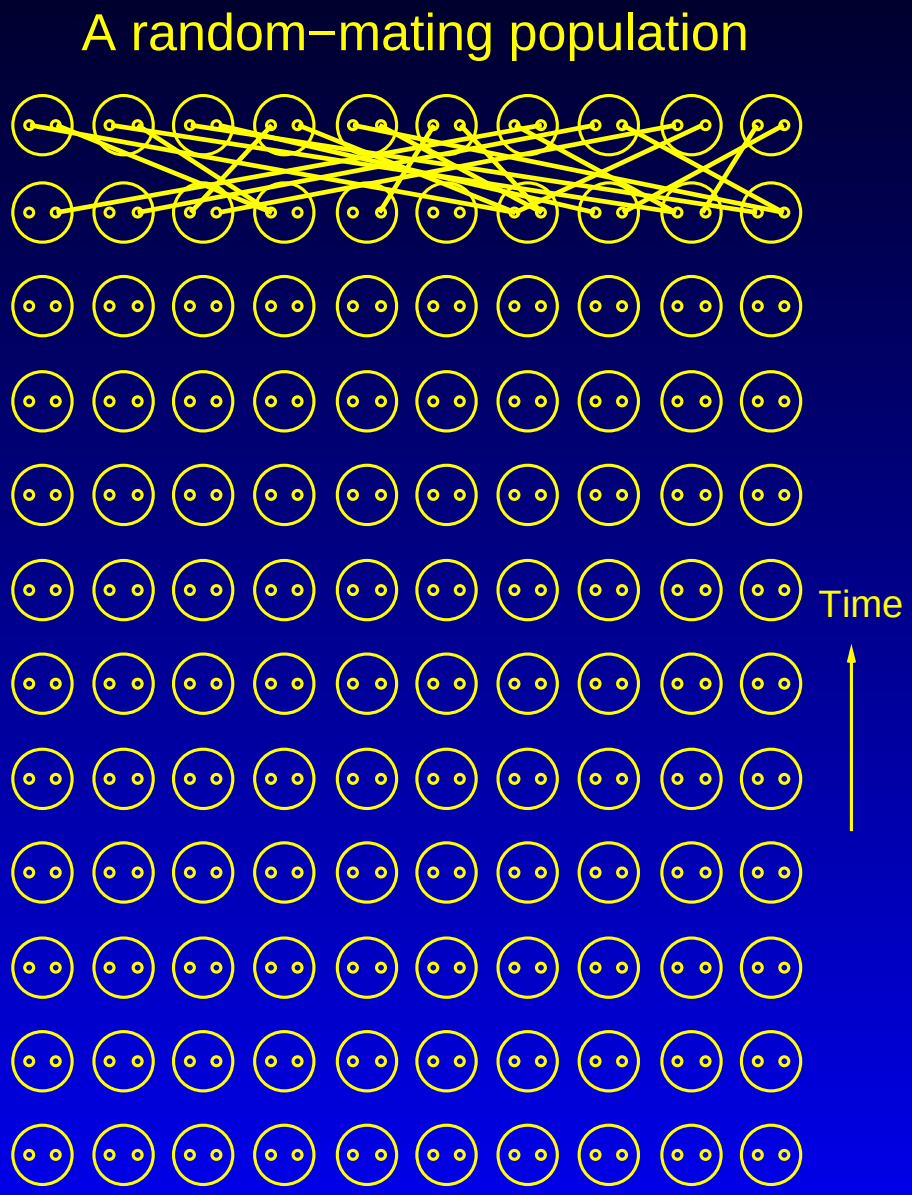
Some of these slides have been borrowed from Dr. Paul Lewis, Dr. Joe Felsenstein. Thanks!

Paul has many great tools for teaching phylogenetics at his web site:

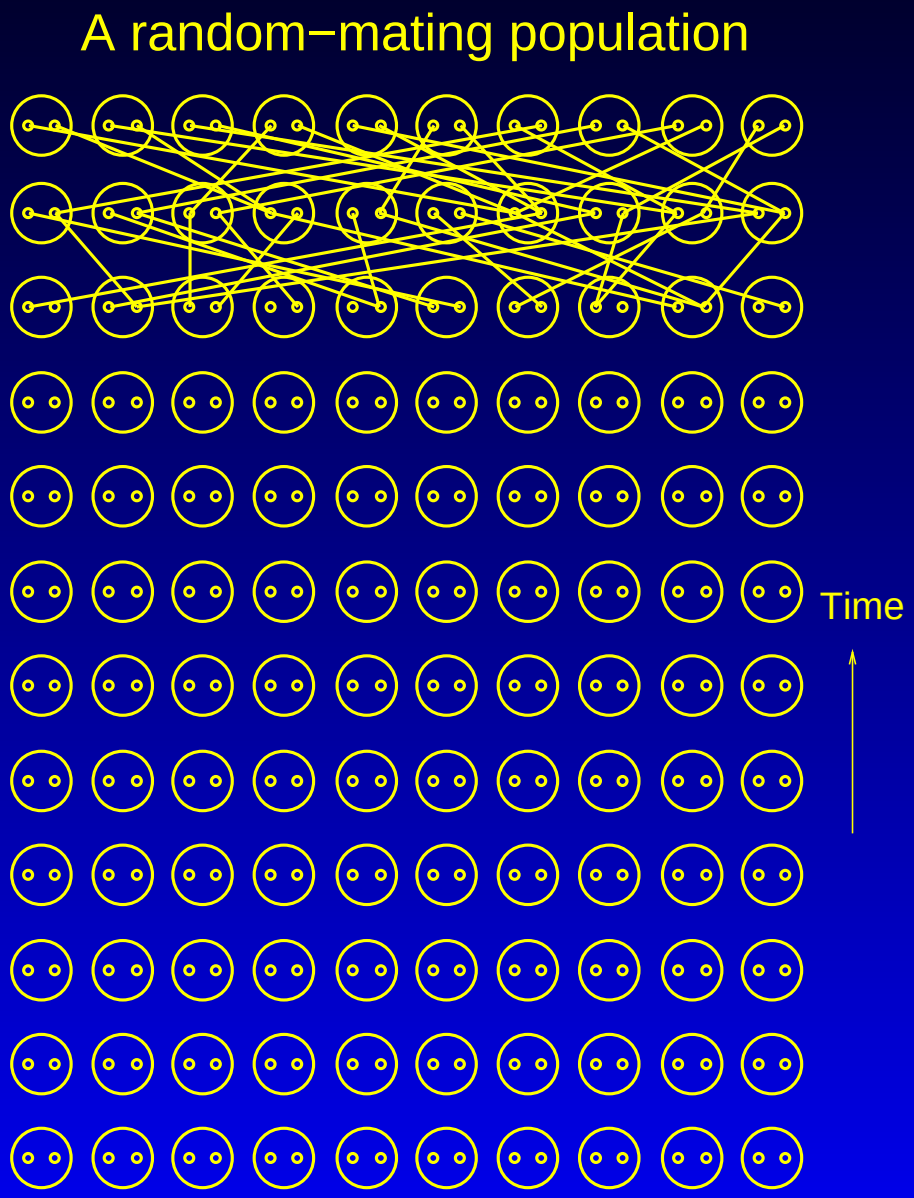
http://hydrodictyon.eeb.uconn.edu/people/plewis

