

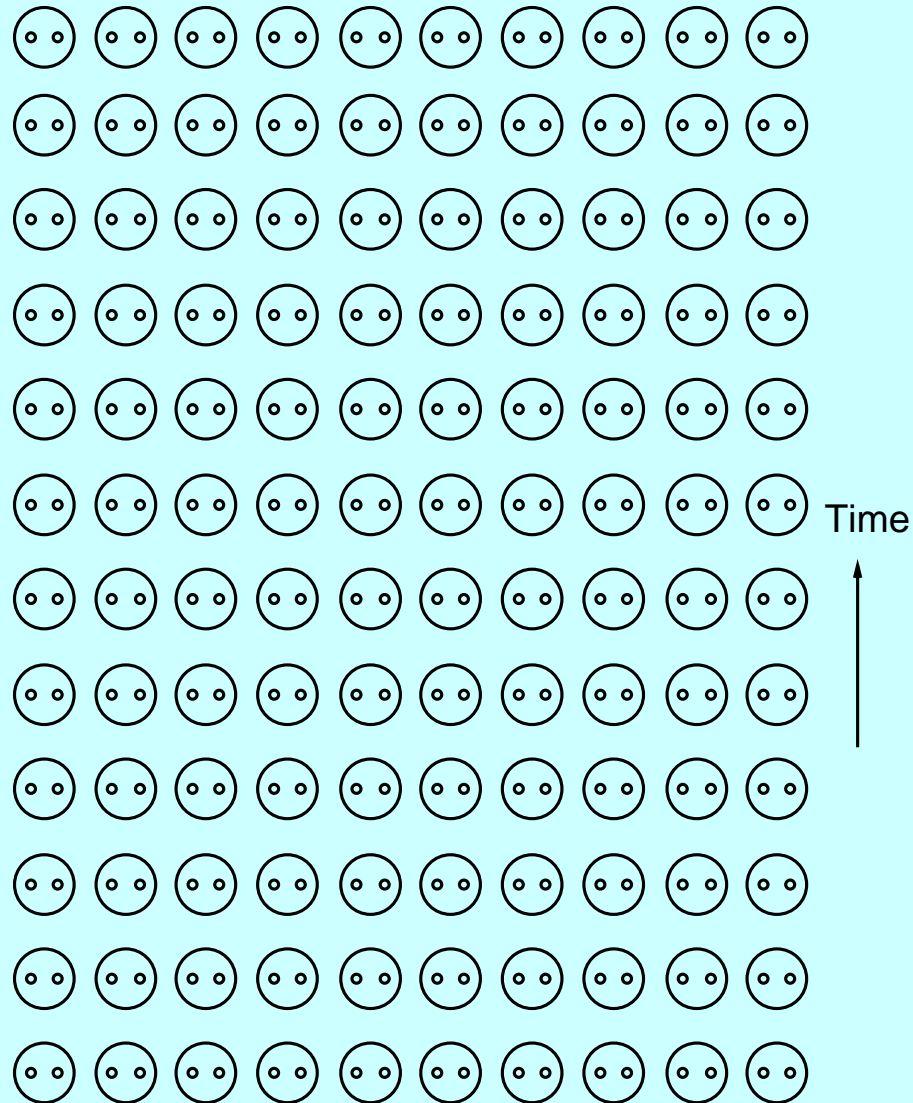
Week 9a: coalescents

Genome 562

March, 2017

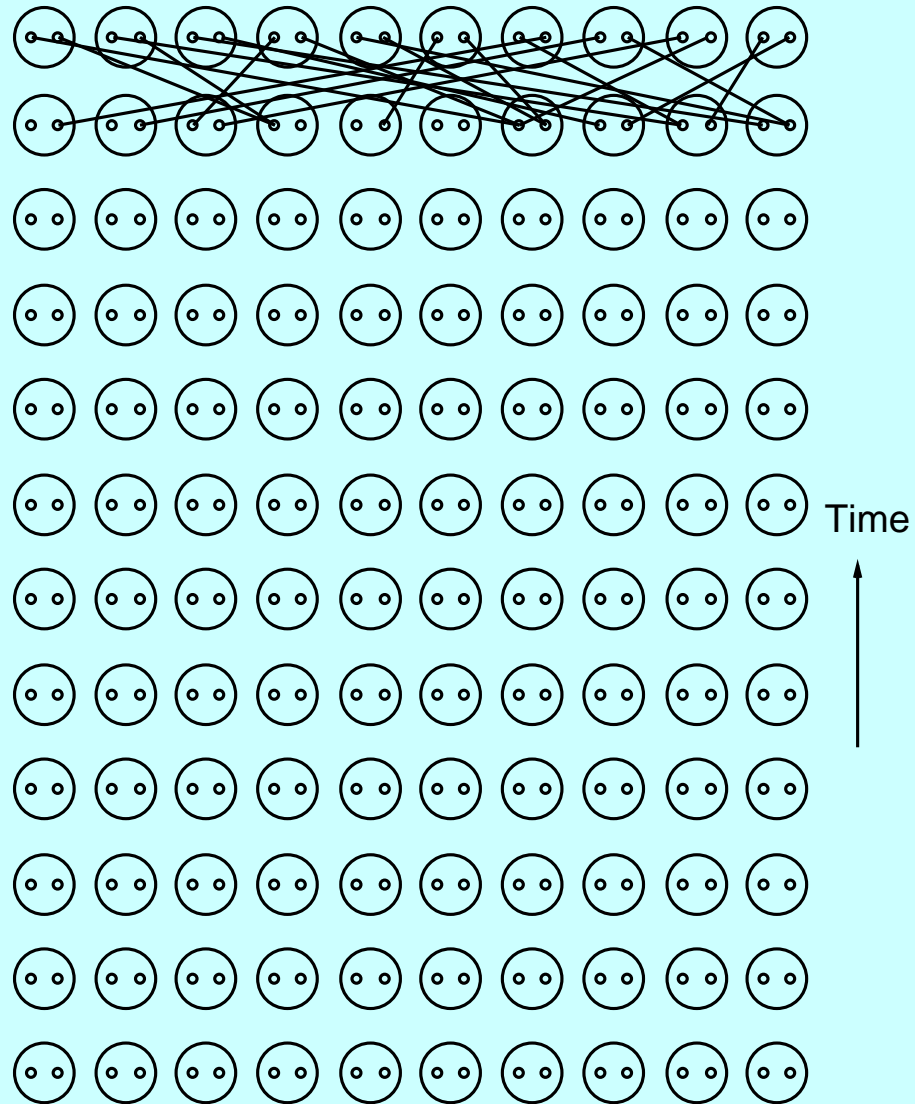
Gene copies in a population of 10 individuals

