Week 10: Coalescents, Consensus trees, etc.

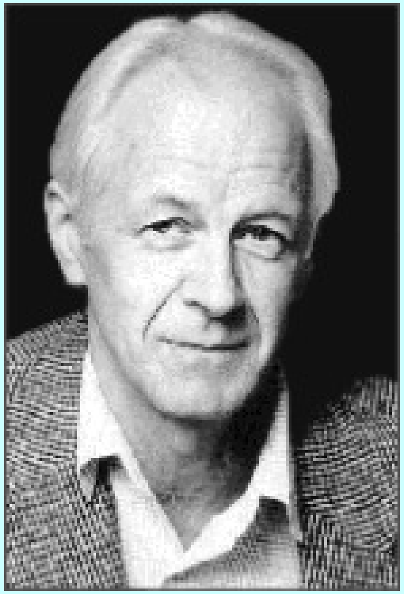
Genome 570

March, 2016

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

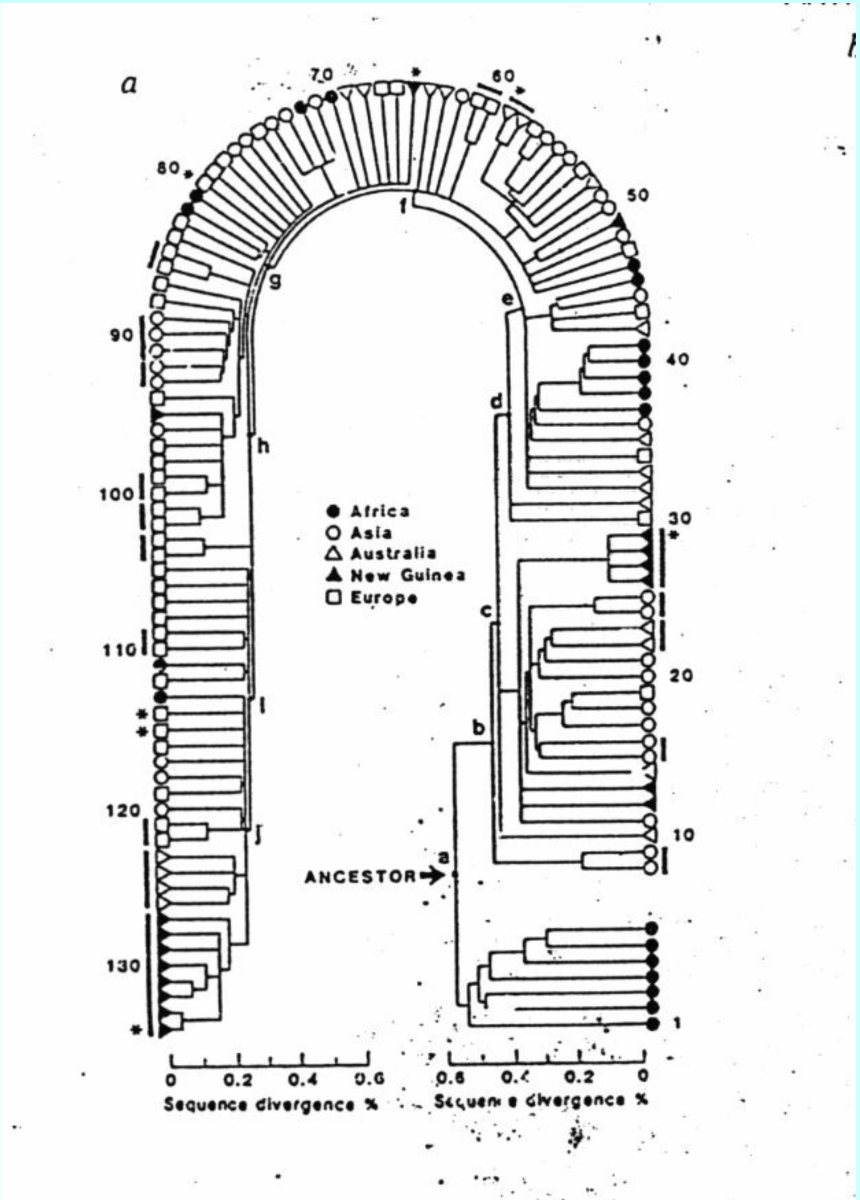
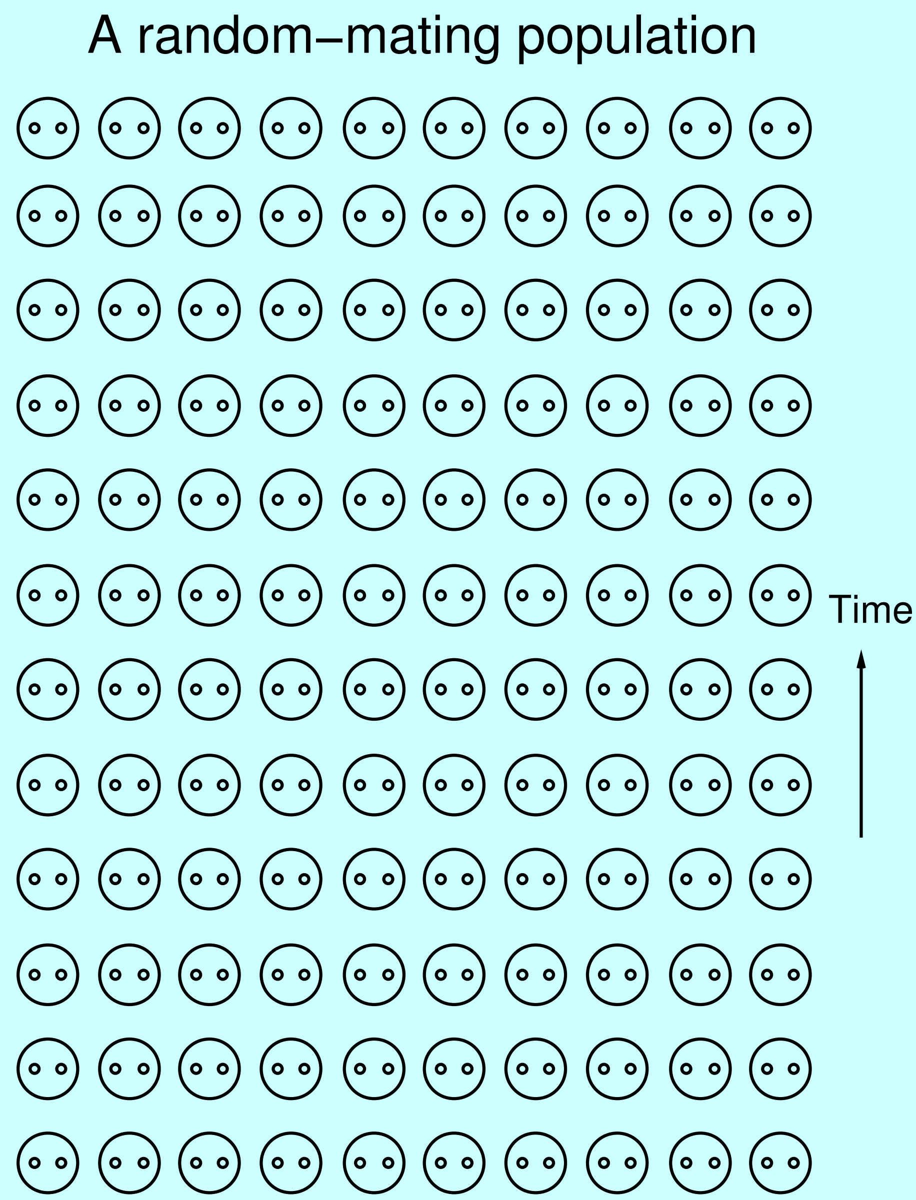
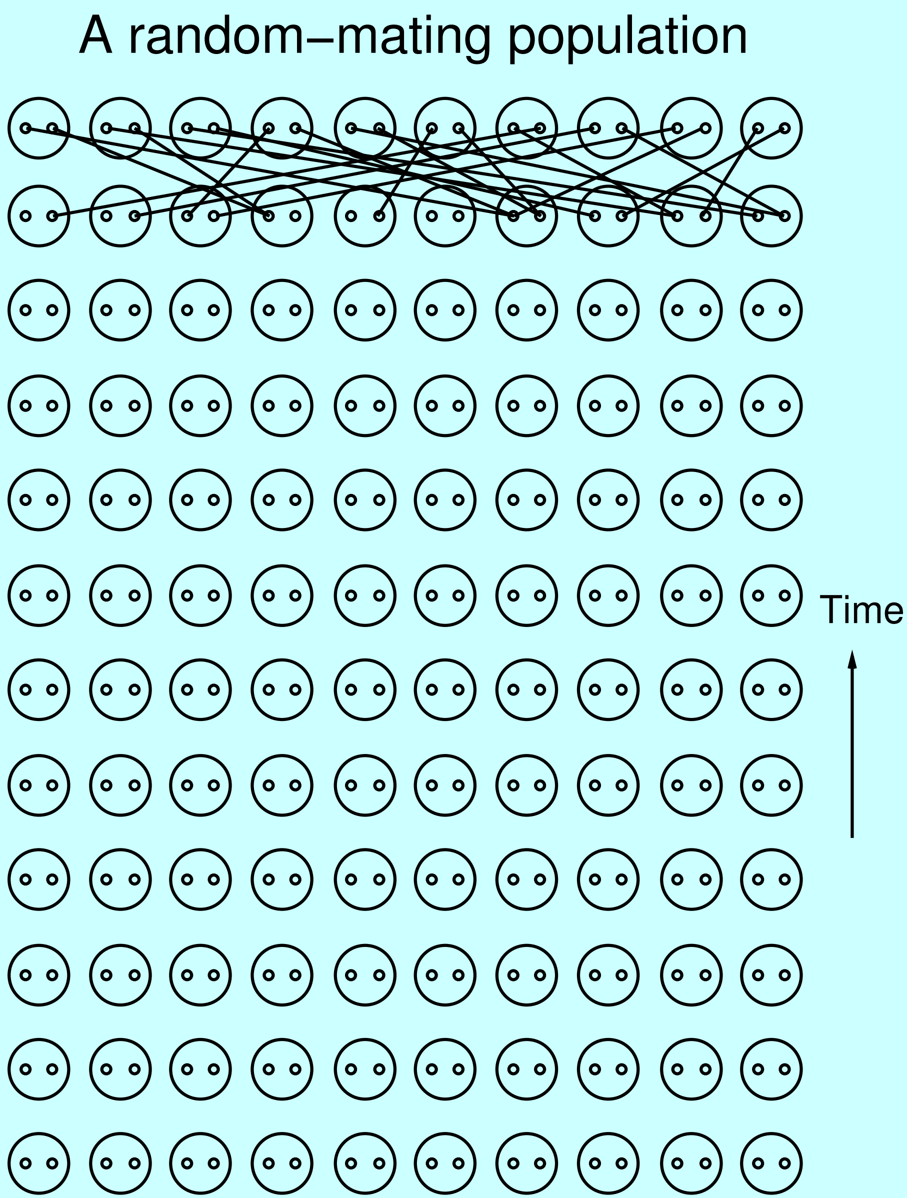


Fig. 3 *a*, Genealogical tree for 134 typ… of human mtDNA (133 restric
restriction sites used. The tree accounts f… the site differences observ

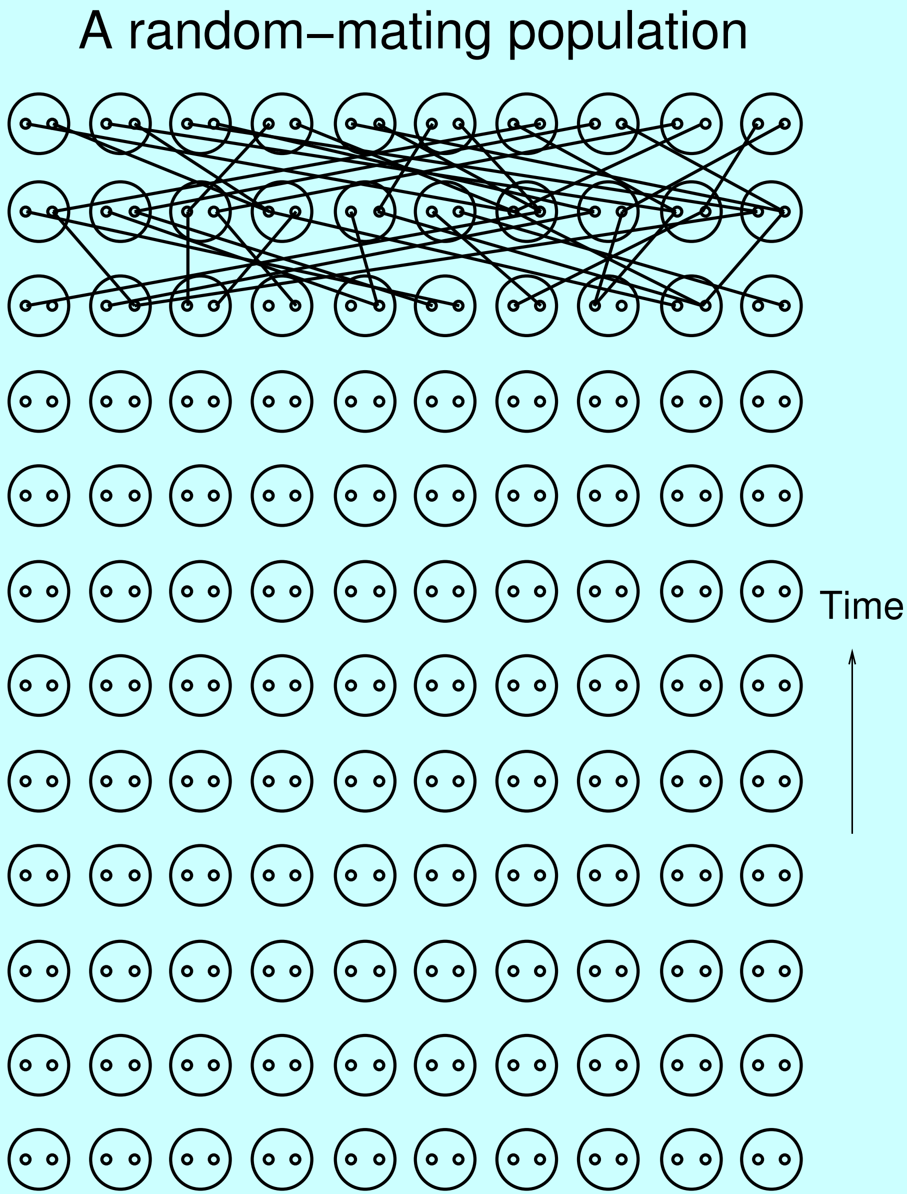
Gene copies in a population of 10 individuals