Gene copies in a population of 10 individuals

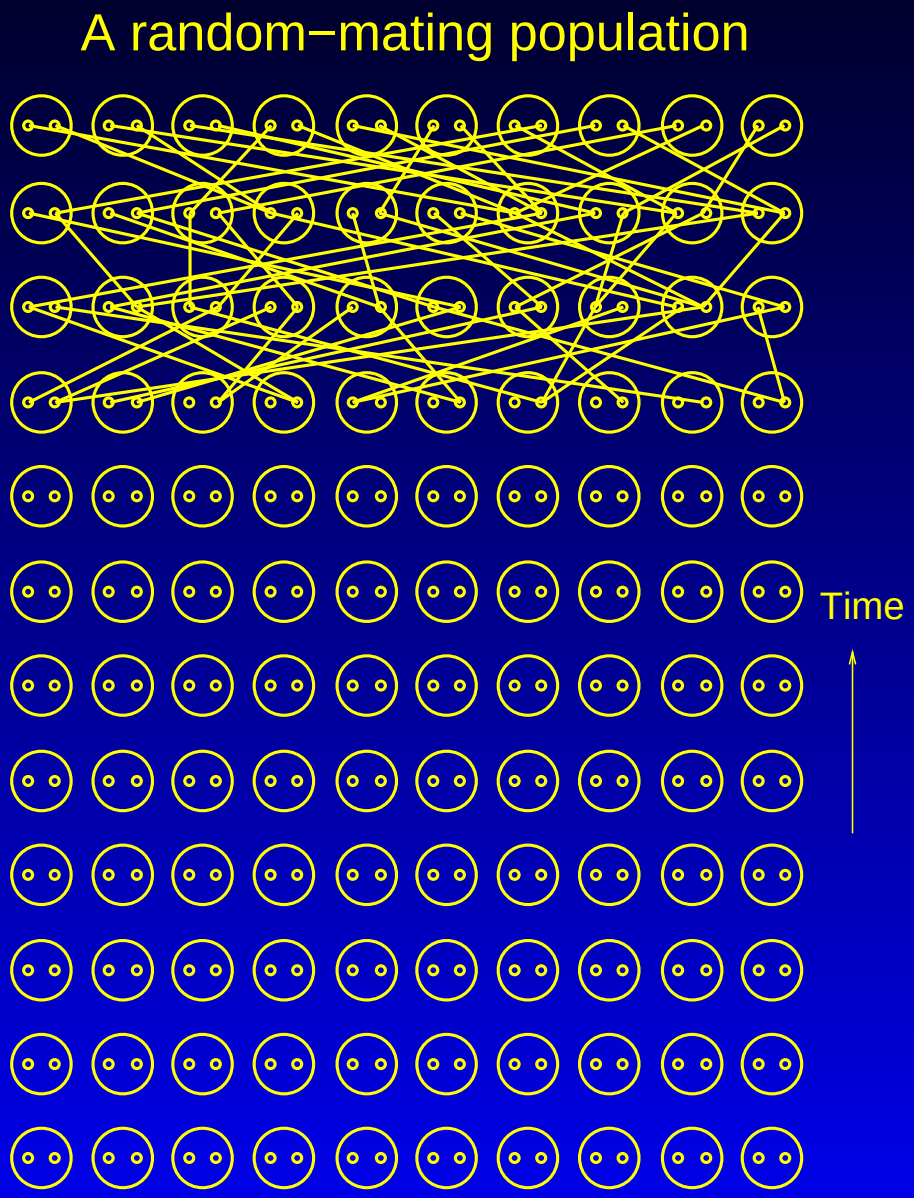
Going back one generation



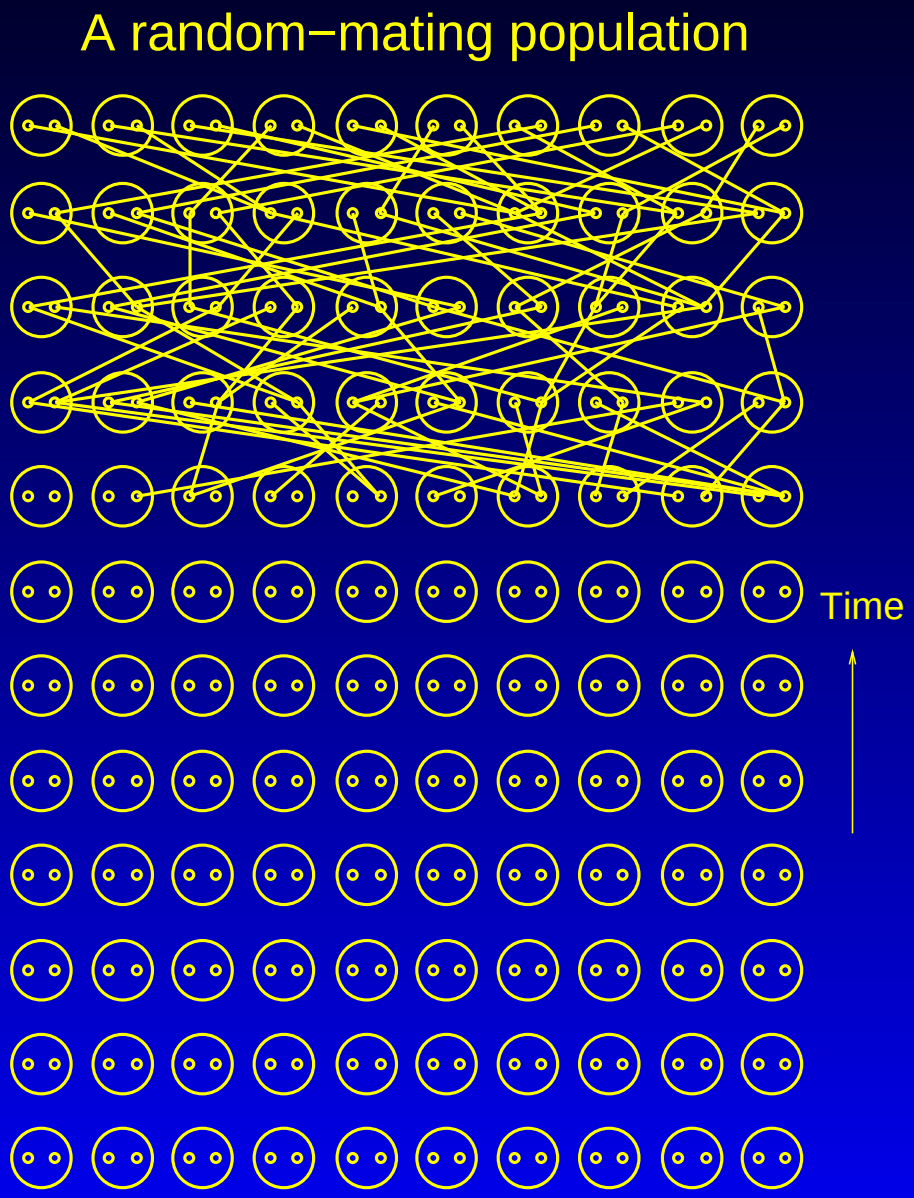
... and one more



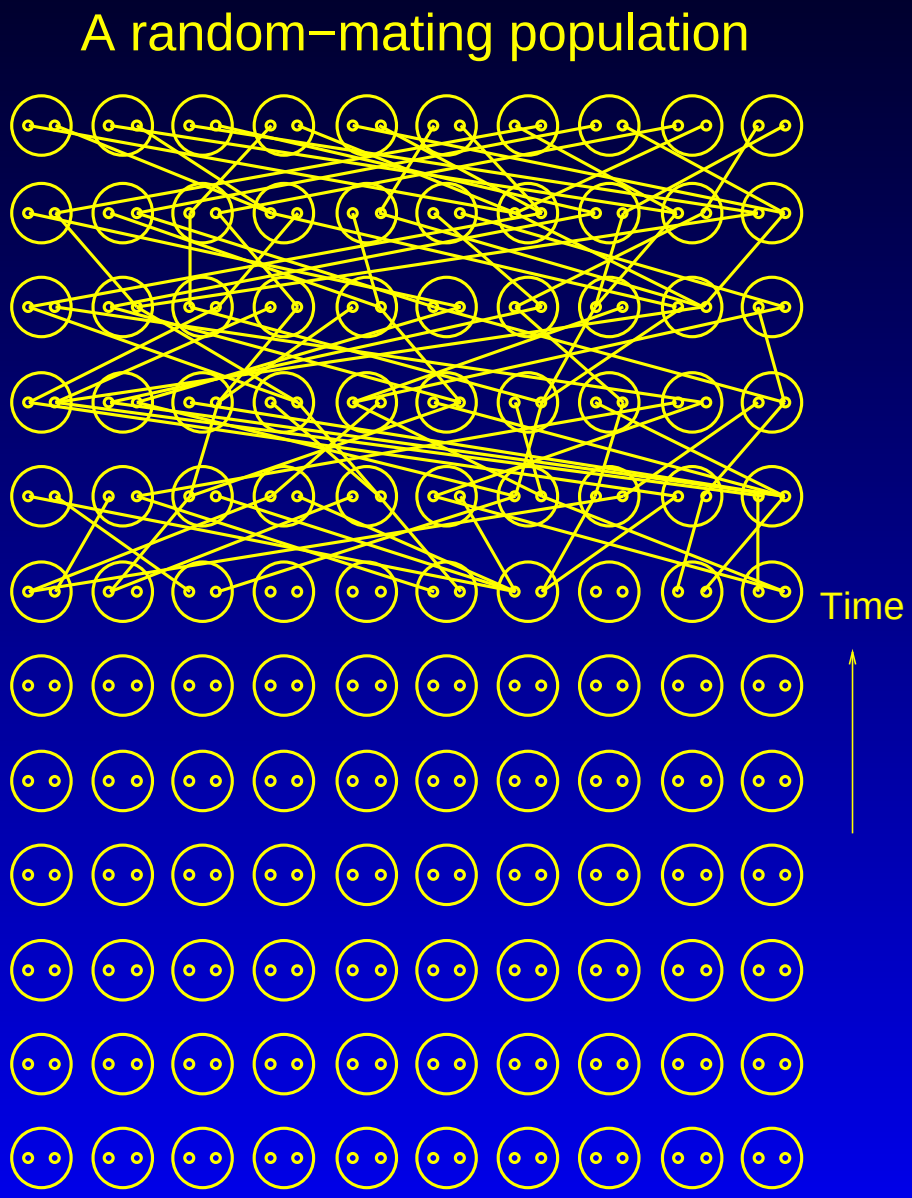
... and one more



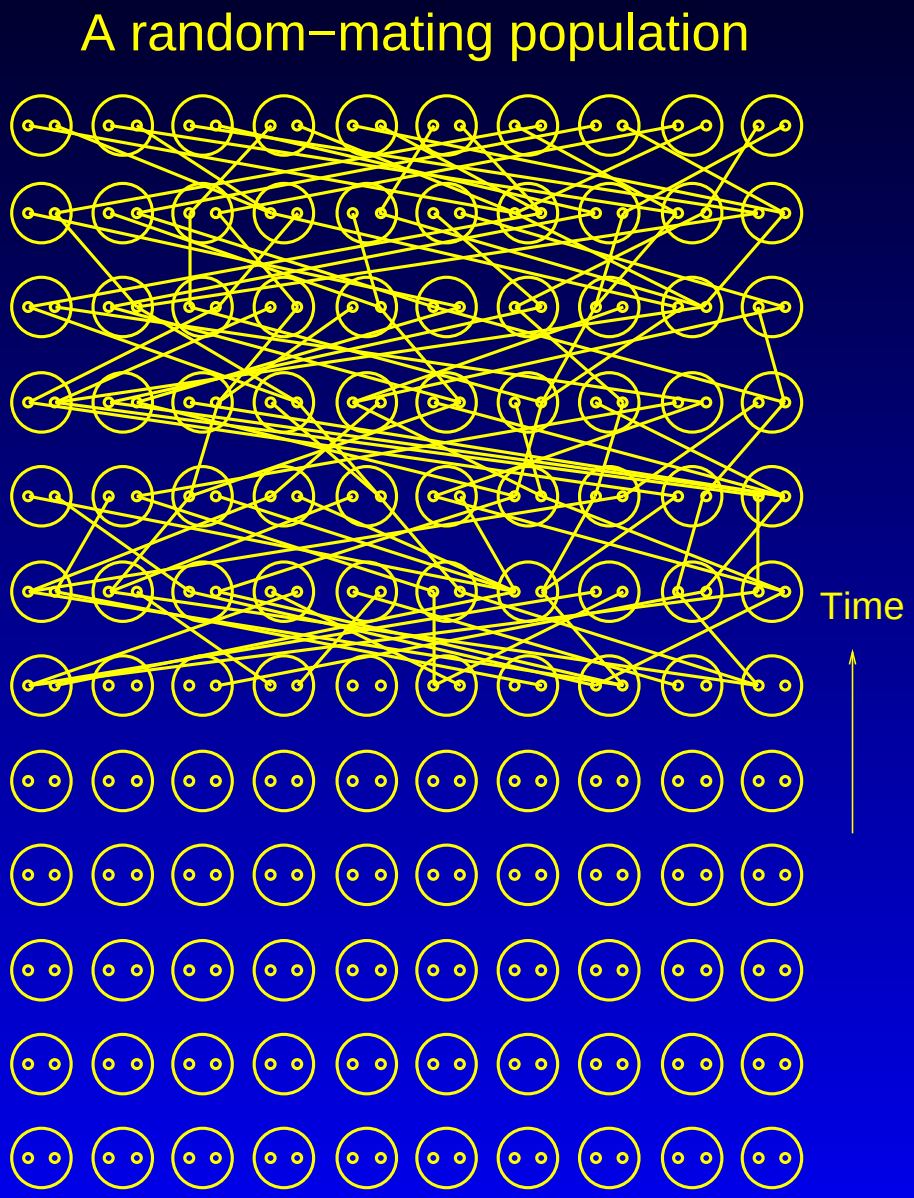