A random-mating population



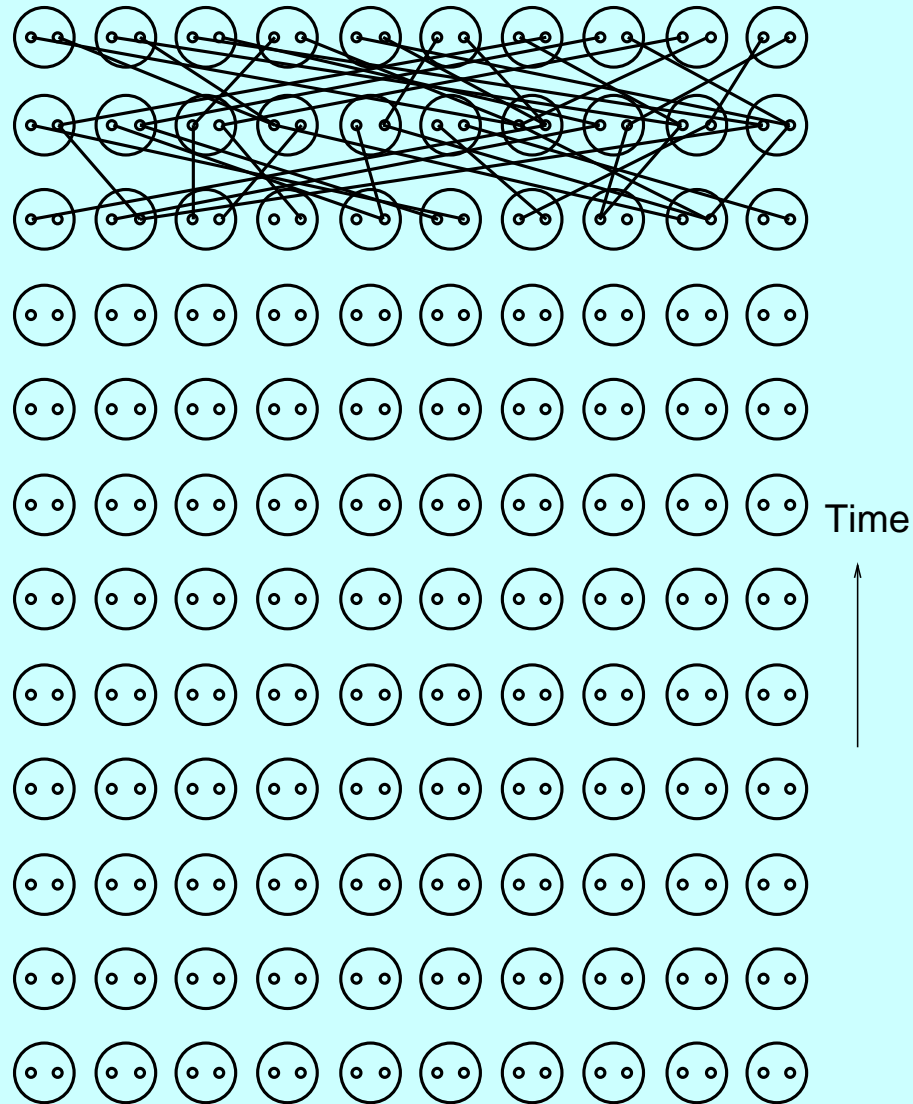
Going back one generation

A random-mating population



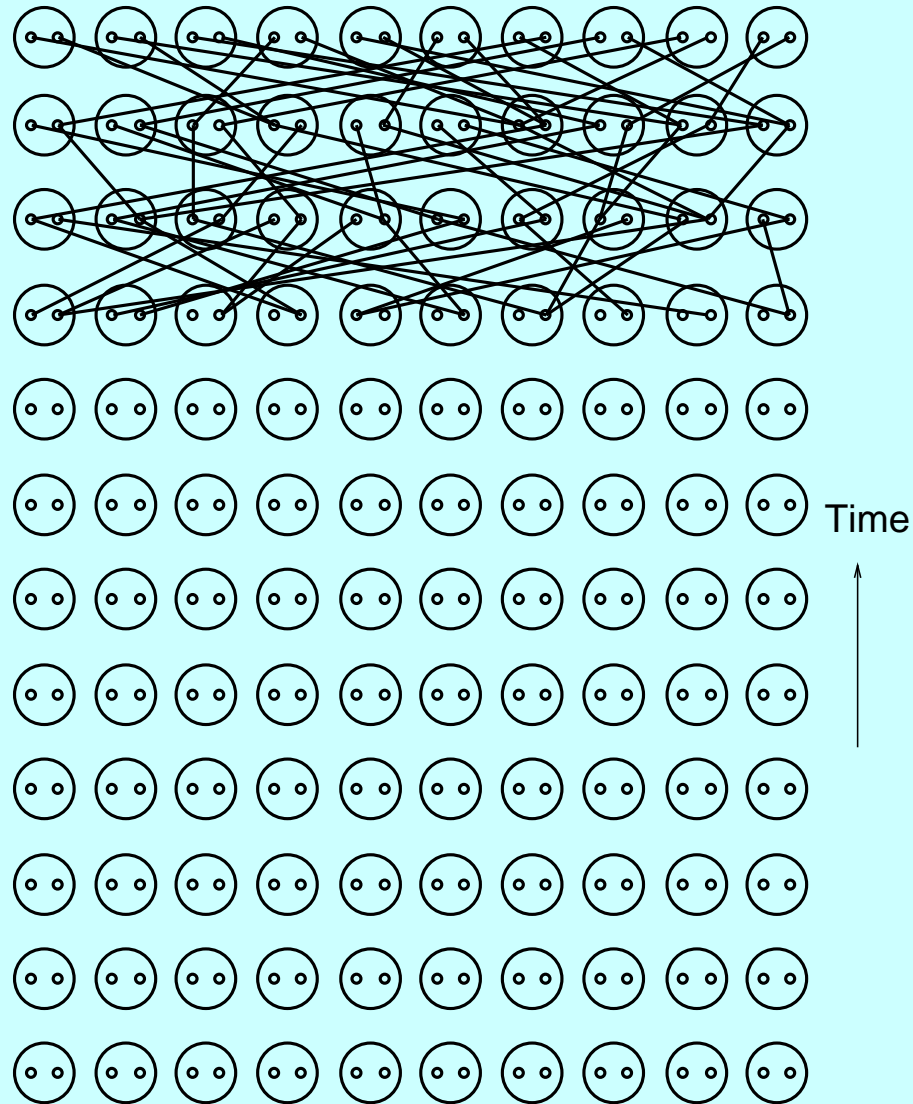
... and one more

A random-mating population



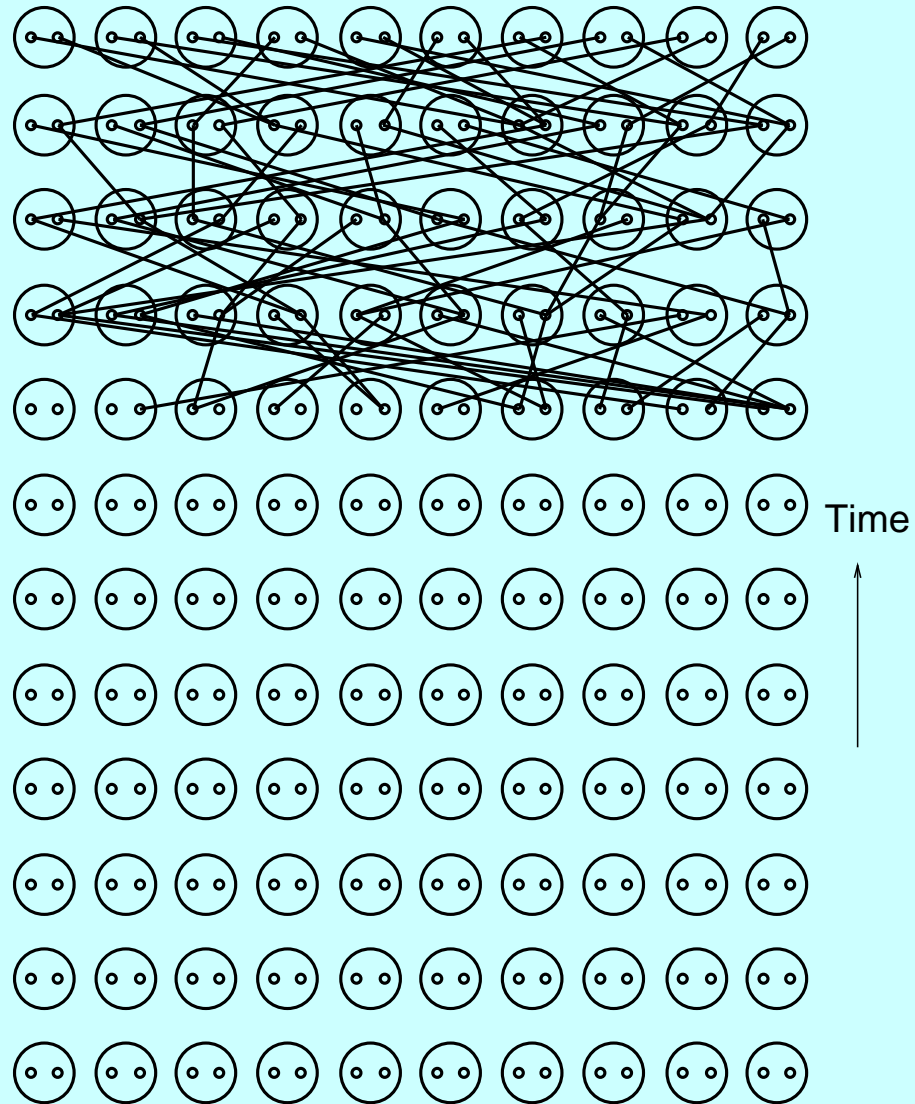
... and one more

A random-mating population



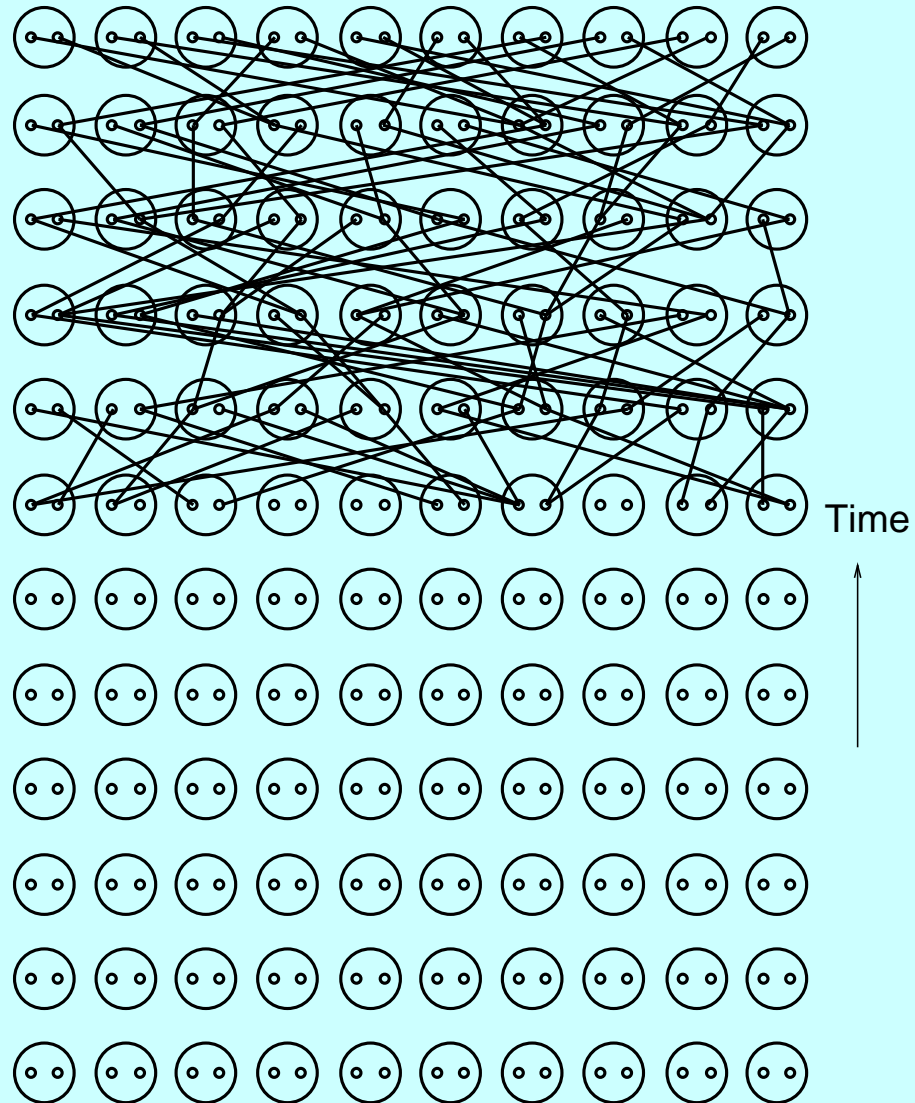
... and one more

A random-mating population



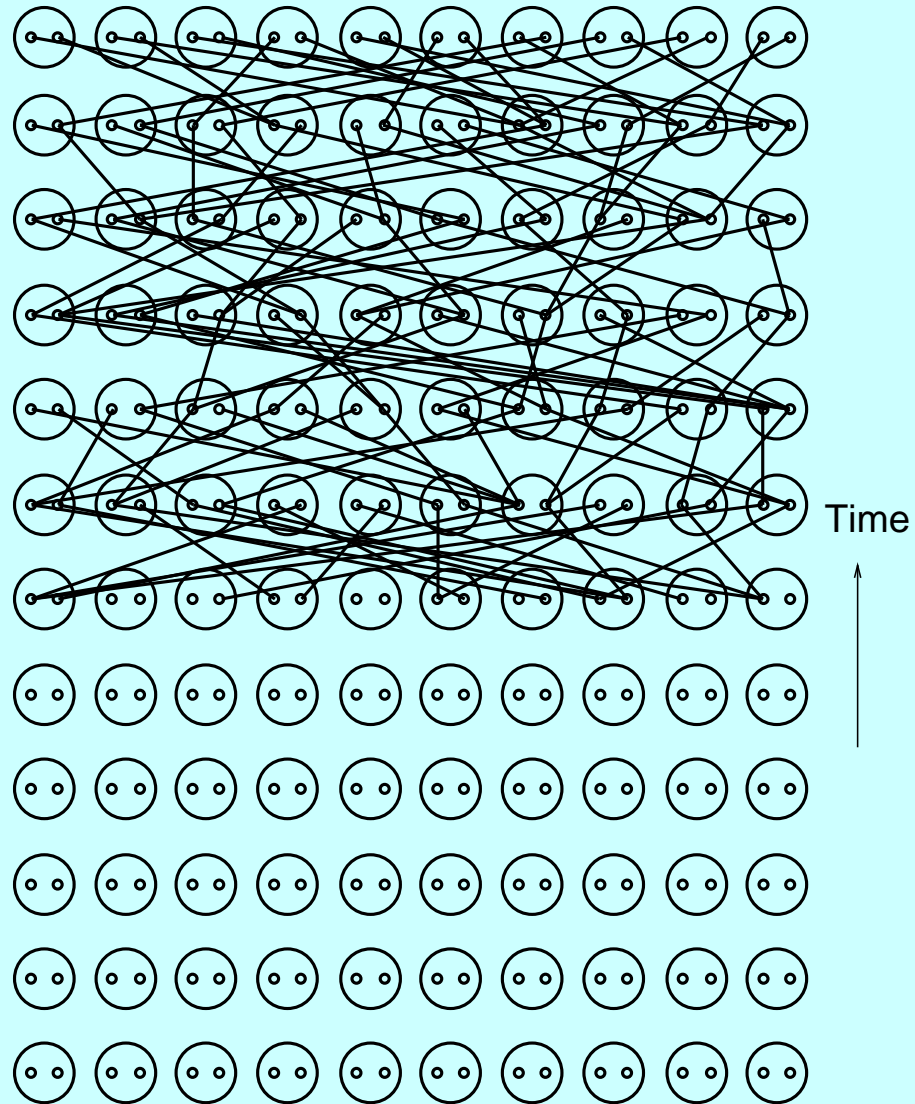
... and one more

A random-mating population



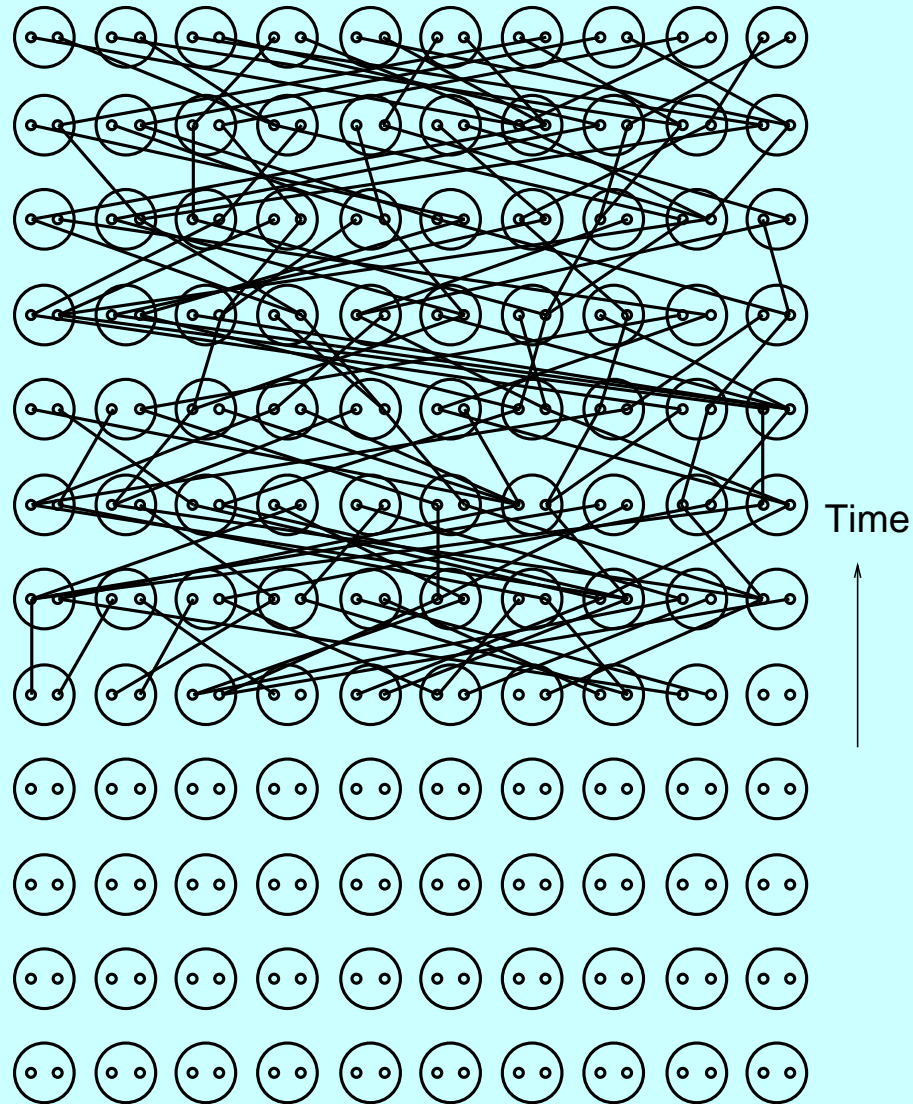
... and one more

A random-mating population



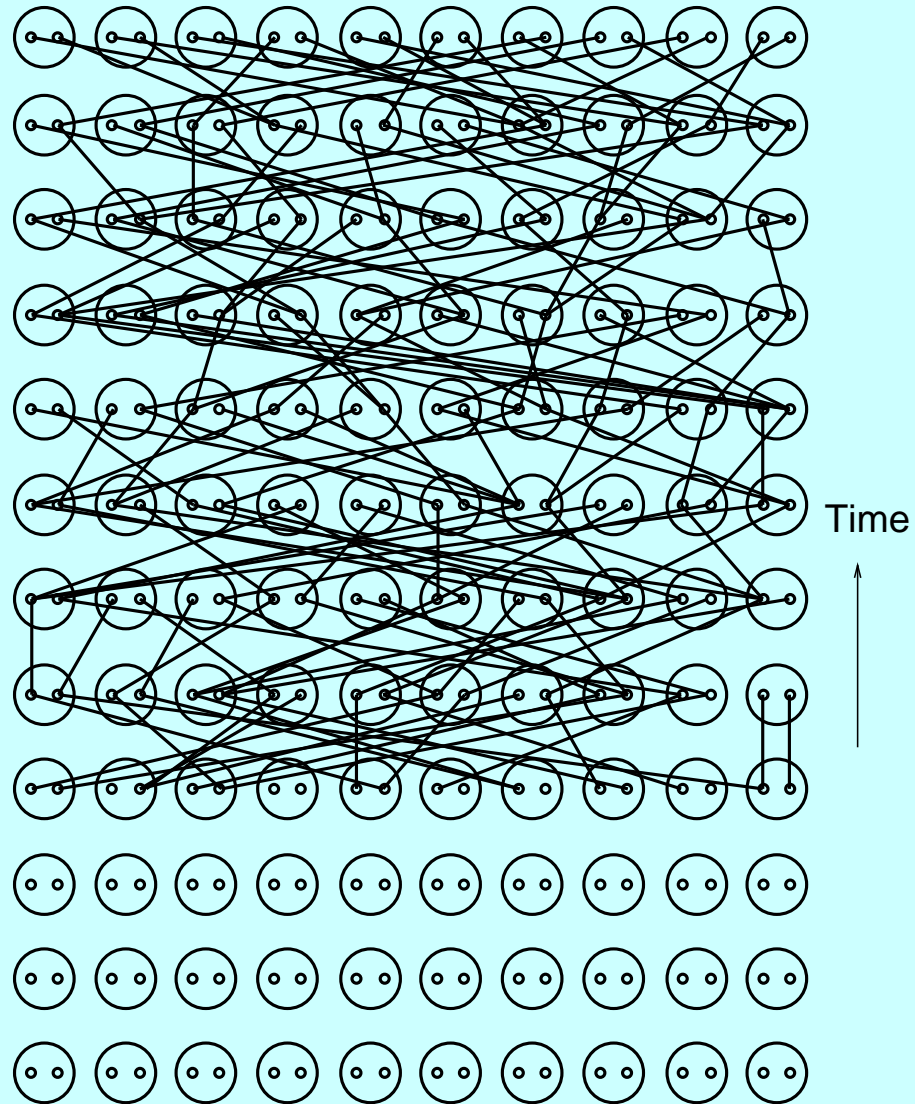
... and one more

A random-mating population



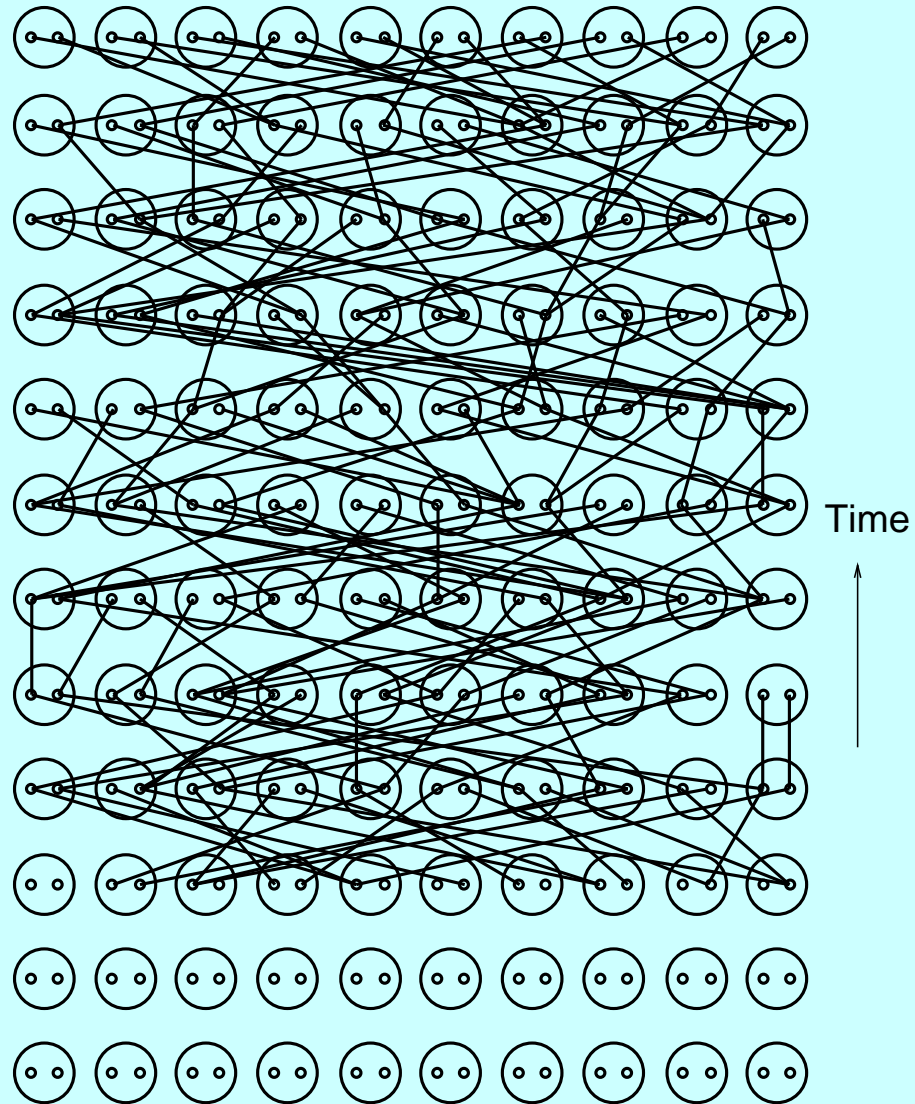
... and one more

A random-mating population



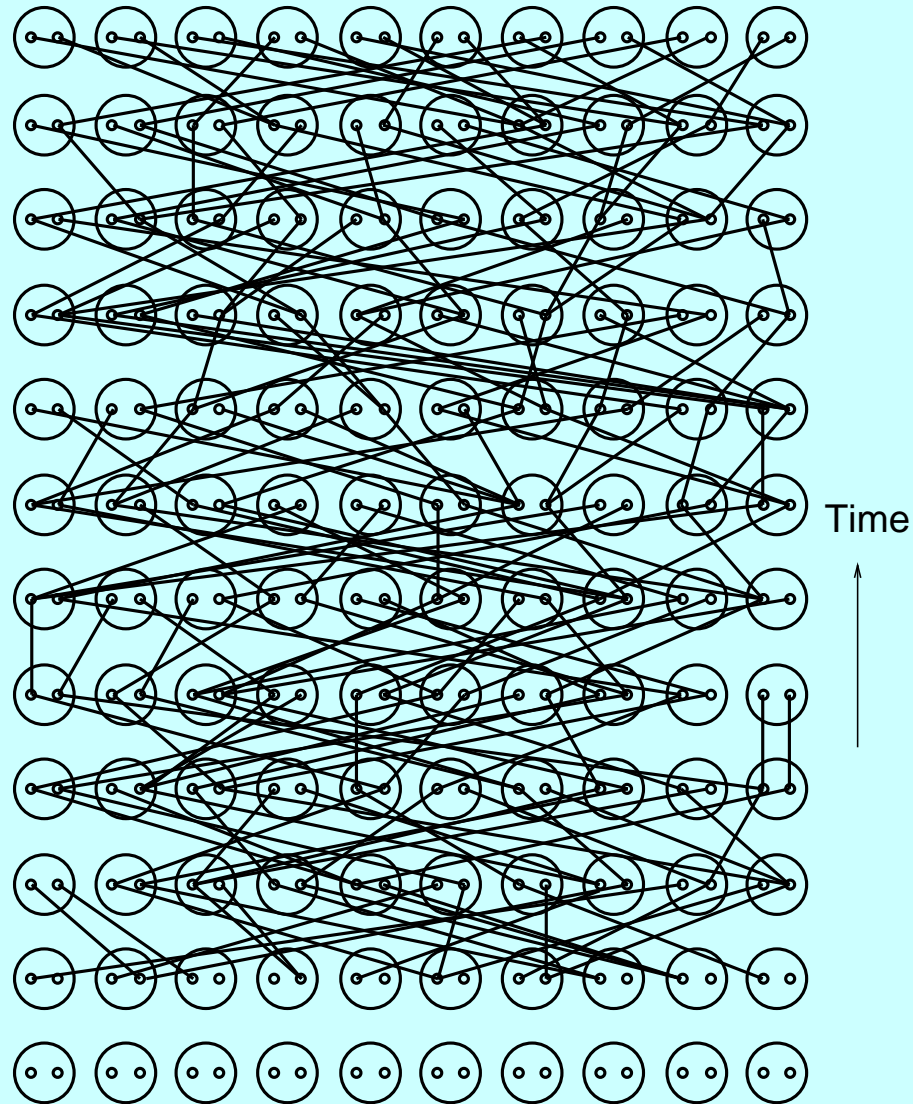
... and one more

A random-mating population



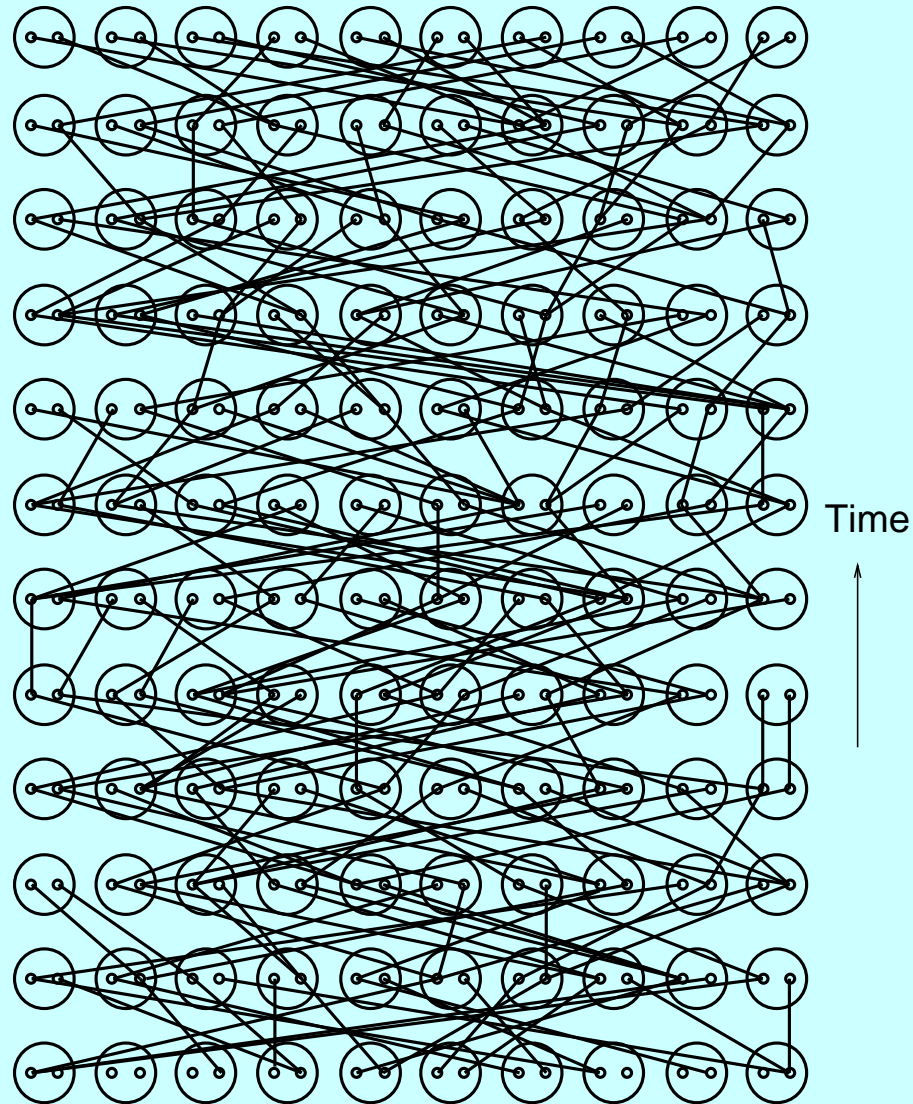
... and one more

A random-mating population



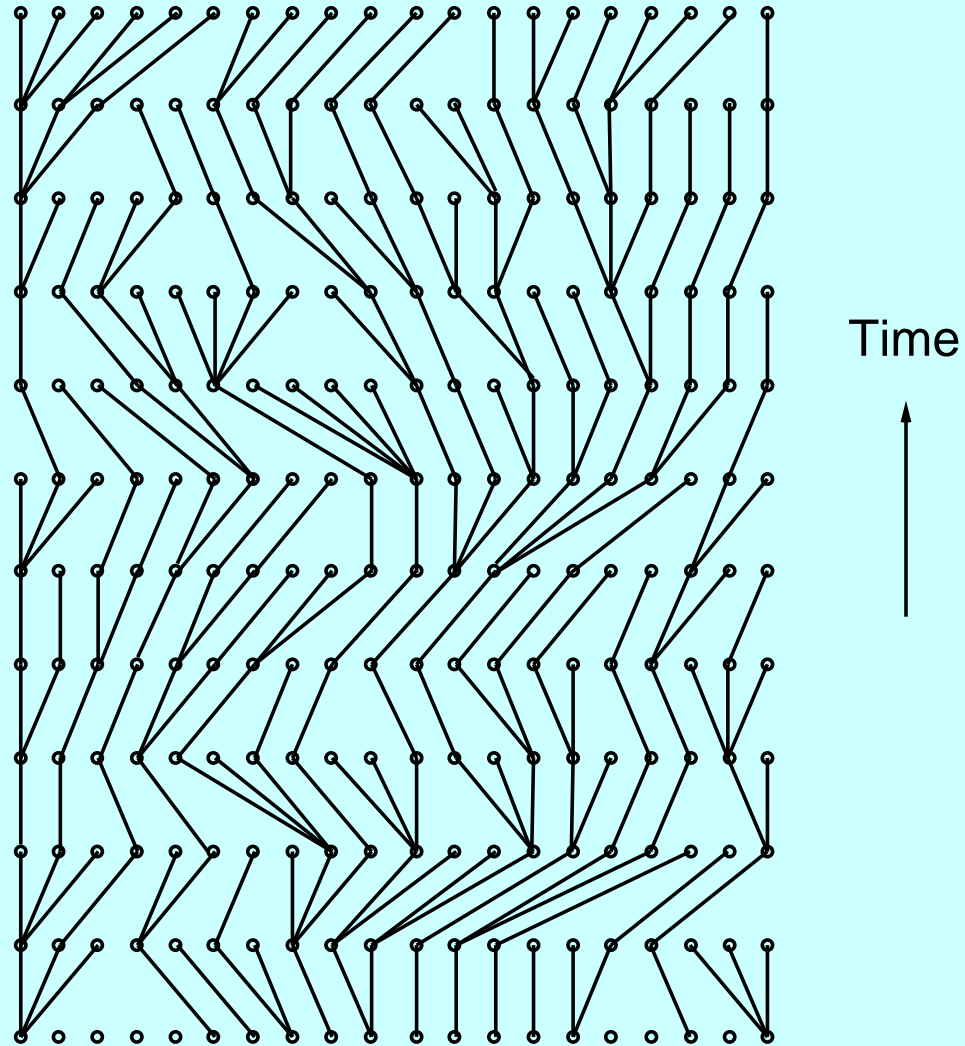
... and one more

A random-mating population



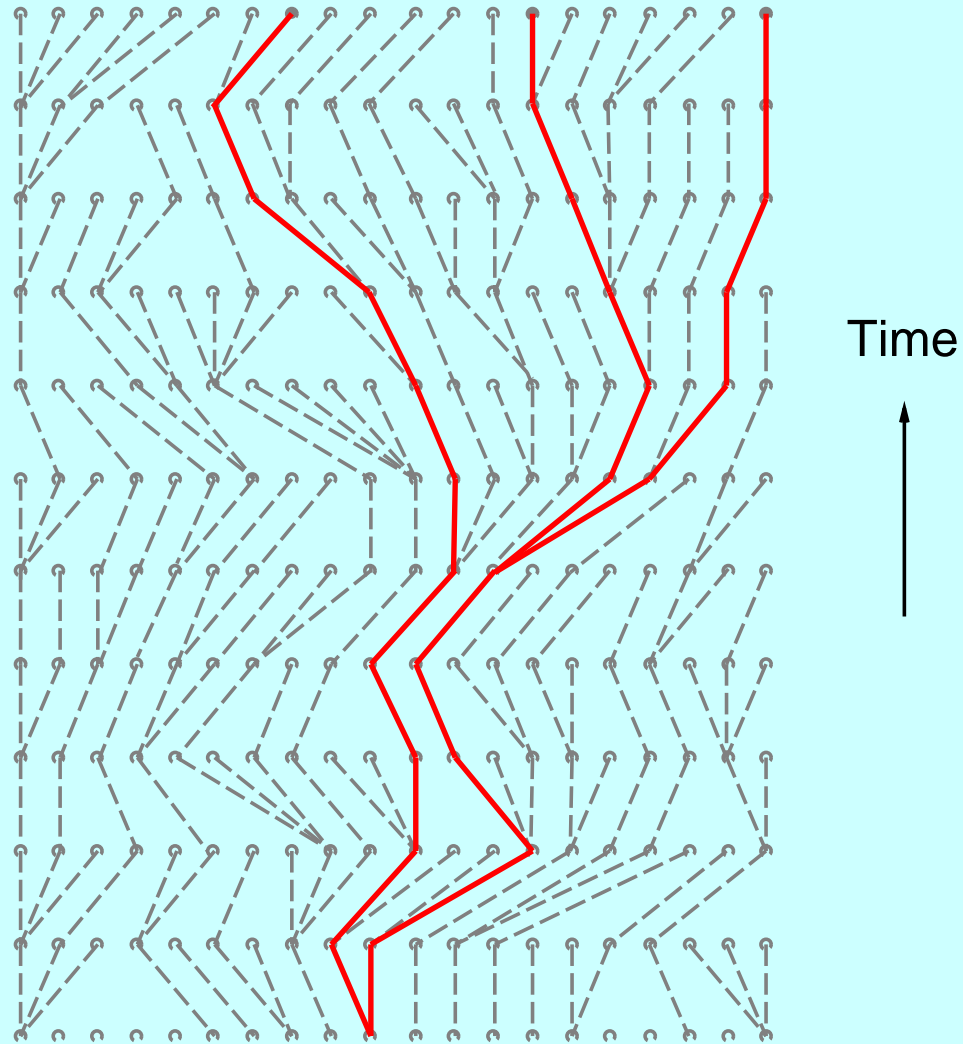
The genealogy of gene copies is a tree

Genealogy of gene copies, after reordering the copies

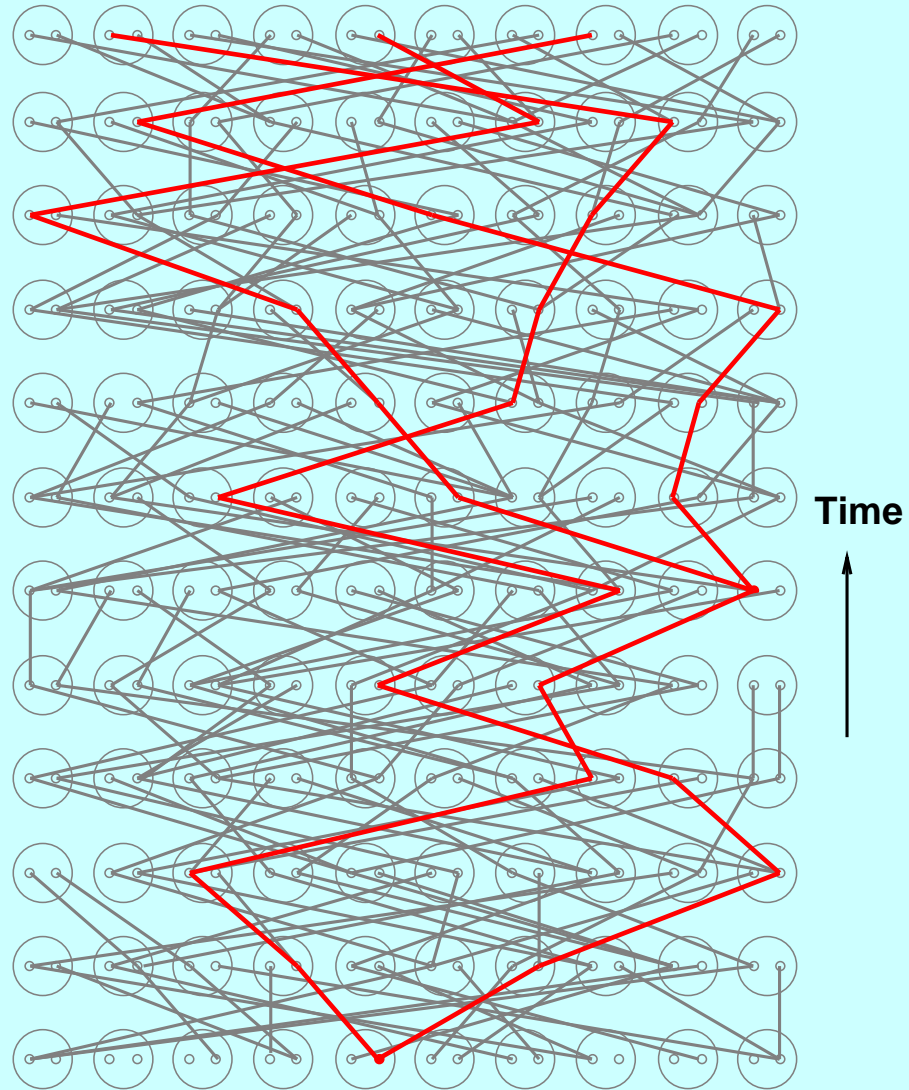


Ancestry of a sample of 3 copies

Genealogy of a small sample of genes from the population



Here is that tree of 3 copies in the pedigree



Sir John Kingman



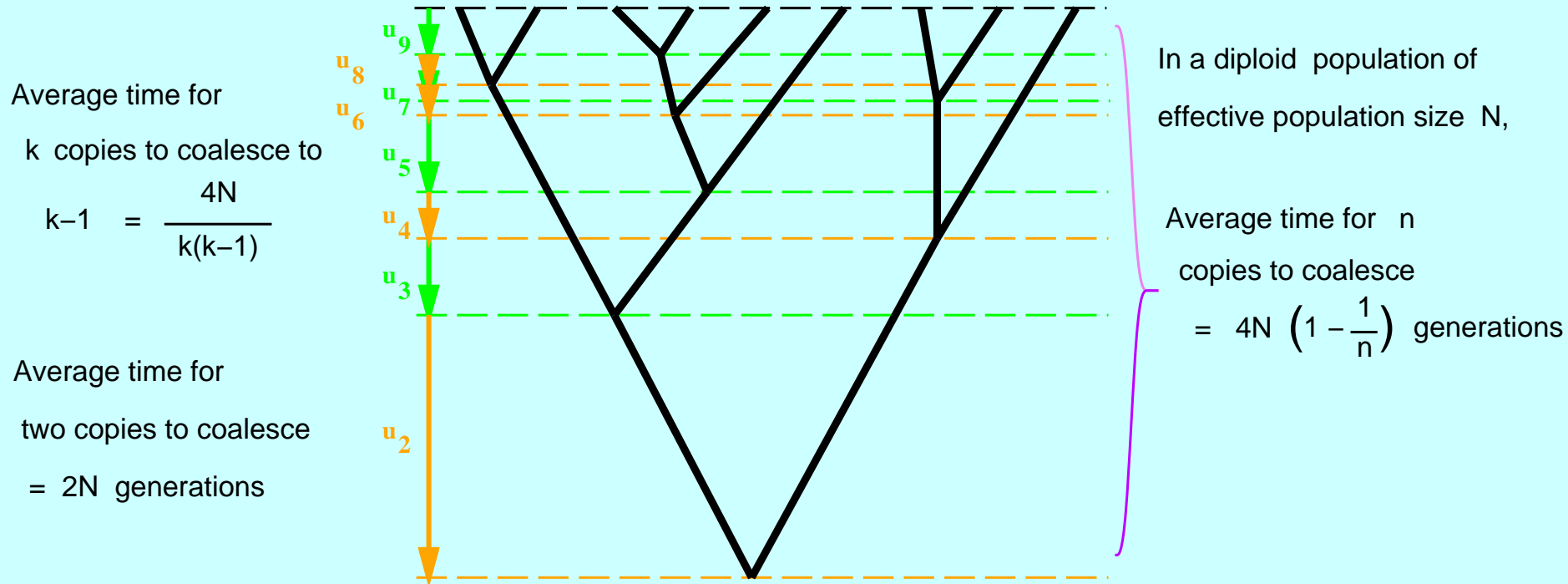
J. F. C. Kingman in about 1983

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



What's misleading about this diagram: the lineages that coalesce are random pairs, not necessarily ones that are next to each other in a linear order.

The coalescent – a derivation

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

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

(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; the sum is:

$$\begin{aligned} & k(k - 1)/2 \times 1 \times \frac{1}{2N} \times \left(1 - \frac{1}{2N}\right) \\ & \times \left(1 - \frac{2}{2N}\right) \times \cdots \times \left(1 - \frac{k-2}{2N}\right) \end{aligned}$$

so that the total probability that a pair coalesces is

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

Probabilities of two or more lineages coalescing

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

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

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

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

and since

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

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 $1/N$ are the same, 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.

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.

How deep is the common ancestor?

Take expected sizes of coalescents with $n, n - 1, \dots$ lineages down to 2.

$$4N \times \left(\frac{1}{n(n-1)} \right) = 4N \times \left(\frac{1}{n-1} - \frac{1}{n} \right)$$

$$4N \times \left(\frac{1}{(n-1)(n-2)} \right) = 4N \times \left(\frac{1}{n-2} - \frac{1}{n-1} \right)$$

and so on until 2:

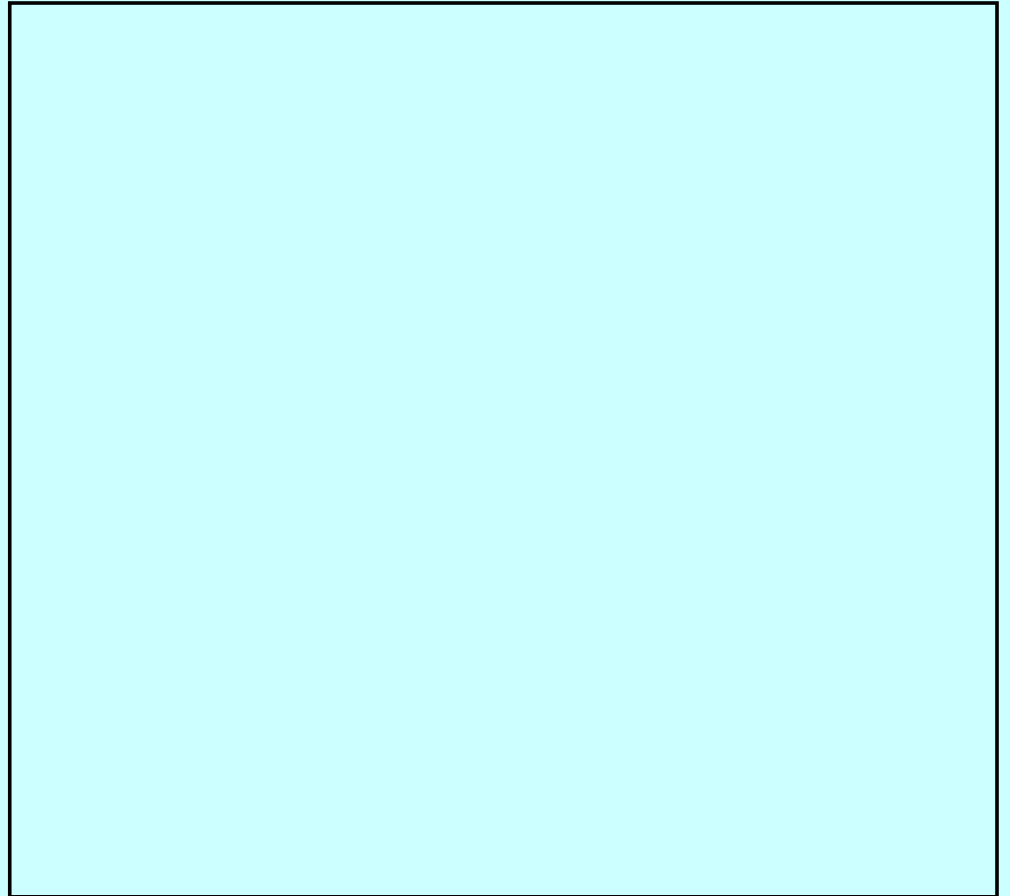
$$4N \times \left(\frac{1}{2 \times 1} \right) = 4N \times \left(\frac{1}{1} - \frac{1}{2} \right)$$

and cancelling lots of terms in the sum of these

$$4N \left(1 - \frac{1}{n} \right)$$

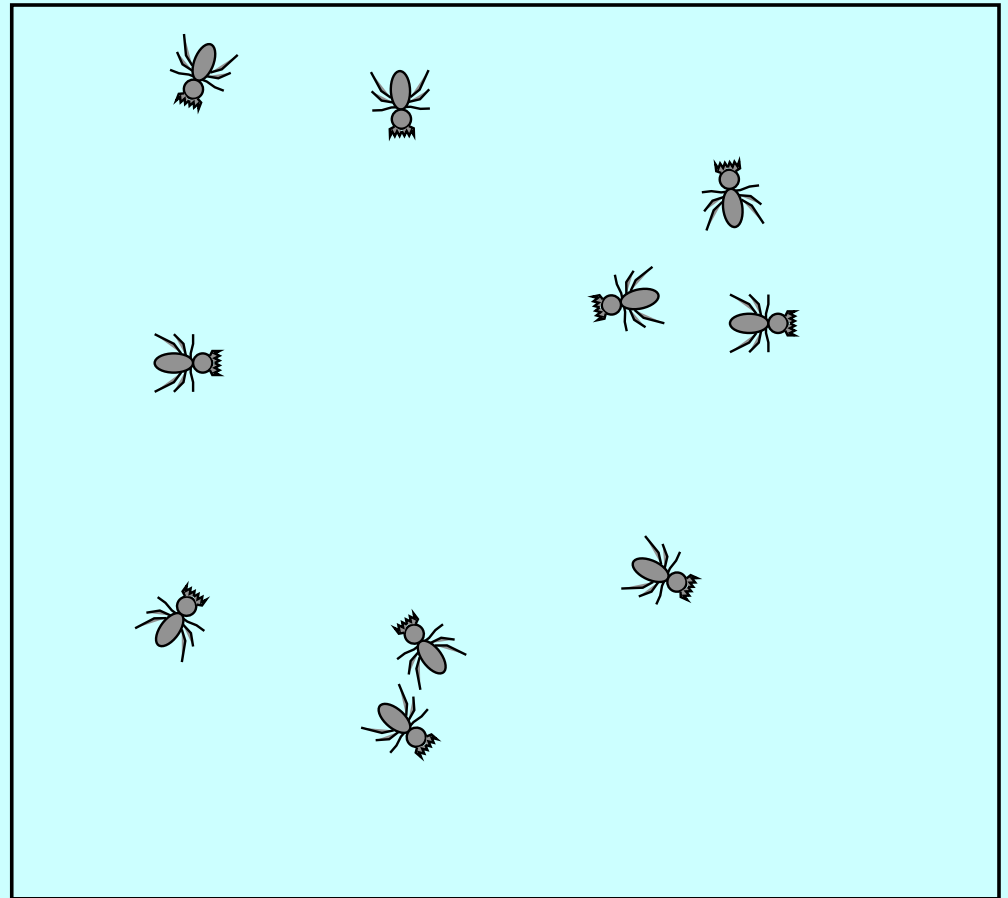
An accurate analogy: Bugs In A Box

There is a box ...



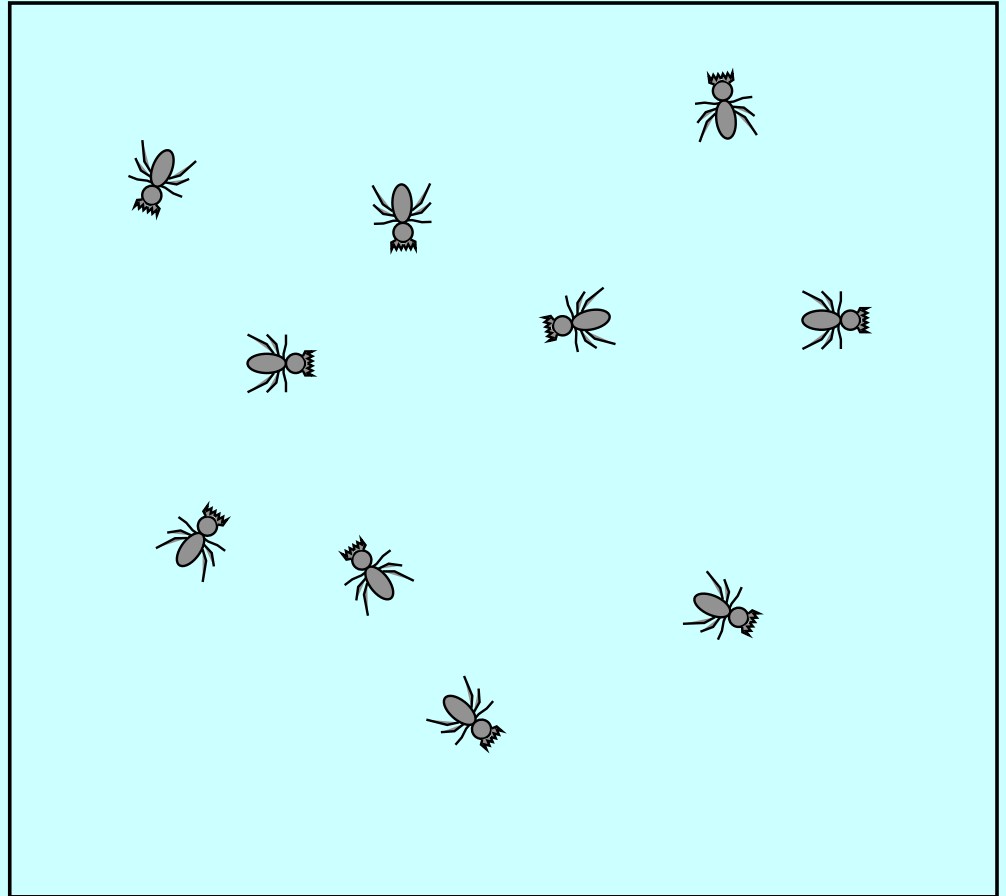
An accurate analogy: Bugs In A Box

with bugs that are ...



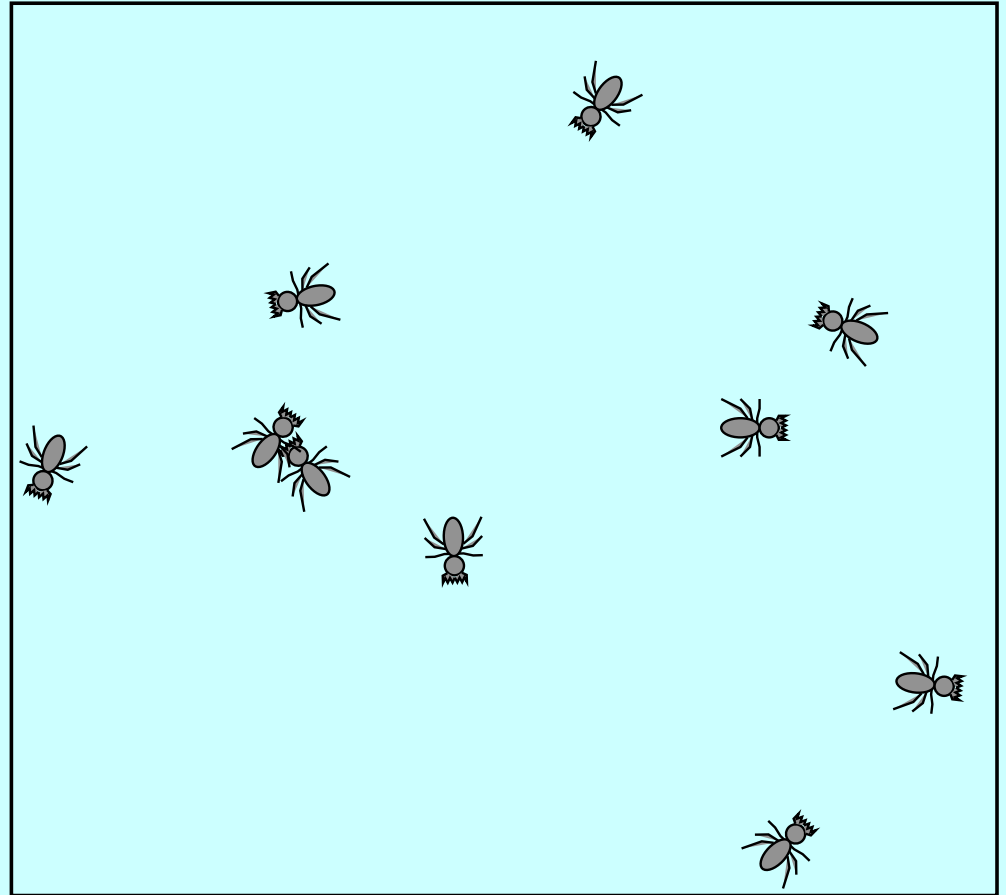
An accurate analogy: Bugs In A Box

hyperactive, ...



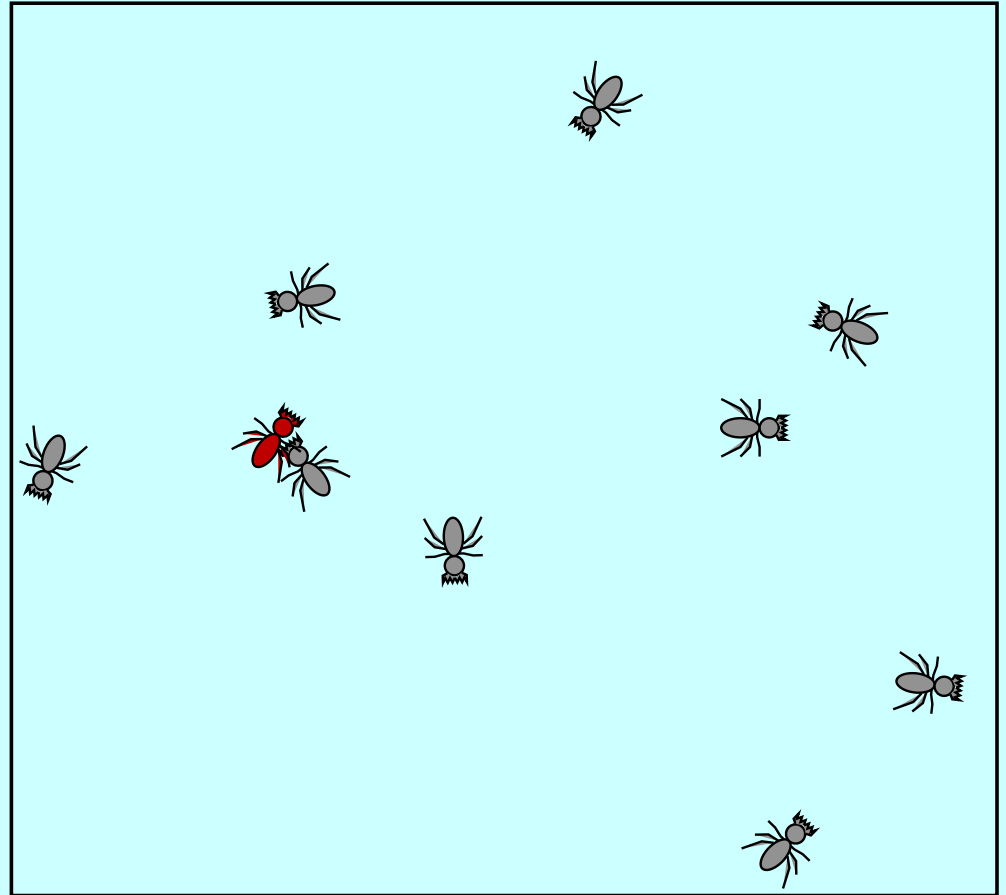
An accurate analogy: Bugs In A Box

indiscriminate, ...