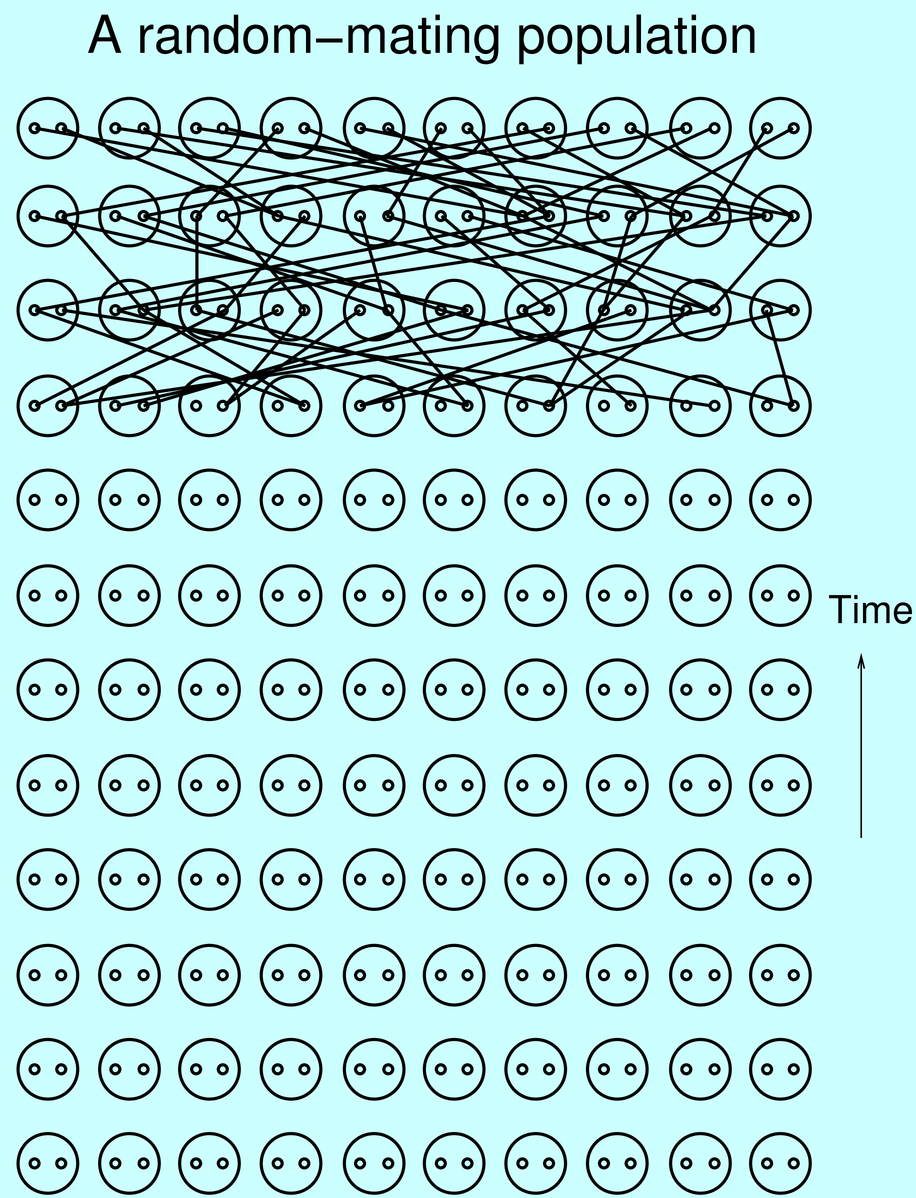
Going back one generation



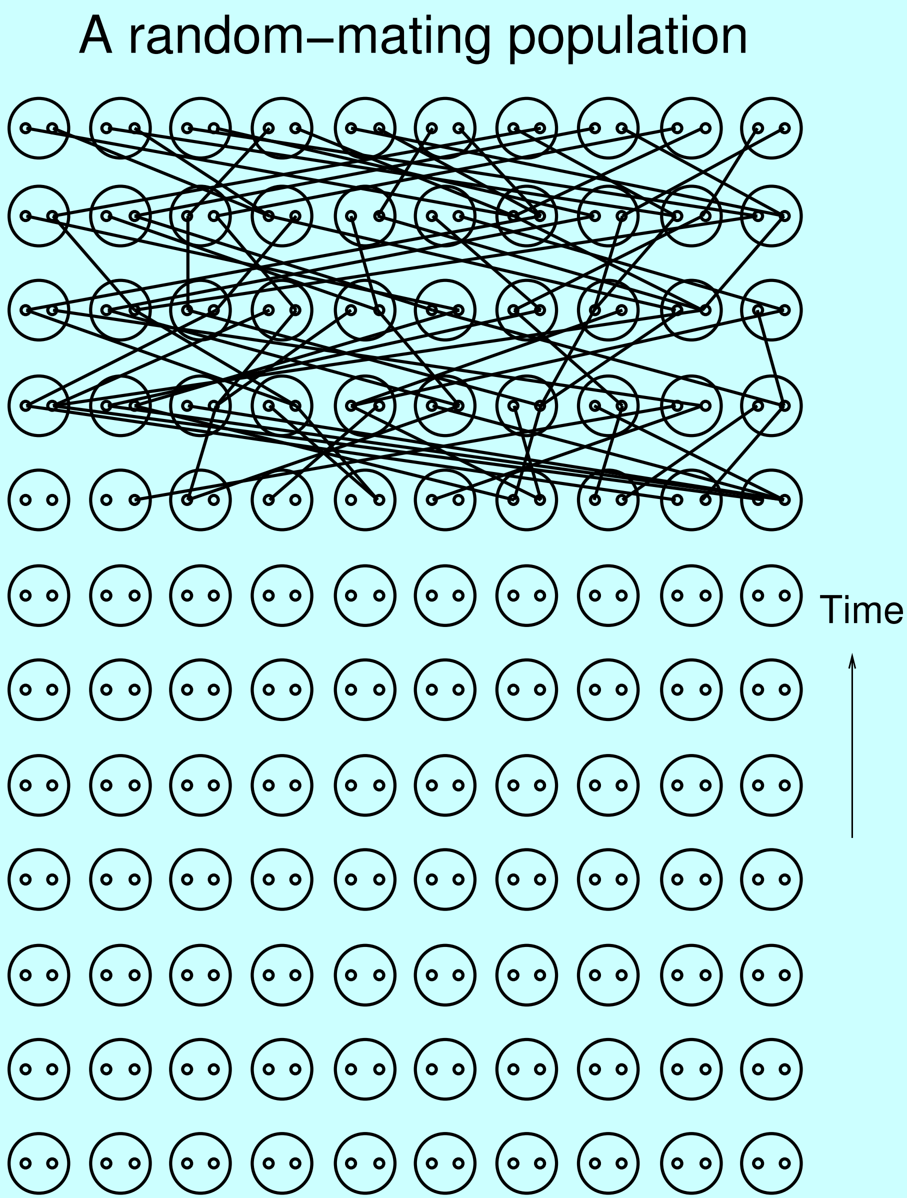
... and one more



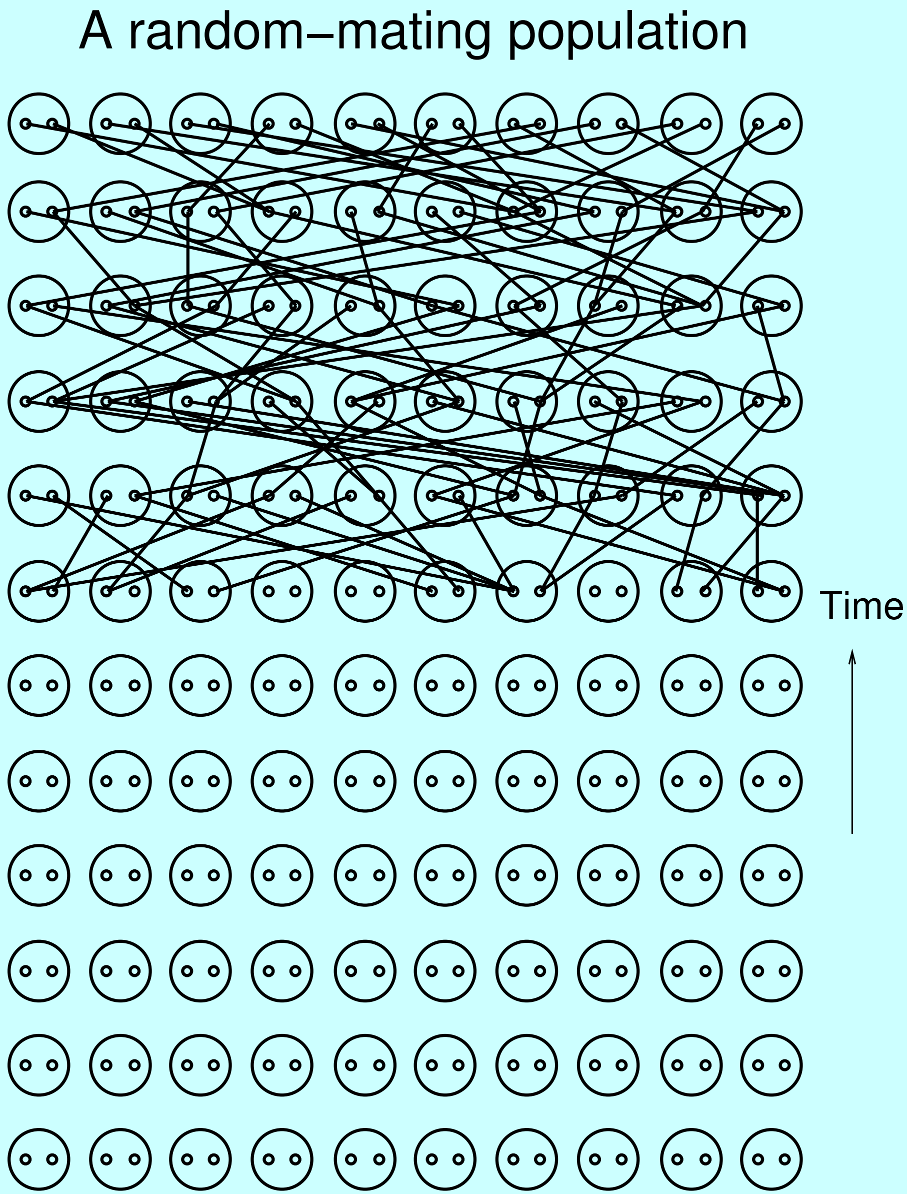
... and one more



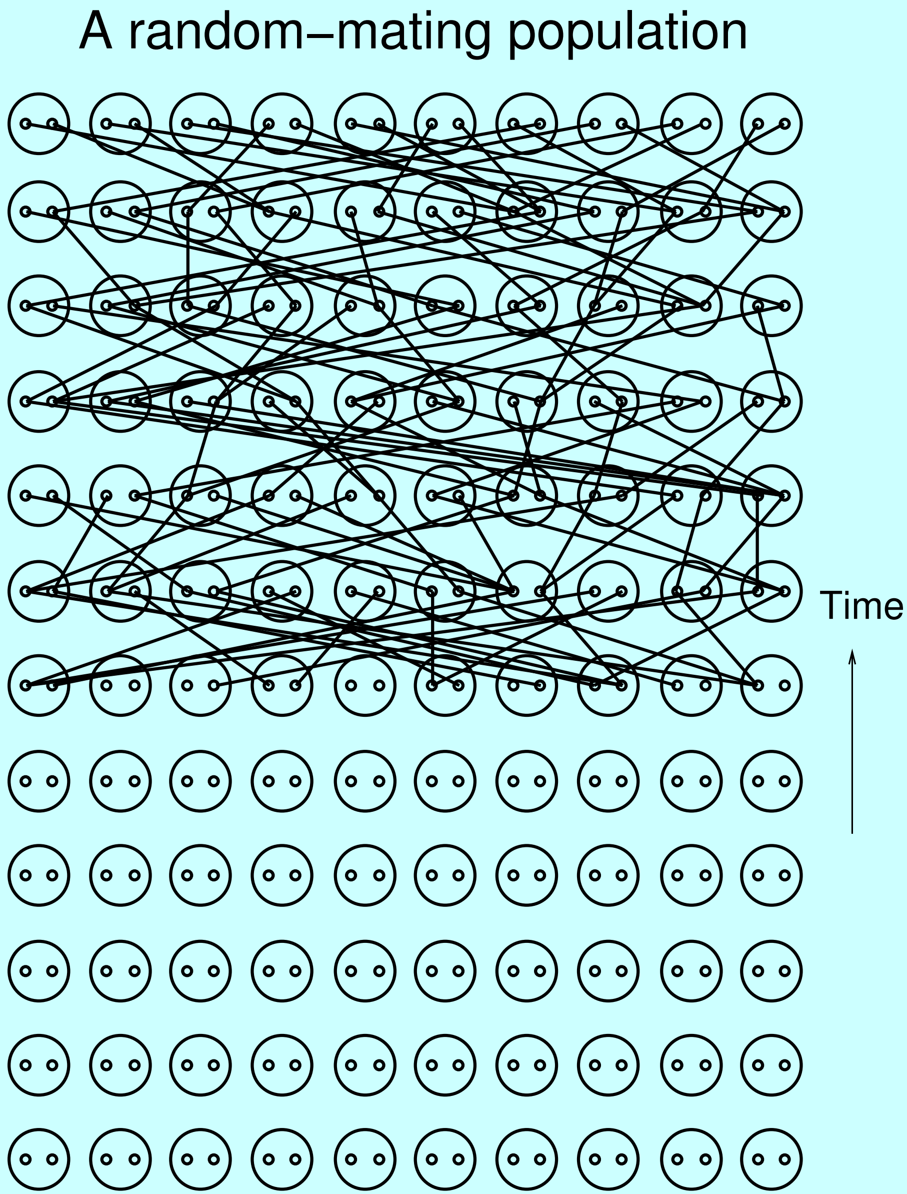
... and one more



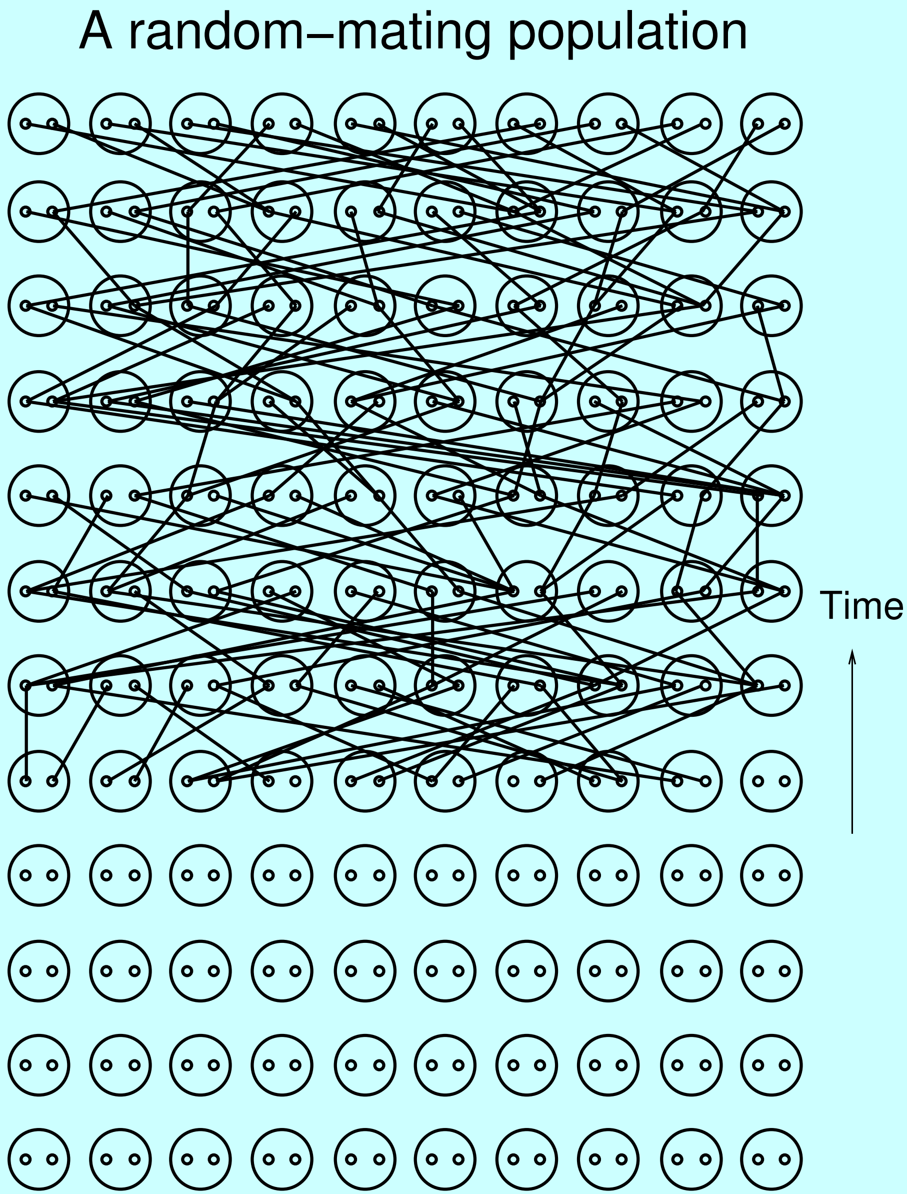
... and one more



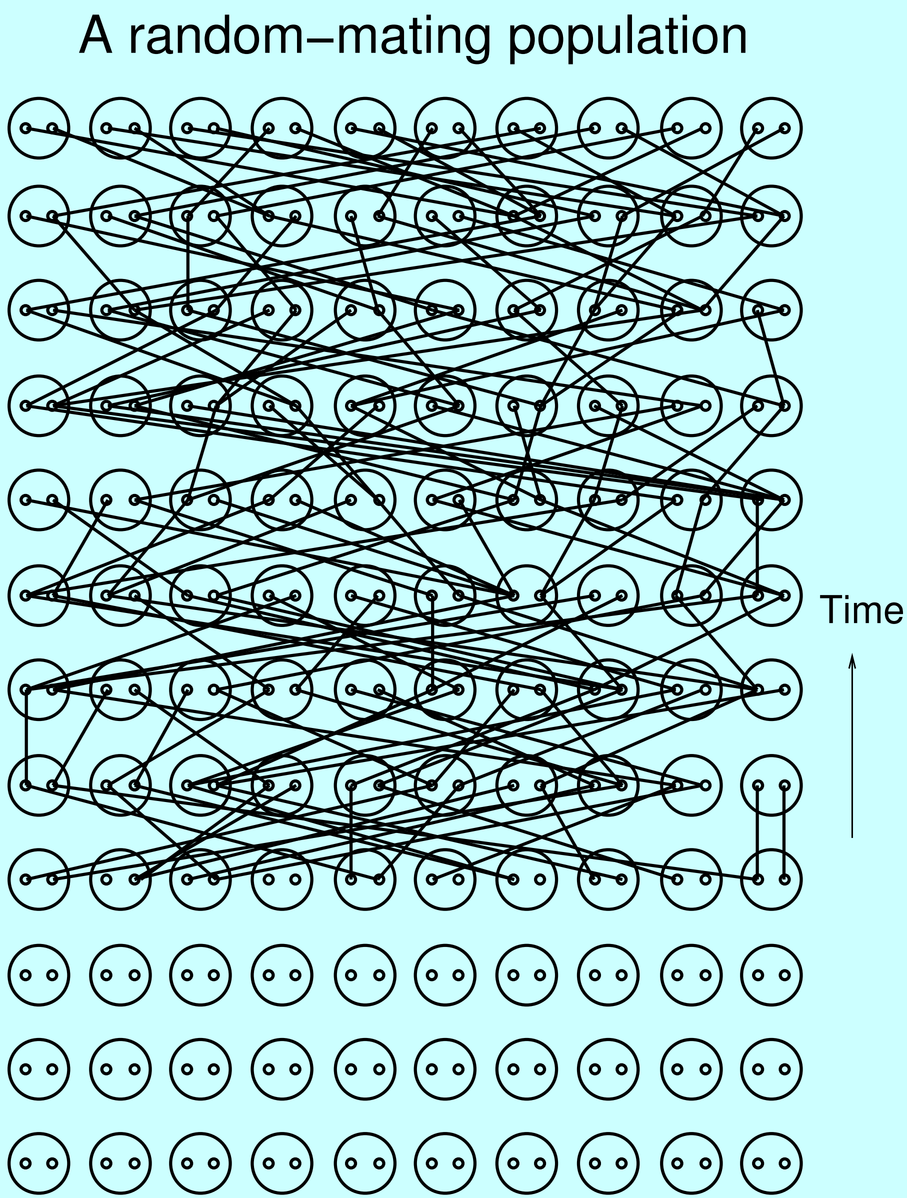
... and one more



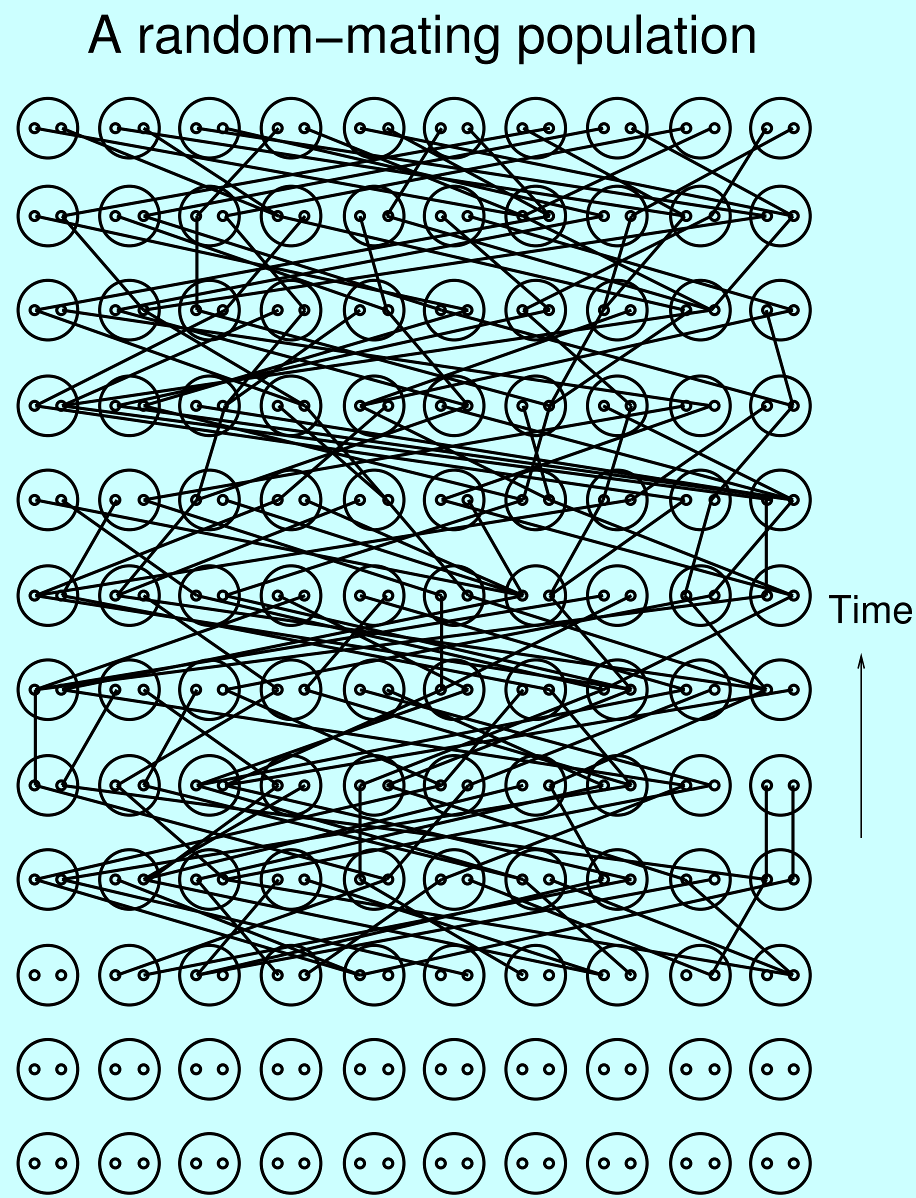
... and one more



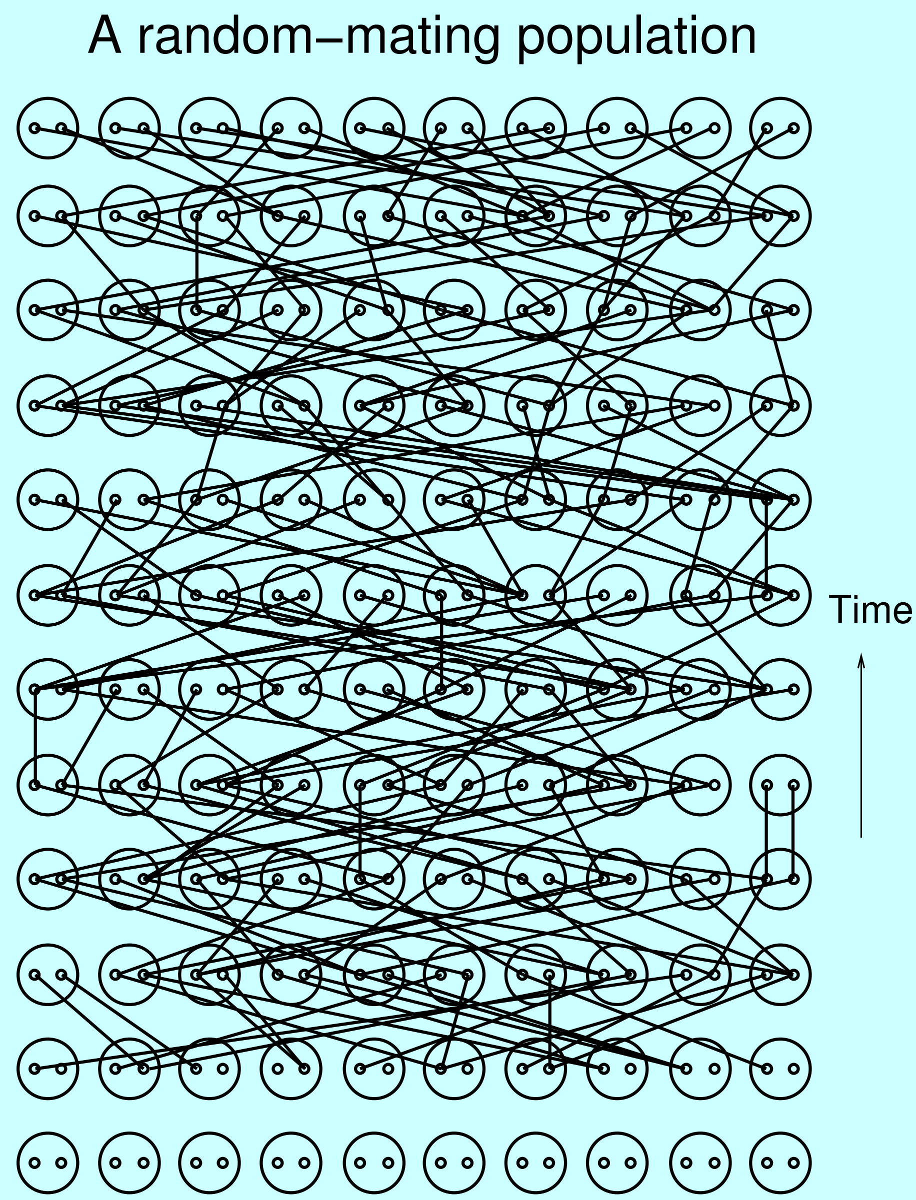
... and one more



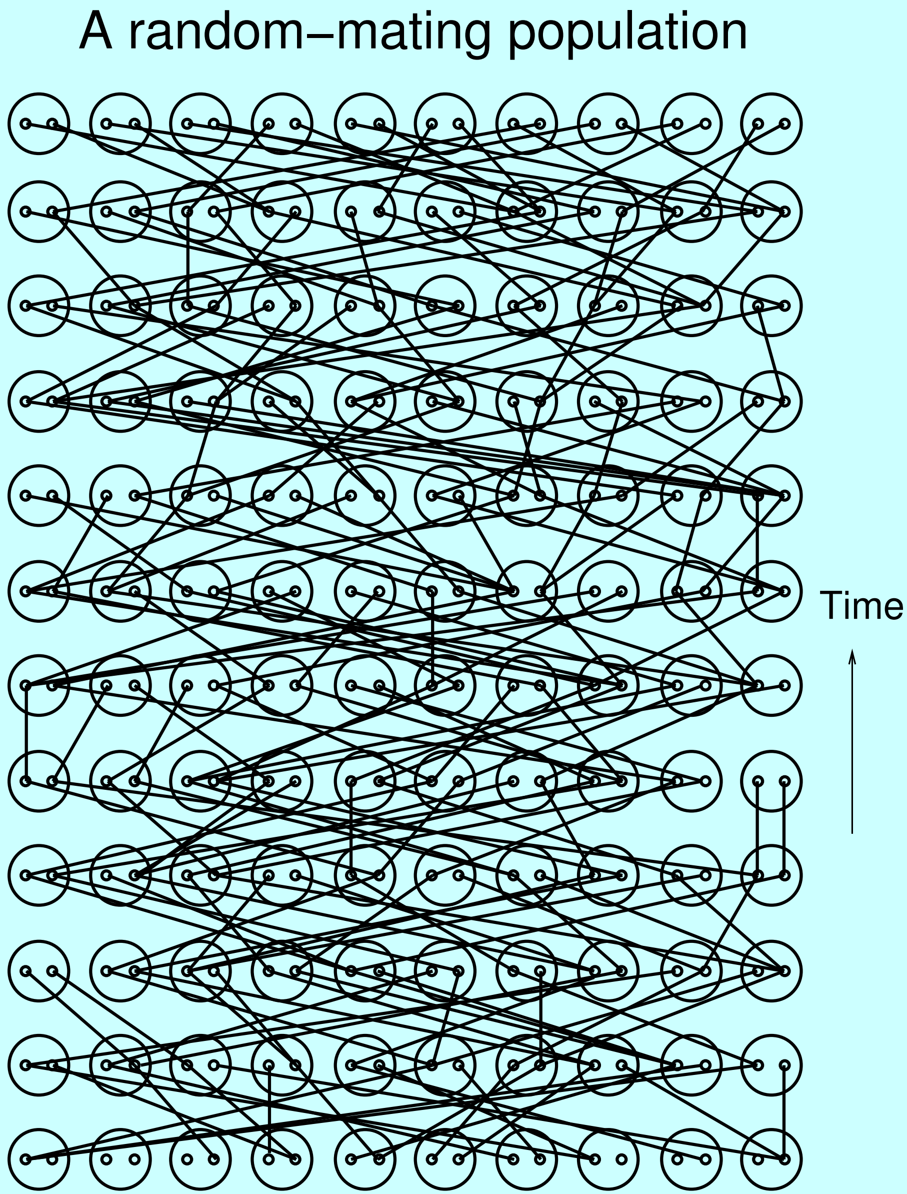
... and one more



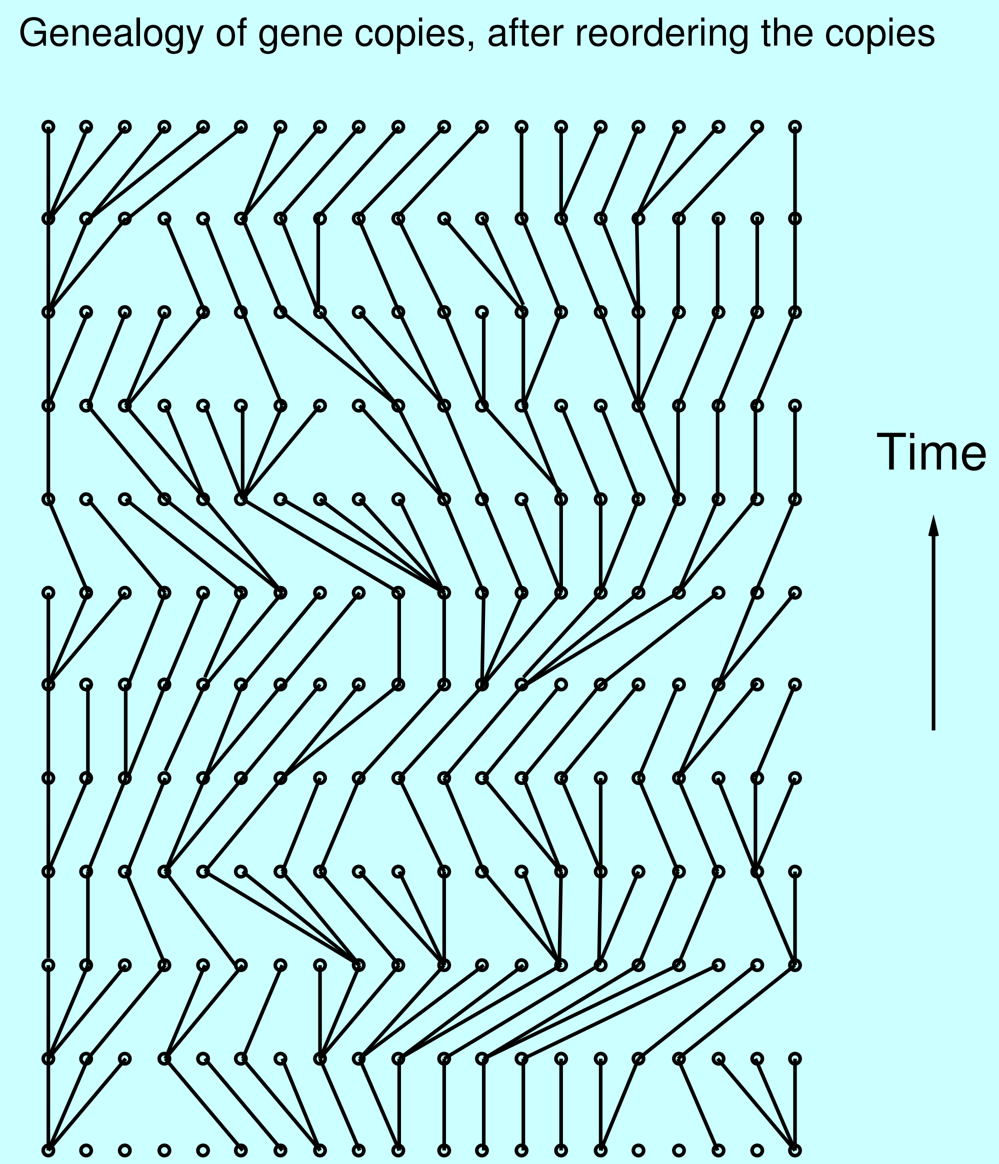
... and one more



... and one more

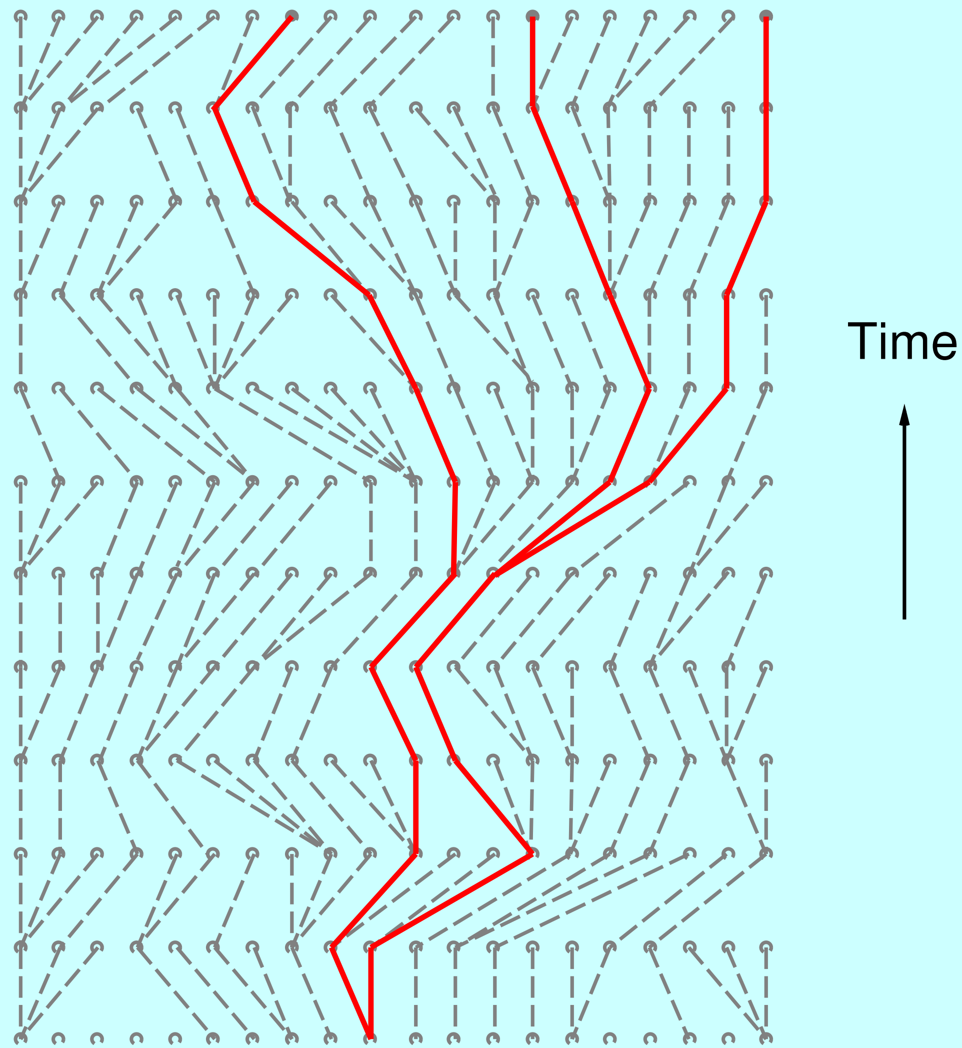


The genealogy of gene copies is a tree

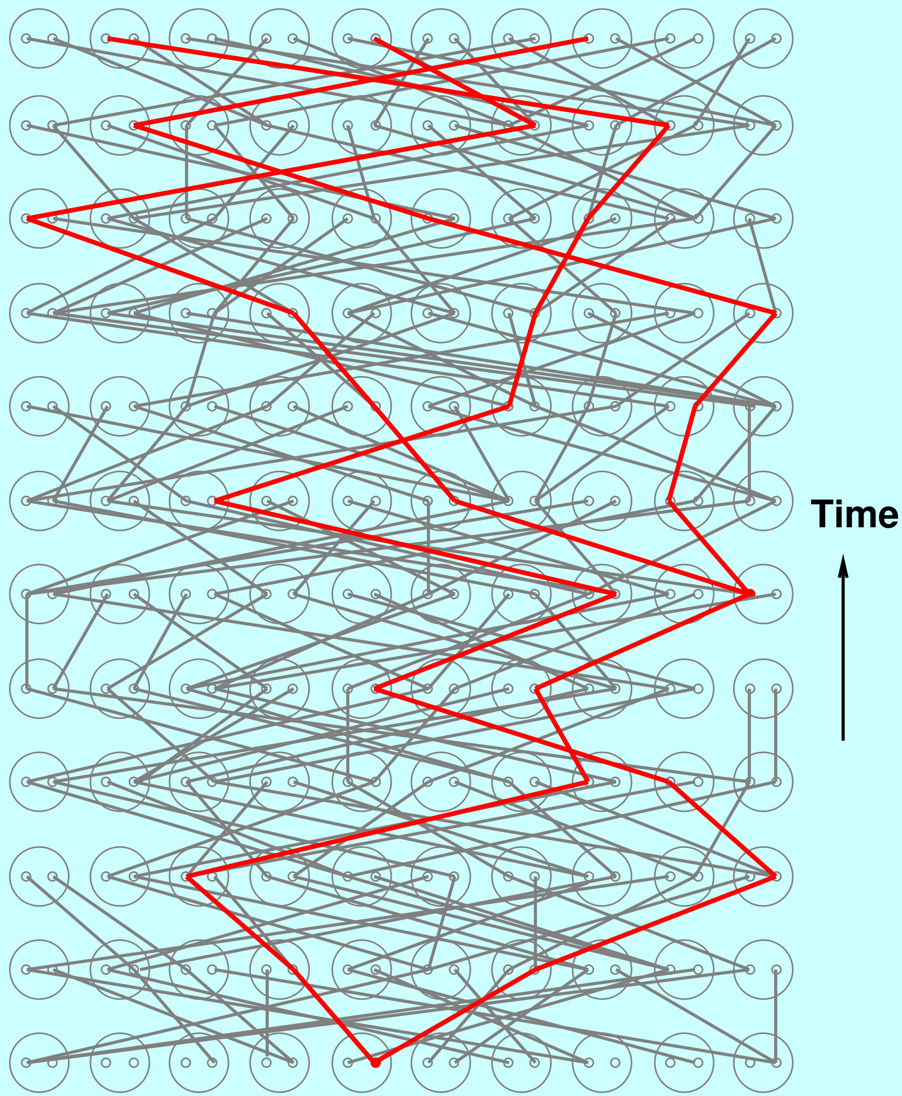


Ancestry of a sample of 3 copies

Genealogy of a small sample of genes from the population



Here is that tree of 3 copies in the pedigree



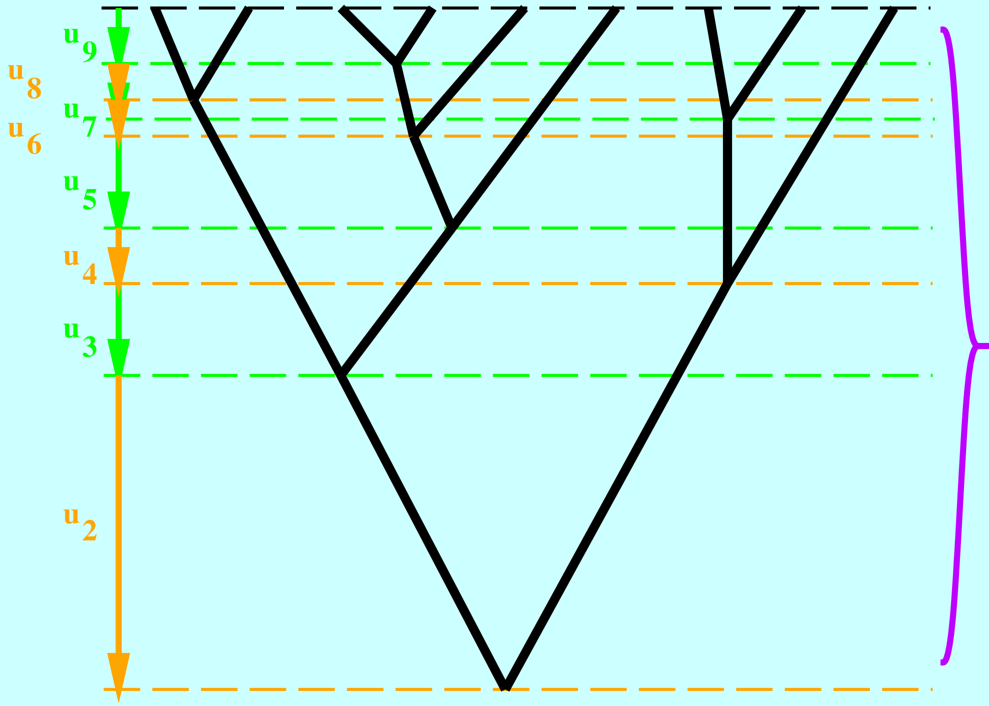
Kingman's coalescent

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

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 $= 4N\left(1 - \dfrac{1}{n}\right)$ generations

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 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 $k-1$ one generation earlier turns out to be (as each lineage "chooses" its ancestor independently):

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