... and one more



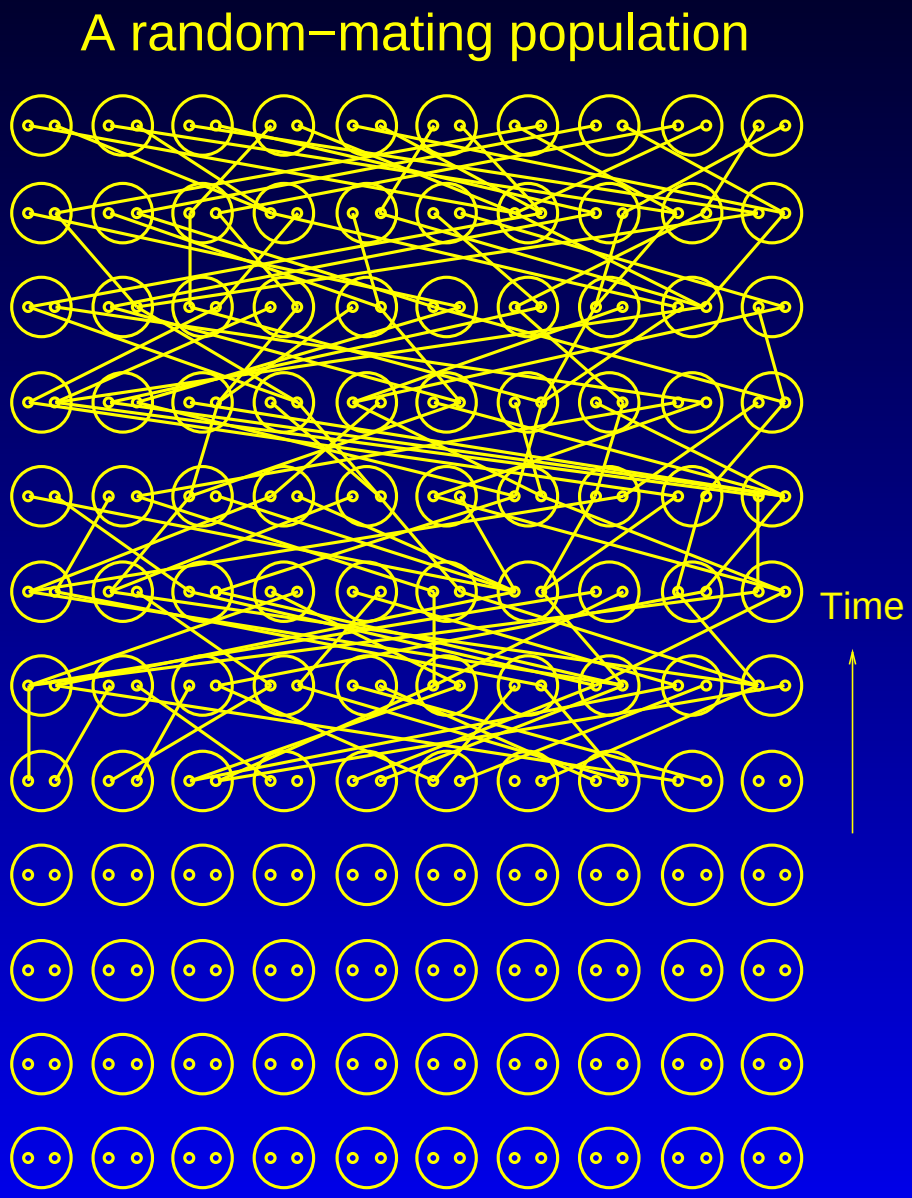
... and one more



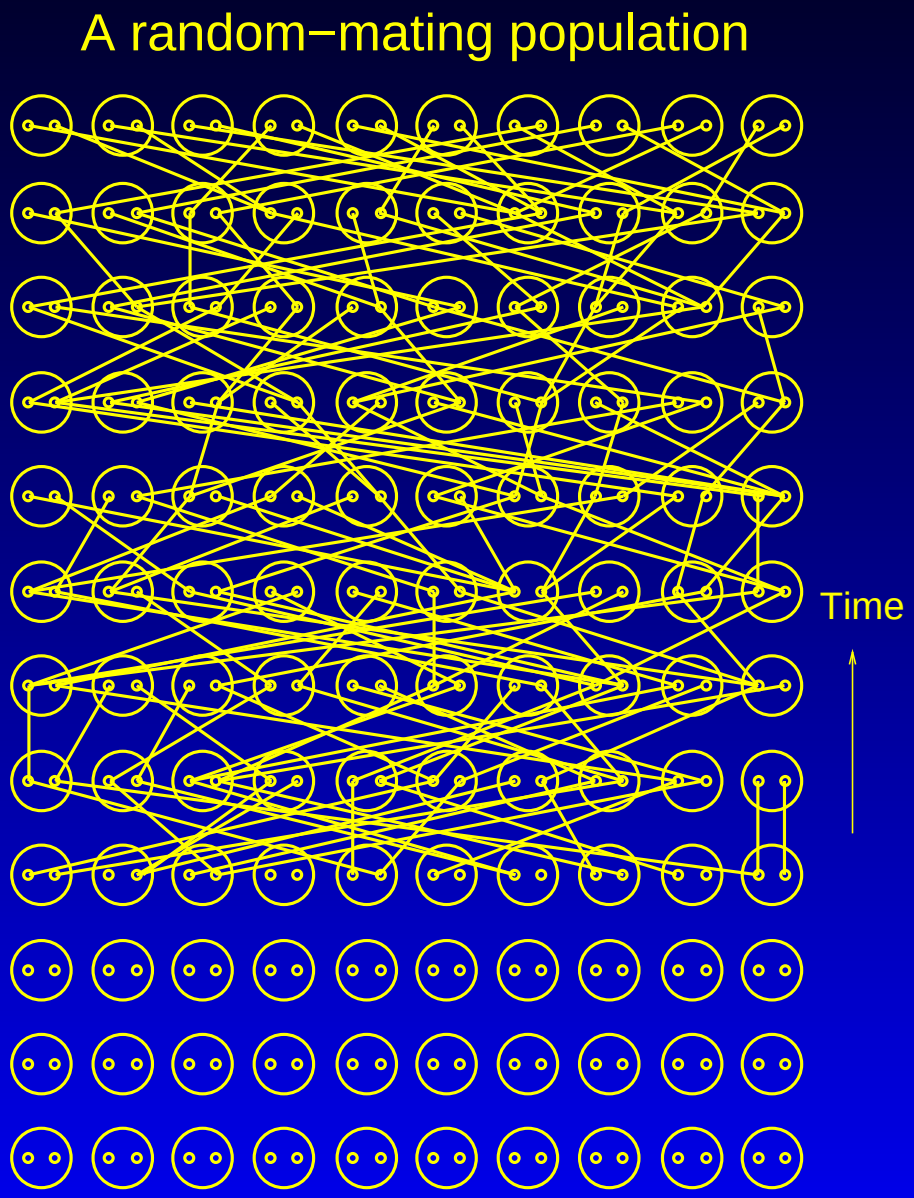
... and one more



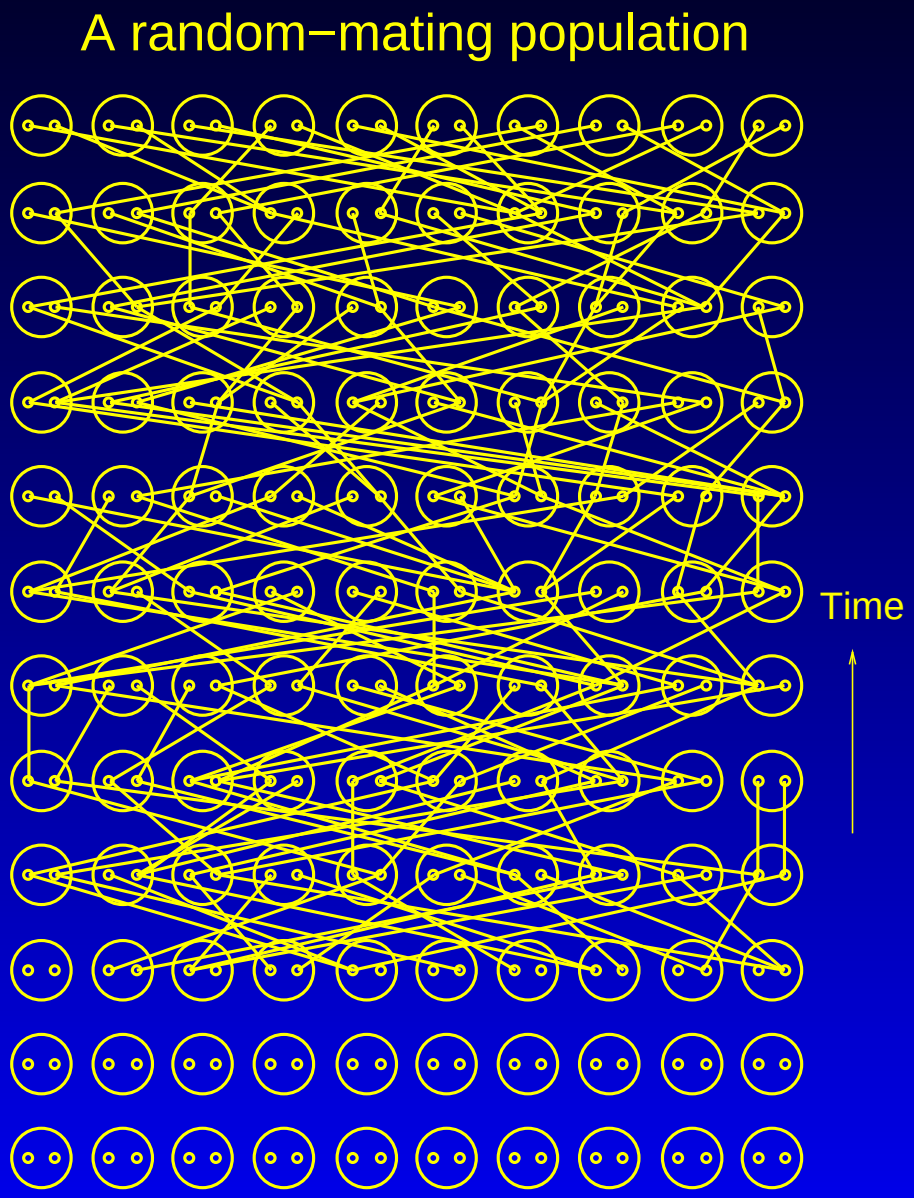
... and one more



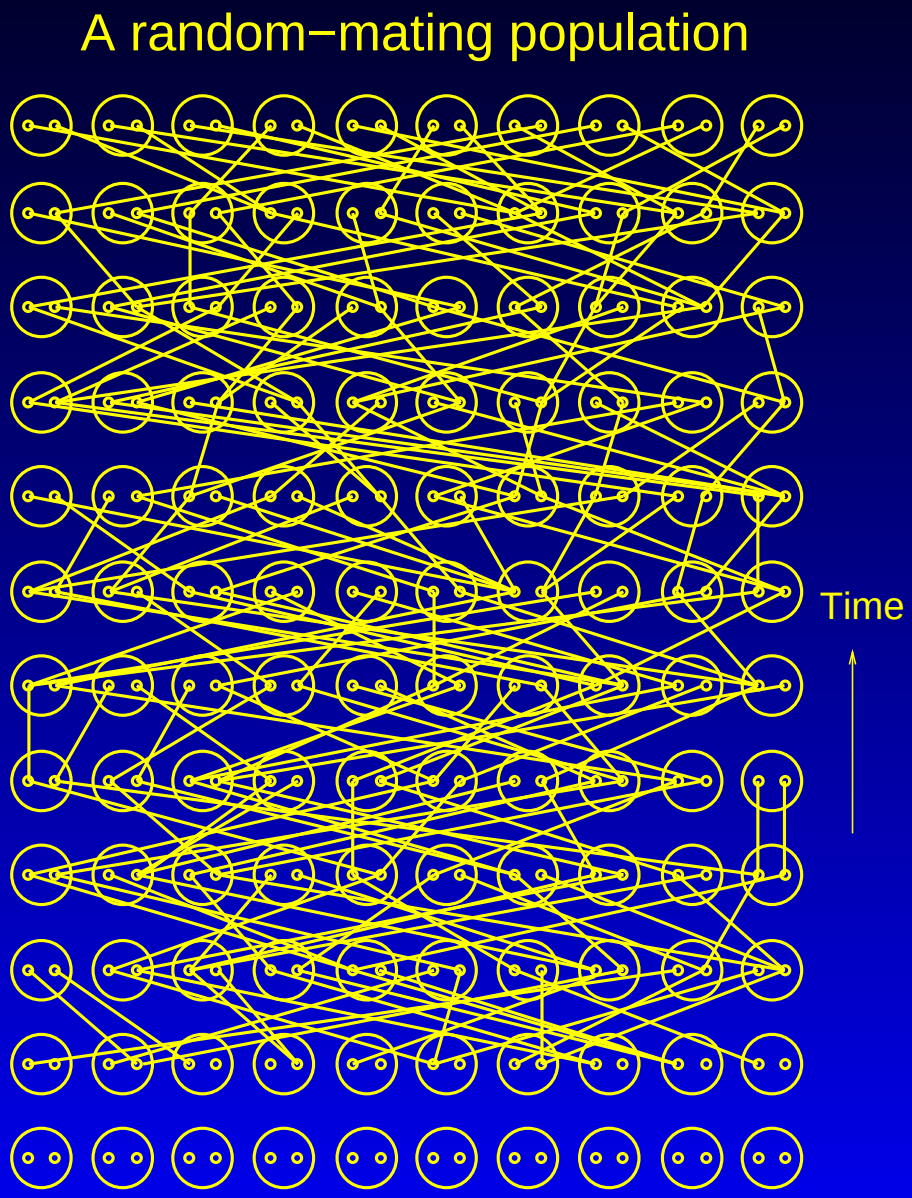