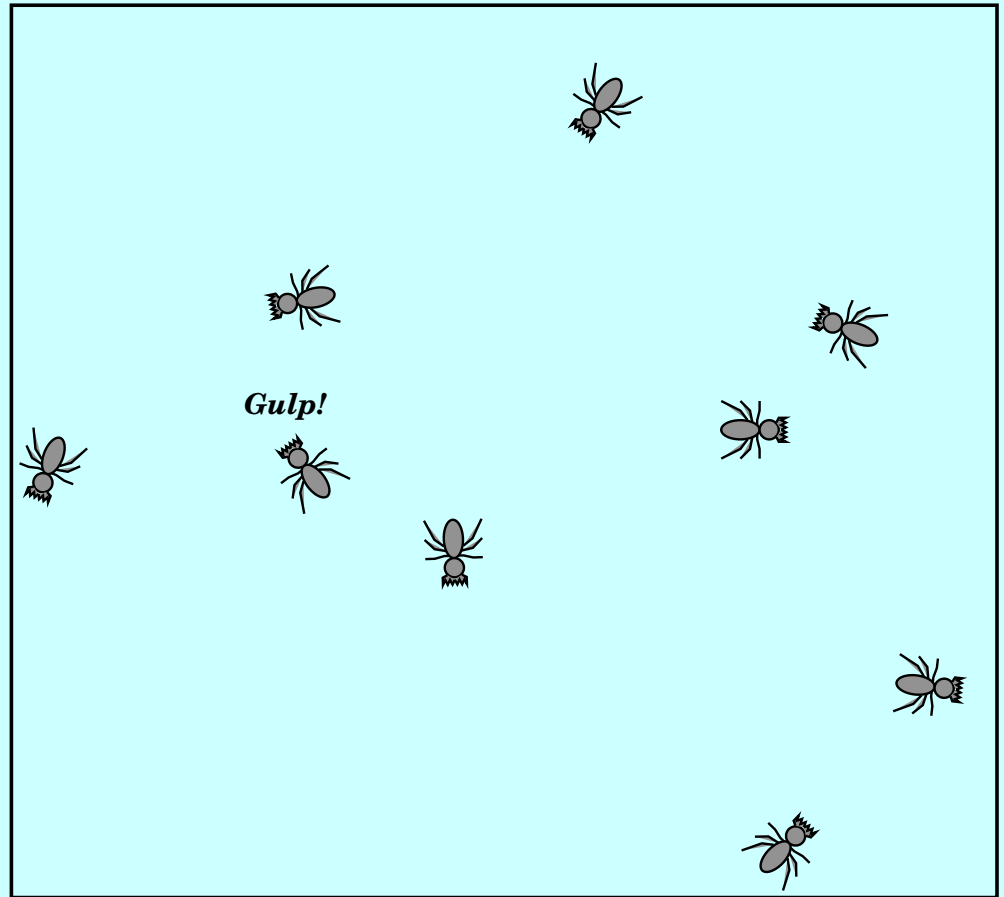
An accurate analogy: Bugs In A Box

voracious ...



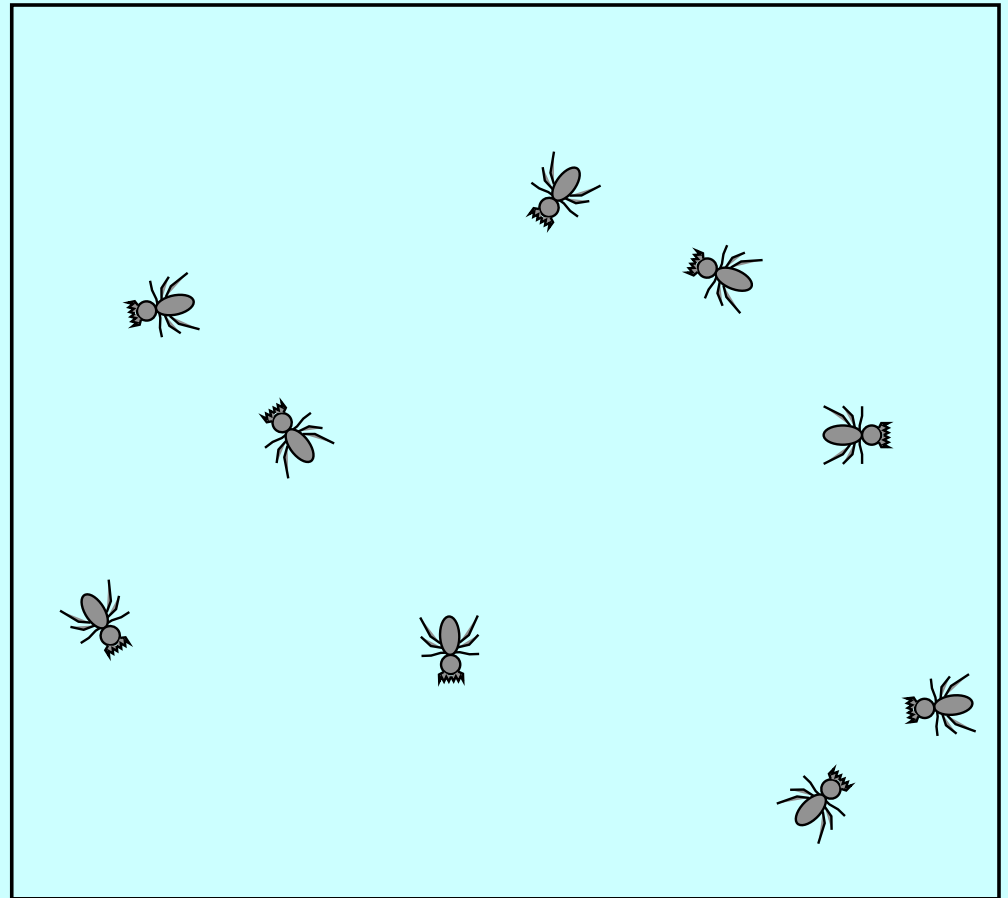
An accurate analogy: Bugs In A Box

(eats other bug) ...

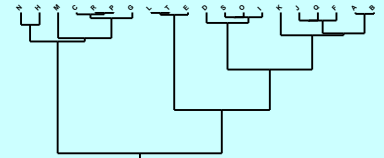
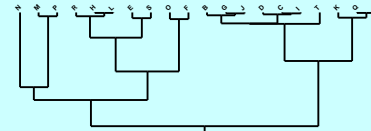
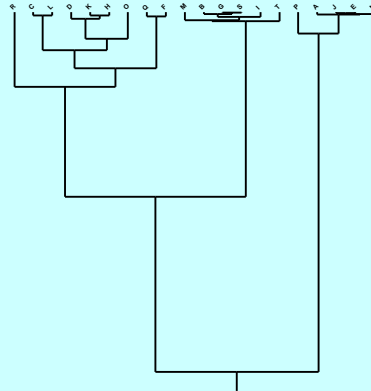
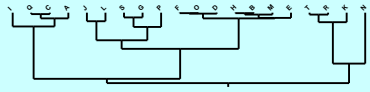
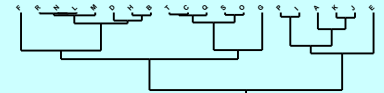
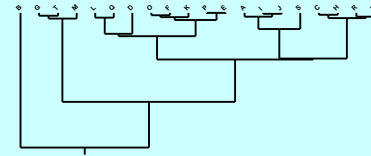
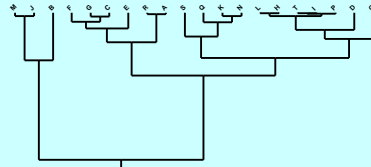
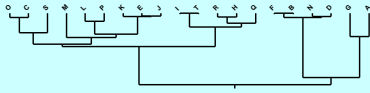


An accurate analogy: Bugs In A Box

and insatiable.

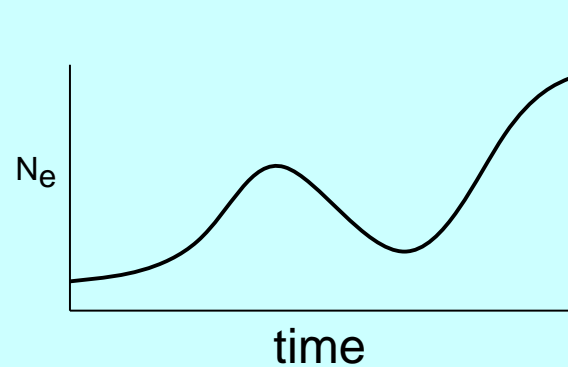


Random coalescent trees with 16 lineages



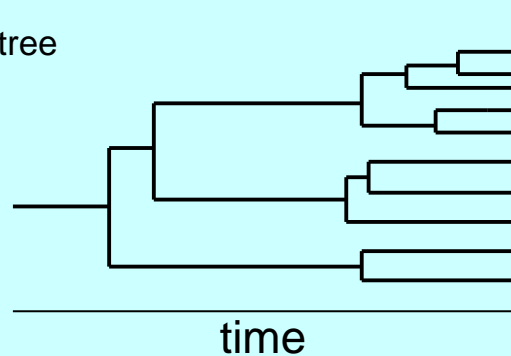
Coalescence is faster in small populations

Change of population size and coalescents

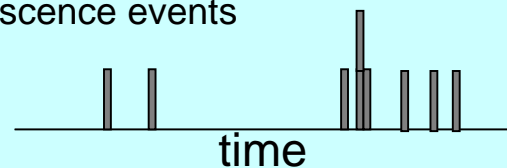


the changes in population size will produce waves of coalescence

the tree

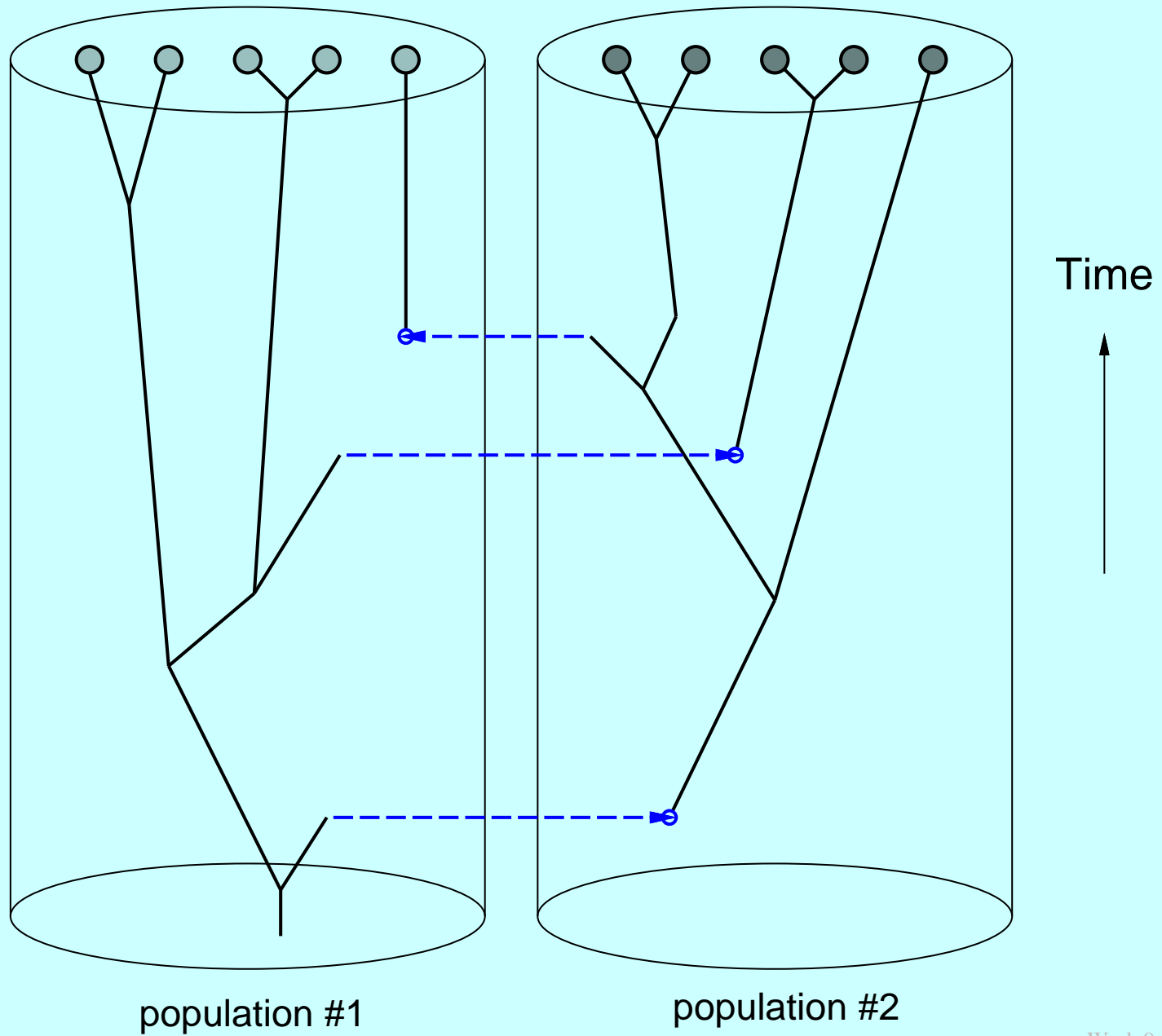


Coalescence events

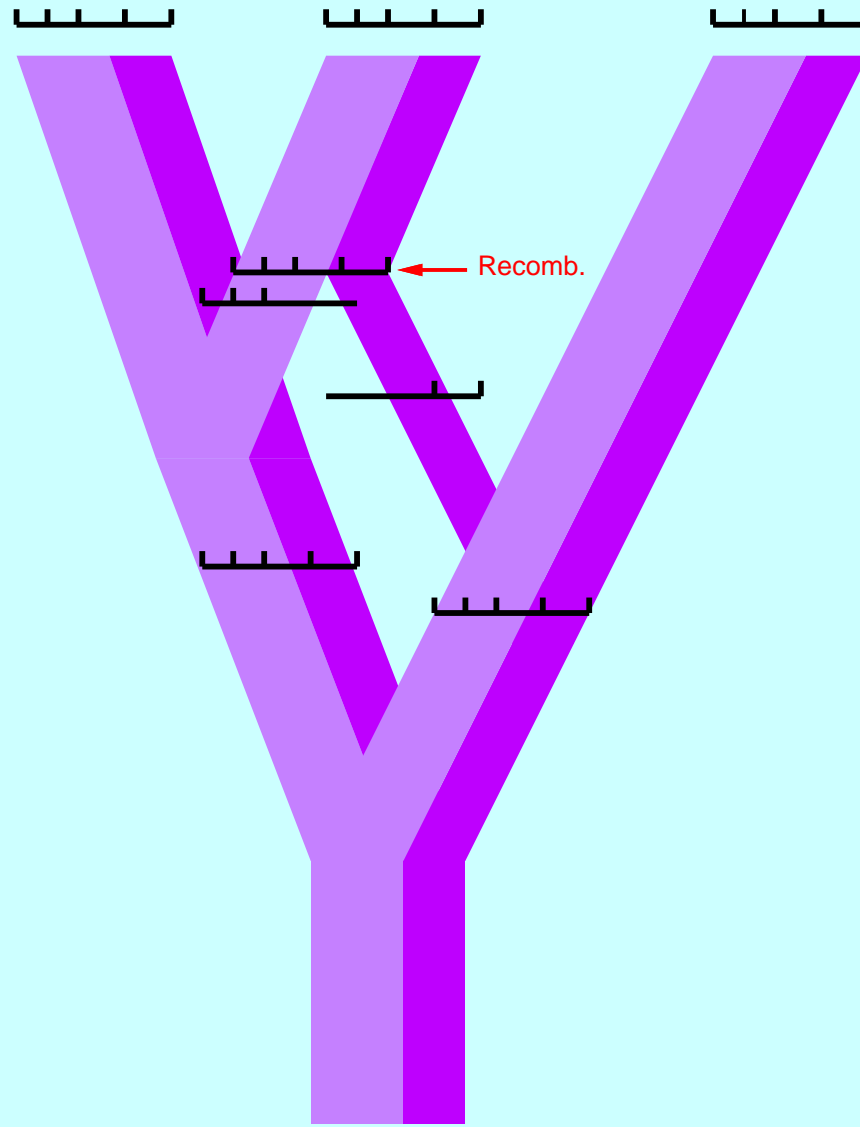


The parameters of the growth curve for N_e can be inferred by likelihood methods as they affect the prior probabilities of those trees that fit the data.

Migration can be taken into account



Recombination creates loops



Different markers have slightly different coalescent trees

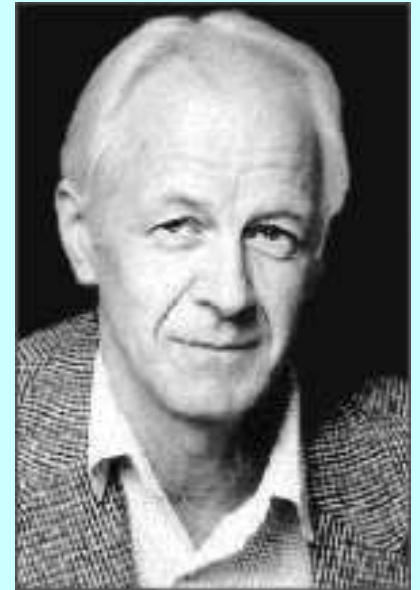
Cann, Stoneking, and Wilson



Becky Cann



Mark Stoneking



the late Allan Wilson

Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* **325**:a 31-36.

Mitochondrial Eve

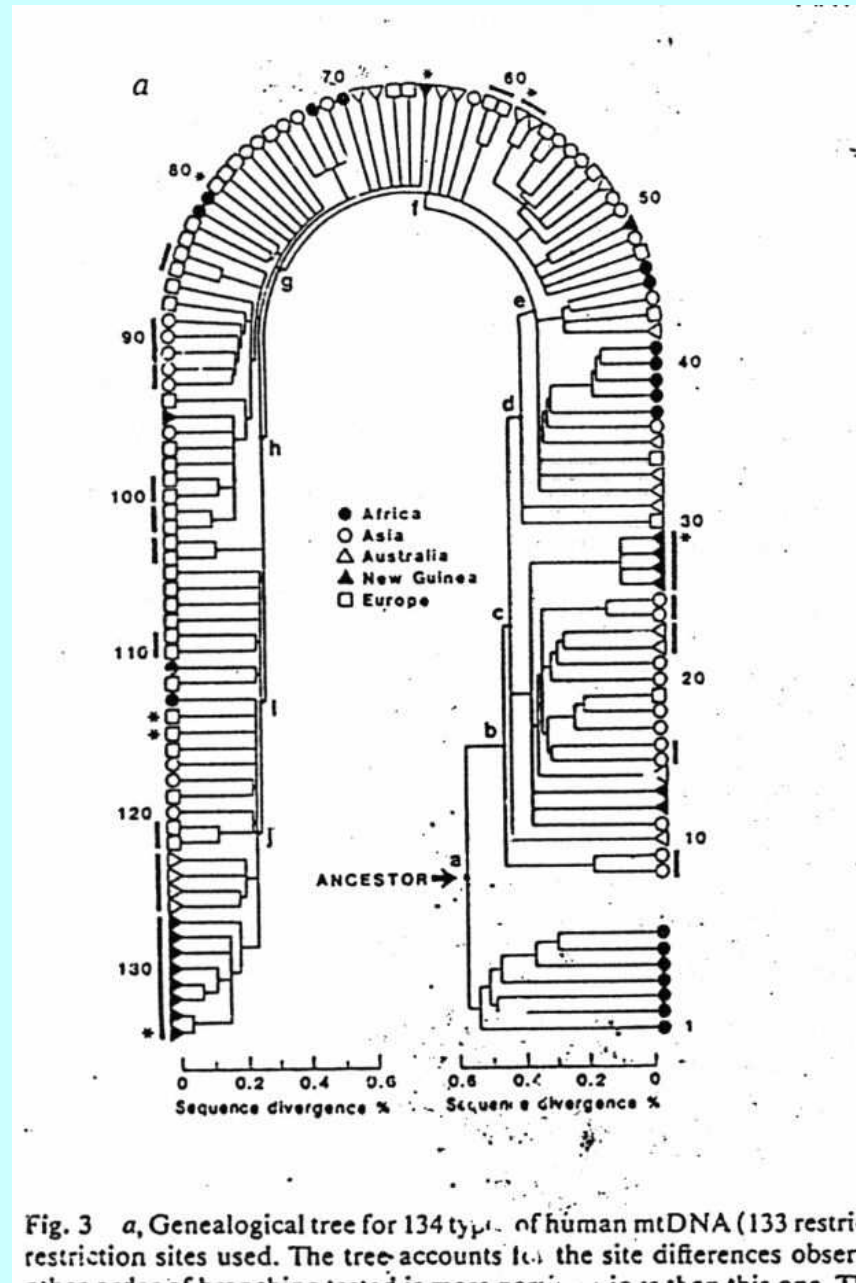
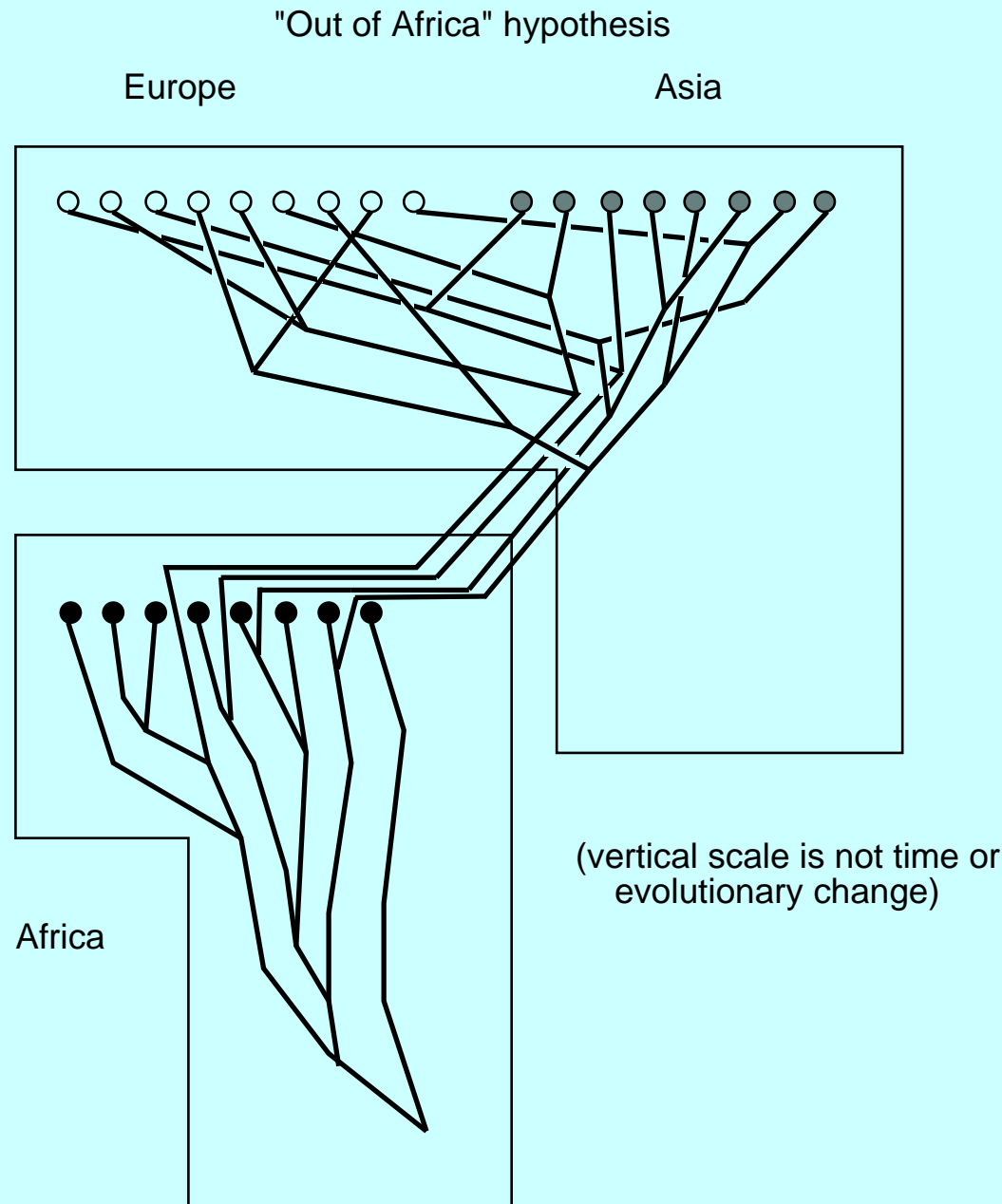


