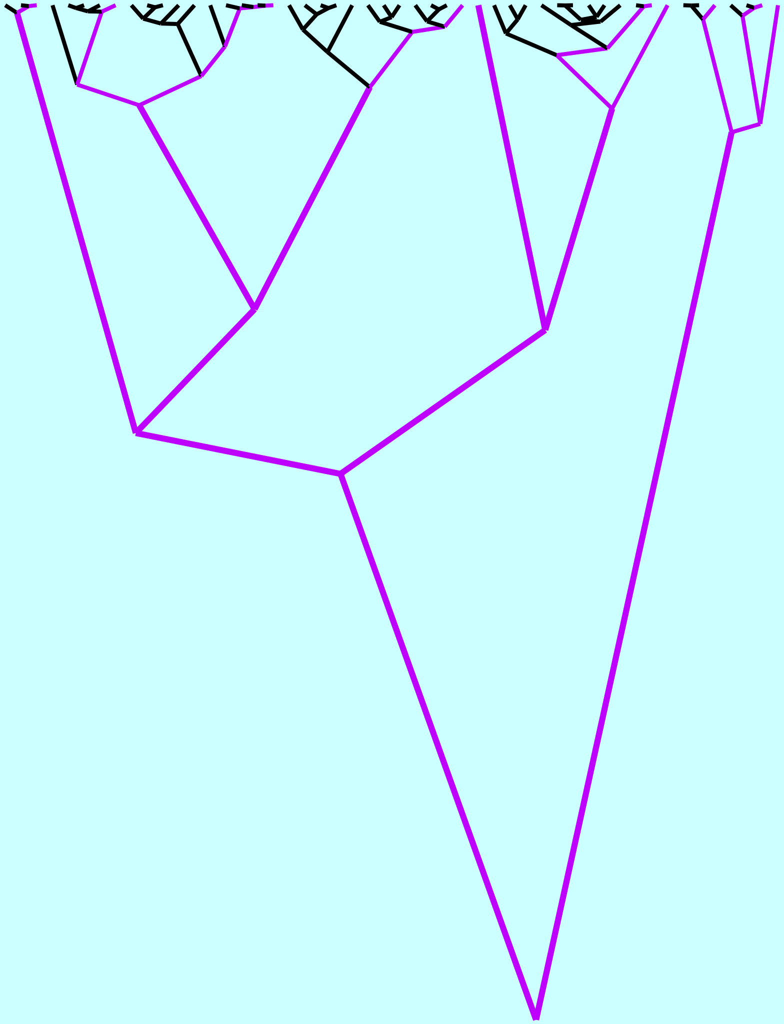
50−gene sample in a coalescent tree