... and one more



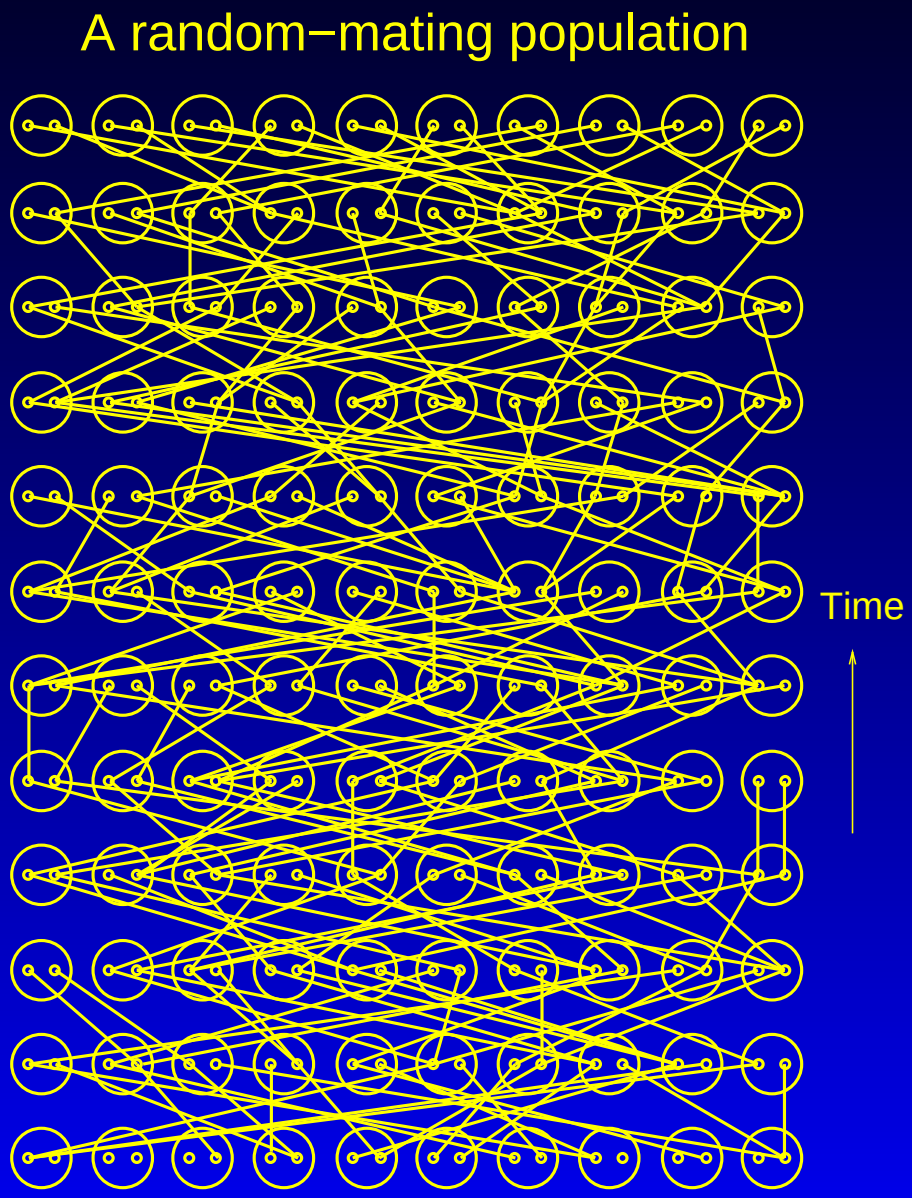
... and one more



... and one more

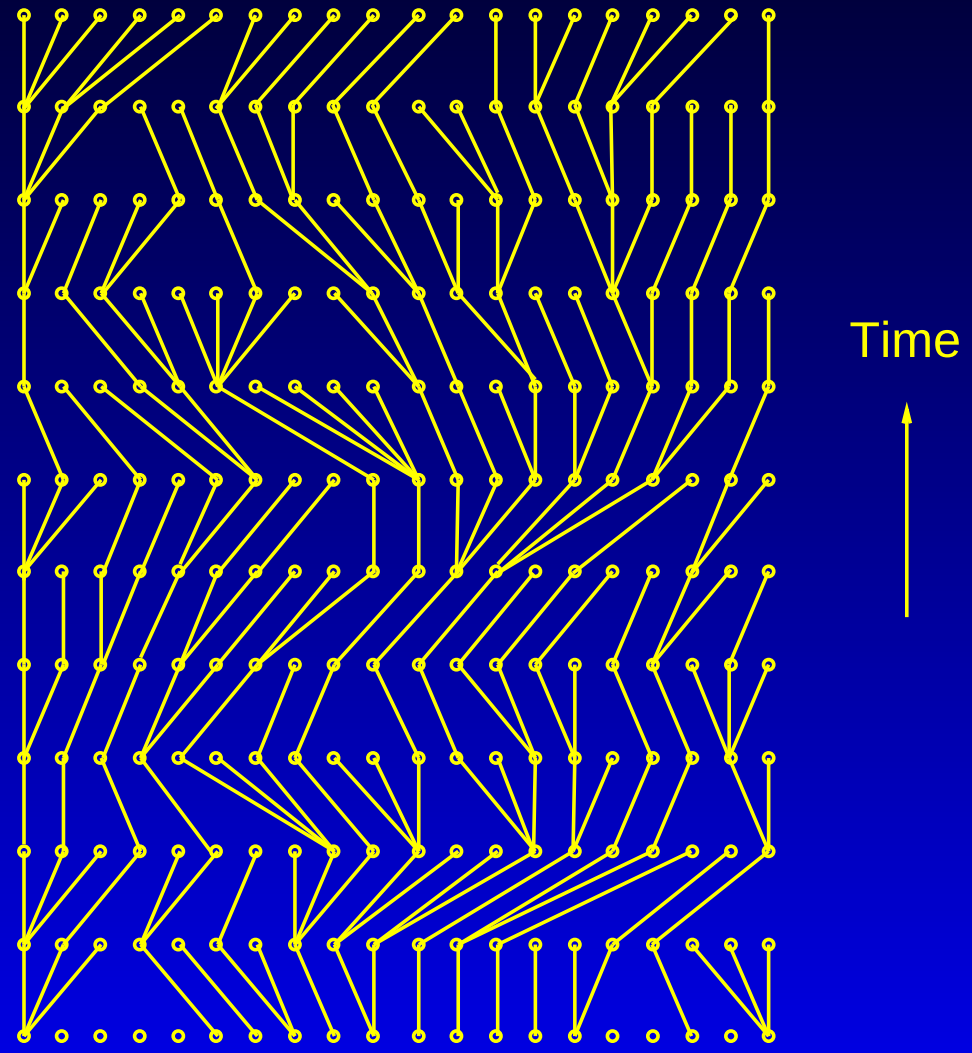


... and one more



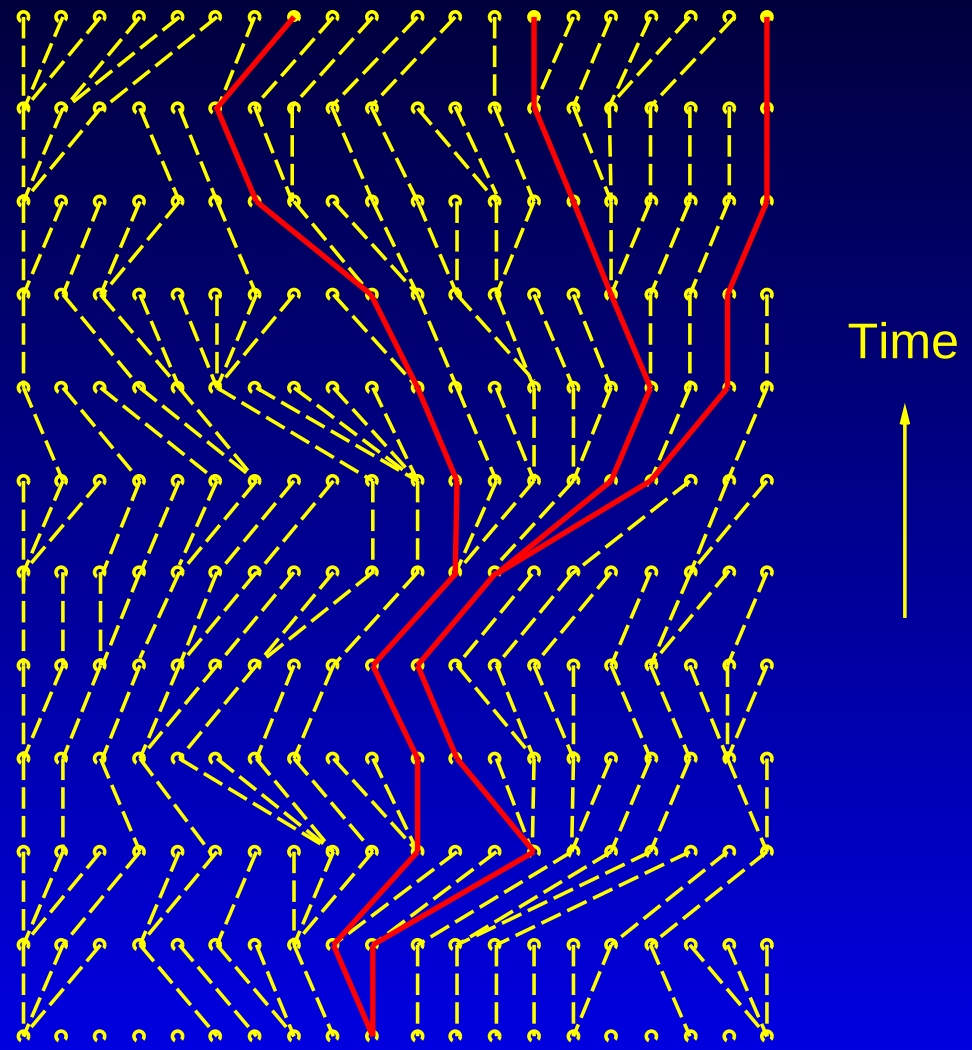
The genealogy of gene copies is a tree