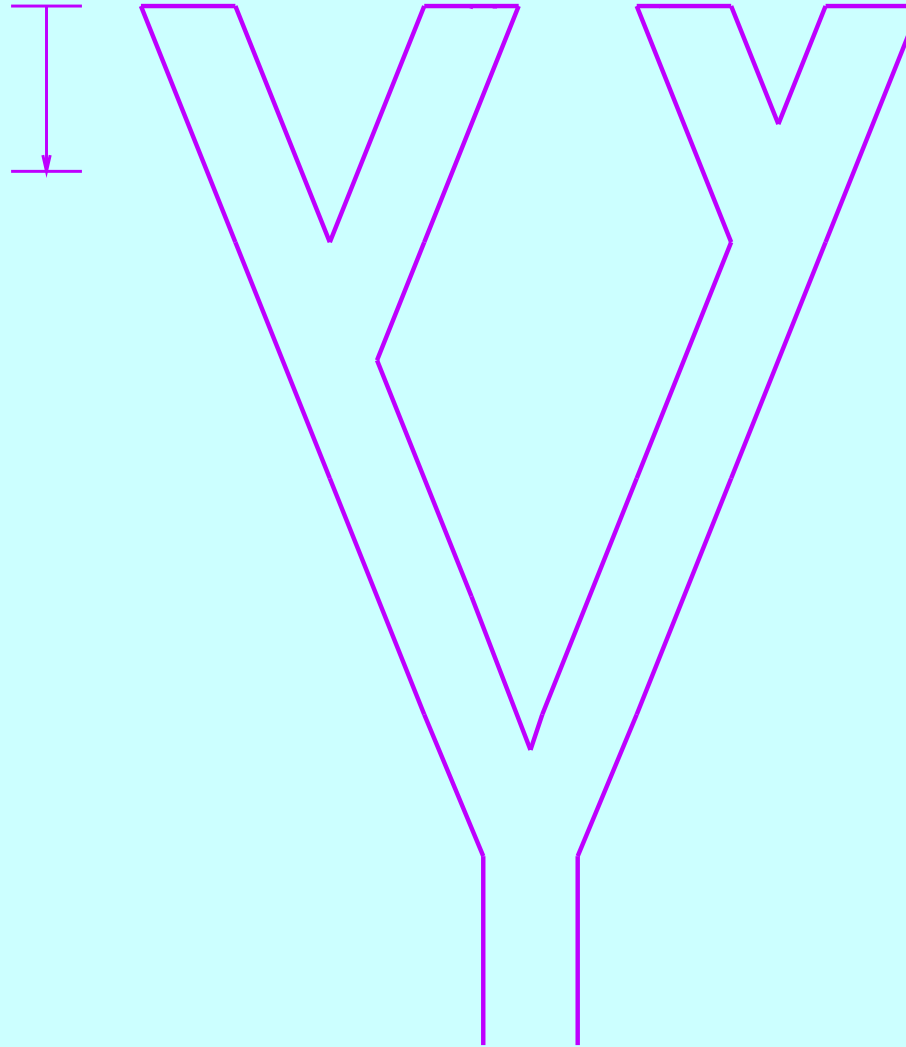
Fig. 3 a, Genealogical tree for 134 types of human mtDNA (133 restriction sites used). The tree accounts for the site differences observed between the mtDNA types. The tree is based on the data of...

We want to be able to analyze human evolution



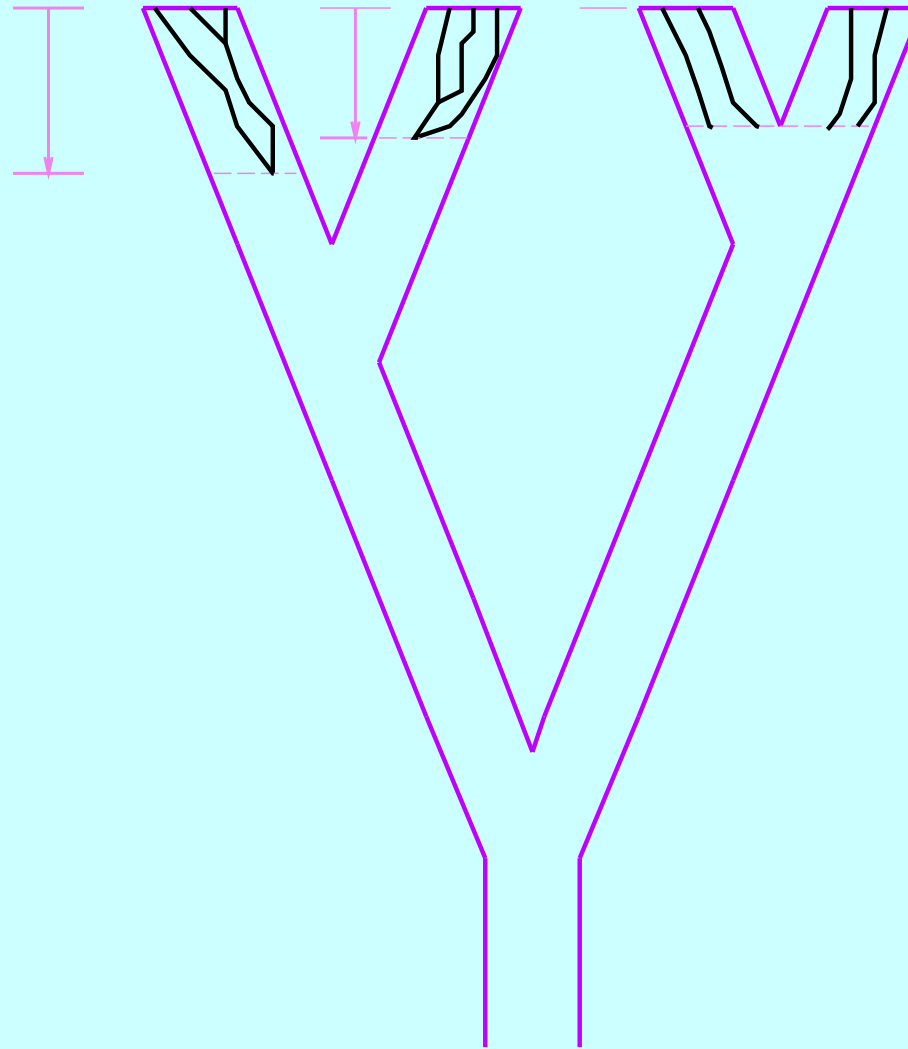
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



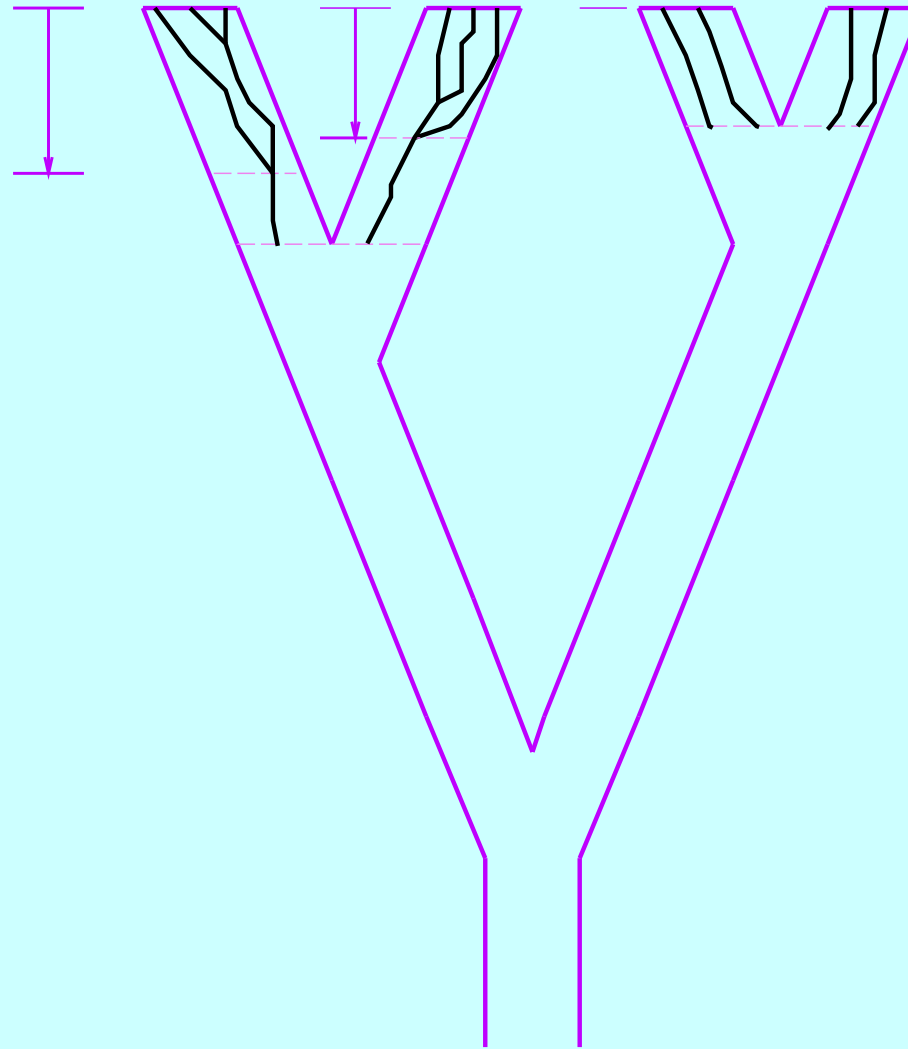
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



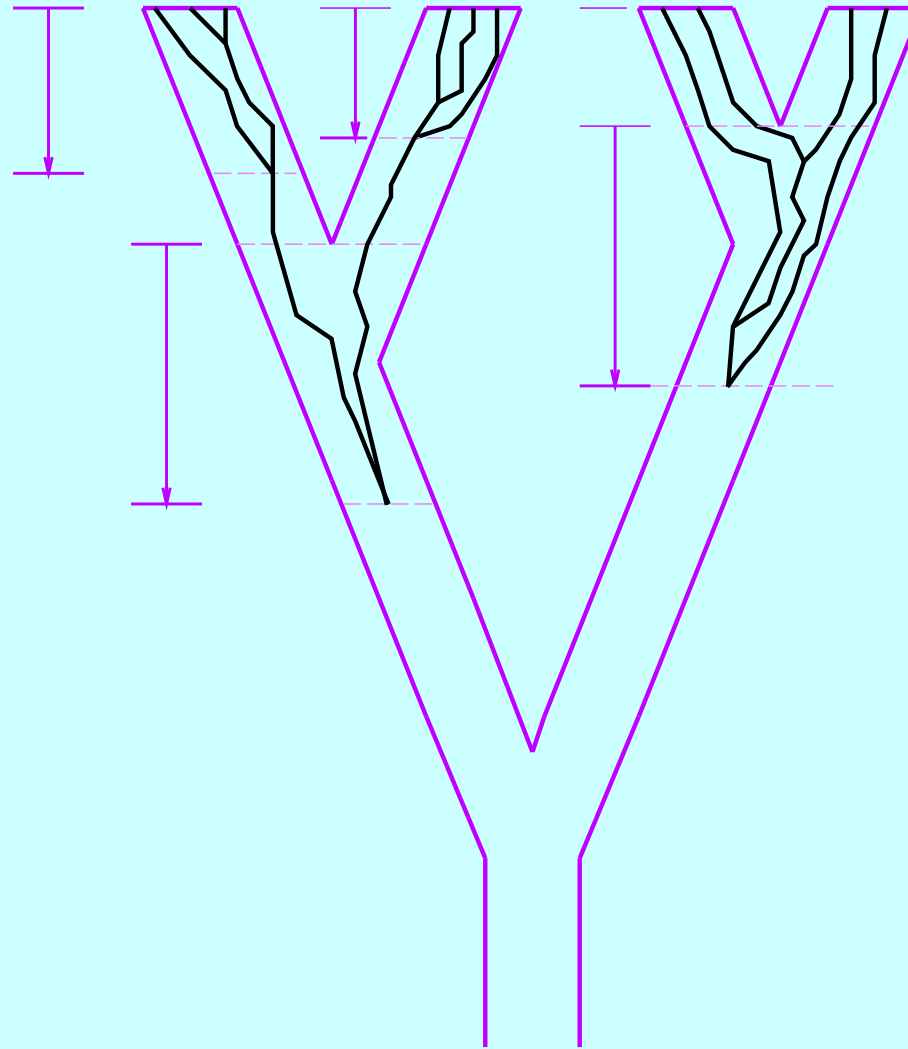
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



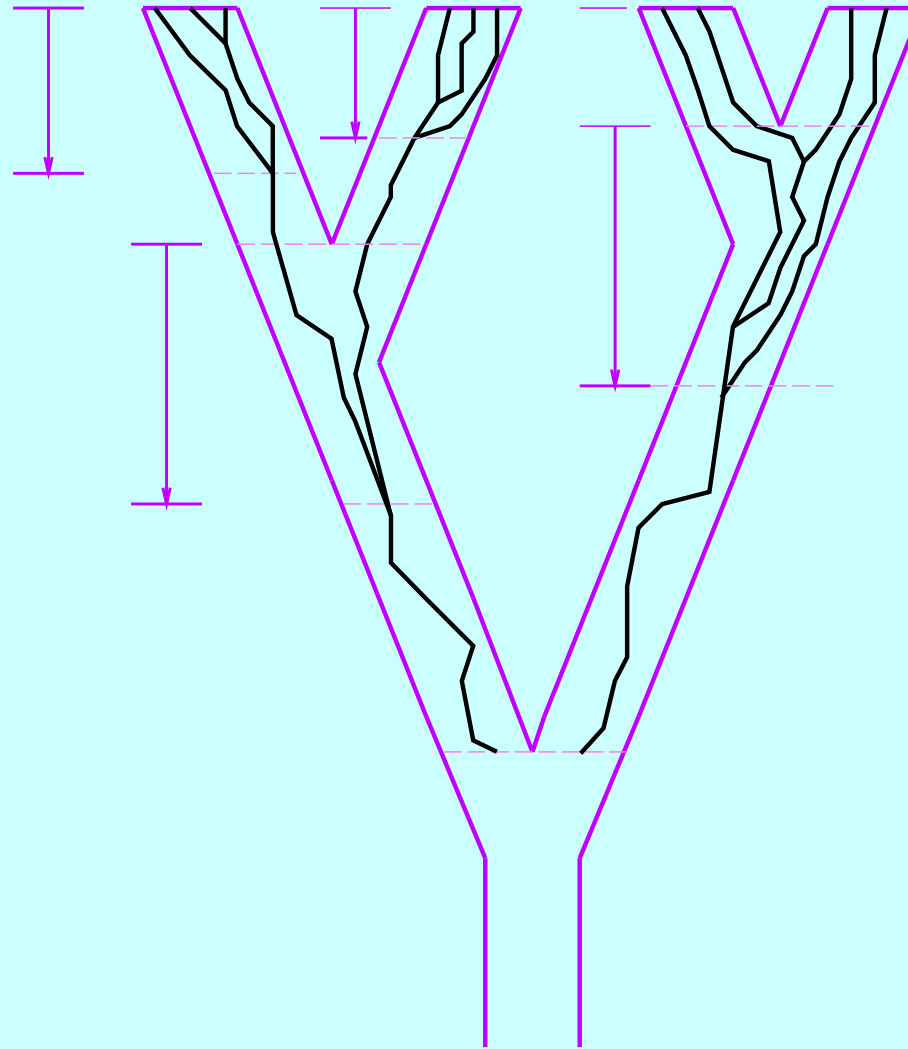
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



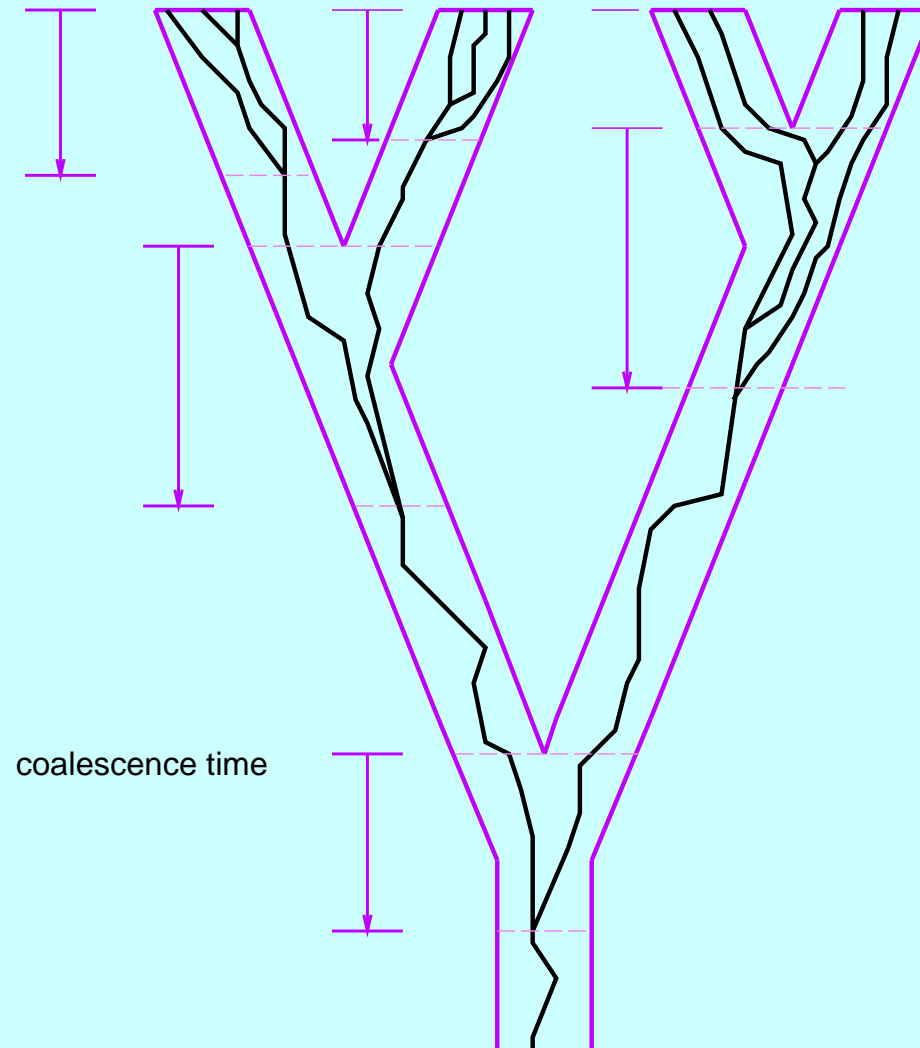
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