(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:

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

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

so that the total probability that a pair coalesces is

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

Can probabilities of two or more lineages coalescing

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

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

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

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

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 $1/N$ 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.

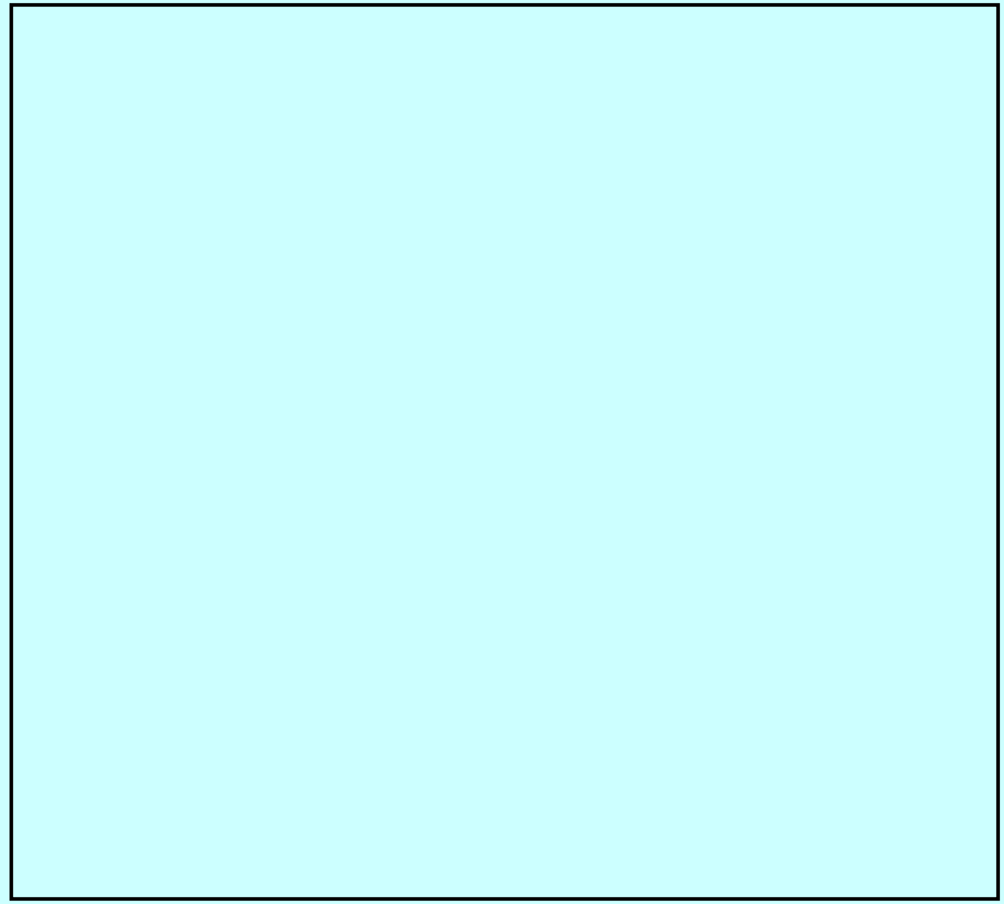
The coalescent

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.

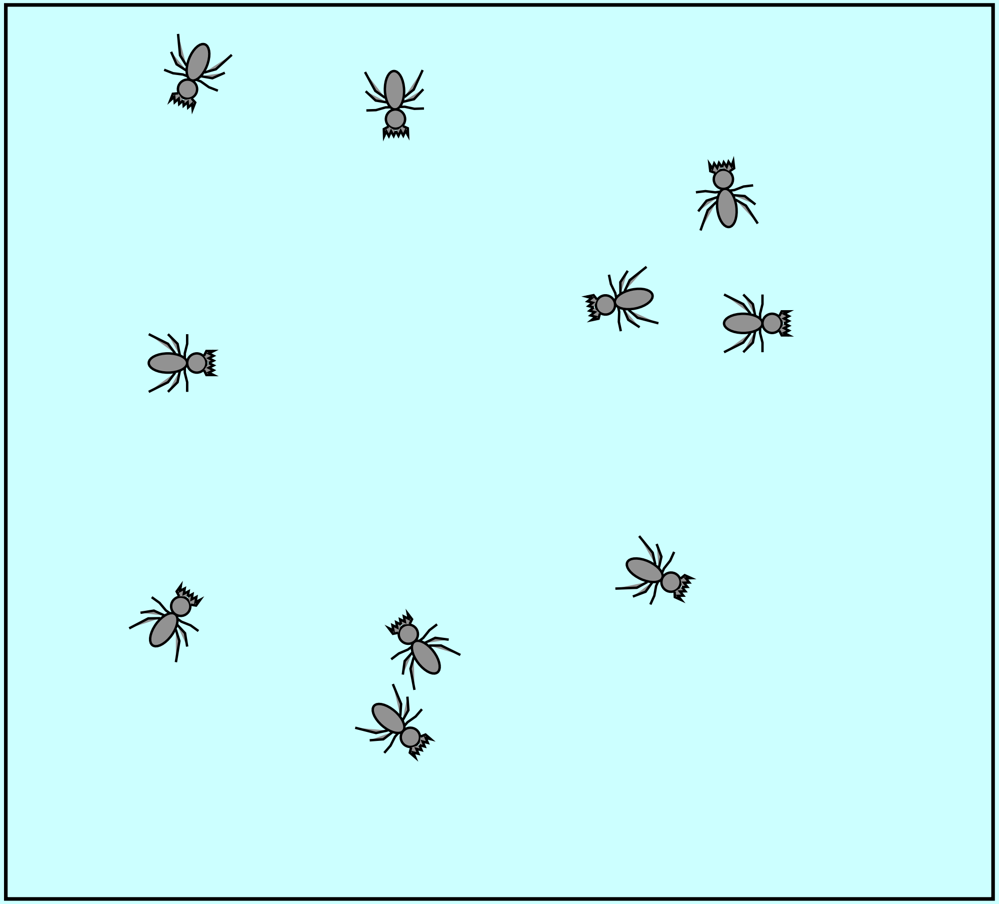
An accurate analogy: Bugs In A Box

There is a box ...

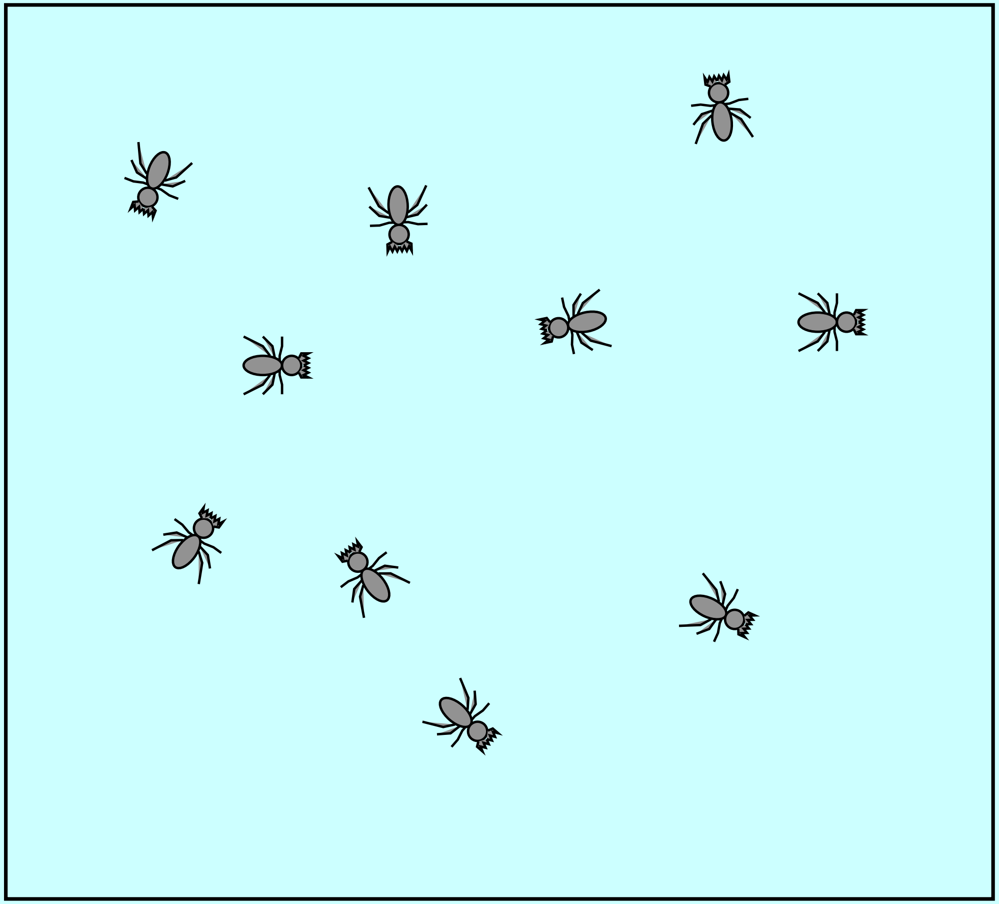


An accurate analogy: Bugs In A Box

with bugs that are ...



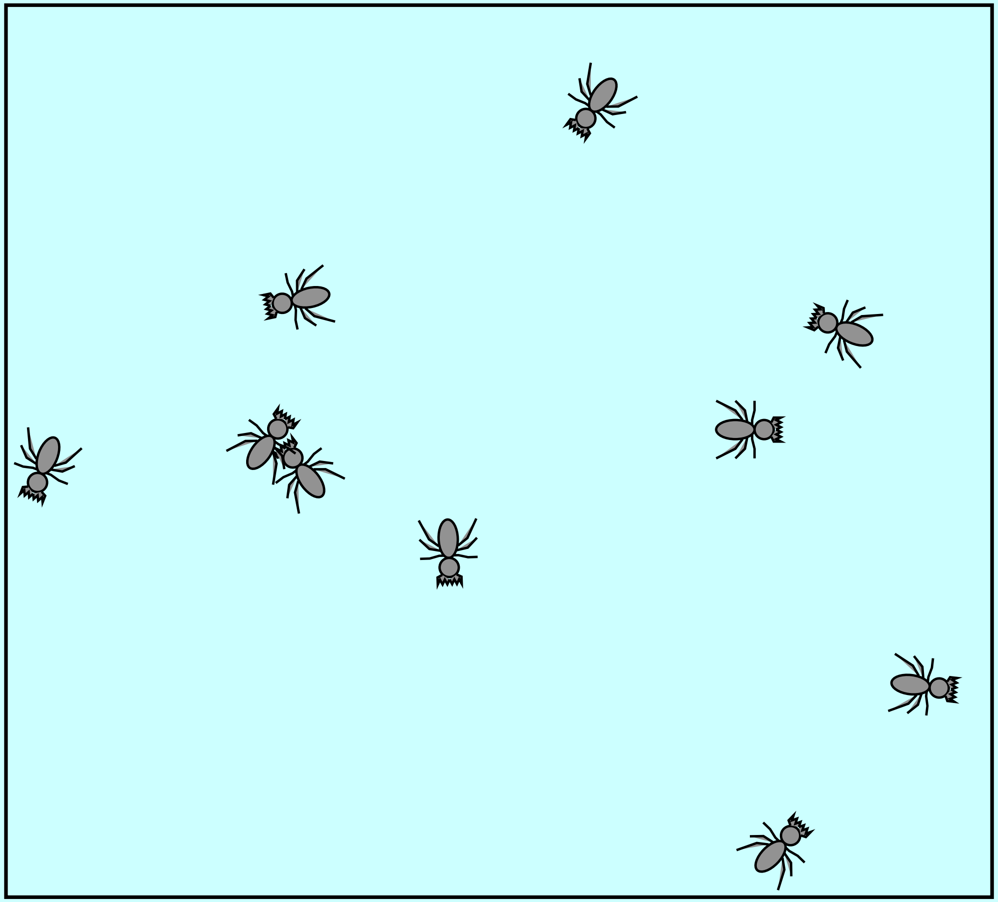
An accurate analogy: Bugs In A Box



hyperactive, ...

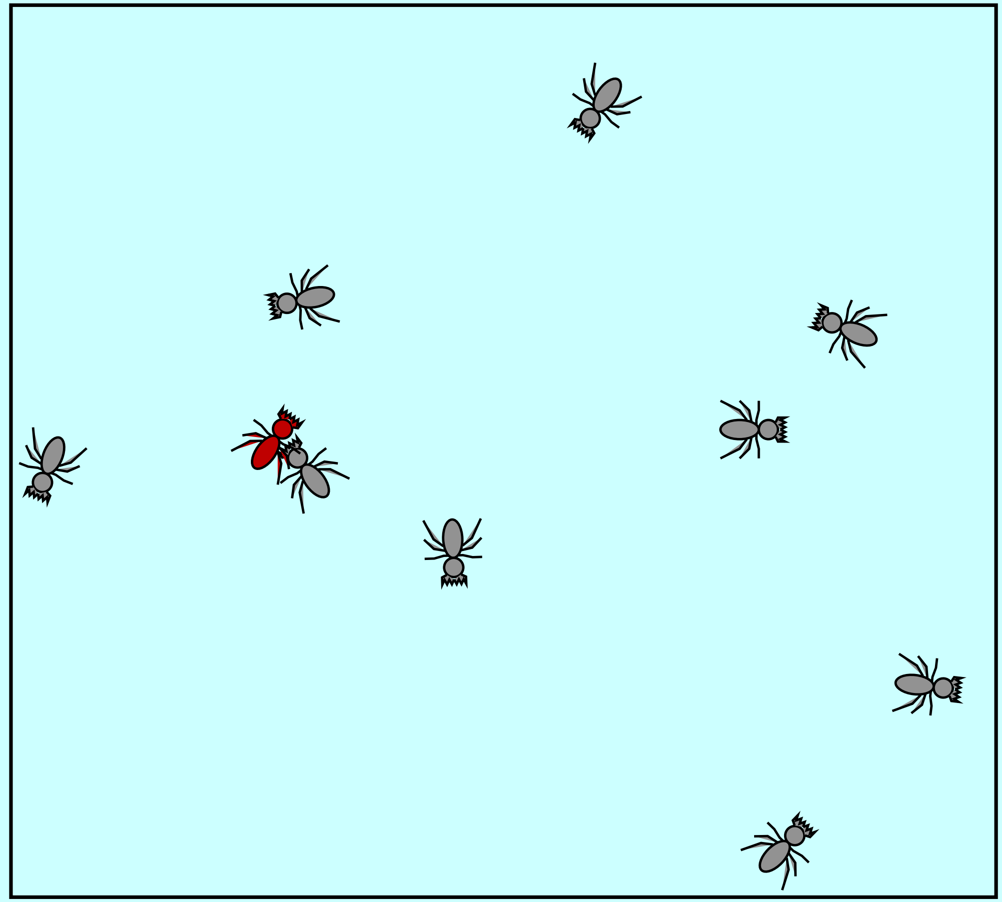
An accurate analogy: Bugs In A Box

indiscriminate, ...



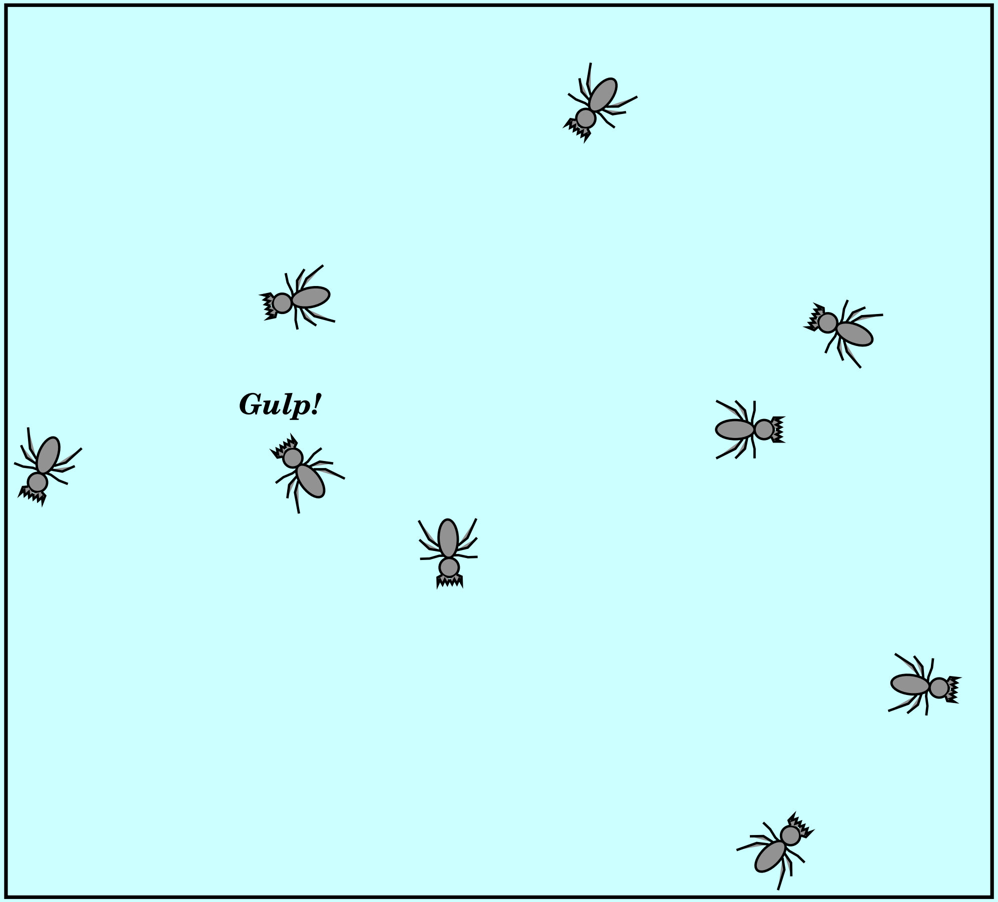
An accurate analogy: Bugs In A Box

voracious ...



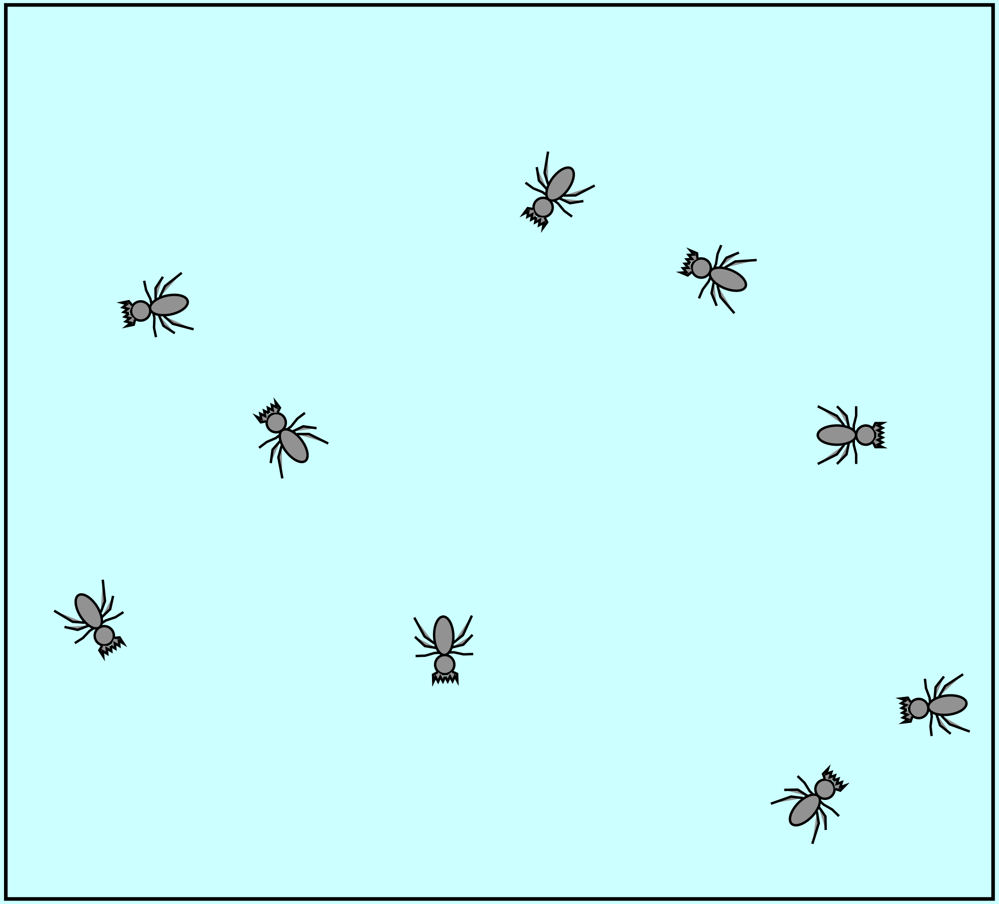
An accurate analogy: Bugs In A Box

(eats other bug) ...