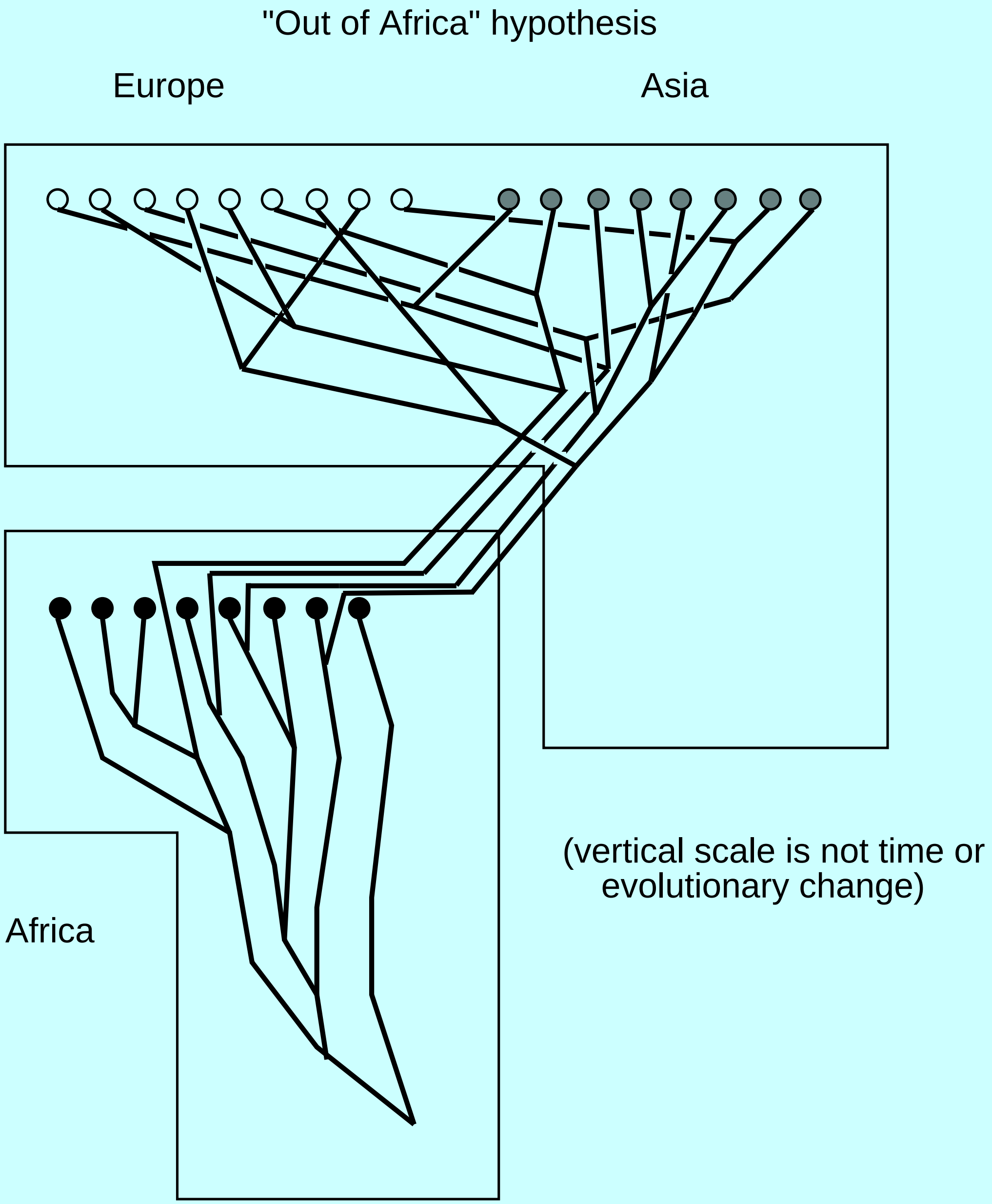
All copies, ancestry of first 10 in purple