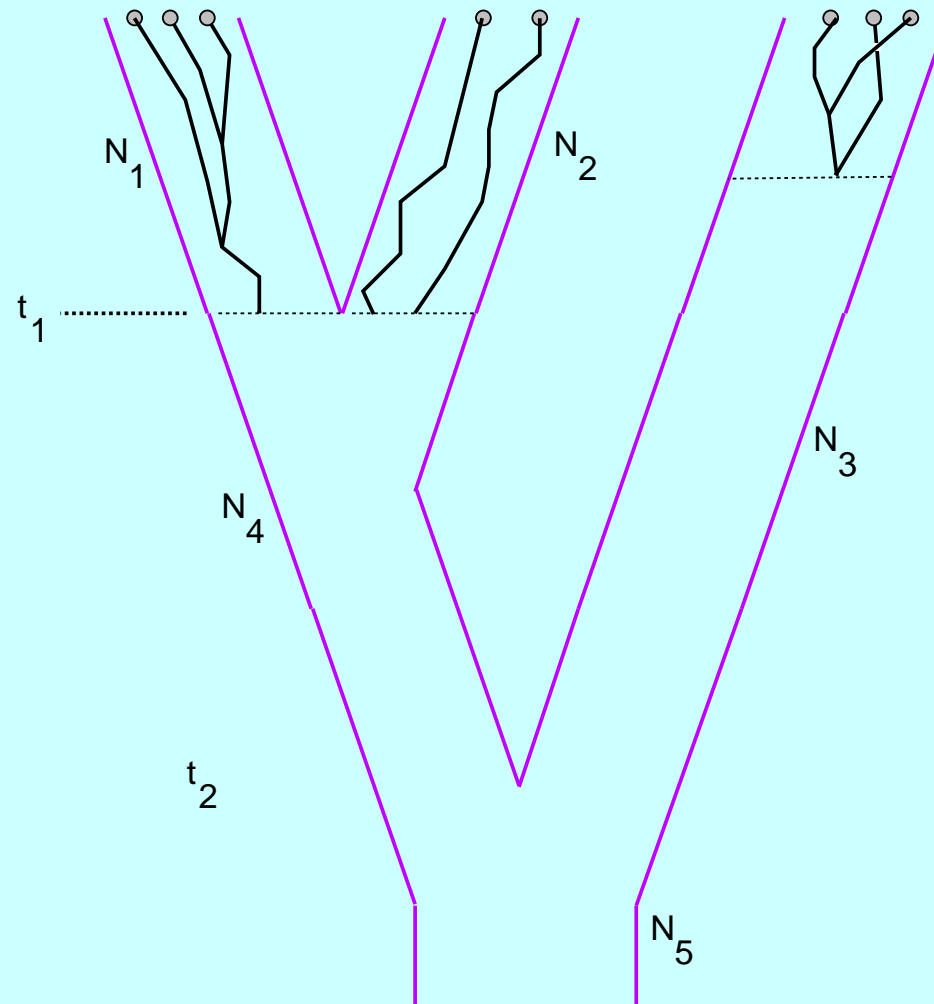
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



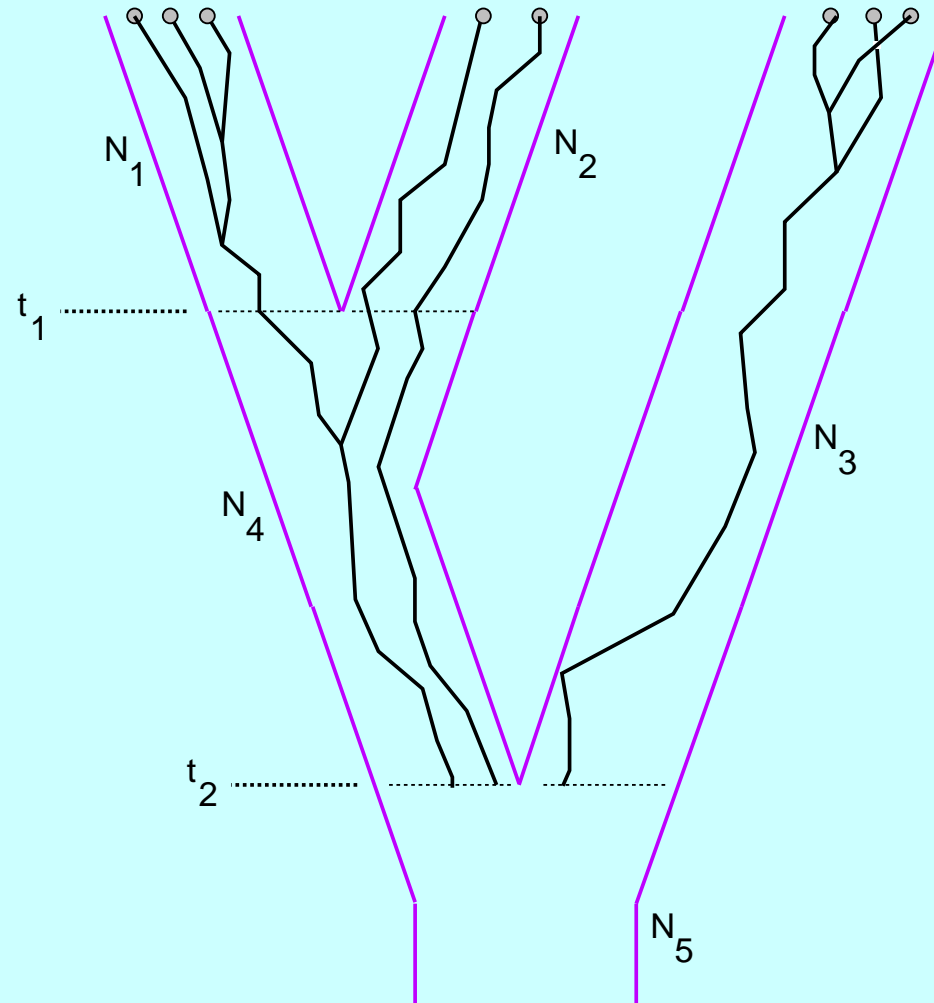
If the branch is more than N_e generations long ...

Gene tree and Species tree



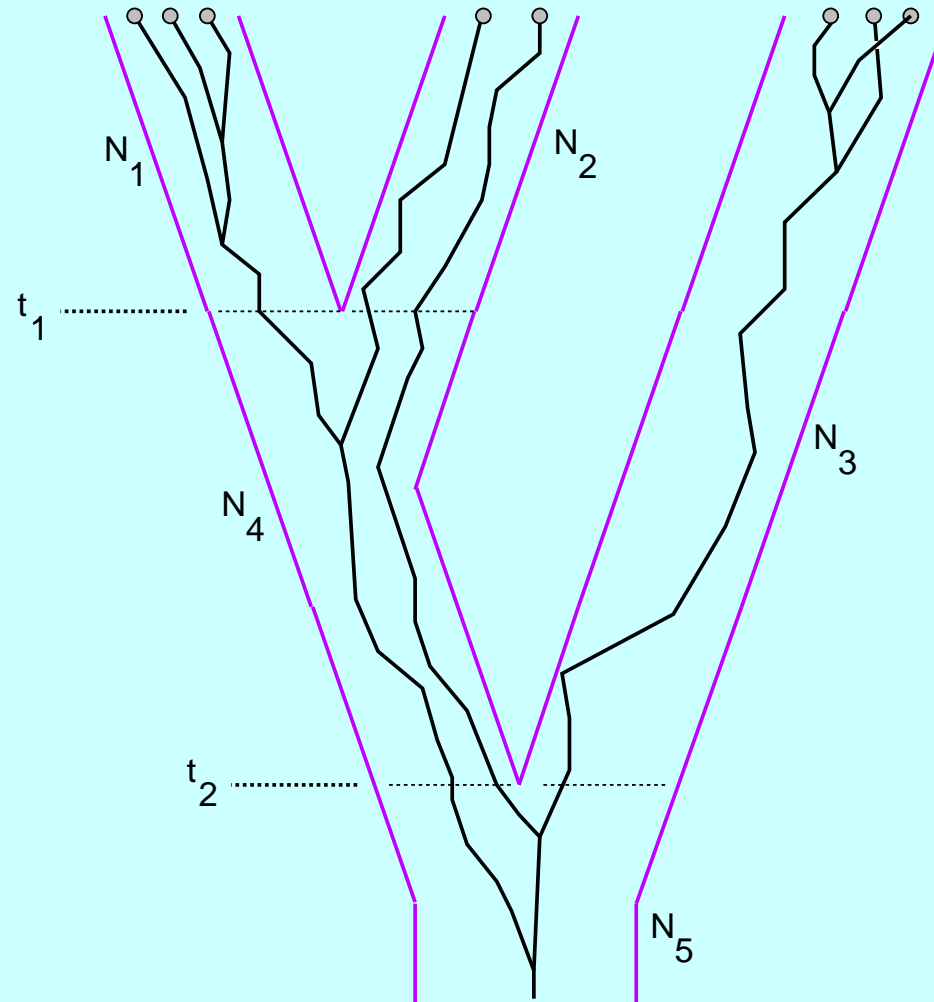
If the branch is more than N_e generations long ...

Gene tree and Species tree

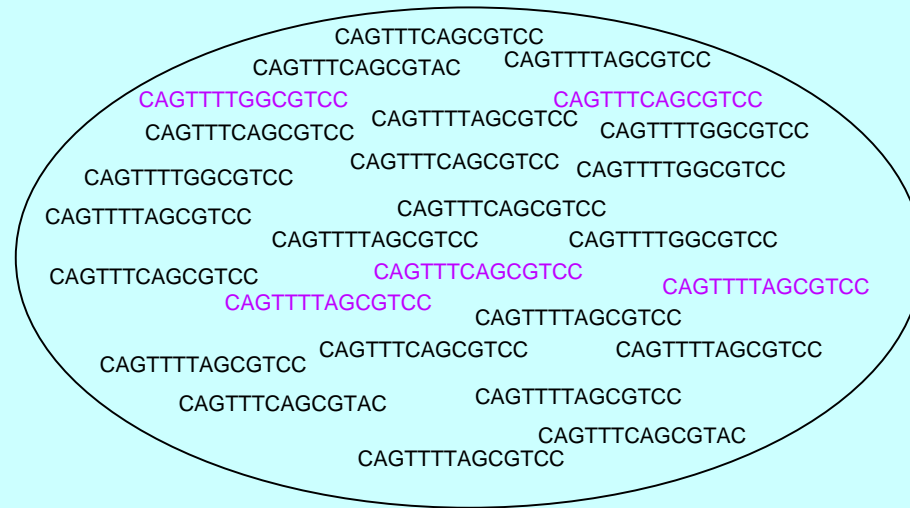


If the branch is more than N_e generations long ...

Gene tree and Species tree

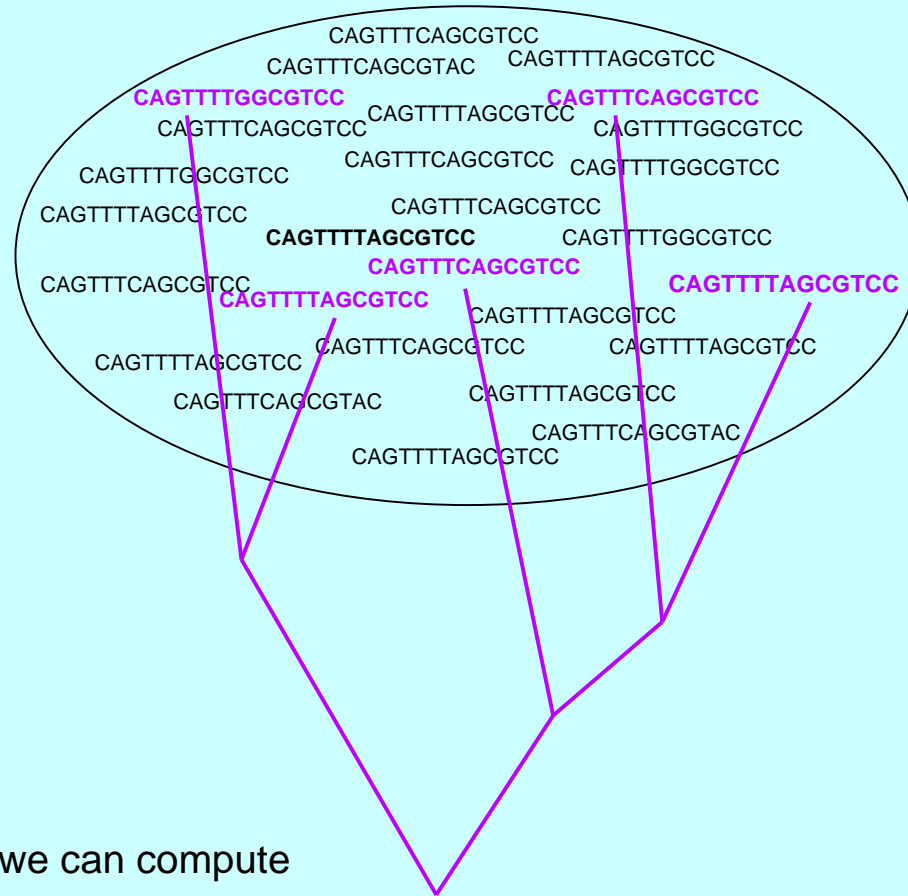


How do we compute a likelihood for a population sample?



$$L = \text{Prob}(\text{CAGTTTCAGCGTCC}, \text{CAGTTTCAGCGTCC}, \dots) = ??$$

If we have a tree for the sample sequences, we can



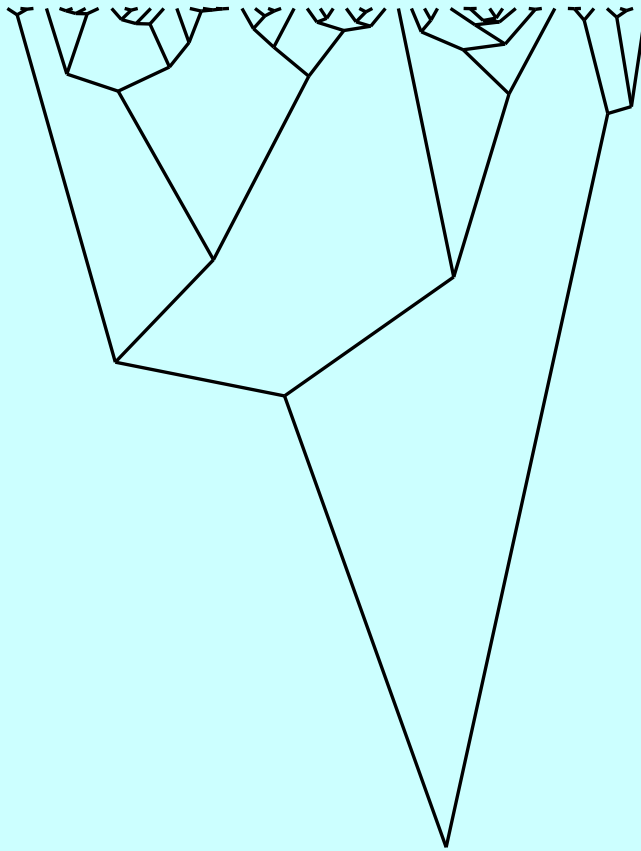
so we can compute

$\text{Prob}(\text{CAGTTTCAGCGTCC}, \text{CAGTTTCAGCGTCC}, \dots \mid \text{Genealogy})$

but how to computer the overall likelihood from this?

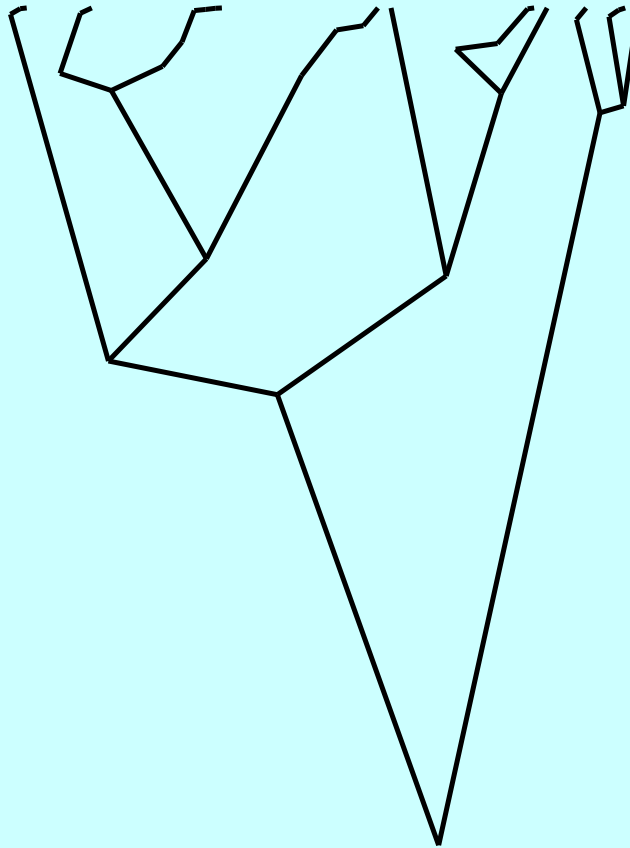
If we have a sample of 50 copies

50-gene sample in a coalescent tree



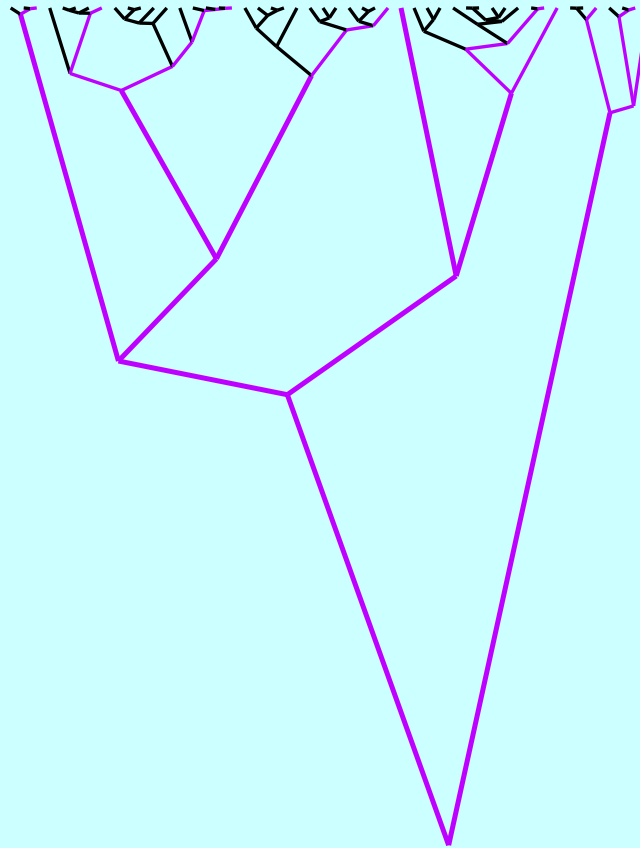
The first 10 account for most of the branch length

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)

The basic equation for coalescent likelihoods

In the case of a single population with parameters


N_e effective population size

μ mutation rate per site

and assuming G' stands for a coalescent genealogy and D for the sequences,

$$L = \text{Prob}(D \mid N_e, \mu)$$

$$= \sum_{G'} \text{Prob}(G' \mid N_e) \text{Prob}(D \mid G', \mu)$$


Kingman's prior likelihood of tree

Rescaling the branch lengths

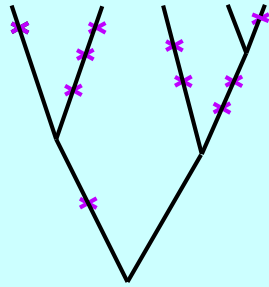
Rescaling branch lengths of G' so that branches are given in expected mutations per site, $G = \mu G'$, we get (if we let $\Theta = 4N_e\mu$)

$$L = \sum_G \text{Prob}(G | \Theta) \text{Prob}(D | G)$$

as the fundamental equation. For more complex population scenarios one simply replaces Θ with a vector of parameters.