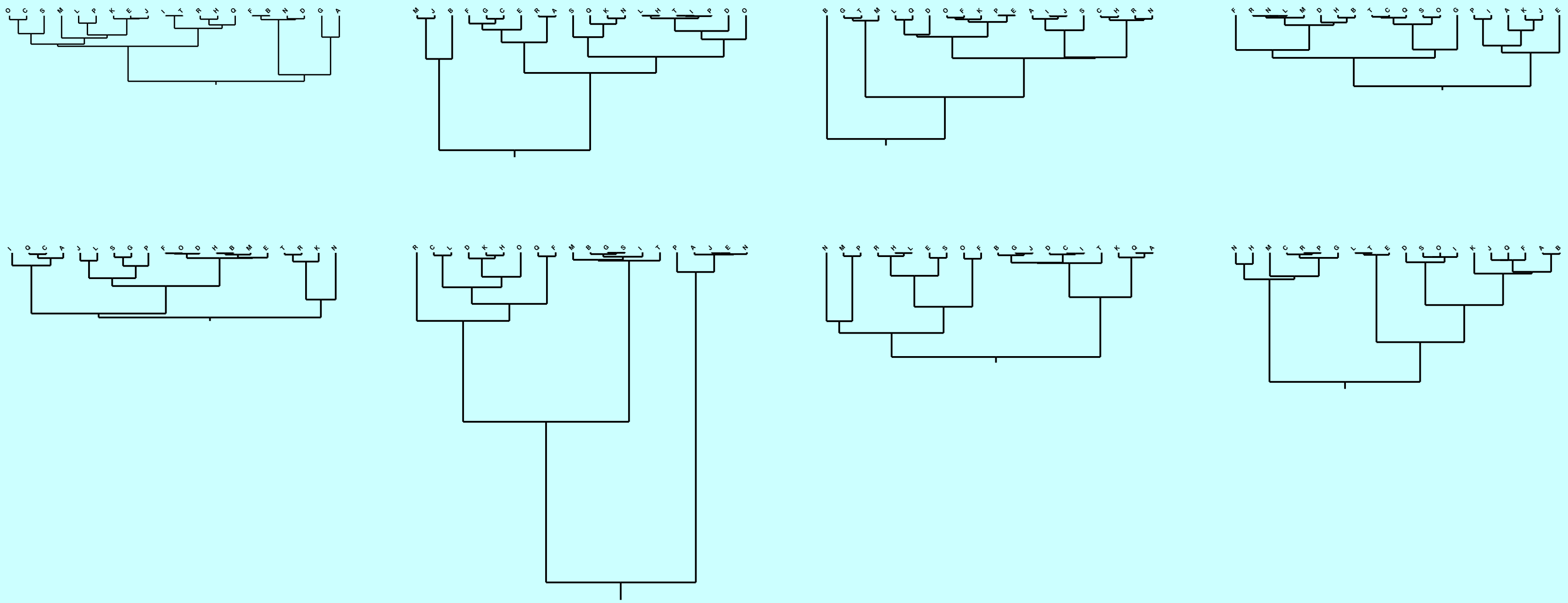


An accurate analogy: Bugs In A Box

and insatiable.

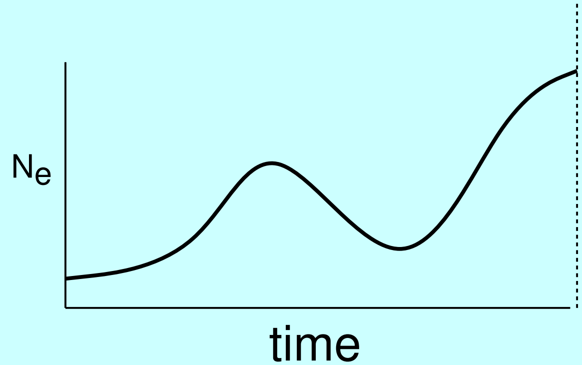


Random coalescent trees with 16 lineages

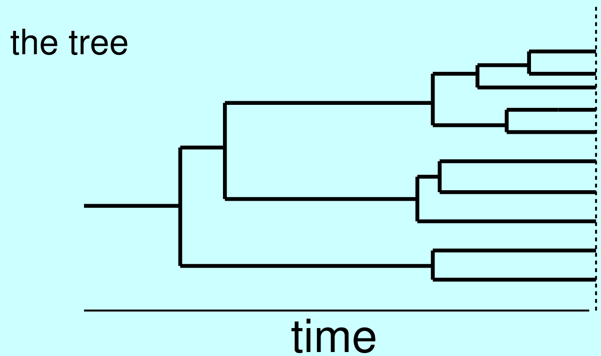


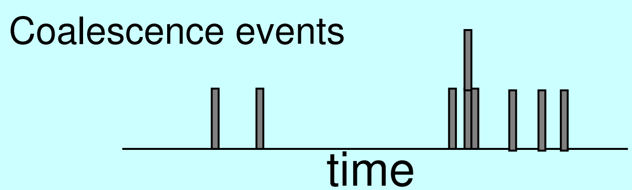
Coalescence is faster in small populations

Change of population size and coalescents



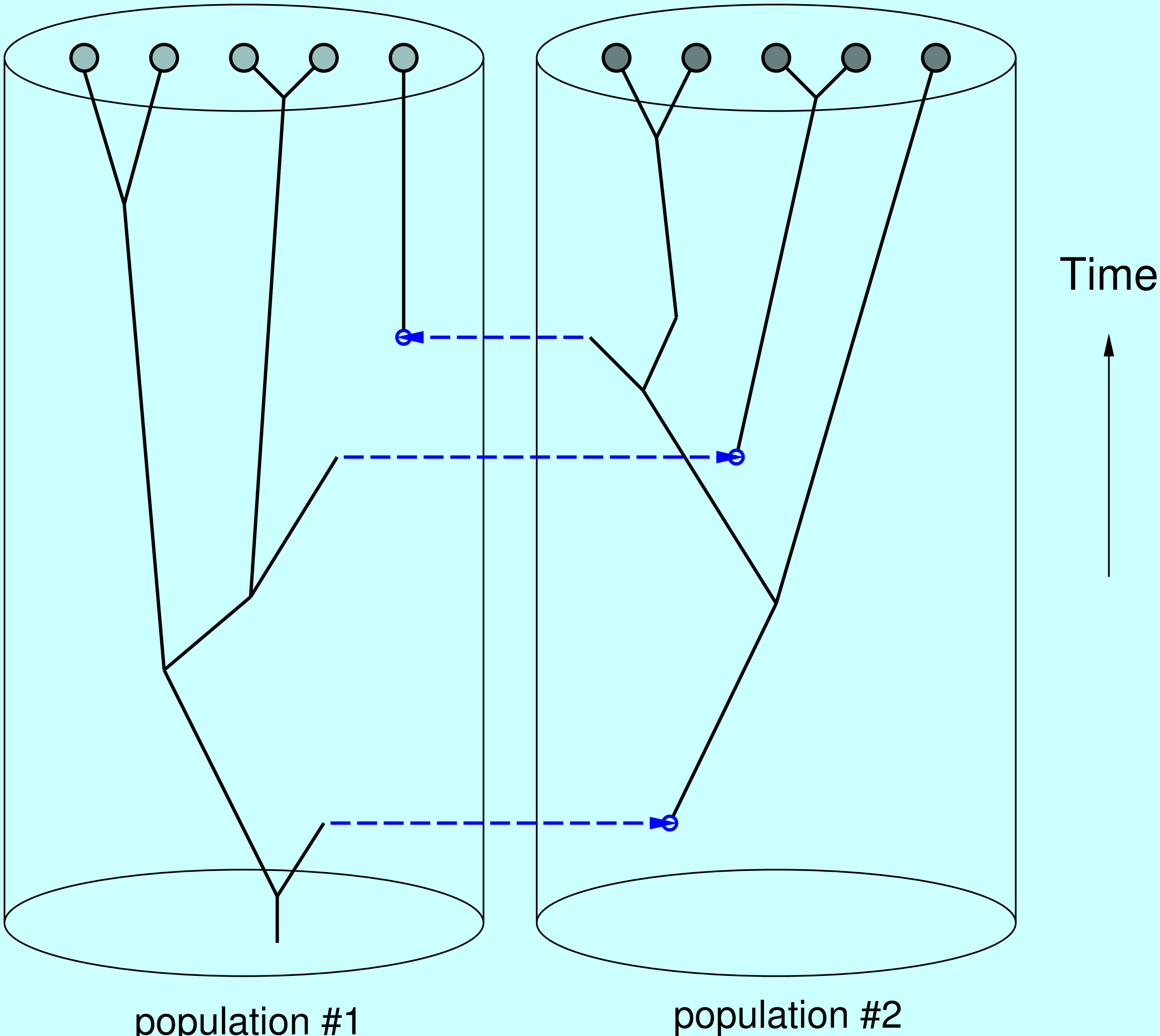
the changes in population size will produce waves of coalescence



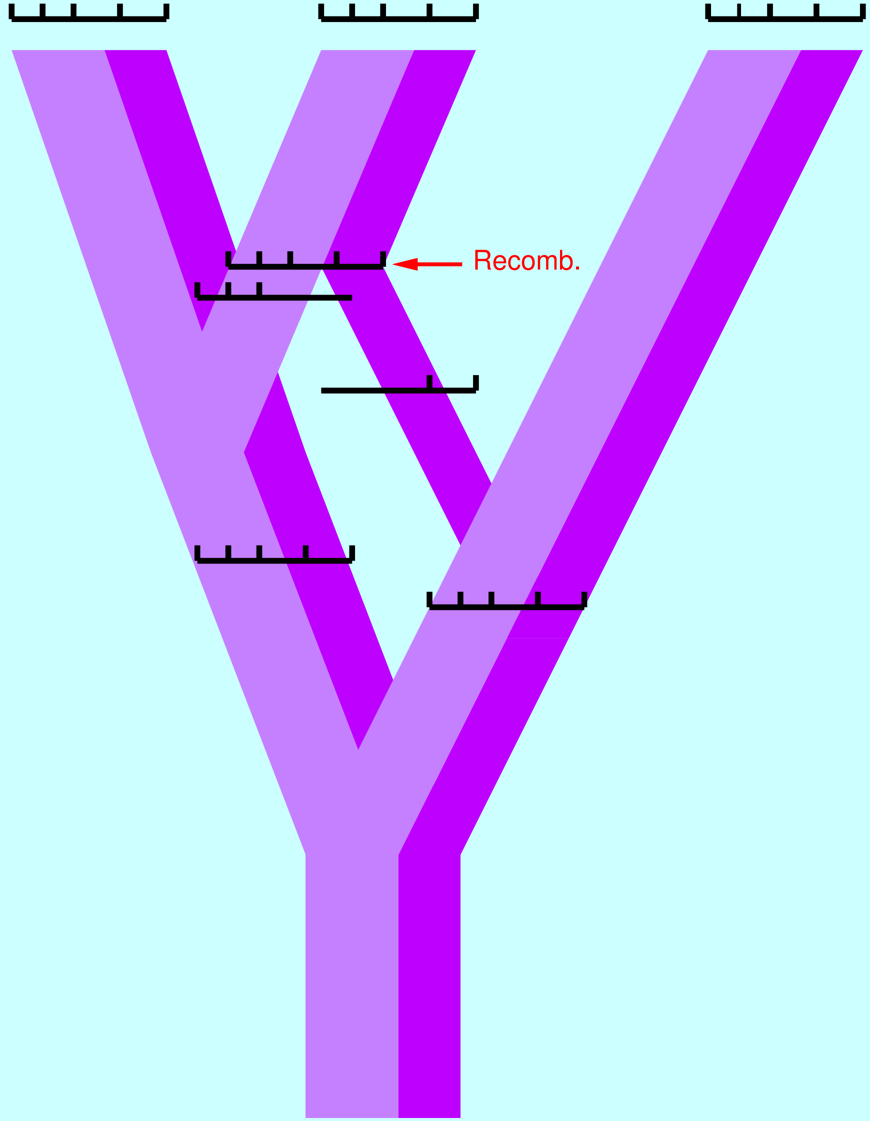


The parameters of the growth curve for N_e can be inferred by likelihood methods as they affect the prior probabilities of those trees that fit the data.

Migration can be taken into account

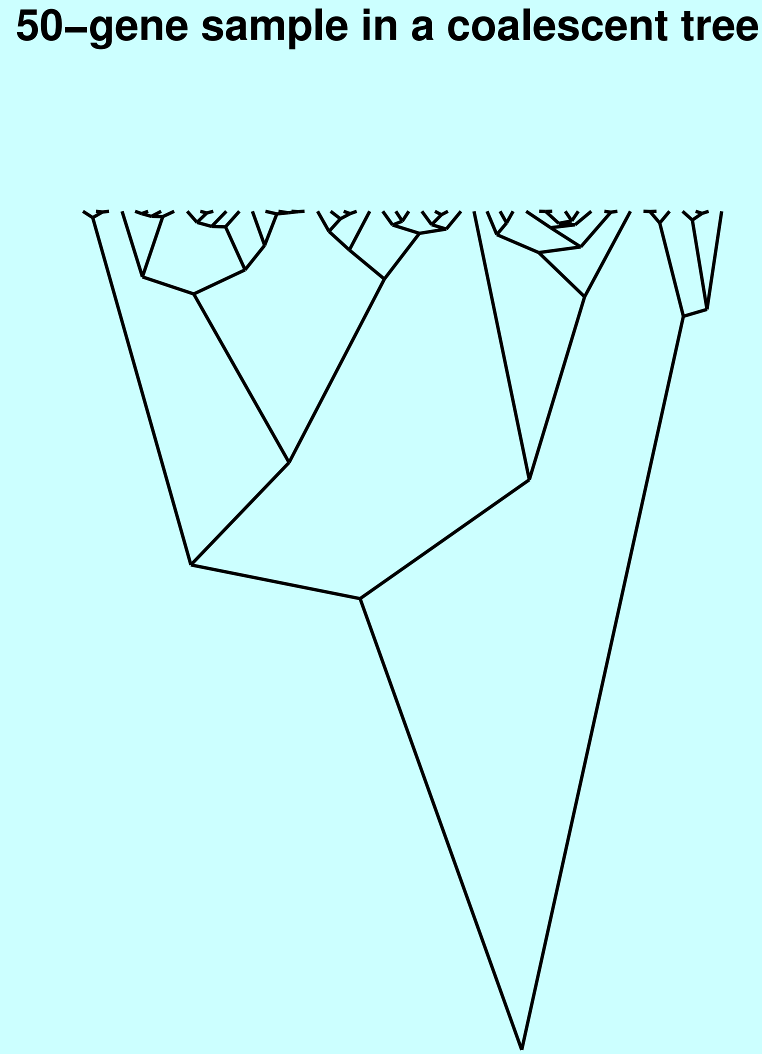


Recombination creates loops

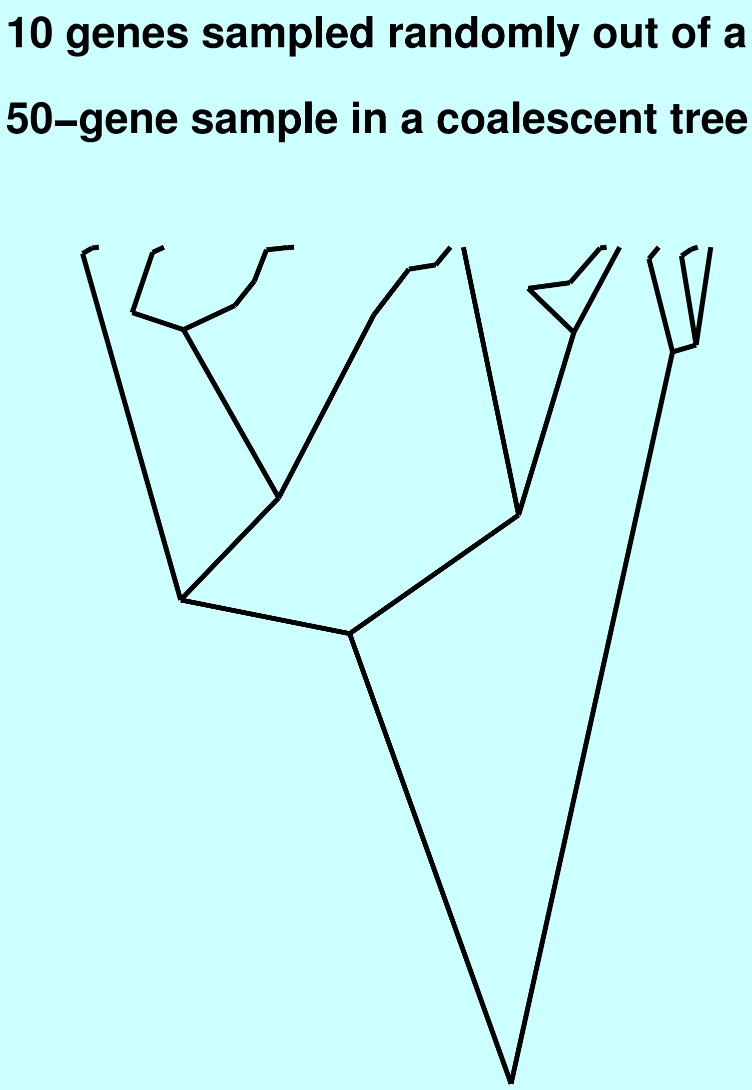


Different markers have slightly different coalescent trees

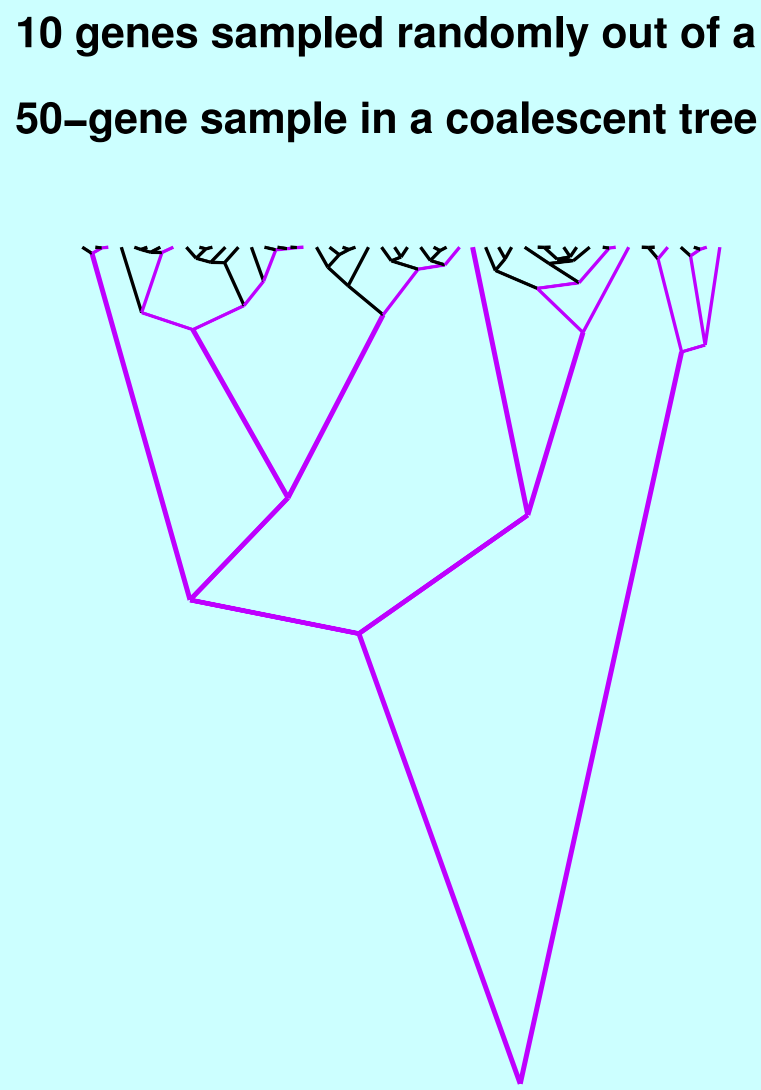
If we have a sample of 50 copies



The first 10 account for most of the branch length

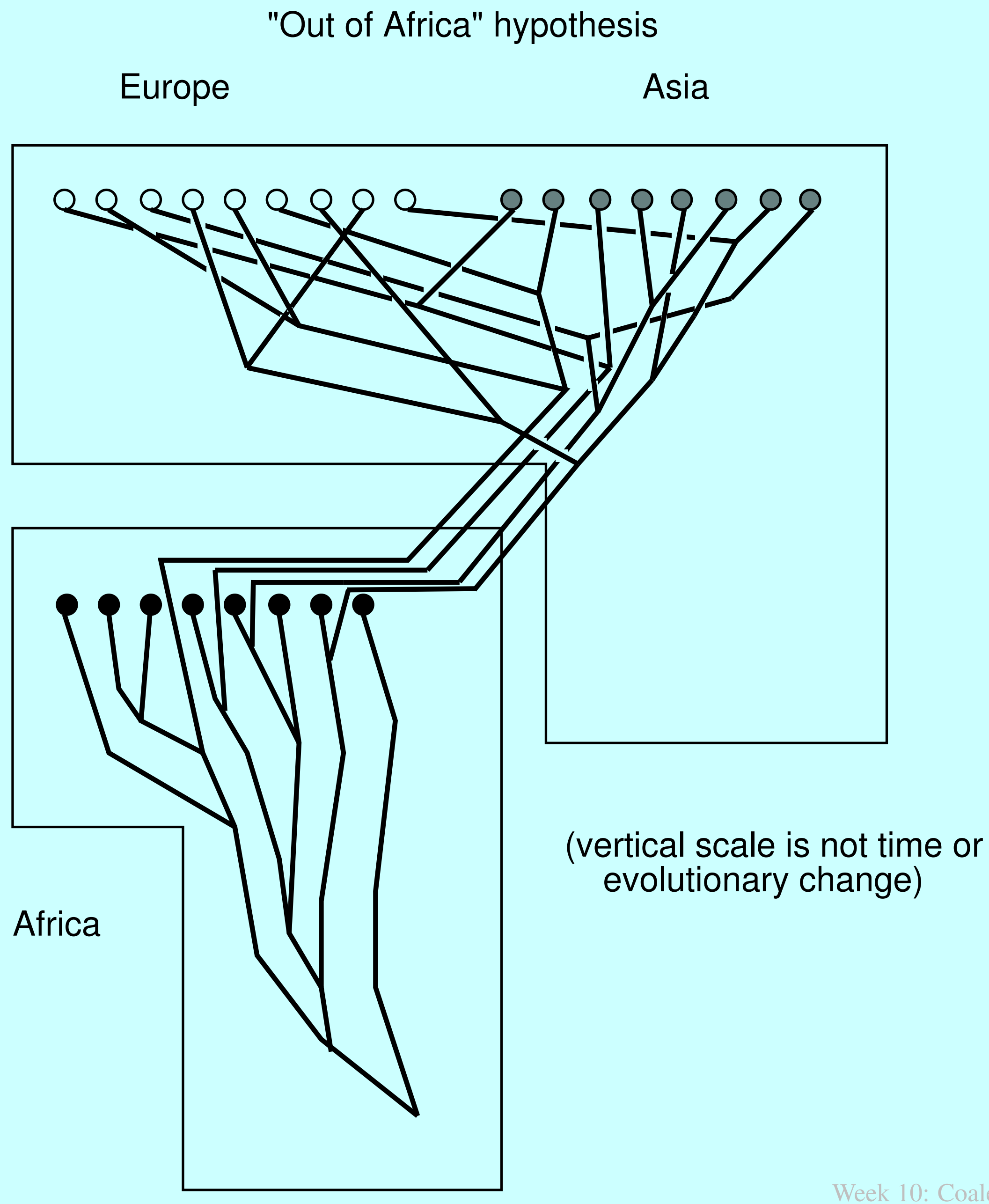


... and when we add the other 40 they add less length



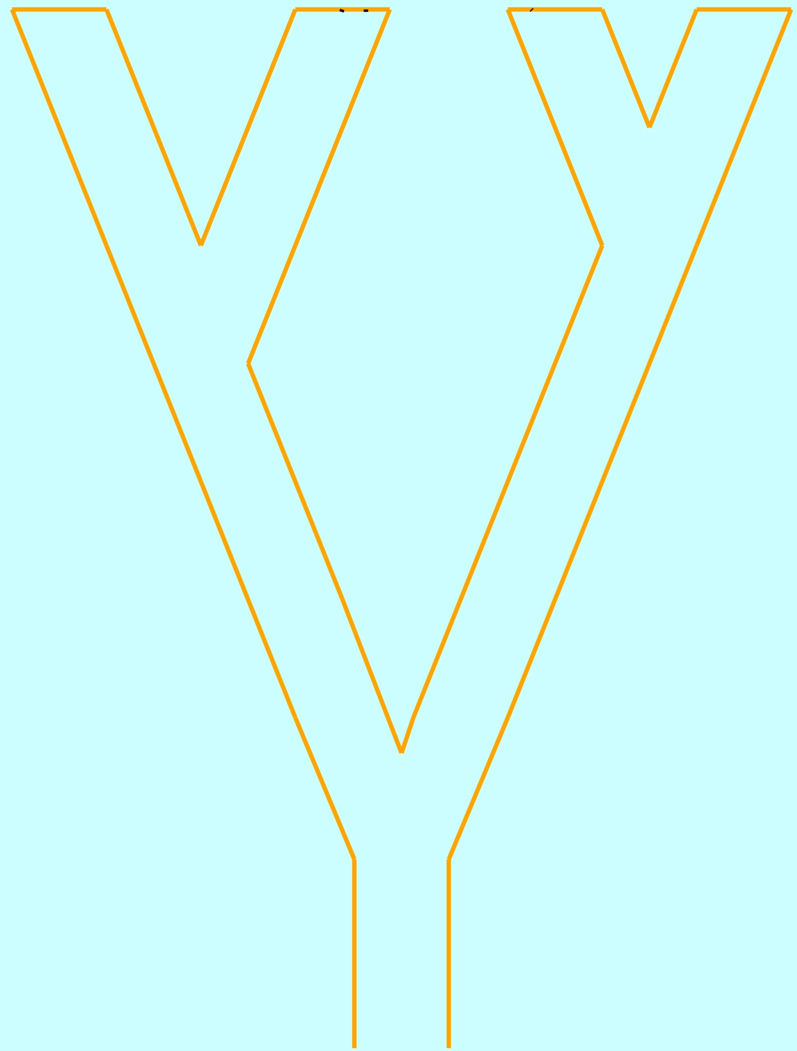
(purple lines are the 10−gene tree)