Genealogy of gene copies, after reordering the copies

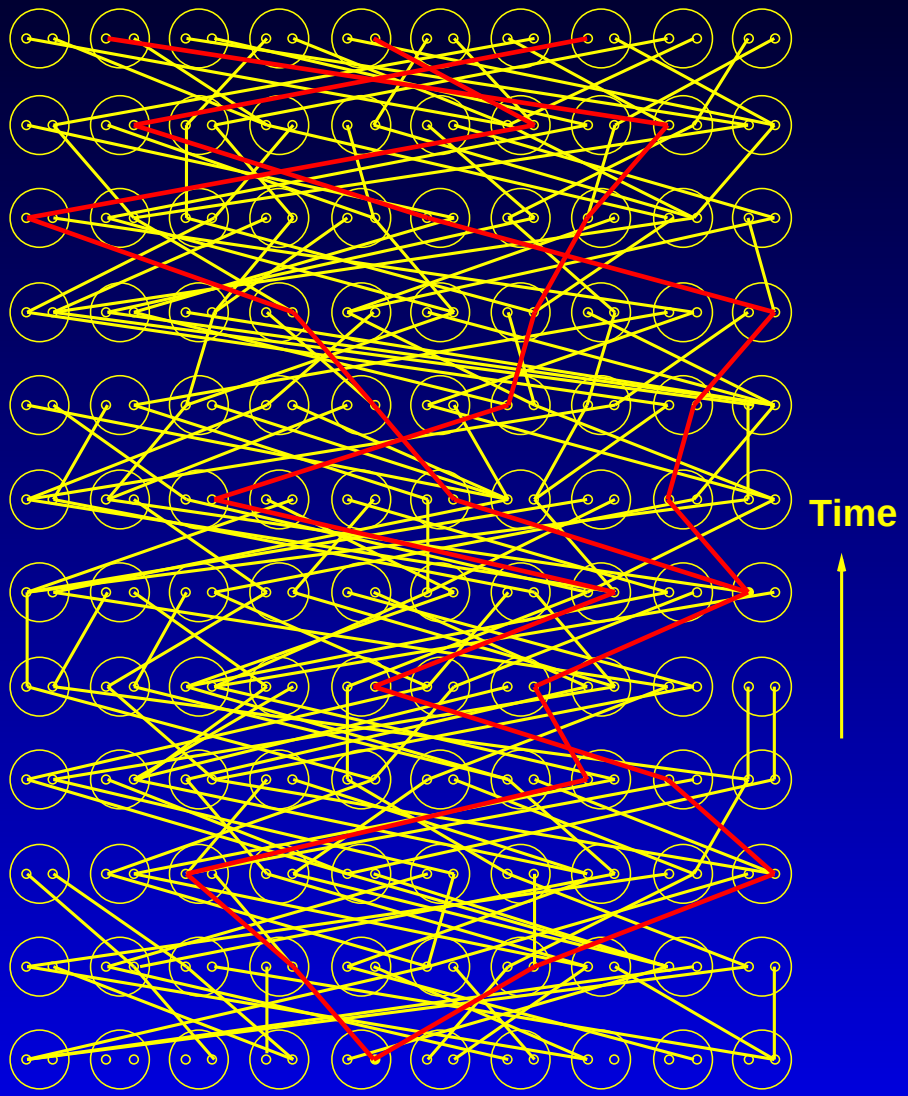


Ancestry of a sample of 3 copies

Genealogy of a small sample of genes from the population



Here is that tree of 3 copies in the pedigree



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