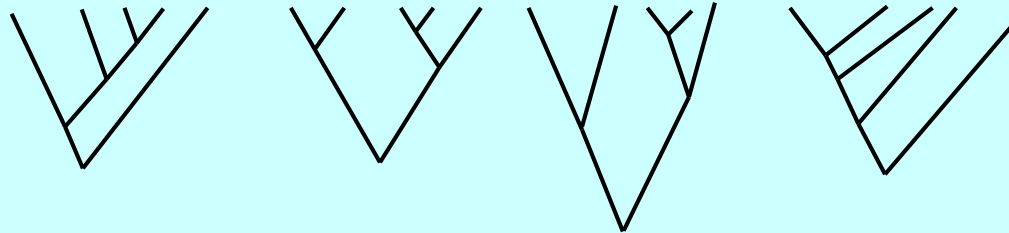
The variability comes from two sources

(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

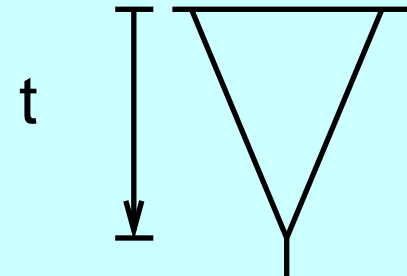
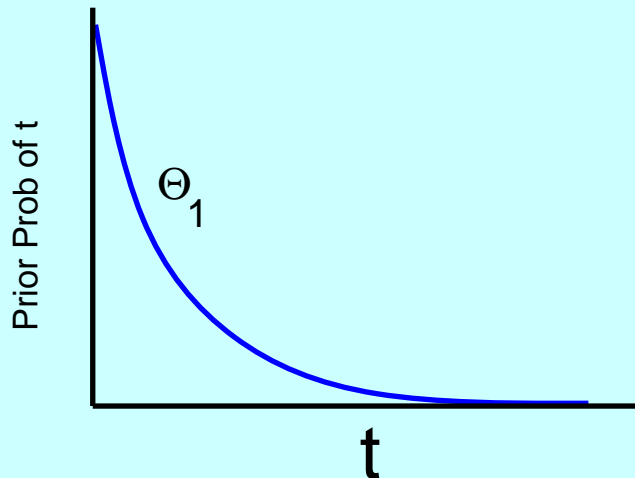
can reduce variability by looking at

- (i) more gene copies, or
- (ii) more loci

Computing the likelihood: averaging over coalescents

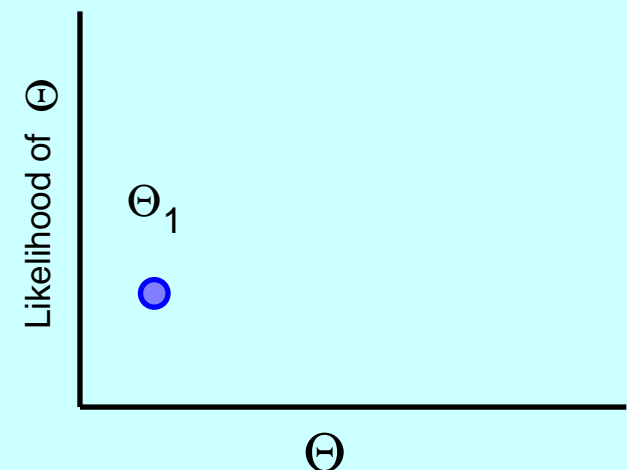
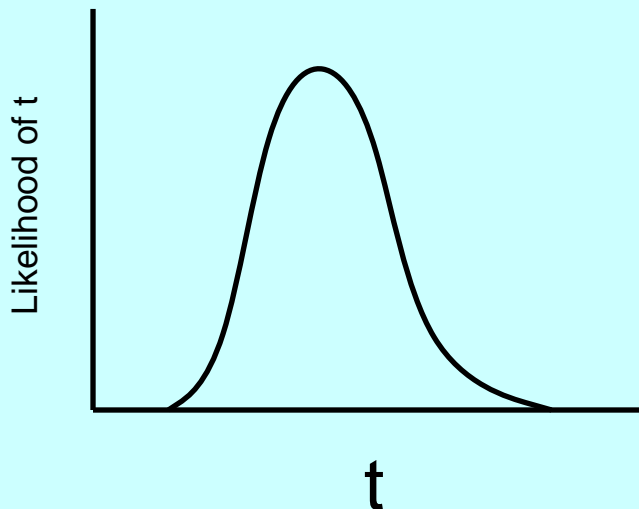
The likelihood calculation in a sample of two gene copies

The product of the prior on t ,



when integrated over all possible t 's, gives the likelihood for the underlying parameter Θ

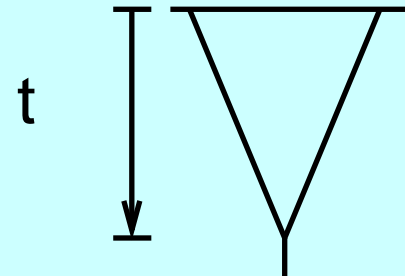
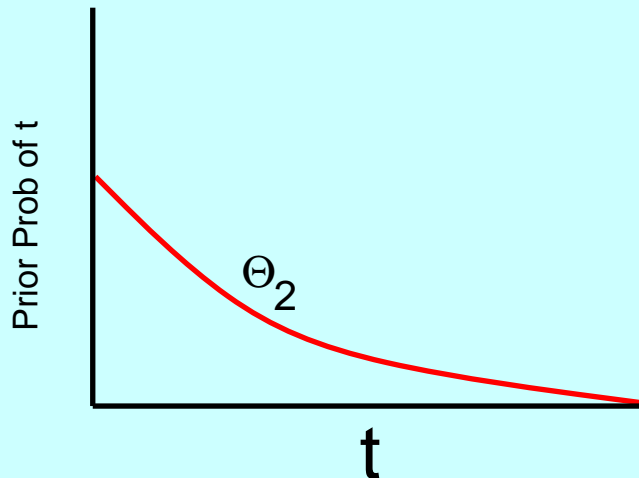
times the likelihood of that t from the data,



Computing the likelihood: averaging over coalescents

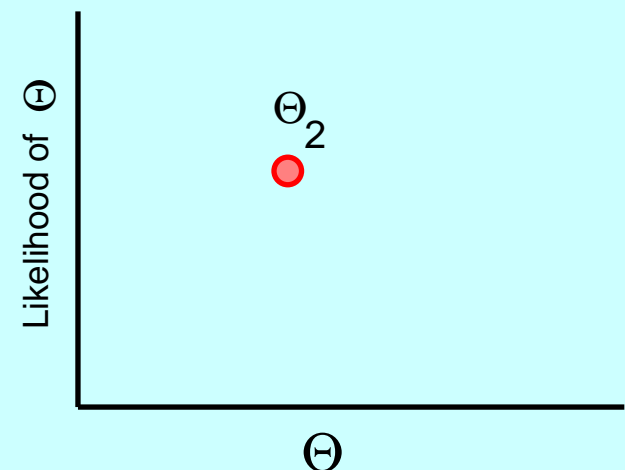
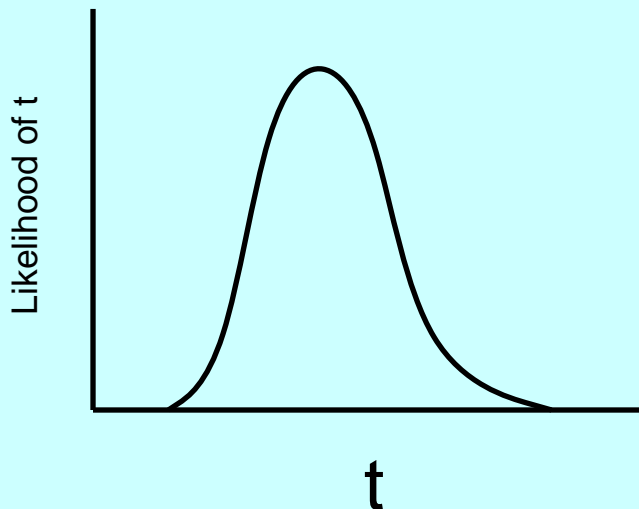
The likelihood calculation in a sample of two gene copies

The product of the prior on t ,



when integrated over all possible t 's, gives the likelihood for the underlying parameter Θ

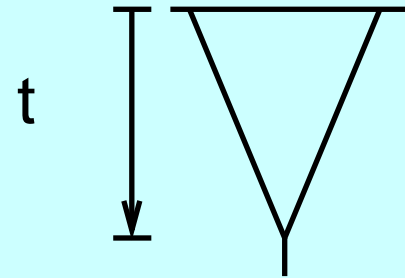
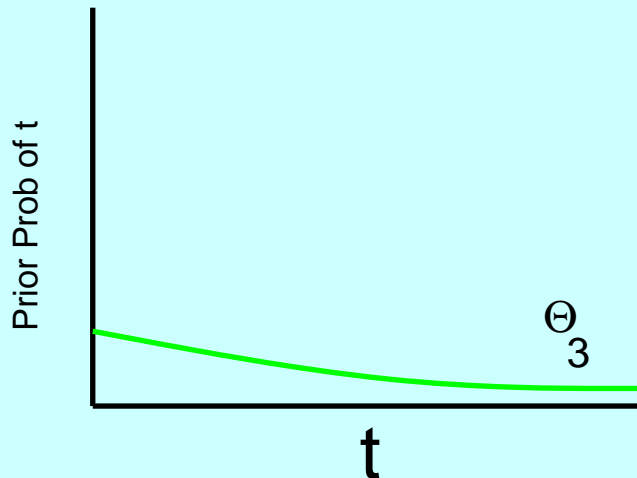
times the likelihood of that t from the data,



Computing the likelihood: averaging over coalescents

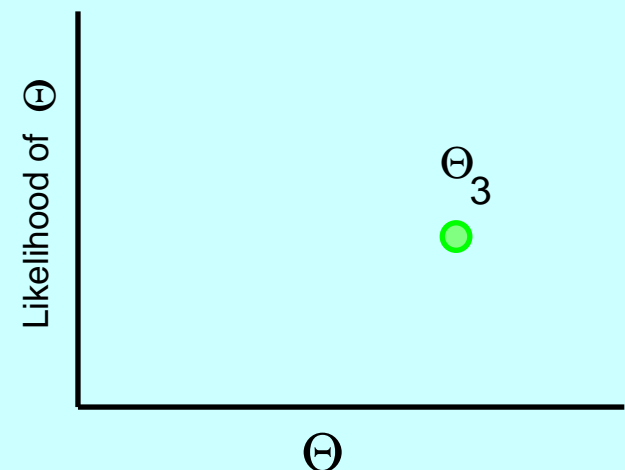
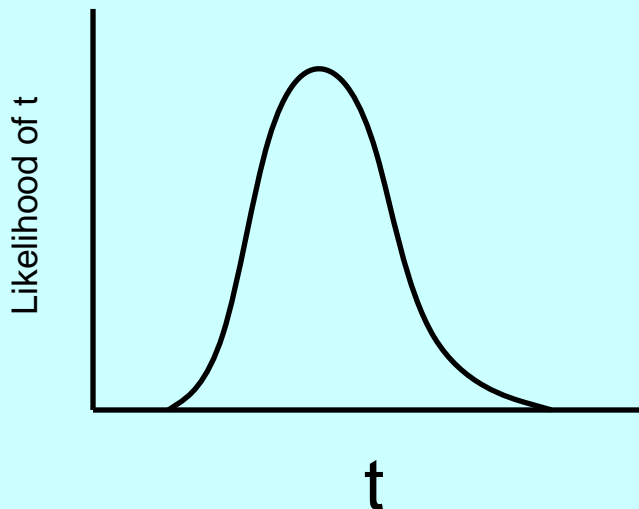
The likelihood calculation in a sample of two gene copies

The product of the prior on t ,



when integrated over all possible t 's, gives the likelihood for the underlying parameter Θ

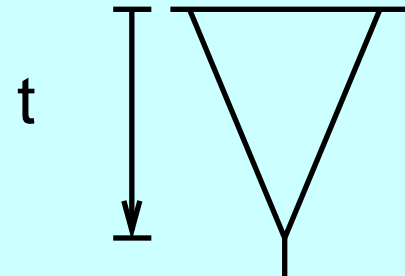
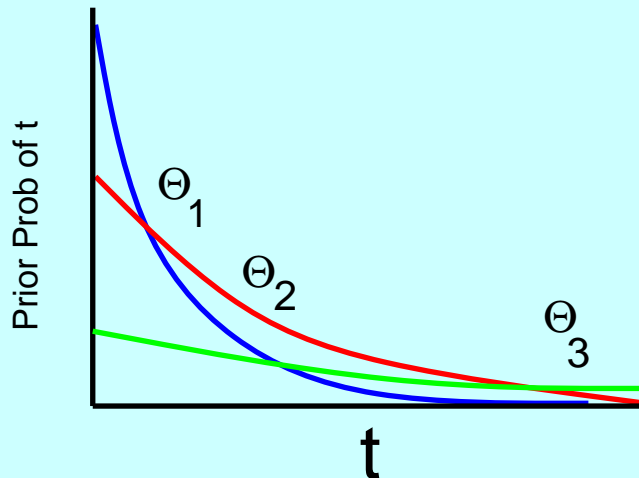
times the likelihood of that t from the data,



Computing the likelihood: averaging over coalescents

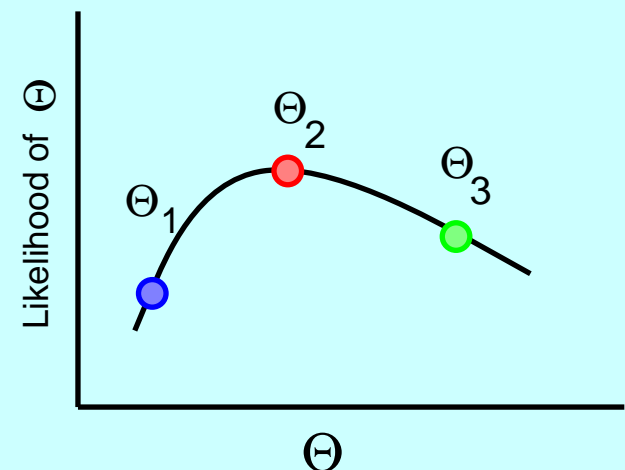
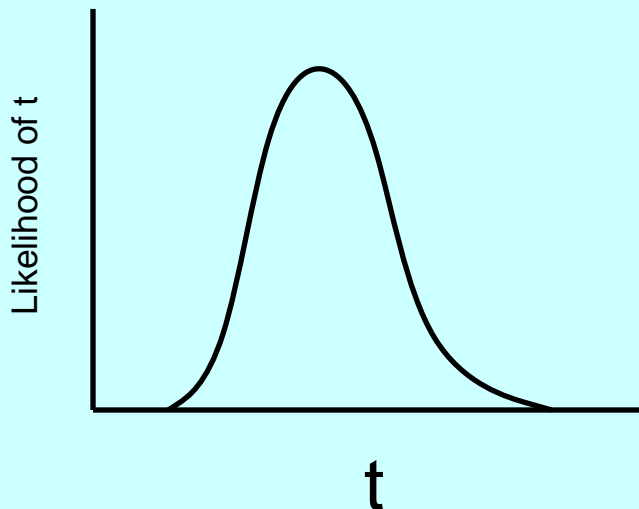
The likelihood calculation in a sample of two gene copies

The product of the prior on t ,



when integrated over all possible t 's, gives the likelihood for the underlying parameter Θ

times the likelihood of that t from the data,

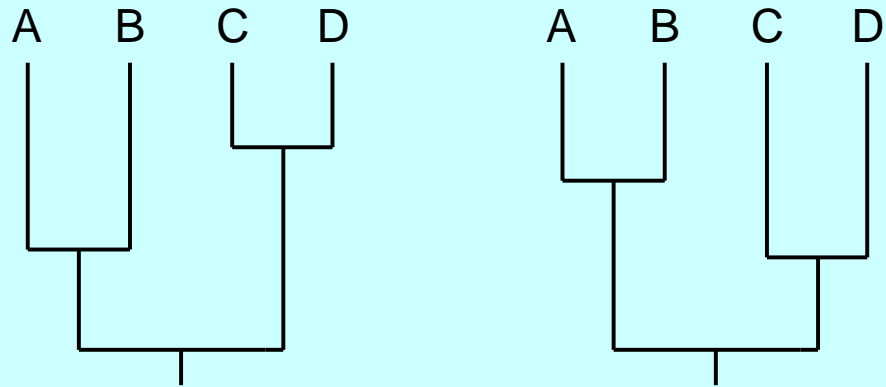


Labelled histories

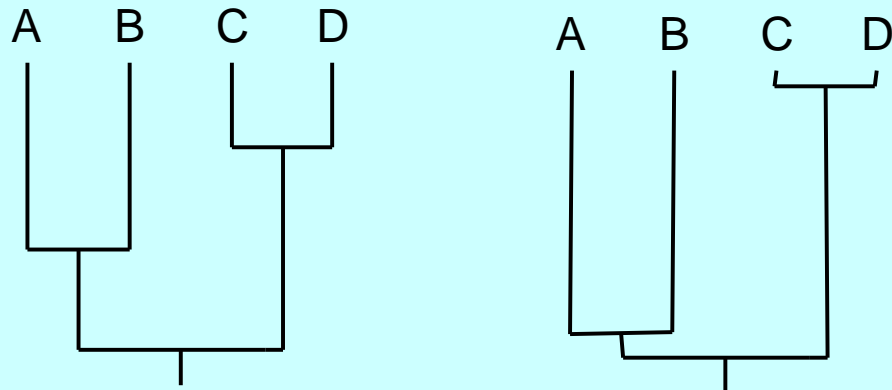
Labelled Histories (Edwards, 1970; Harding, 1971)

Trees that differ in the time-ordering of their nodes

These two are different:



These two are the same:



Sampling approaches to coalescent likelihood



Bob Griffiths



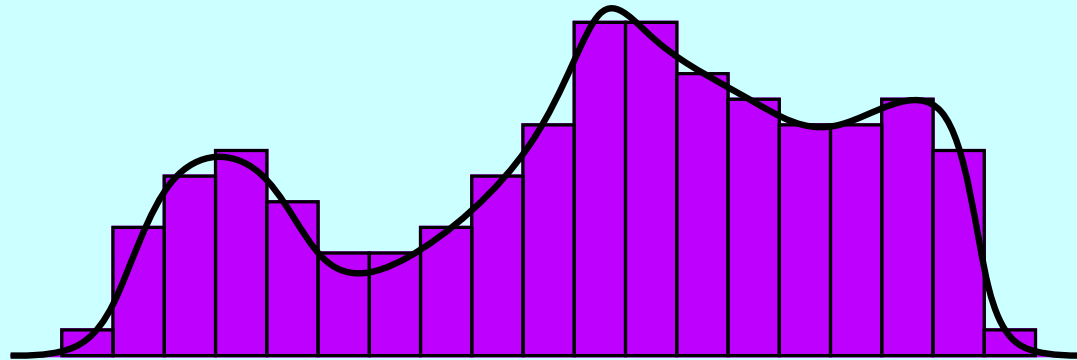
Simon Tavaré



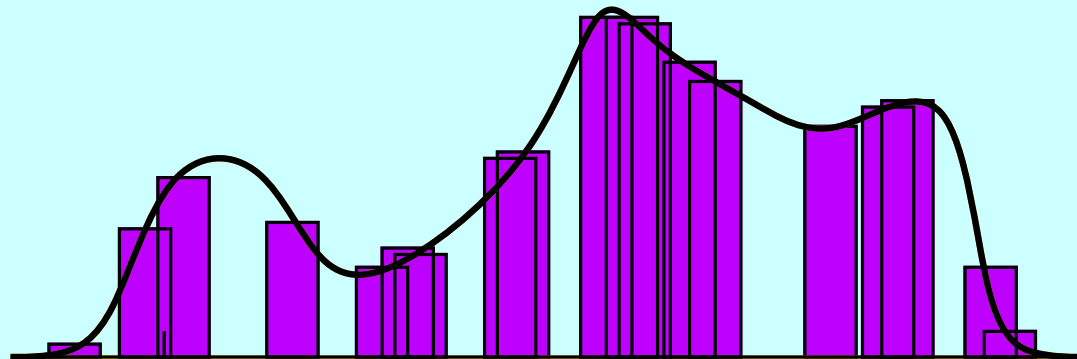
Mary Kuhner and Jon Yamato

Monte Carlo integration

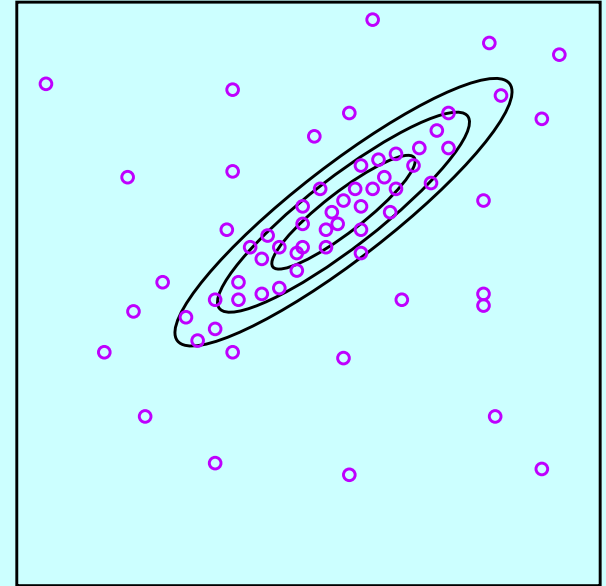
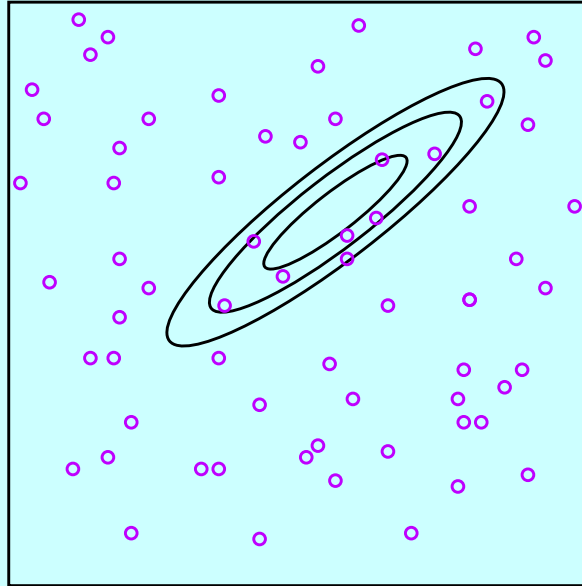
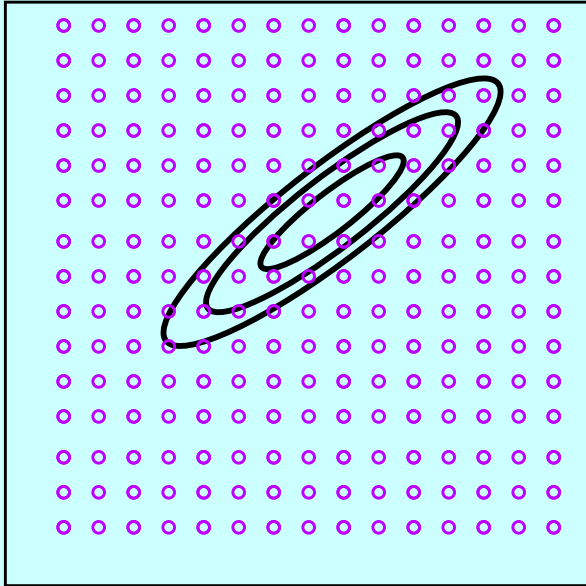
To get the area under a curve, we can either evaluate the function ($f(x)$) at a series of grid points and add up heights \times widths:



or we can sample at random the same number of points, add up height \times width:



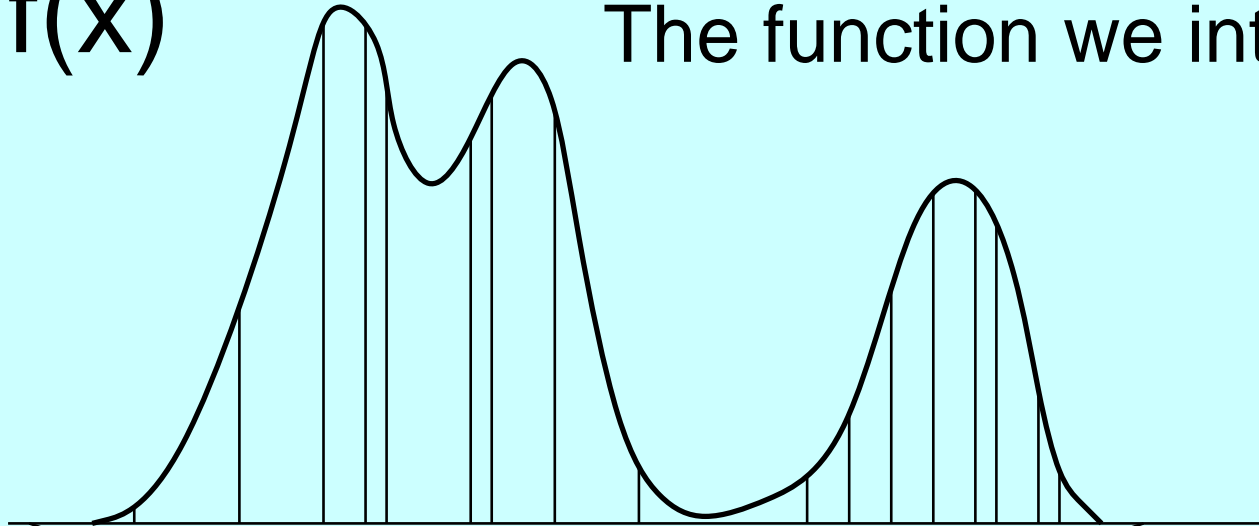
Importance sampling



Importance sampling

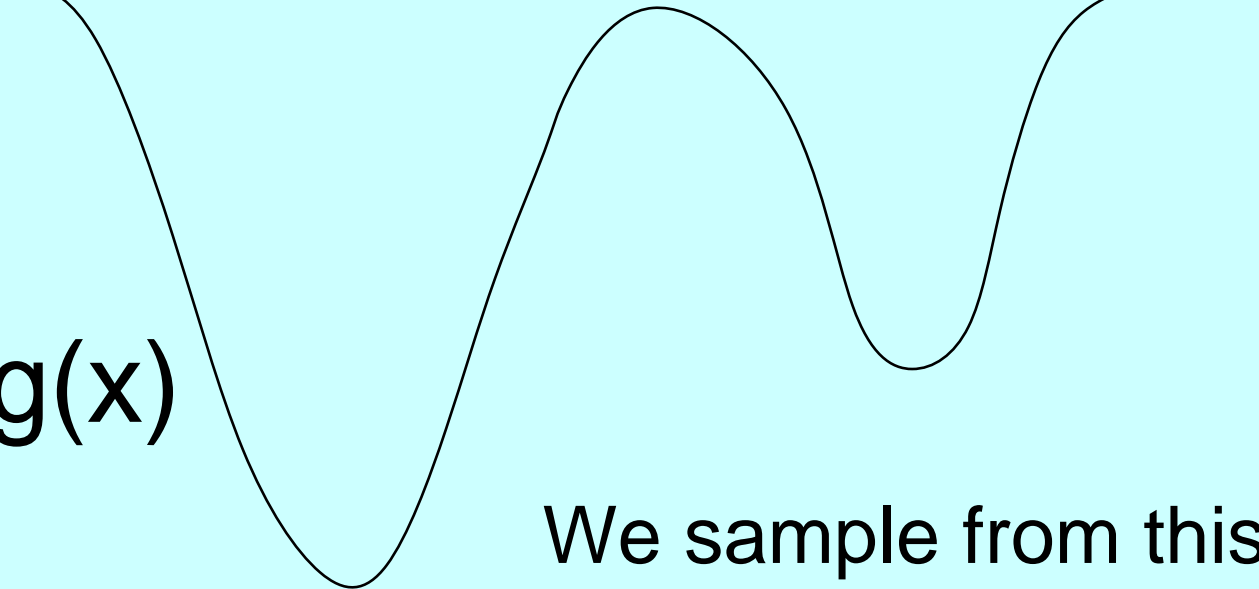
$f(x)$

The function we integrate



$g(x)$

We sample from this density



The math of importance sampling

$$\begin{aligned}\int f(x) dx &= \int \frac{f(x)}{g(x)} g(x) dx \\ &= \mathbb{E}_g \left[\frac{f(x)}{g(x)} \right]\end{aligned}$$

which is the expectation for points sampled from $g(x)$ of the ratio $\frac{f(x)}{g(x)}$.