We want to be able to analyze human evolution



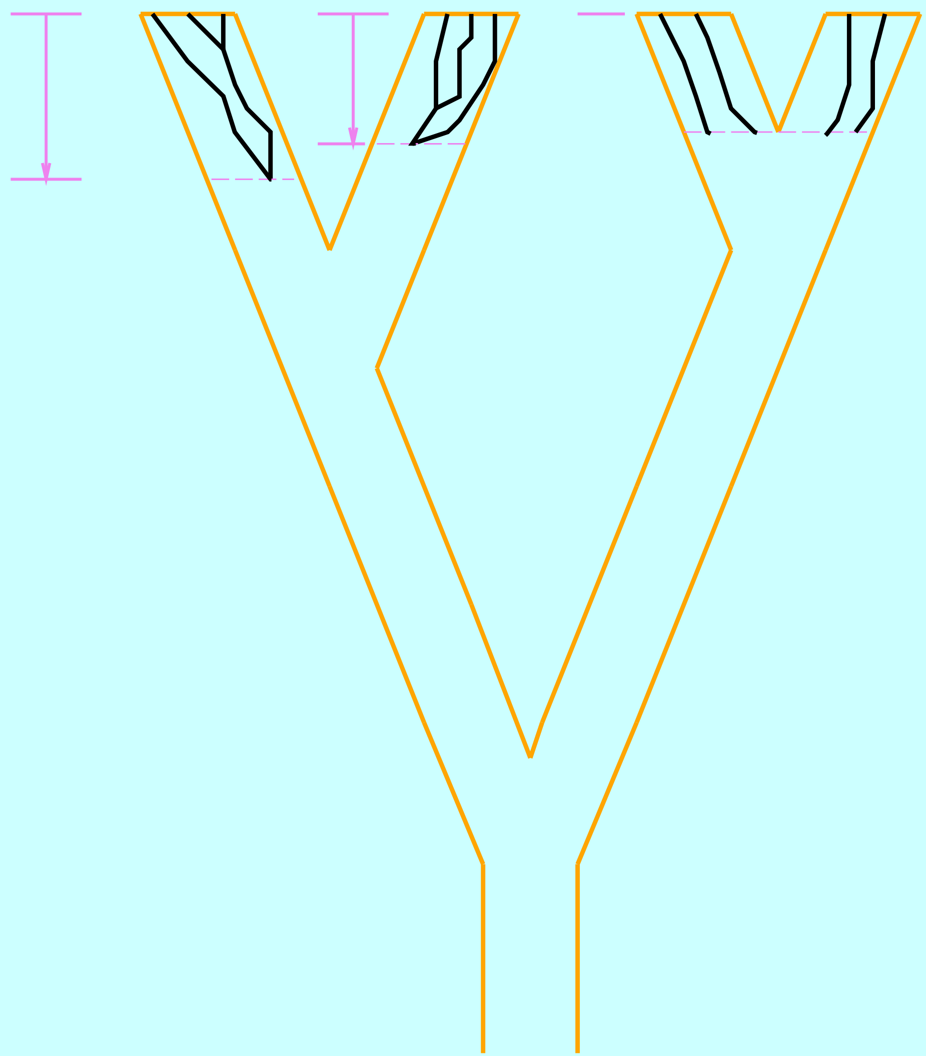
coalescent and “gene trees” versus species trees

The species tree



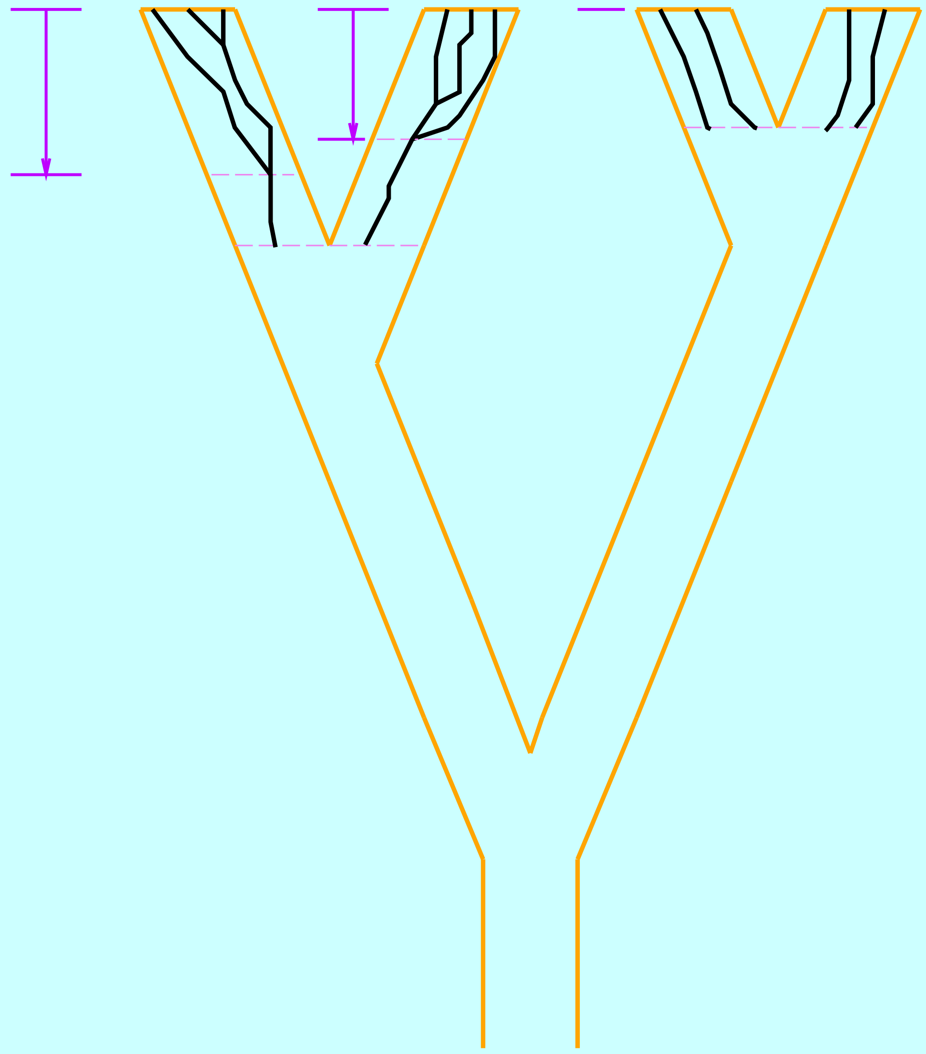
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



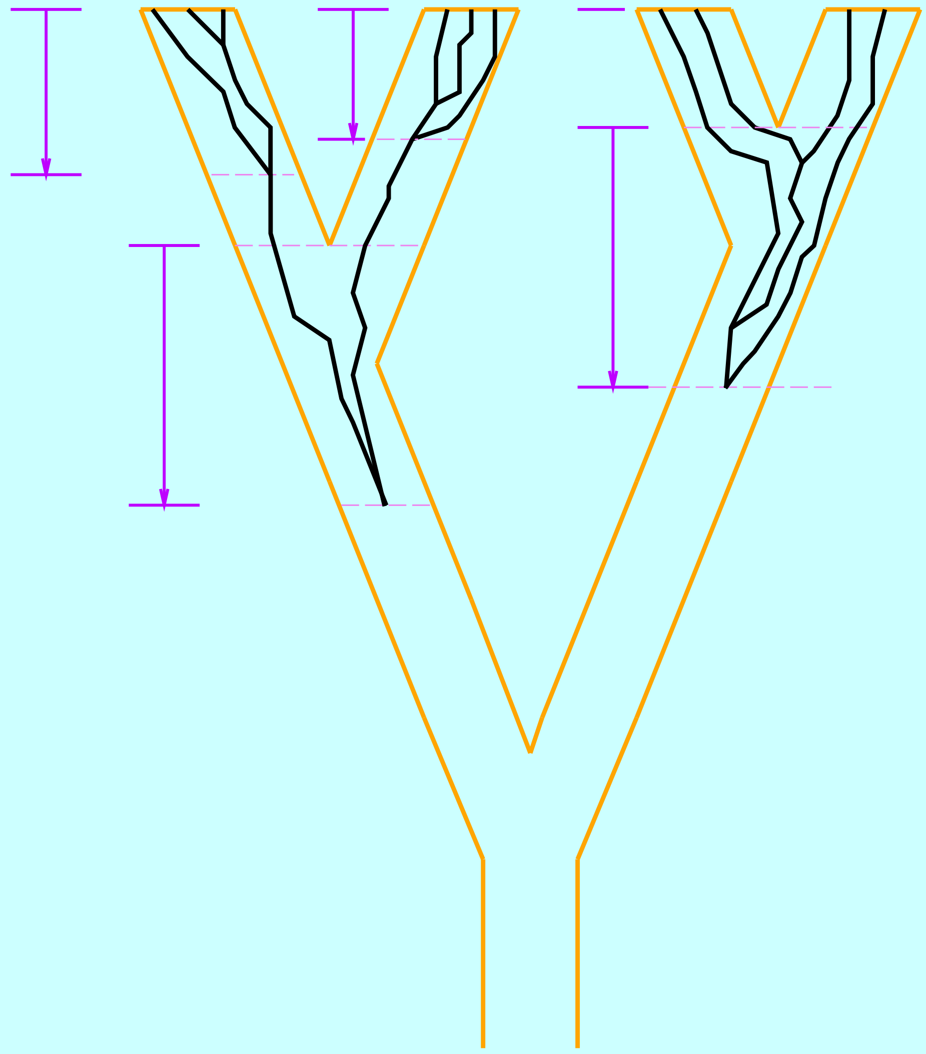
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



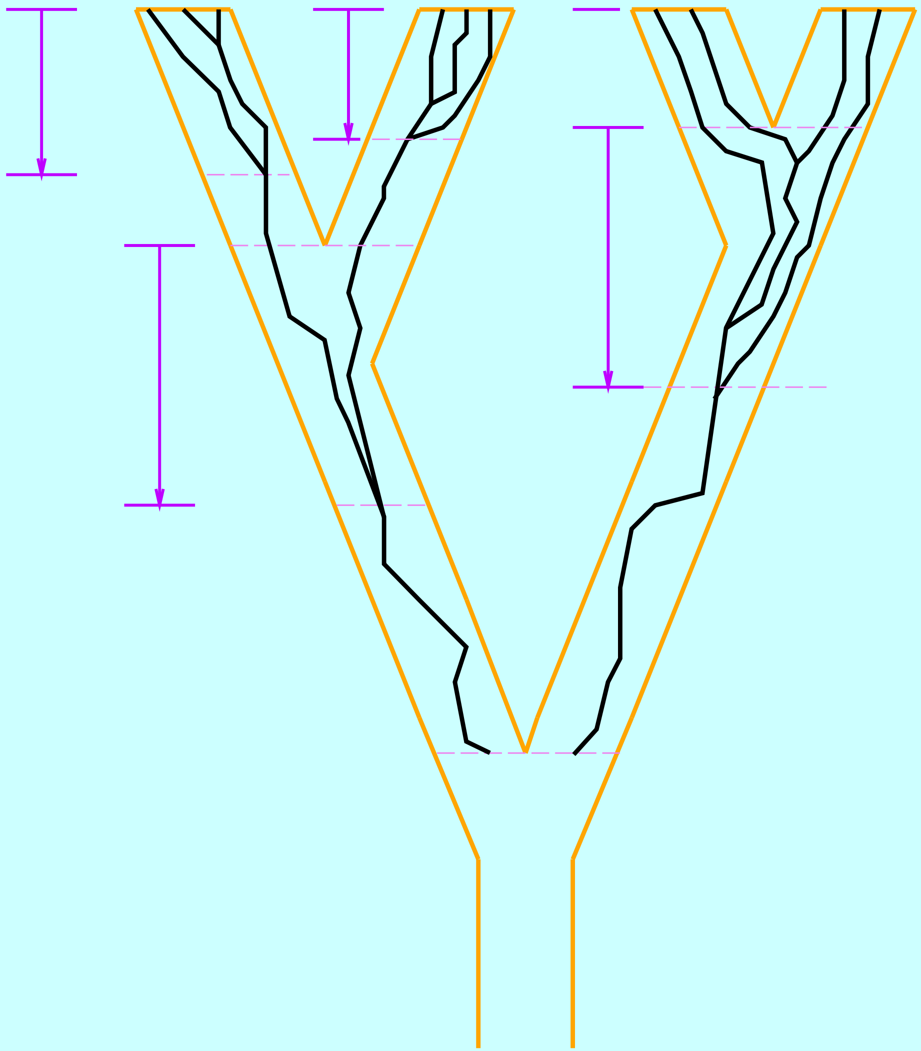
coalescent and "gene trees" versus species trees

Consistency of gene tree with species tree



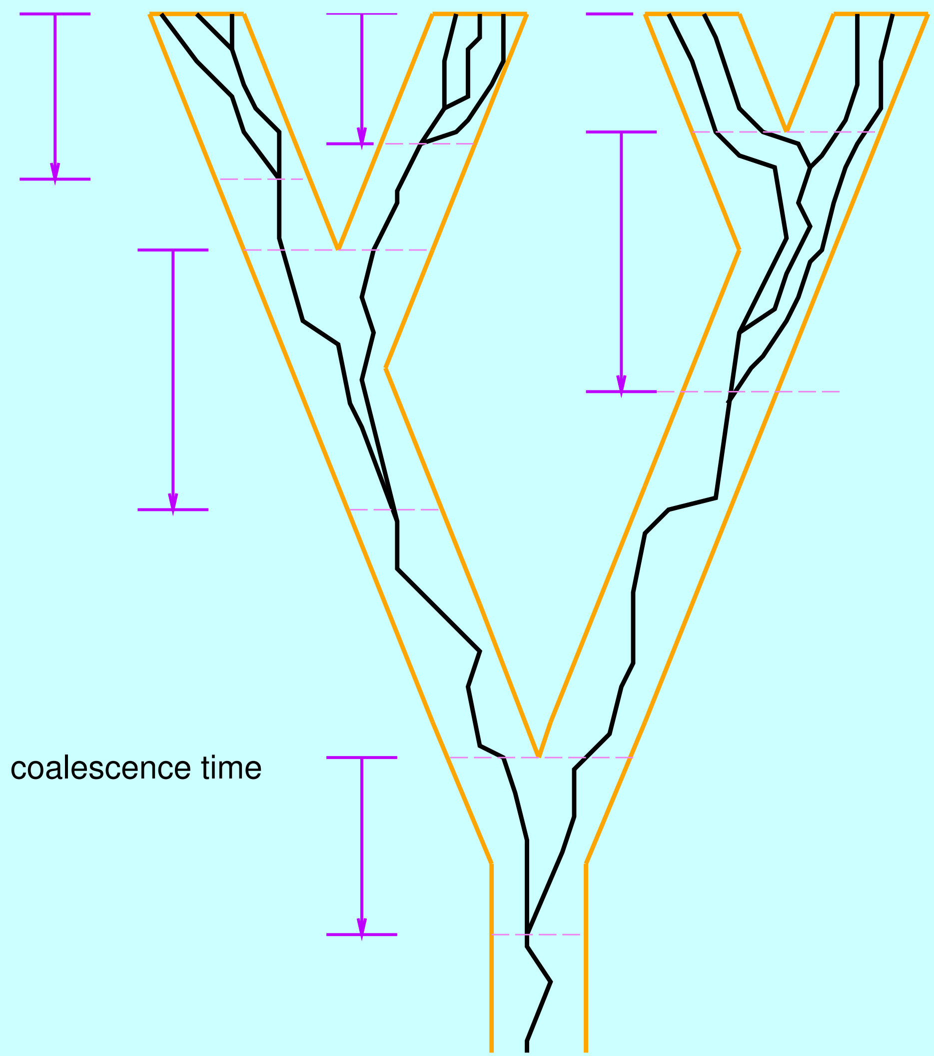
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree

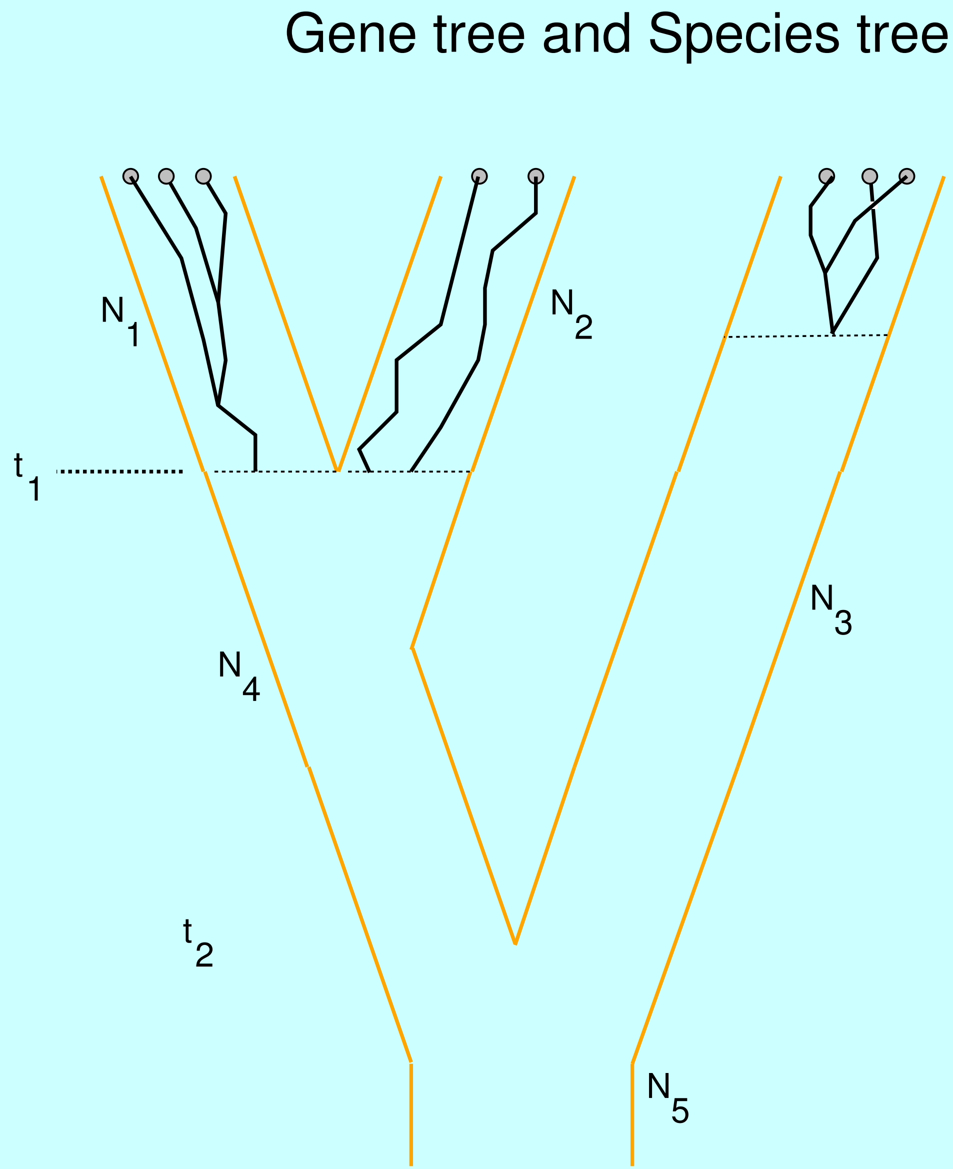


coalescent and “gene trees” versus species trees

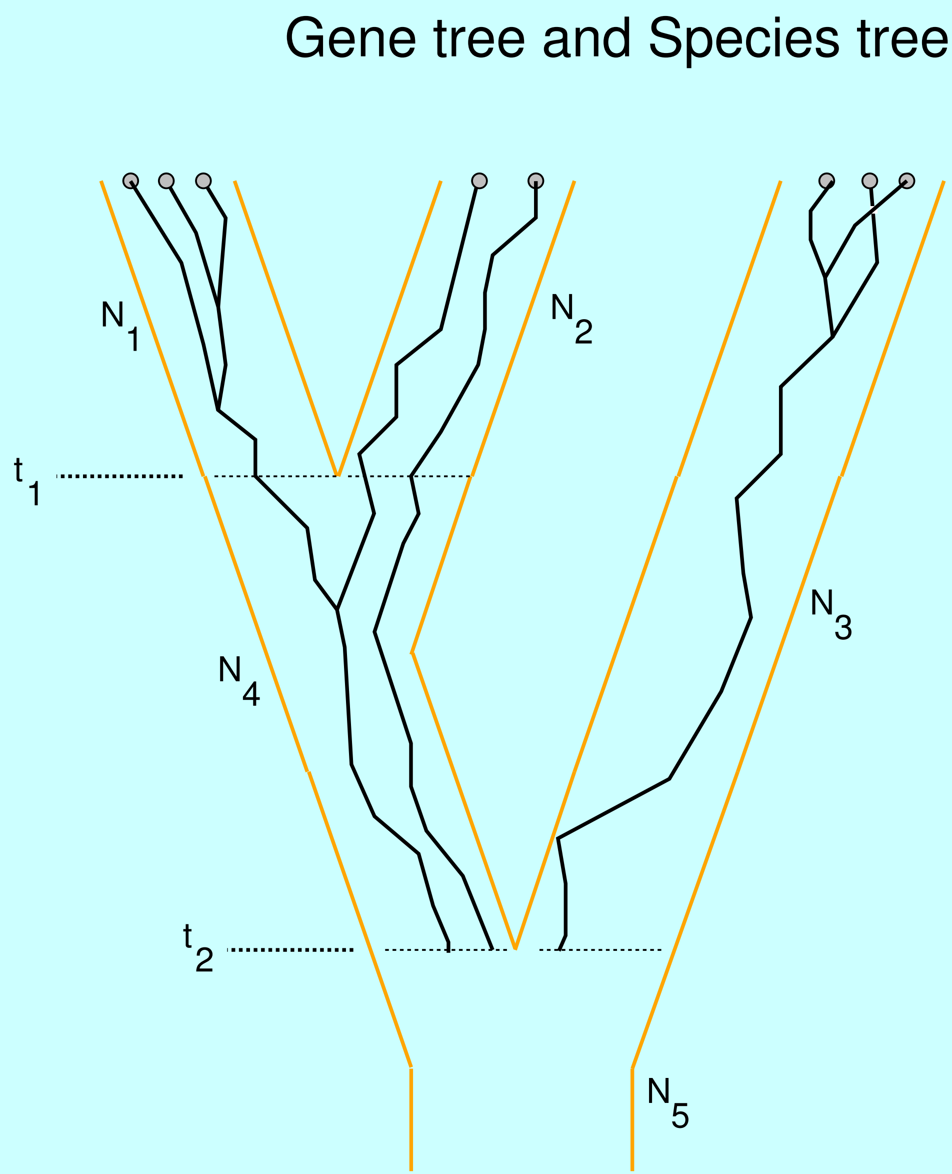
Consistency of gene tree with species tree



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

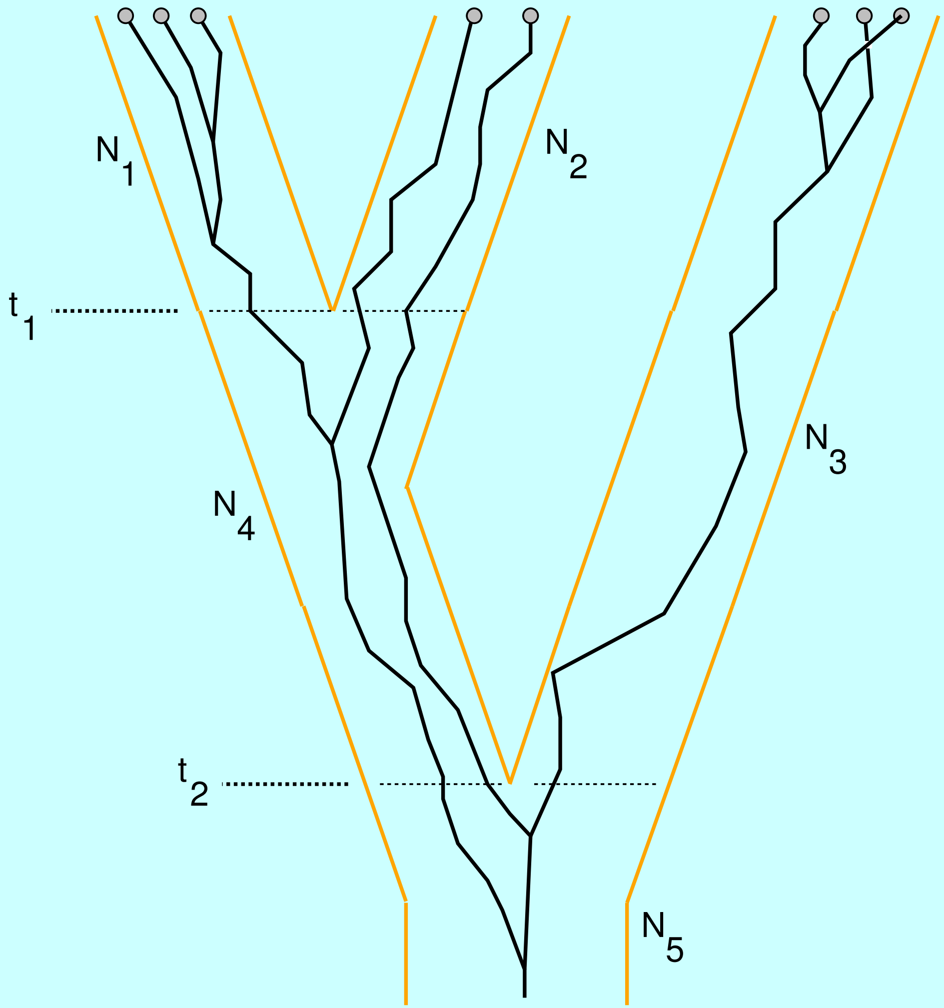


If the branch is more than N_e generations long ...

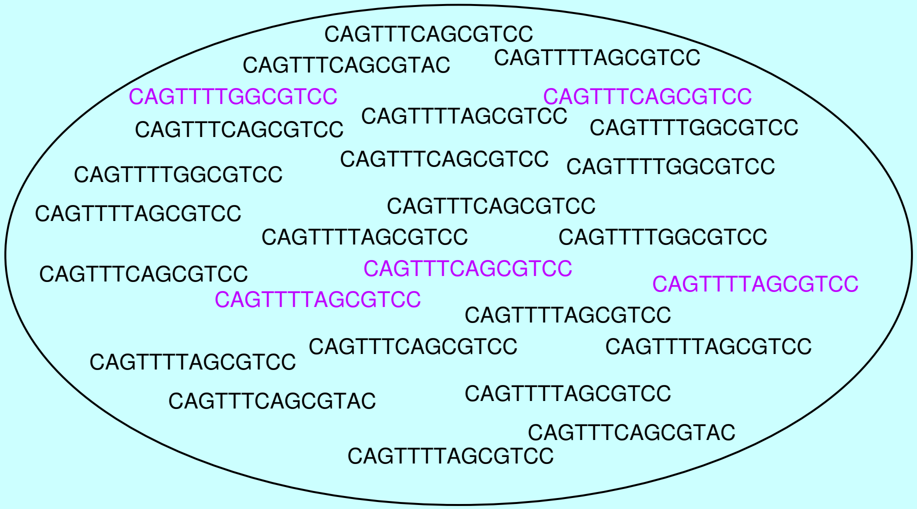


If the branch is more than N_e generations long ...