10 genes sampled randomly out of a

50−gene sample in a coalescent tree

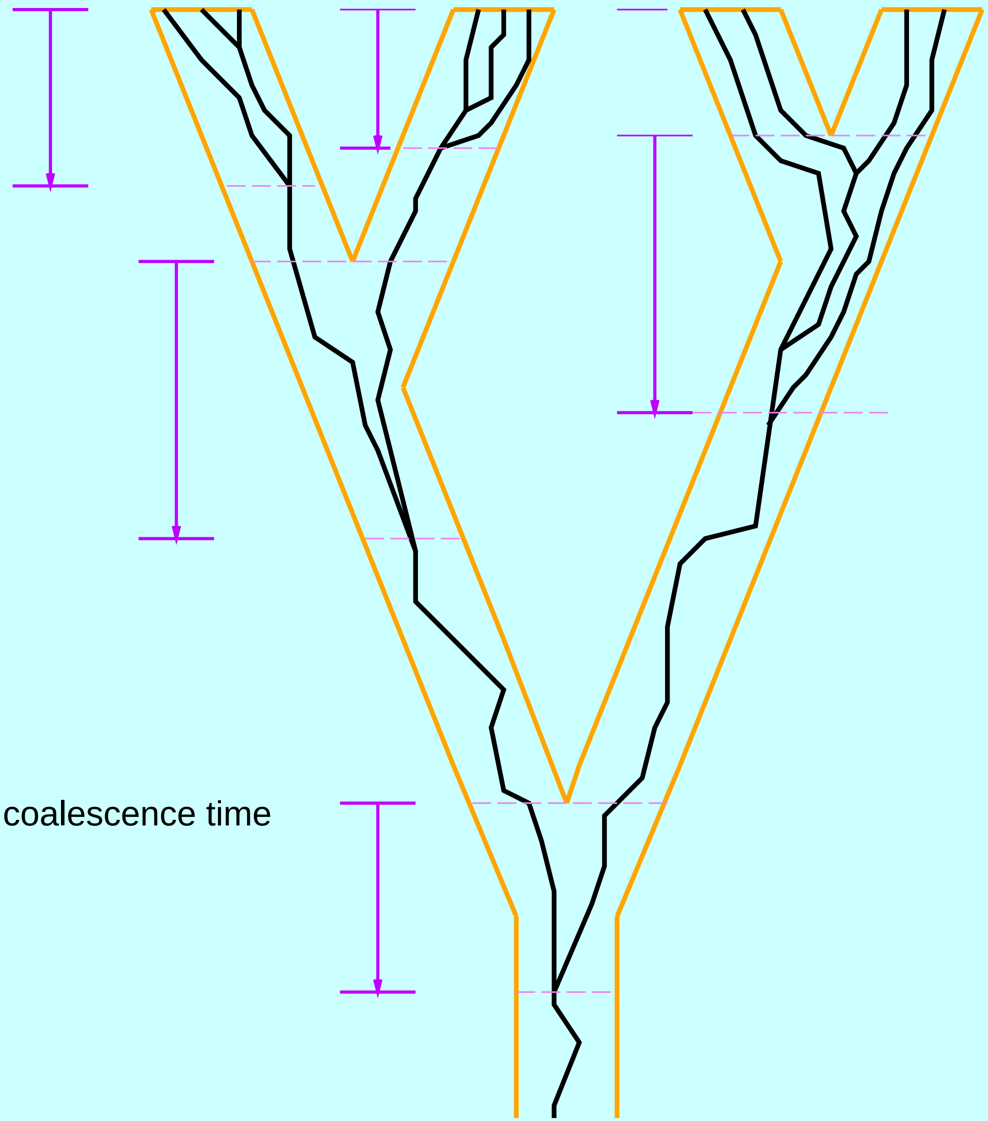


We ultimately want to treat this case



coalescents in related species

Consistency of gene tree with species tree



References

Kingman, J. F. C. 1982a. The coalescent. *Stochastic Processes and Their Applications* **13:** 235-248. [One of the papers in which the coalescent is introduced]

Kingman, J. F. C. 1982b. On the genealogy of large populations. *Journal of Applied Probability* **19A:** 27-43. [One of the other papers in which the coalescent is introduced]

Takahata, N. 1988. The coalescent in two partially isolated diffusion populations. *Genetical Research* **52:** 213-222. [Coalescents with migration]

Hudson, R. R. and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111:** 147-164. [Coalescent with recombination]

Felsenstein, J. 1971. The rate of loss of multiple alleles in finite haploid populations. *Theoretical Population Biology* **2:** 391-403. [Can be used to derive rates of coalescence]

more references

Hein, J., M. Schierup, and C. Wiuf. 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, Oxford. [One of two books so far on coalescents. Light on estimation issues.]

Wakeley, J. 2008. *Coalescent Theory*. Roberts and Co., Greenwood Village, Colorado. [One of two books so far on coalescents. Light on estimation issues.]

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts. [Chapters 26, 27, and 28 are a (very) good introduction to coalescents and inferences using them.]

Wakeley, J. 2008. *Coalescent Theory. An Introduction.* Roberts and Co. Publishers, Greenwood Village, Colorado. [A book on coalescents, comparable to Hein et al. Due in 2008, probably also light on estimation issues.]

Krone, S. M. and C. Neuhauser. 1997. Ancestral processes with selection. *Theoretical Population Biology* **51:** 210-237. [A very original extension of the coalescent to allow selection]

Neuhauser, C., and S. M. Krone. 1997. The genealogy of samples in models with selection. *Genetics* **145:** 519-534. [A very original extension of the coalescent to allow selection]