This is approximated by sampling a lot (n) of points from $g(x)$ and the computing the average:

$$L = \frac{1}{n} \sum_{i=1}^n \frac{f(x_i)}{g(x_i)}$$

The importance function used in LAMARC

In Mary Kuhner and Jon Yamato's program LAMARC they use as the importance function the probability density of the tree given the data at a set of "driving values" θ_0 of the parameters:

$$f(G) = \frac{\text{Prob}(D | G) \text{Prob}(G | \theta_0)}{\text{Prob}(D | \theta_0)}$$

The denominator is impossible to evaluate but as we will see, isn't really needed.

The resulting likelihood ratio is

$$\frac{L(\Theta)}{L(\Theta_0)} = \frac{1}{n} \sum_{i=1}^n \frac{\text{Prob}(G_i | \Theta)}{\text{Prob}(G_i | \Theta_0)}$$

Markov Chain Monte Carlo (MCMC) methods

To do the importance sampling, MCMC methods are employed (in all programs that do full likelihood or Bayesian analyses).

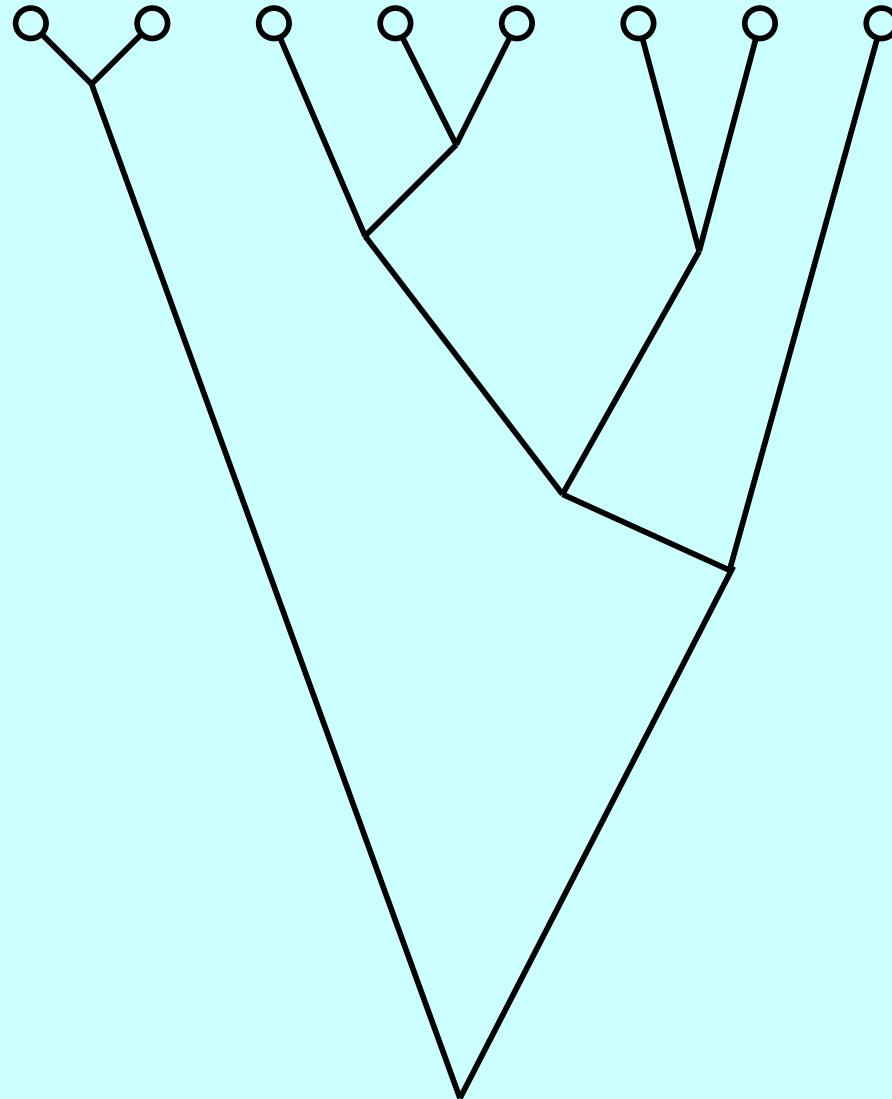
To sample from $f(G)$, start with a tree G_{old} and

1. Have a “proposal distribution” from which you sample a new tree G_{new}
2. Compute the function $f(G_{new})$ (we have that also for the old tree)
3. Draw a random fraction R between 0 and 1
4. If $R < \frac{f(G_{new})}{f(G_{old})}$, accept the new tree. (Note that in that ratio any horrible, but shared, denominators cancel out).

repeat this vast numbers of times (the correct number of times is infinity).

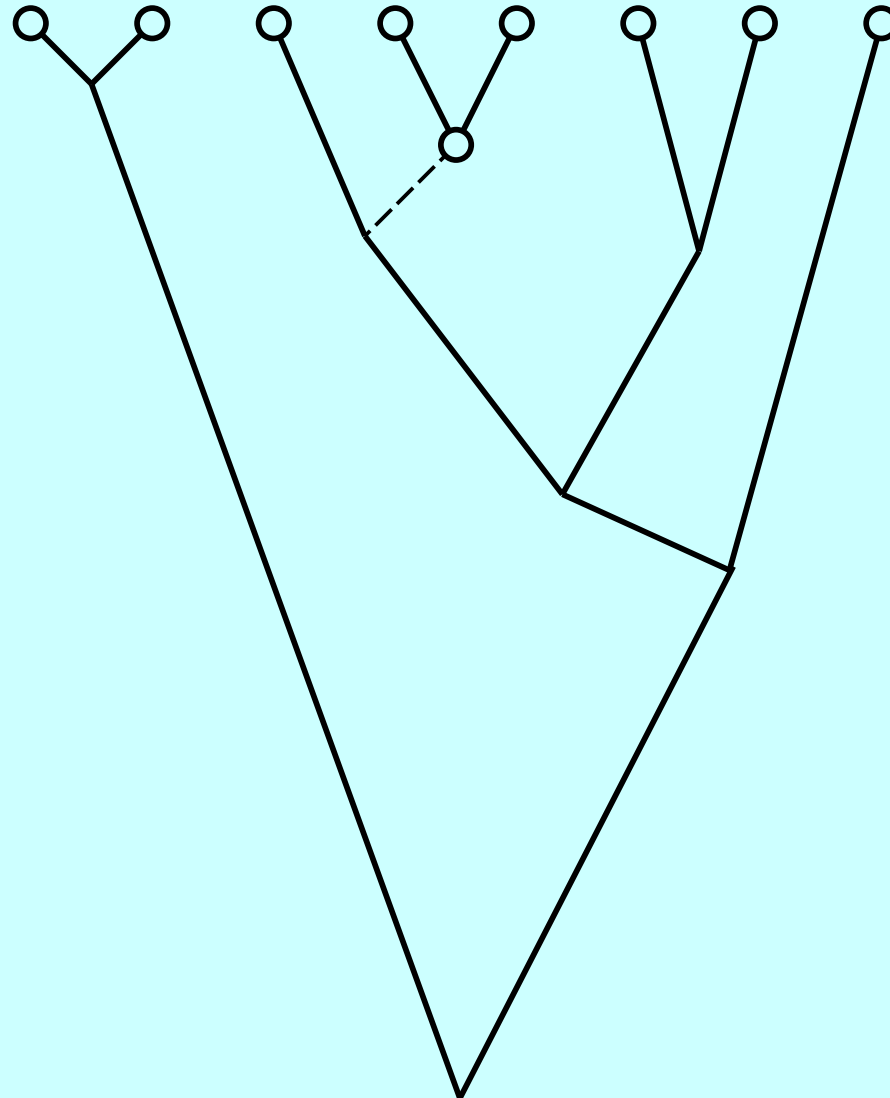
Rearrangement to sample points in tree space

A conditional coalescent rearrangement strategy



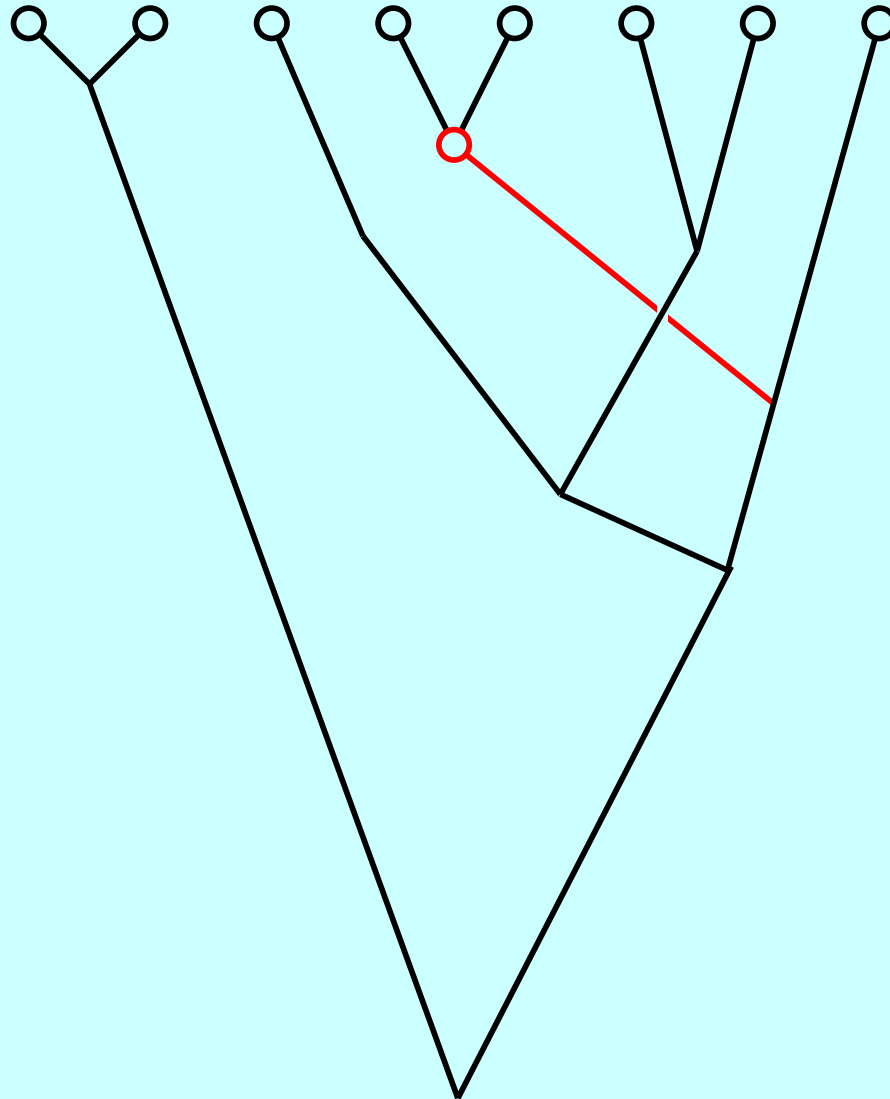
Dissolving a branch and regrowing it backwards

First pick a random node (interior or tip) and remove its subtree



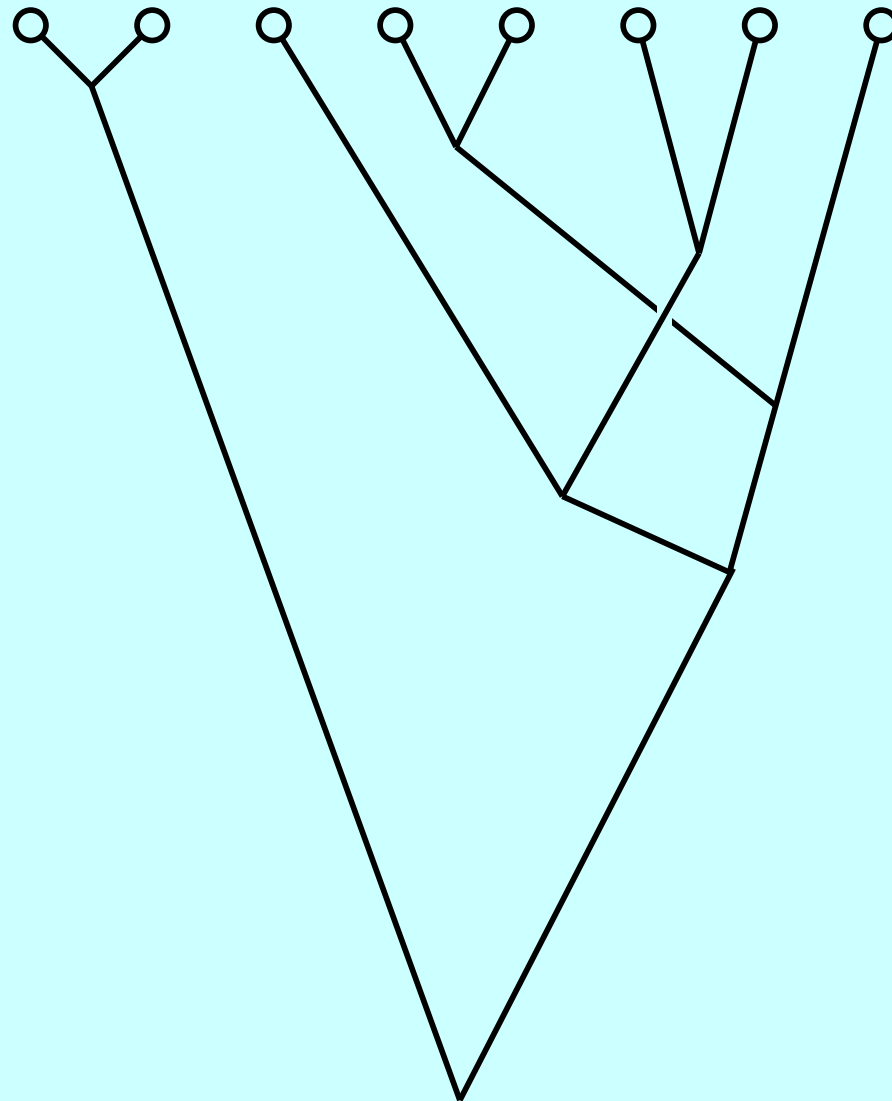
We allow it coalesce with the other branches

Then allow this node to re-coalesce with the tree



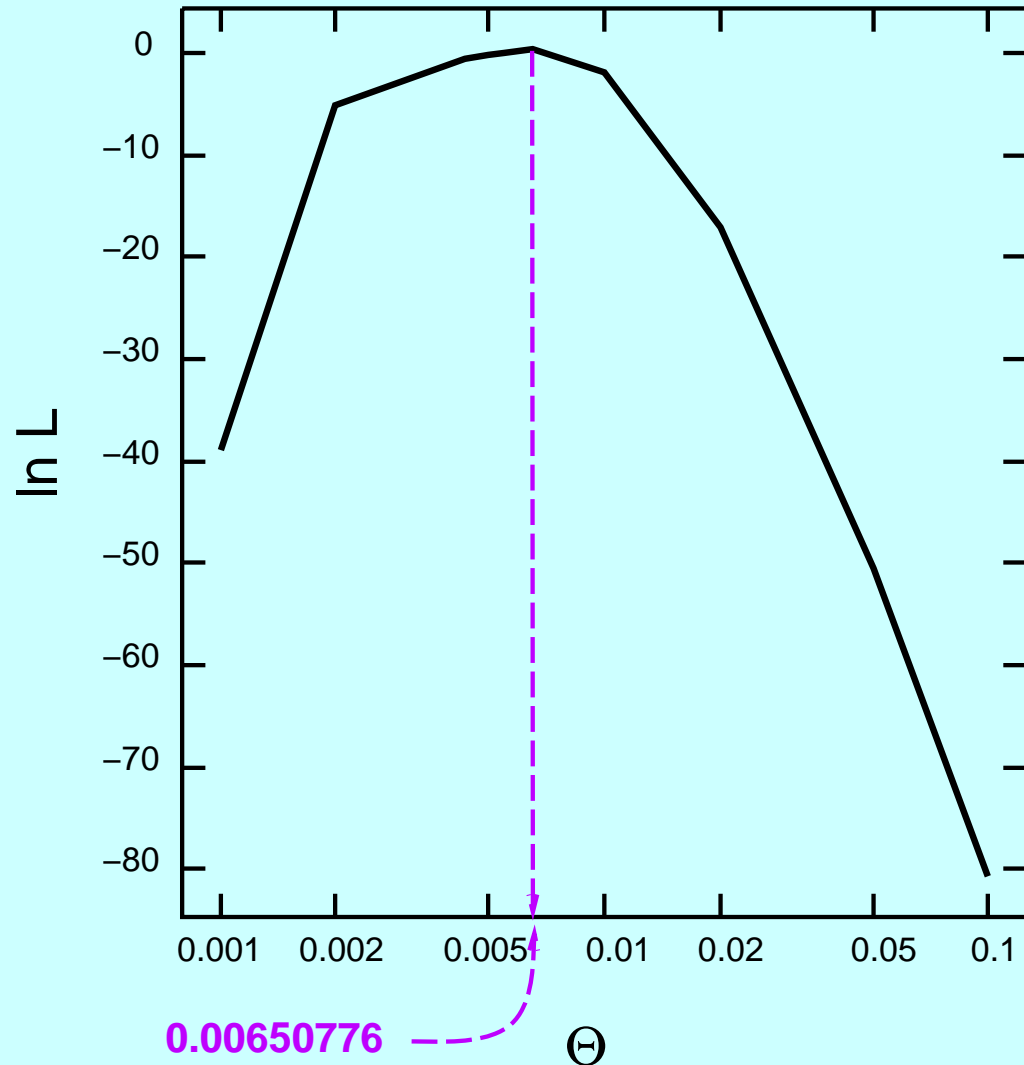
and this gives another coalescent

The resulting tree proposed by this process



An example of an MCMC likelihood curve

Results of analysing a data set with 50 sequences of 500 bases which was simulated with a true value of $\Theta = 0.01$



Major MCMC likelihood or Bayesian programs

- **LAMARC** by Mary Kuhner and Jon Yamato and others. Likelihood inference with multiple populations, recombination, migration, population growth. No historical branching events or serial sampling, yet.
- **BEAST** by Andrew Rambaut, Alexei Drummond and others. Bayesian inference with multiple populations related by a tree. Support for serial sampling. Recently got some support for migration. (No recombination yet).
- **IM** and **IMA2** by Rasmus Nielsen and Jody Hey. Two or more populations allowing both historical splitting and migration after that. No recombination yet.
- **genetree** by Bob Griffiths and Melanie Bahlo. Likelihood inference of migration rates and changes in population size. No recombination or historical branching events.
- **migrate** by Peter Beerli. Likelihood inference with multiple populations and migration rates. No recombination or historical branching events yet.

Approximately Bayesian Computation (ABC) methods

These involve approximating the sampling by computing some “summary statistics” from the data, then finding parameter values that, in a simulation of a tree and data, result in summary statistic values close to these.

They are faster, and very popular now.

But ... they are very dependent on getting the right summary statistics so as not to lose too much power compared to fully-powerful likelihood or Bayesian MCMC methods.