Gene tree and Species tree

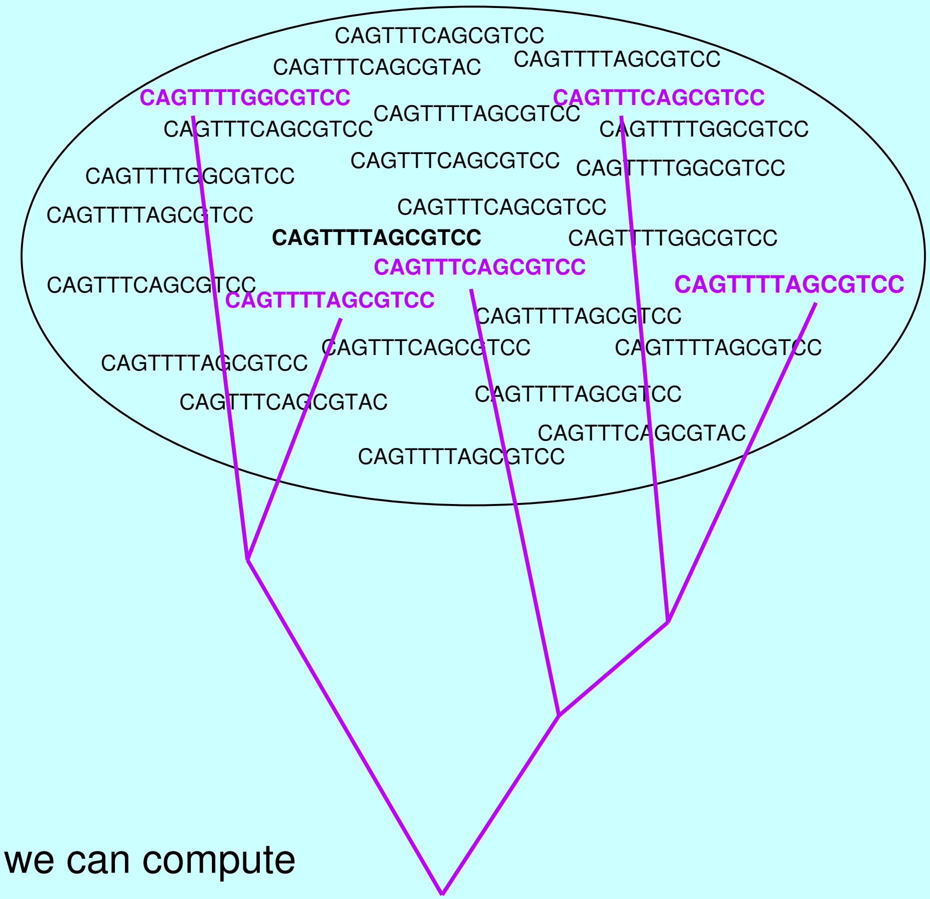


How do we compute a likelihood for a population sample?



L = Prob (CAGTTTCAGCGTCC , CAGTTTCAGCGTCC , ...) = ??

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



so we can compute

Prob(CAGTTTCAGCGTCC , CAGTTTCAGCGTCC , ... | Genealogy)

but how to computer the overall likelihood from this?

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,

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

$$= \sum_{G'} \underbrace{\text{Prob}\,(G' \mid N_e)}_{\text{Kingman's prior}} \underbrace{\text{Prob}\,(D \mid G', \mu)}_{\text{likelihood of tree}}$$

Rescaling the branch lengths

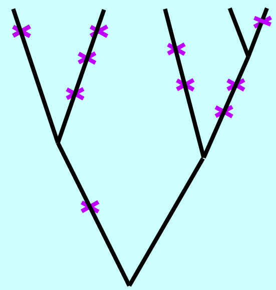
Rescaling branch lengths of G' so that branches are given in expected mutations per site, $G = \mu G'$, we get (if we let $\Theta = 4N_e\mu$)

$$L = \sum_{G} \mathrm{Prob}\,(G \mid \Theta)\,\mathrm{Prob}\,(D \mid G)$$

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

The variability comes from two sources

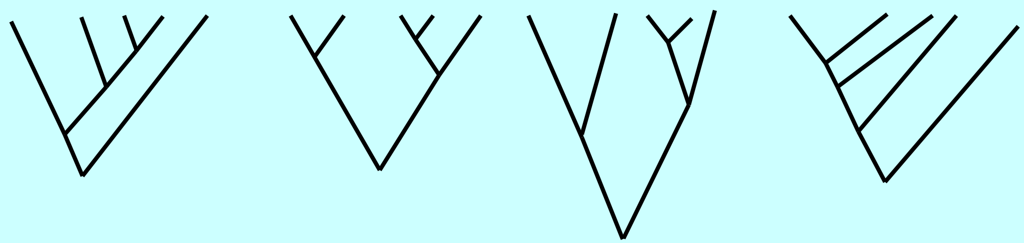
(1) Randomness of mutation



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

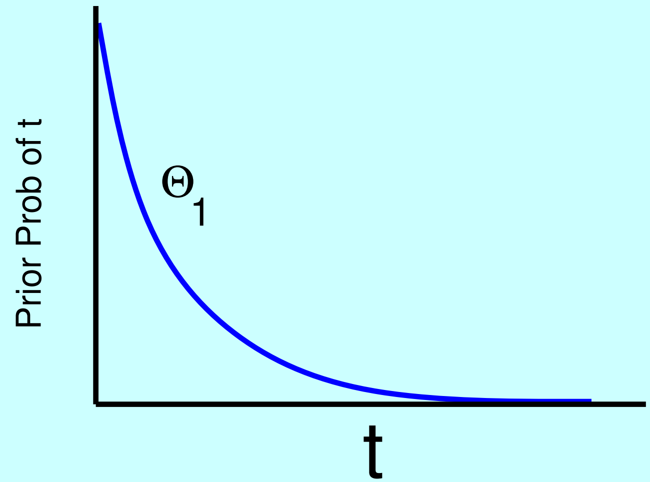
(i) more gene copies, or

(ii) more loci

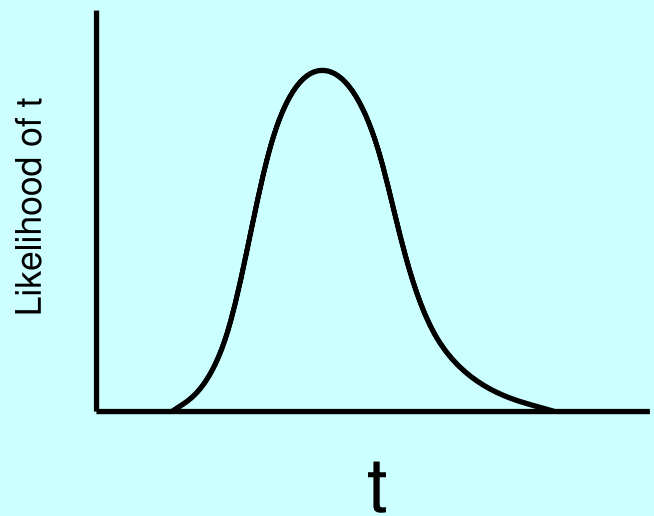
Computing the likelihood: averaging over coalescents

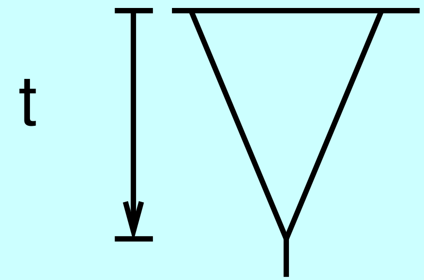
The likelihood calculation in a sample of two gene copies

The product of the prior on t,

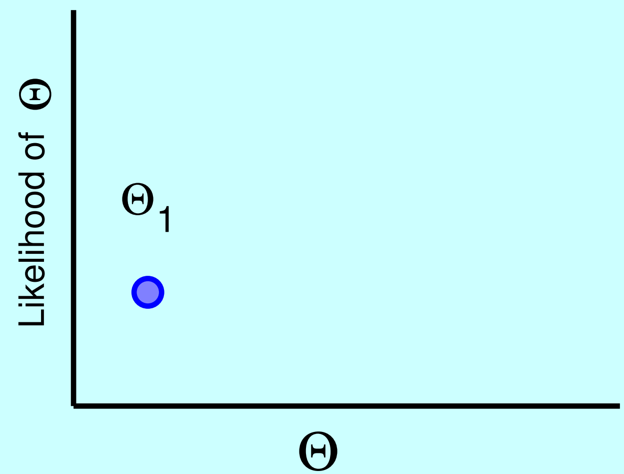


times the likelihood of that t from the data,





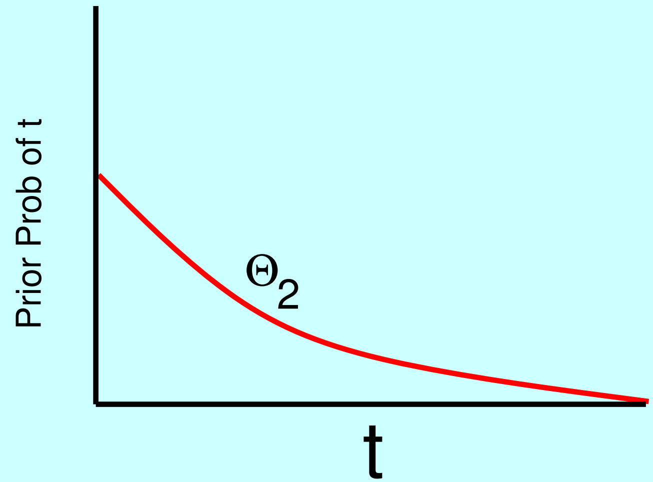
when integrated over all possible t's, gives the likelihood for the underlying parameter Θ

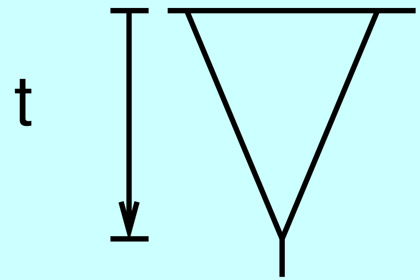


Computing the likelihood: averaging over coalescents

The likelihood calculation in a sample of two gene copies

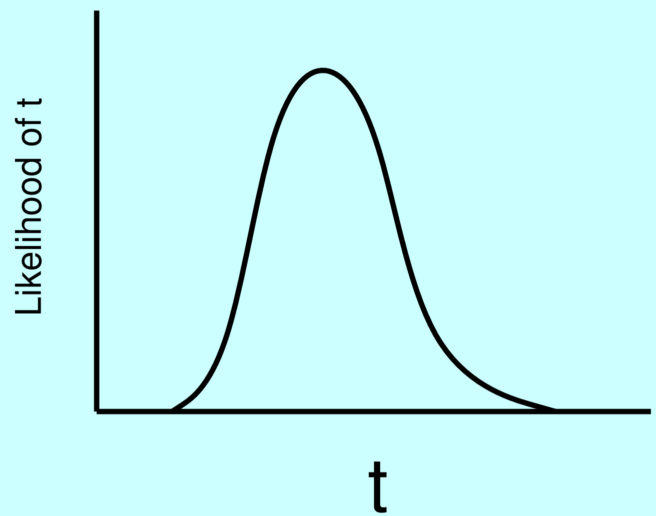
The product of the prior on t,

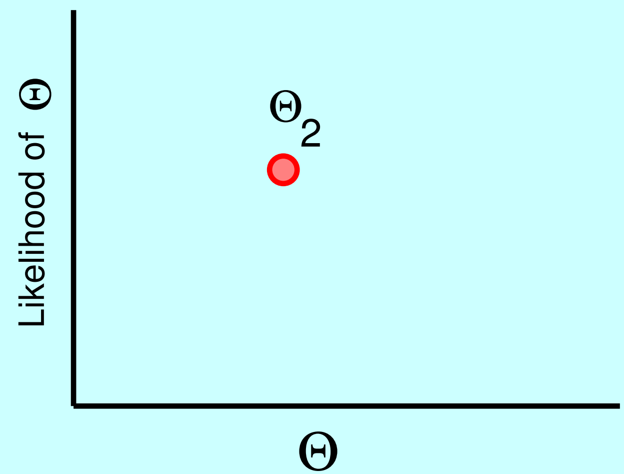




when integrated over all possible t's, gives the likelihood for the underlying parameter Θ

times the likelihood of that t from the data,

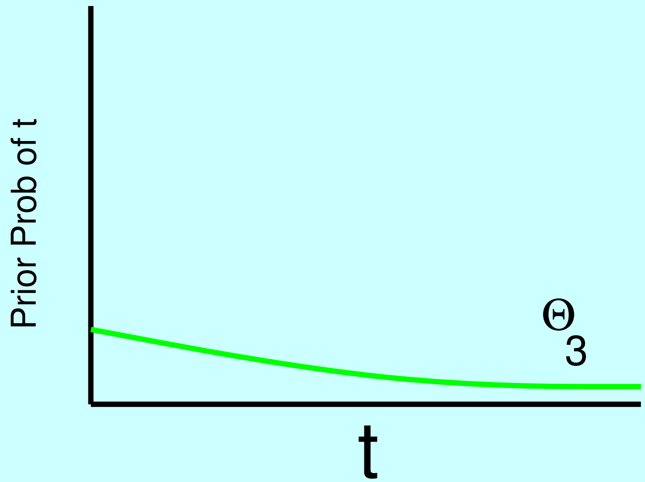


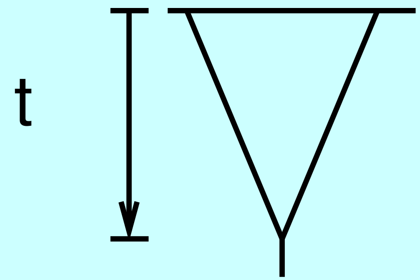


Computing the likelihood: averaging over coalescents

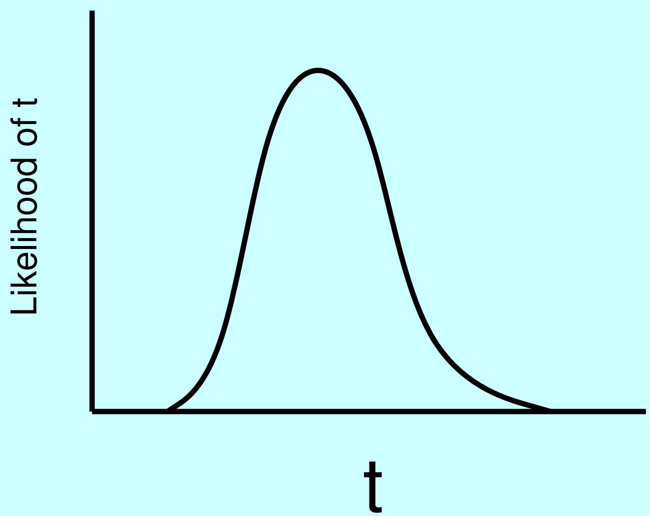
The likelihood calculation in a sample of two gene copies

The product of the prior on t,

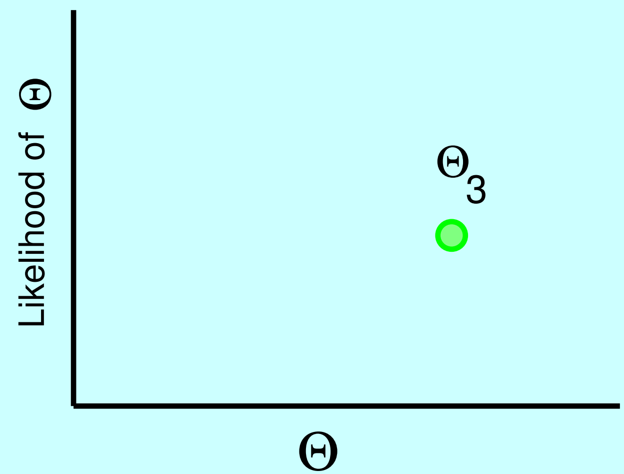




times the likelihood of that t from the data,



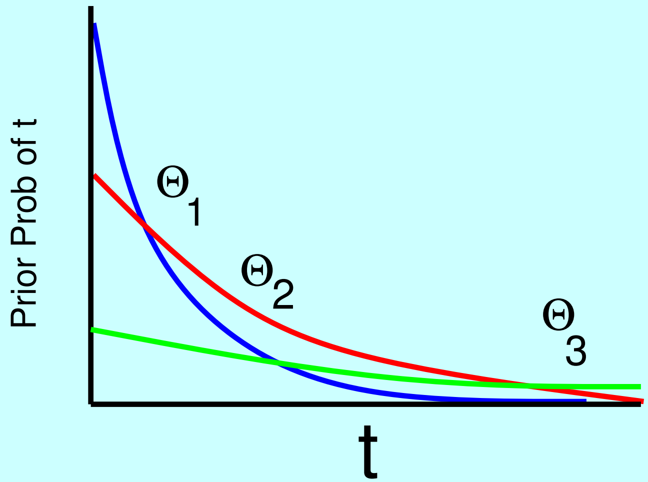
when integrated over all possible t's, gives the likelihood for the underlying parameter Θ

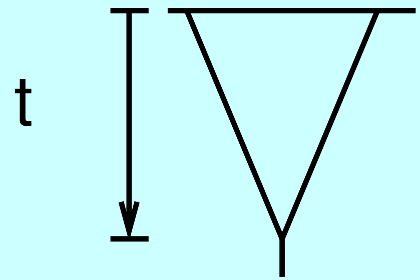


Computing the likelihood: averaging over coalescents

The likelihood calculation in a sample of two gene copies

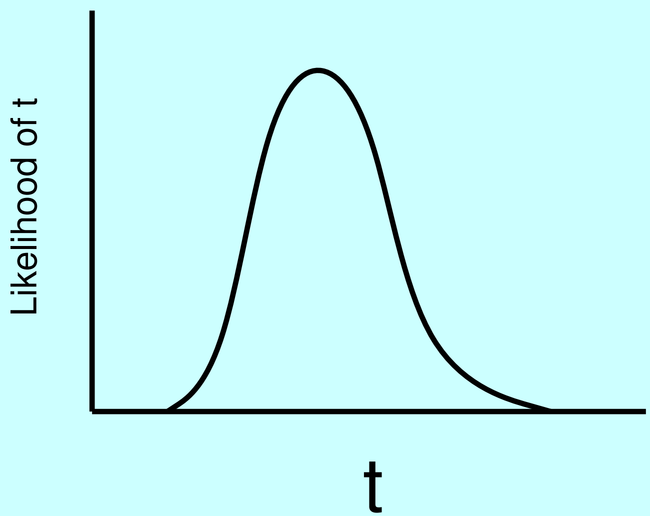
The product of the prior on t,

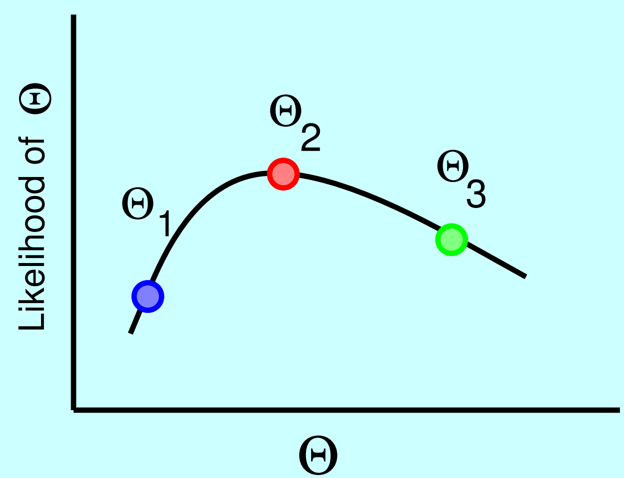




when integrated over all possible t's, gives the likelihood for the underlying parameter Θ

times the likelihood of that t from the data,



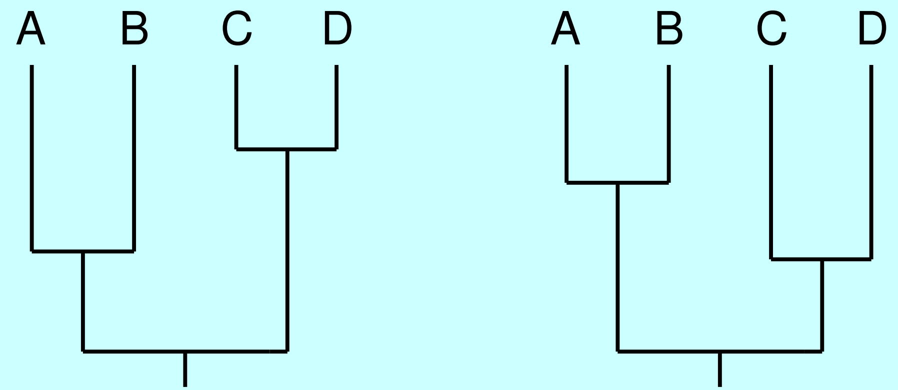


Labelled histories

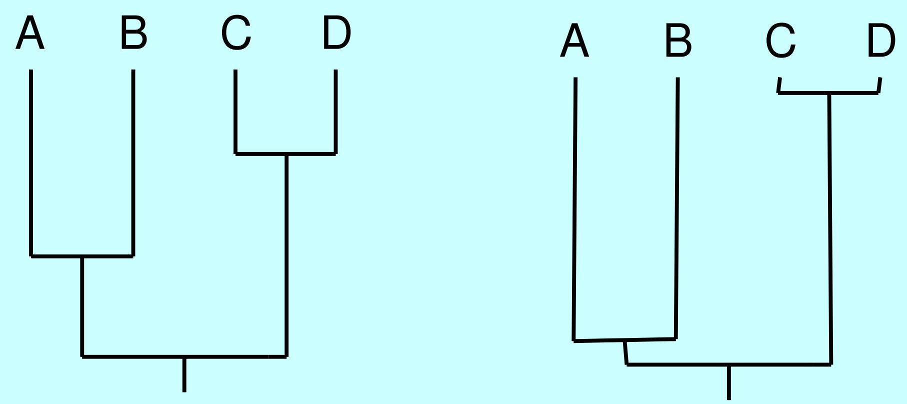
Labelled Histories (Edwards, 1970; Harding, 1971)

Trees that differ in the time−ordering of their nodes

These two are different:

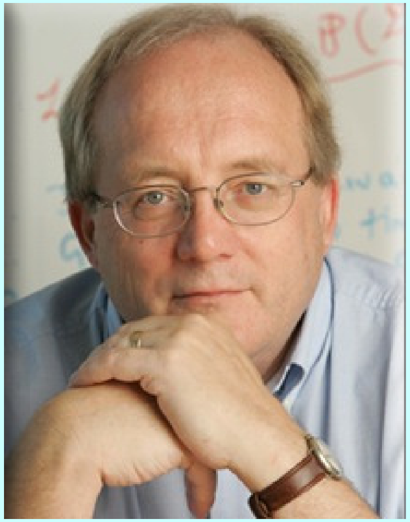


These two are the same:



Sampling approaches to coalescent likelihood

Bob Griffiths

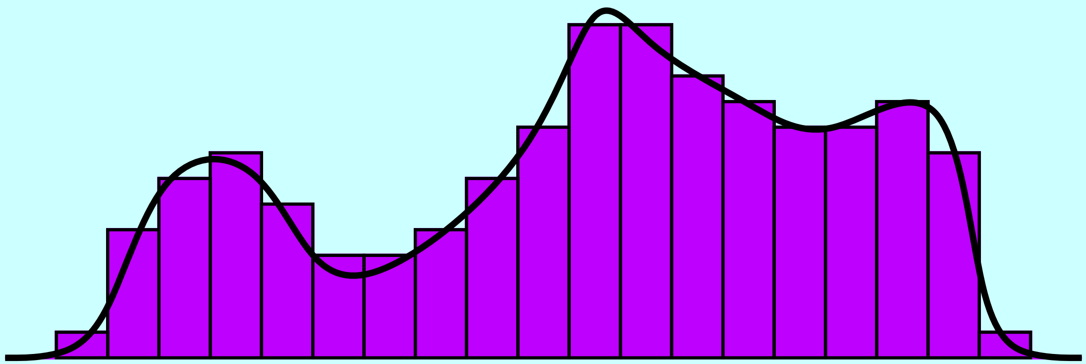


Simon Tavaré

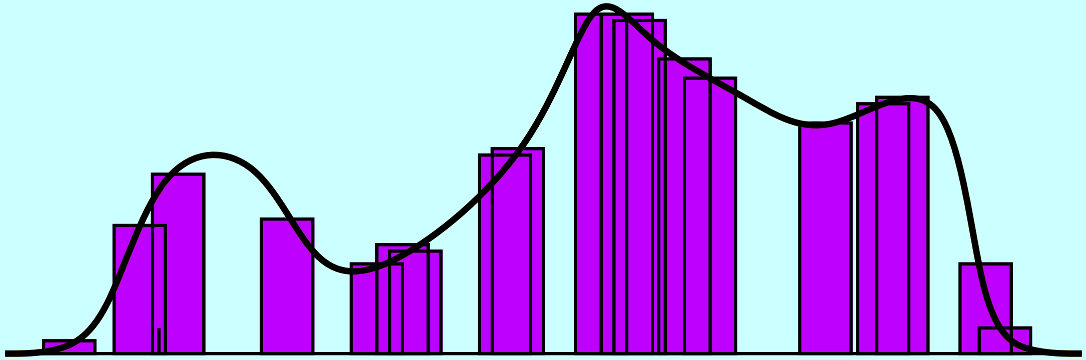
Mary Kuhner and Jon Yamato

Monte Carlo integration

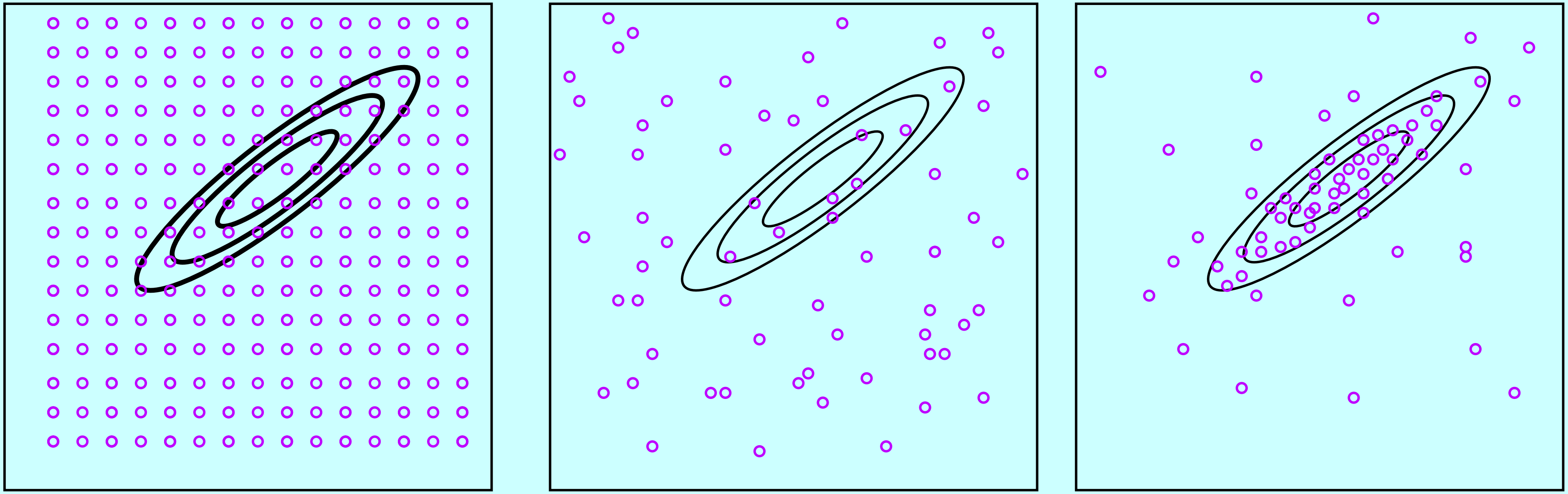
To get the area under a curve, we can either evaluate the function ($f(x)$) at a series of grid points and add up heights $\times$ widths:



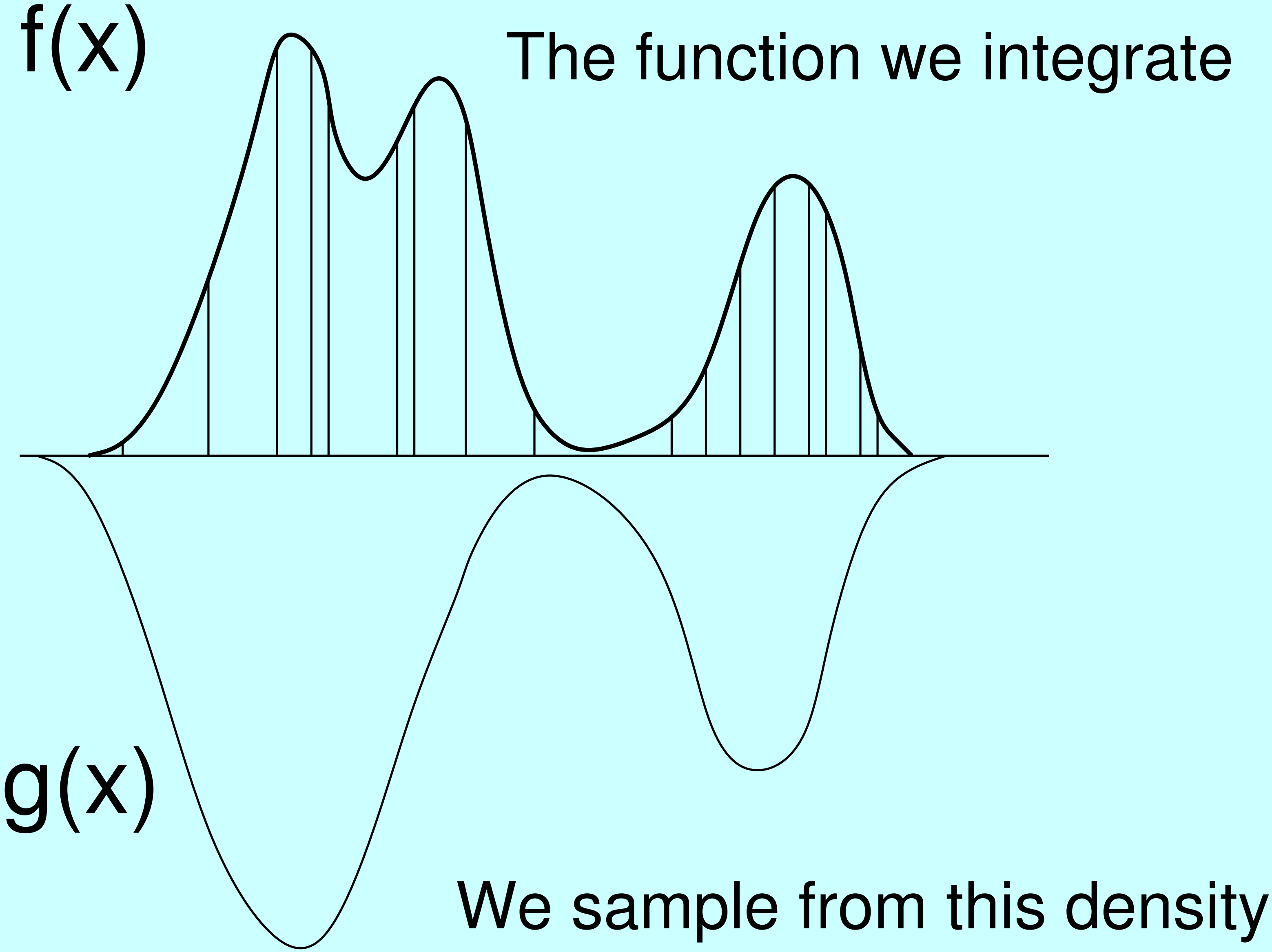
or we can sample at random the same number of points, add up height $\times$ width:



Importance sampling



Importance sampling



The math of importance sampling

$$\int f(x)\, dx = \int \frac{f(x)}{g(x)}\, g(x)\, dx$$

$$= E_g\left[\frac{f(x)}{g(x)}\right]$$

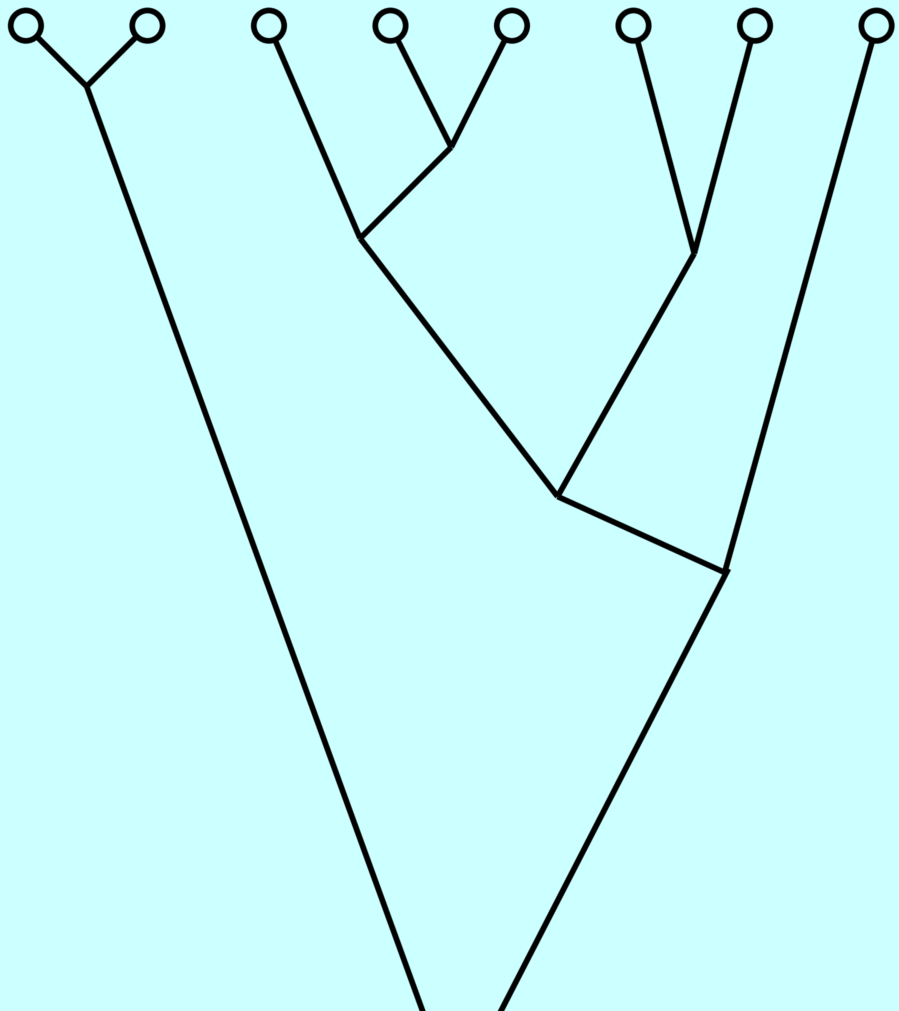
which is the expectation for points sampled from $g(x)$ of the ratio $\frac{f(x)}{g(x)}$.

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

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

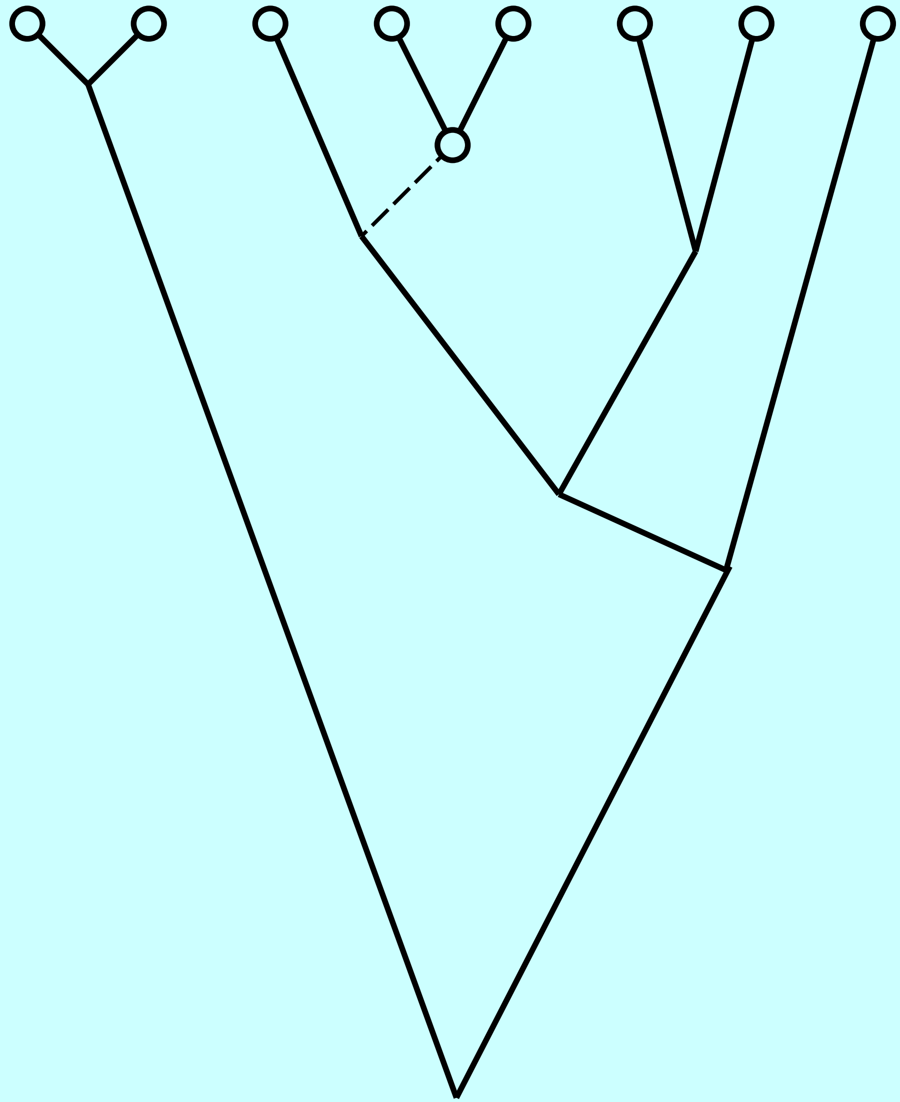
Rearrangement to sample points in tree space

A conditional coalescent rearrangement strategy

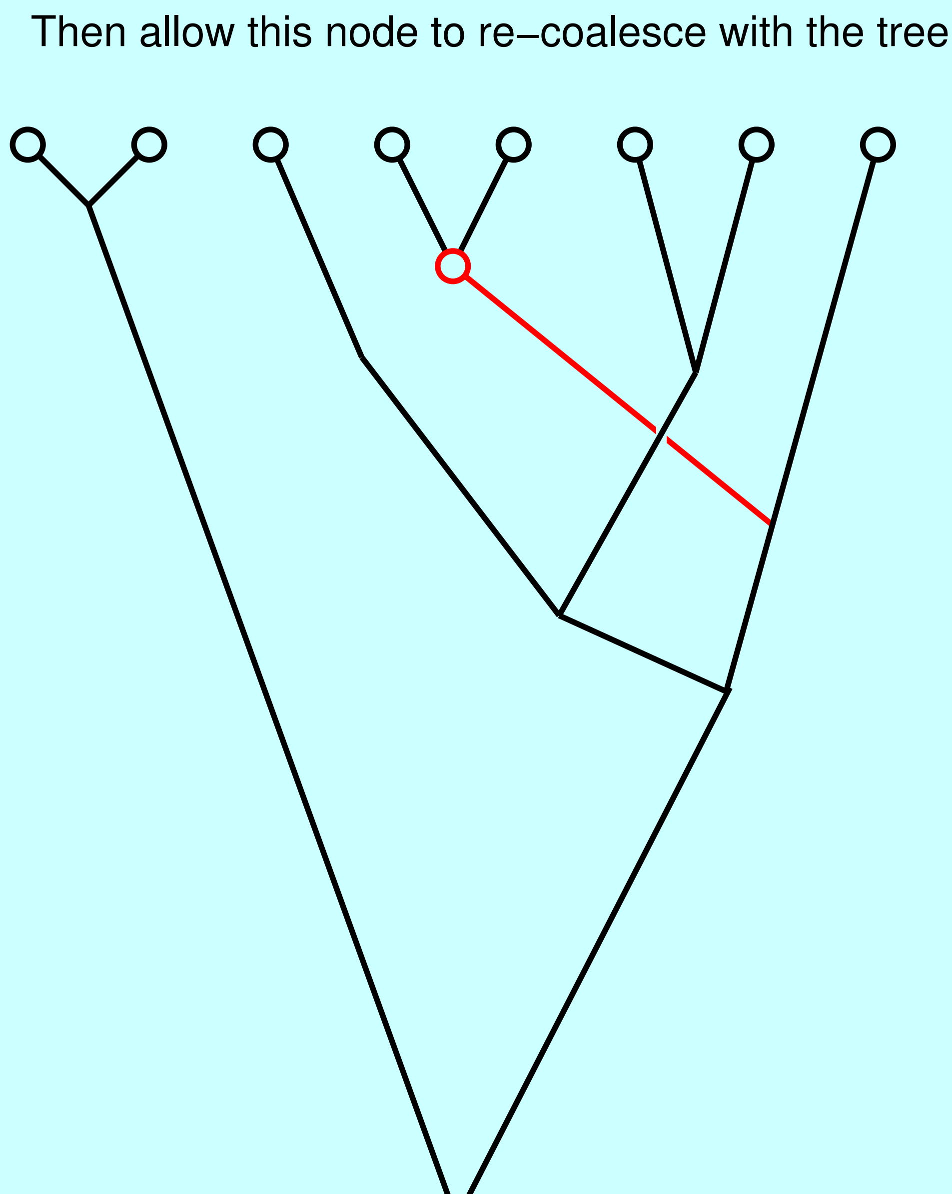


Dissolving a branch and regrowing it backwards

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

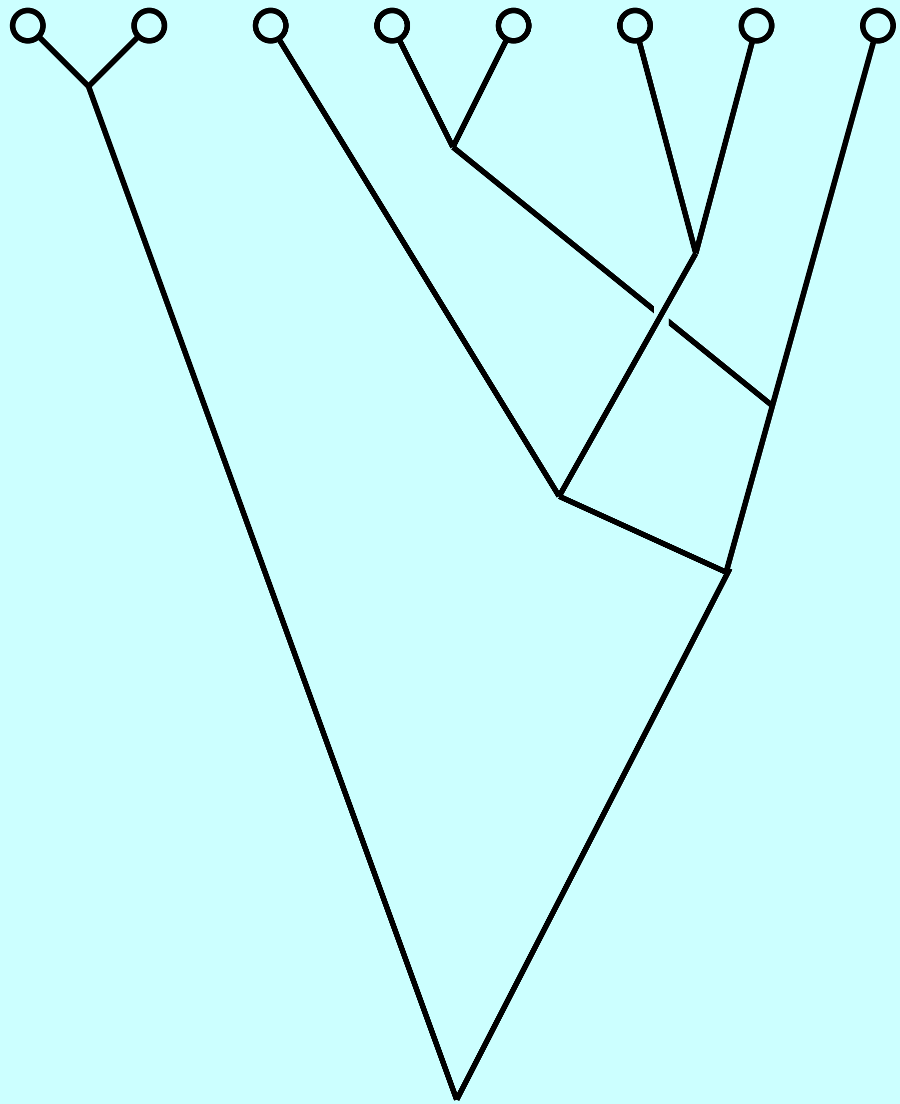


We allow it coalesce with the other branches



and this gives another coalescent

The resulting tree proposed by this process



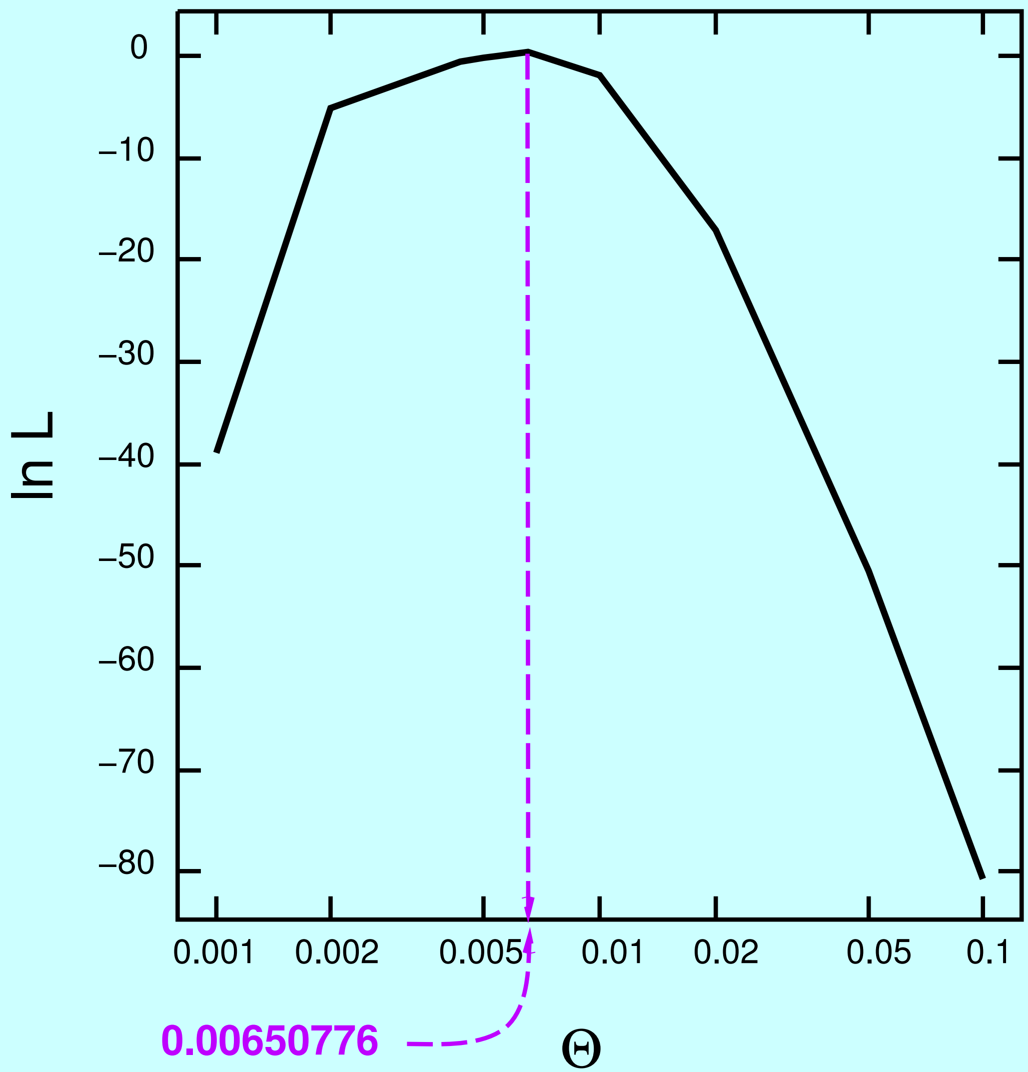
The resulting likelihood ratio is

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

(“Wait a second – where in this expression is the data?”) It’s in the sampling that gives you the G_i: the data biases those samples in the correct way.

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 $\Theta = 0.01$

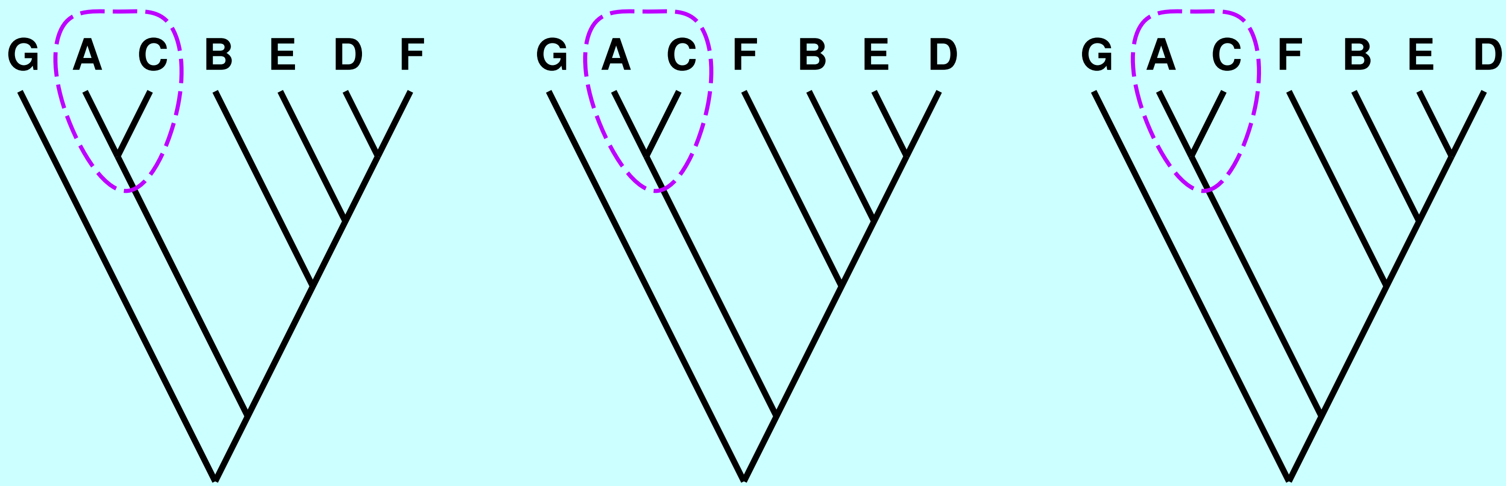


Major MCMC likelihood or Bayesian programs

- **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 (no migration or 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.
- **IM** and **IMa** by Rasmus Nielsen and Jody Hey. Two or more populations allowing both historical splitting and migration after that. No recombination yet.

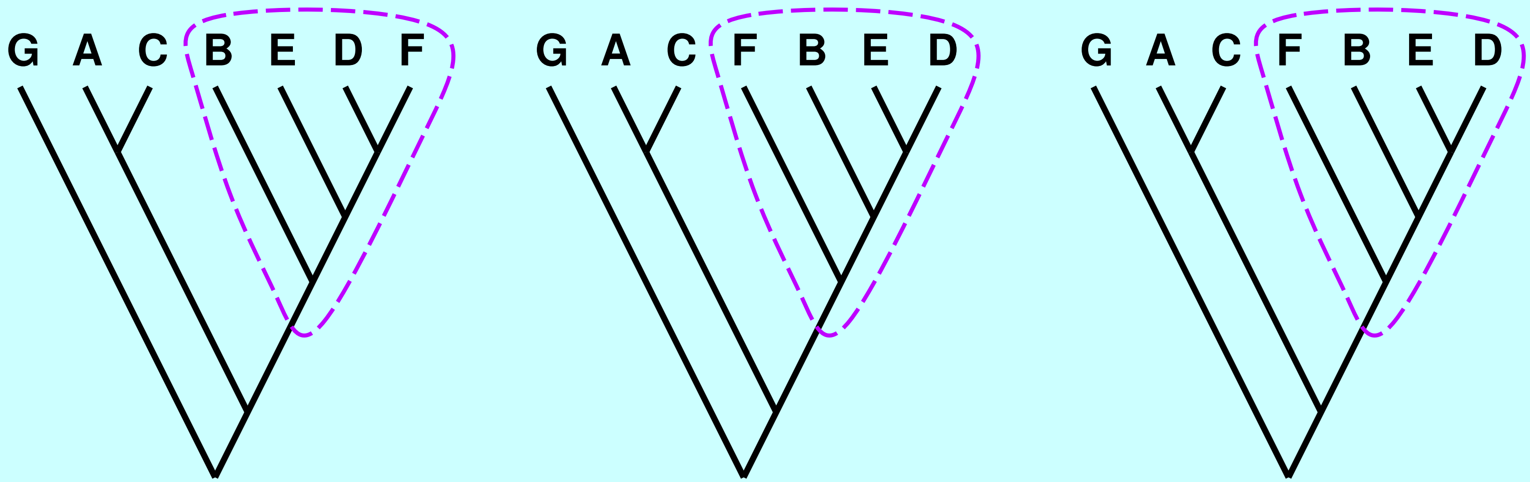
Trees we will use for consensus trees

Trees we will use for consensus trees

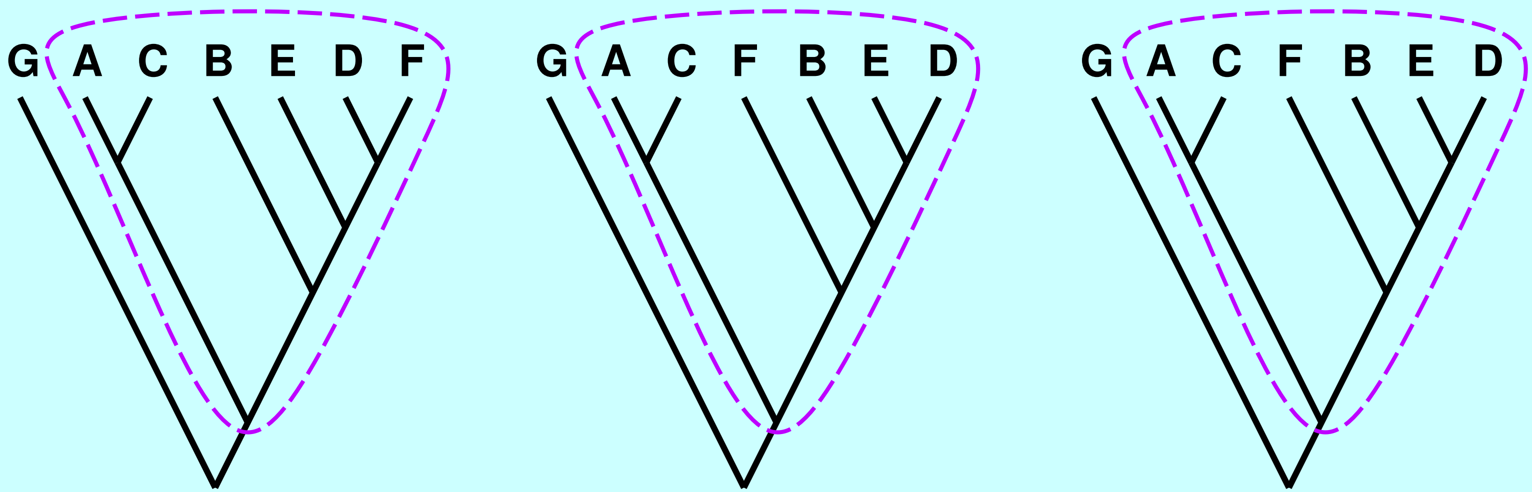


(for unrooted trees we would use partitions induced by branches instead of clades)

Trees we will use for consensus trees

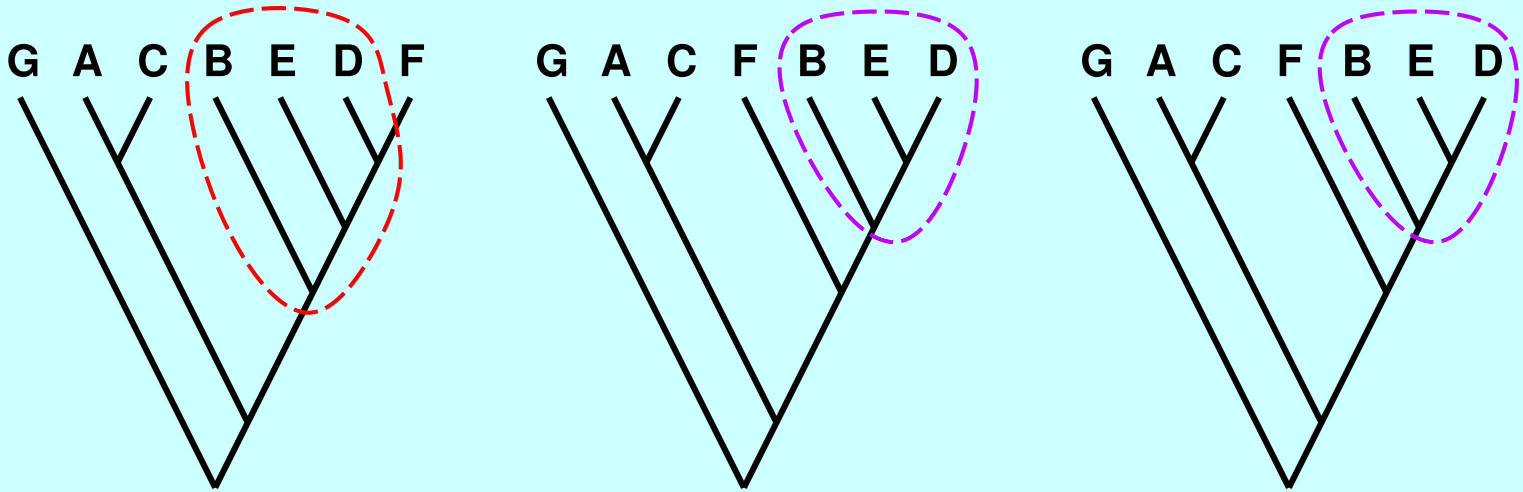


Trees we will use for consensus trees



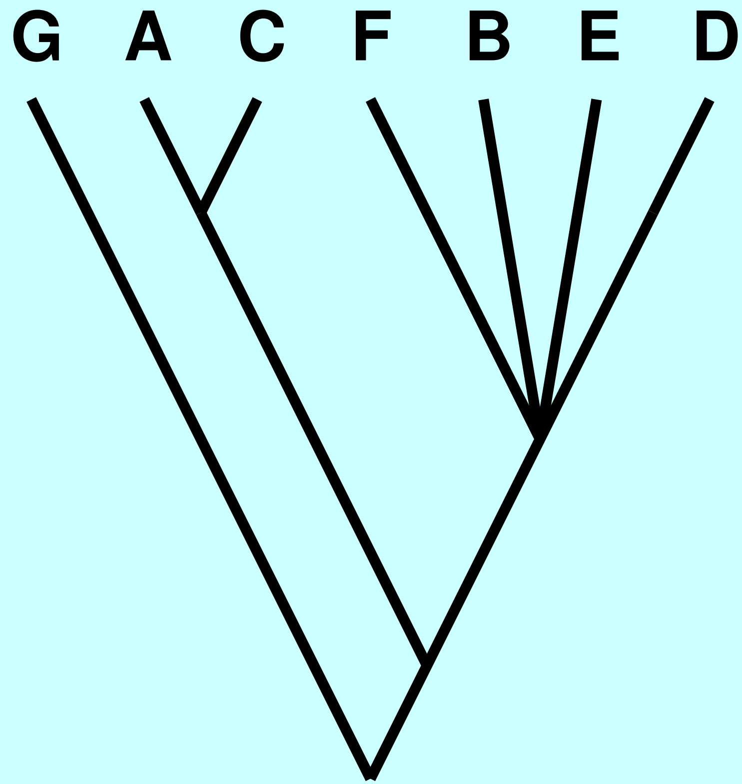
(Do we count this one if the trees are considered rooted? unrooted?)

Trees we will use for consensus trees

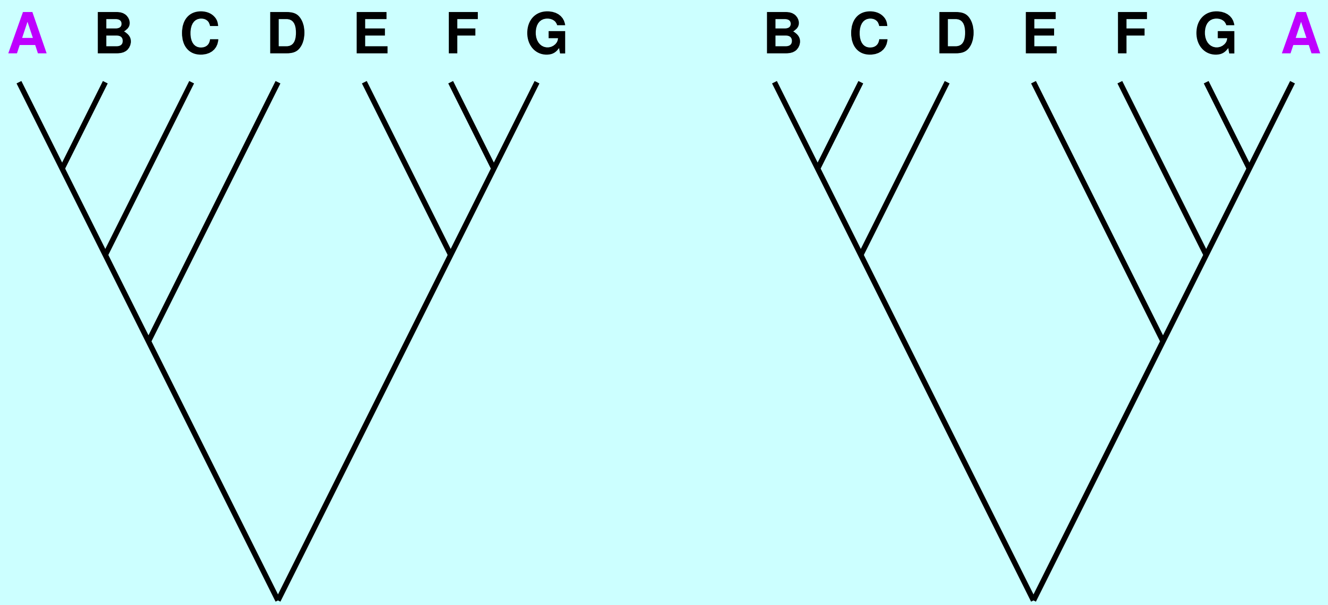


Here is a clade that is found on only two of the trees, so it is not included in the Strict Consensus Tree.

Their strict consensus tree

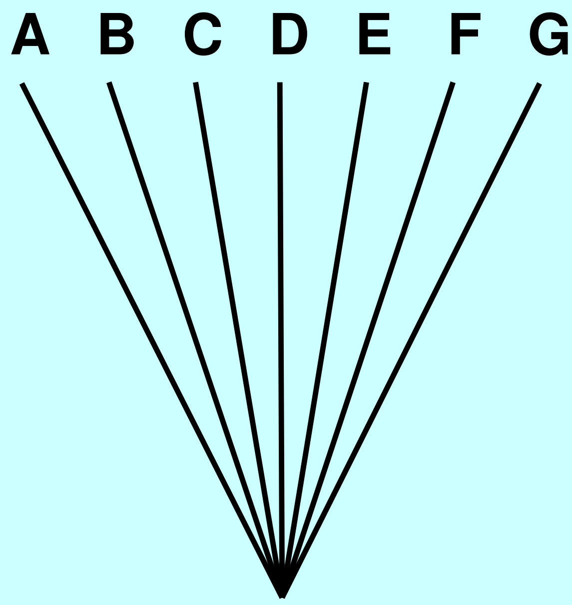


A distressing case for the strict consensus tree



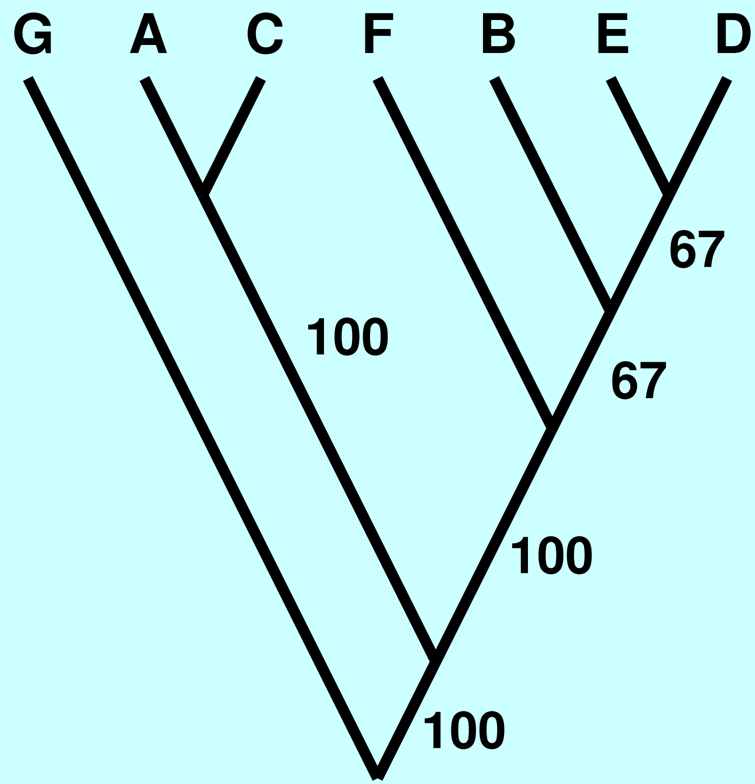
Only one species moves ...

A distressing case for the strict consensus tree



... but the strict consensus tree becomes totally unresolved.

Majority-rule consensus tree



The Adams consensus tree

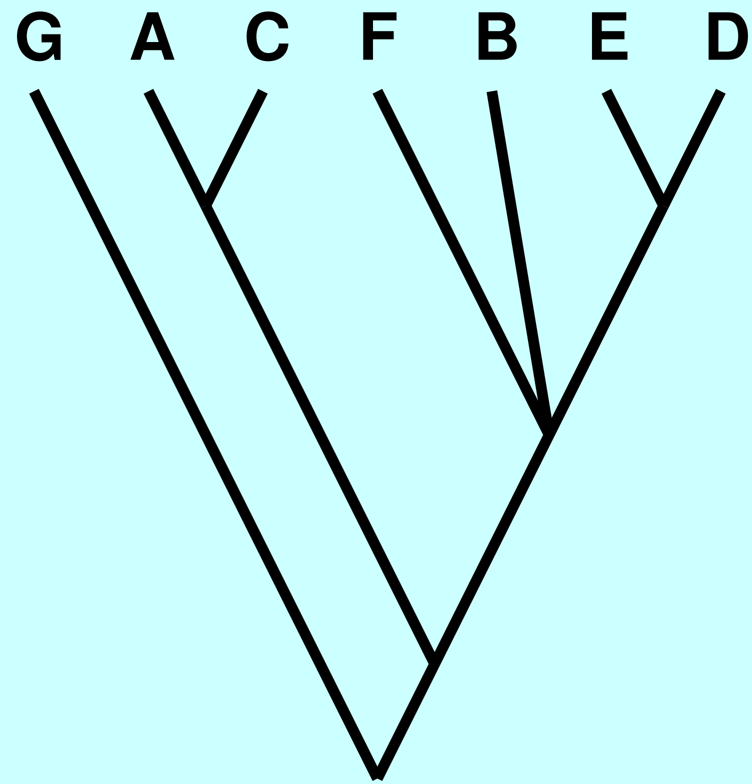
For rooted trees, Adams (1972, 1986) suggested:

1. Take all rooted triples on each tree.
2. Retain those that are not contradicted, where lack of resolution does not count as contradiction.
3. Construct a tree of these.

Two of the possible triples to examine

The green triple shows the same rooted topology on all three trees. The red triple is contradicted and does not get used in the Adams Consensus Tree.

The Adams consensus tree



Steel, Böcker, and Dress's shocking disproof

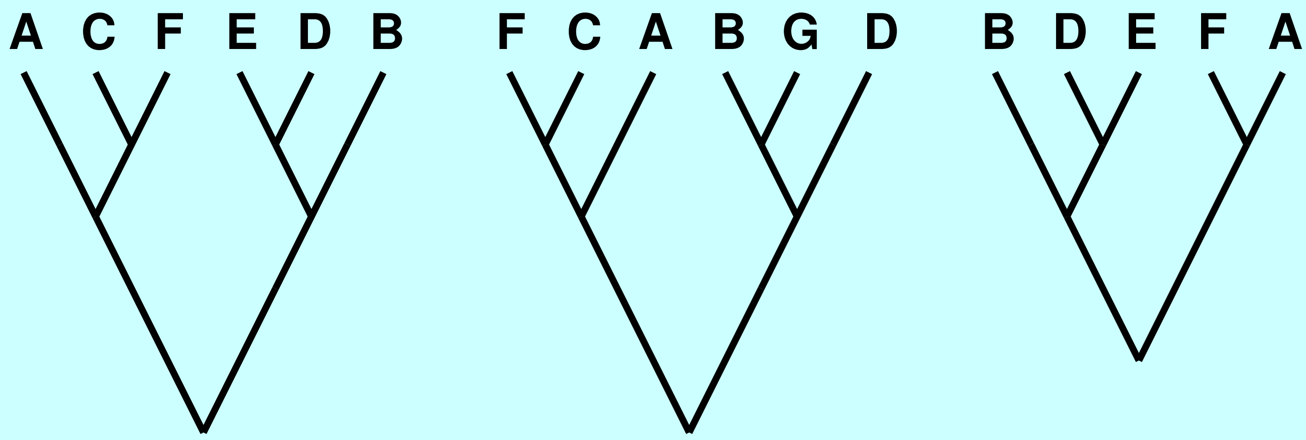
Steel, M., S. Böcker, and A. W. M. Dress. 2000. Simple but fundamental limits for supertree and consensus tree methods. *Systematic Biology* **49(2):** 363-368.

They put forward three minimal requirements for an unrooted Adams-like consensus tree based on observations of quartets, rather than triples. Note that a quartet, like a triple, has three possible topologies, but unrooted ones: ((A,B),(C,D)) and ((A,C),(B,D)) and ((A,D),(B,C)).

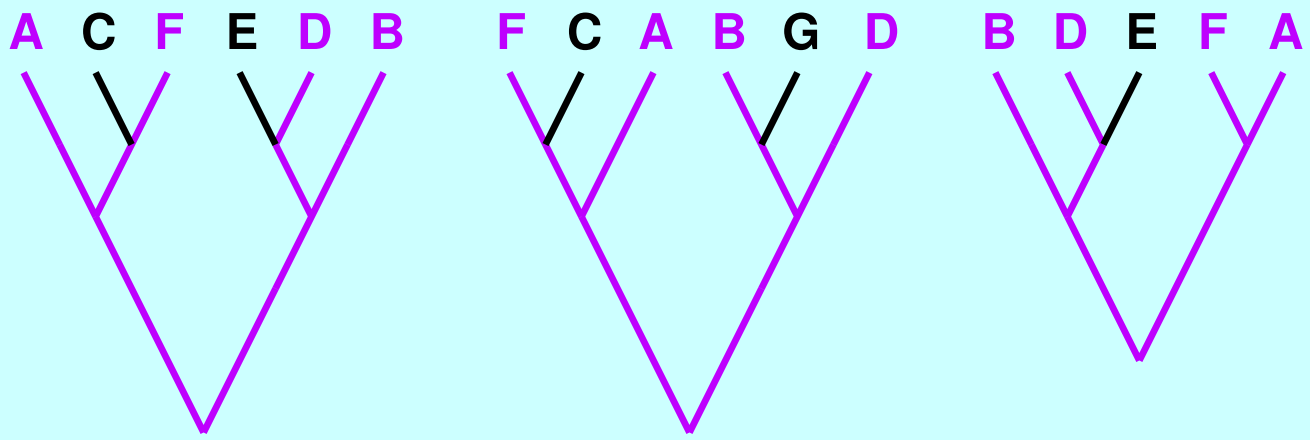
- The result shouldn't be altered by relabelling all the species in a consistent way.
- The result should not depend on the order in which the trees are input.
- If a quartet appears in all trees, it should appear in the consensus.

Alas, they then show there is no consensus tree method for unrooted trees that can satisfy all of these!

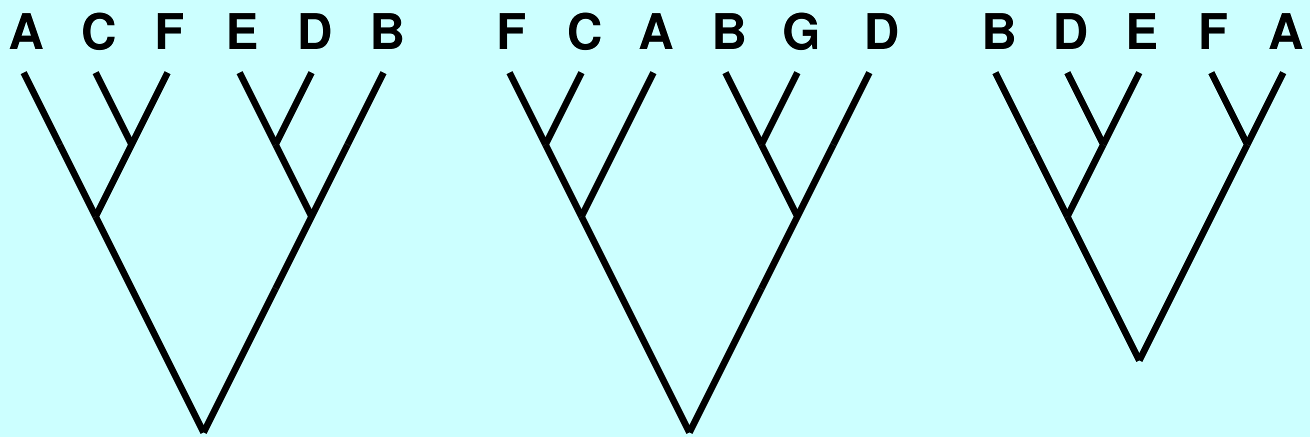
A consensus subtree

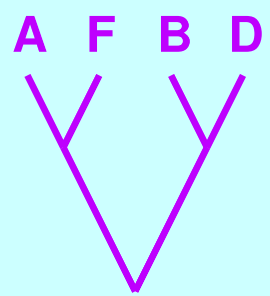


A consensus subtree

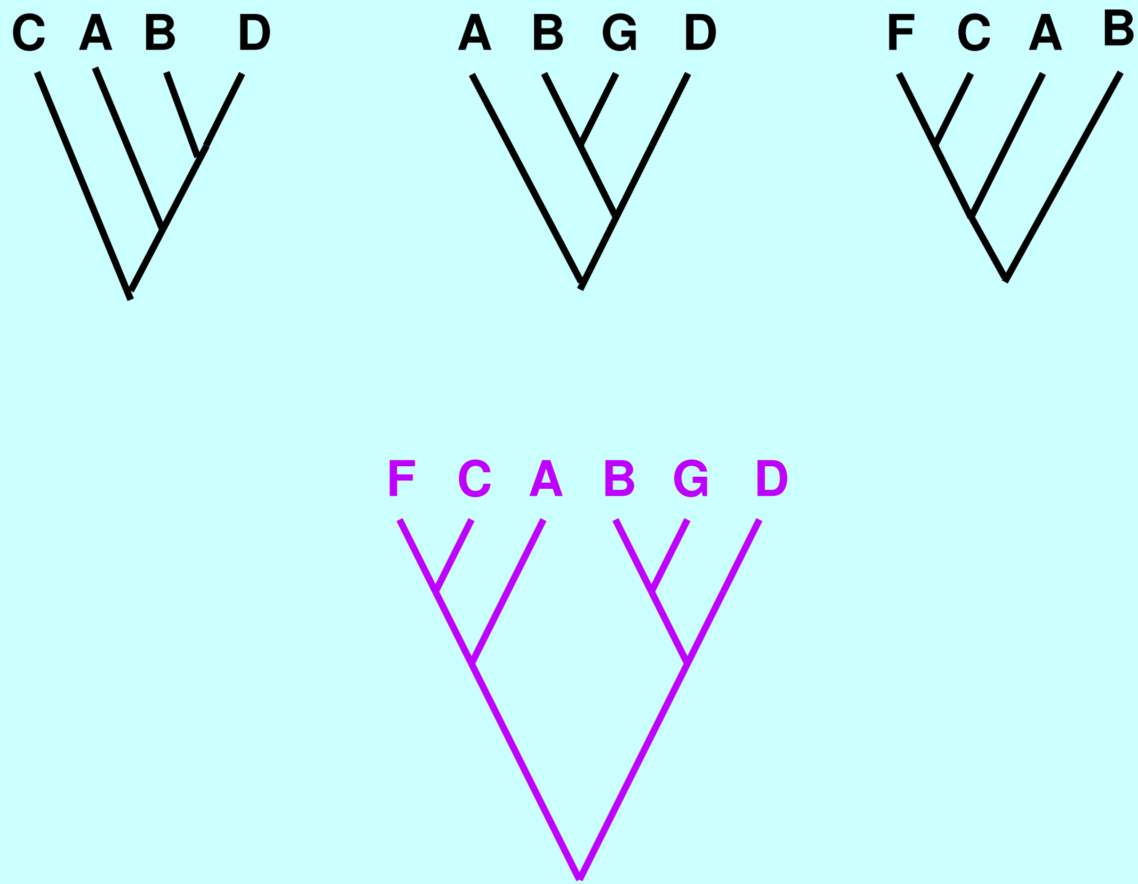


A consensus subtree





A supertree



Construct a tree with all tips, for which each of the smaller trees is a subtree. What to do if there is conflict? There are various suggestions.