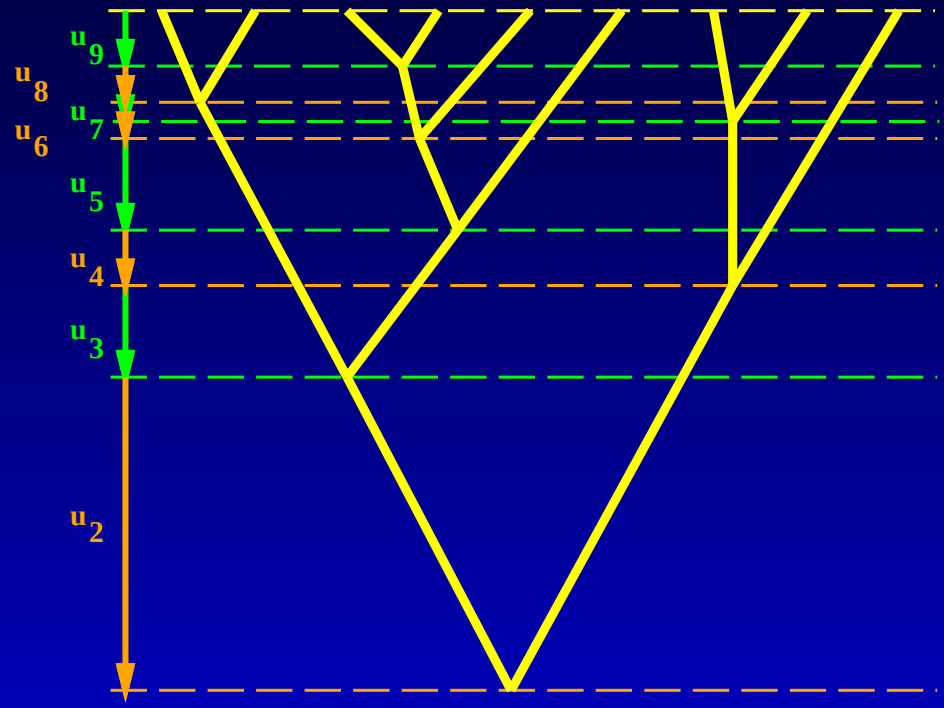
k copies to coalesce to

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

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

$= 4N\left(1-\frac{1}{n}\right)$ 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 $k-1$ one generation earlier is (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:

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

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

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

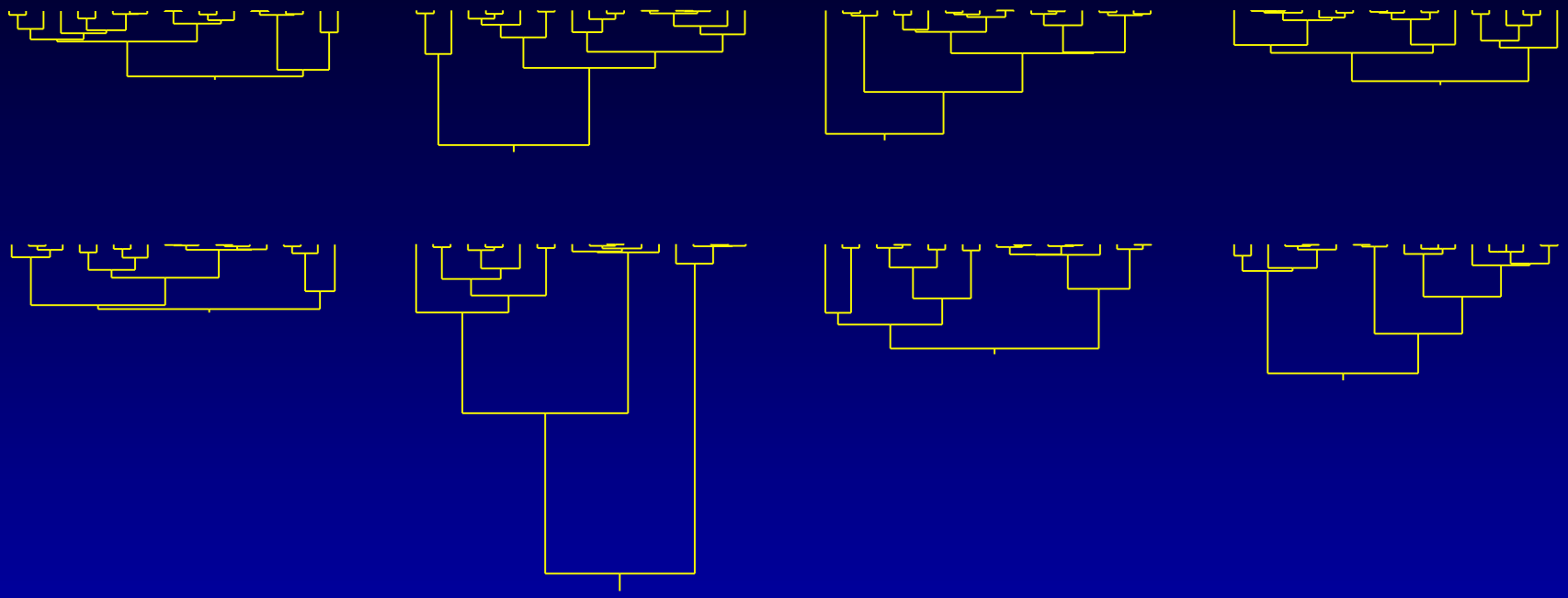
Note that increasing the population size by a factor of 10 reduces the coalescent rate for pairs by about 10-fold, but reduces the rate for triples (or more) by about 100-fold.

The coalescent

To simulate a random genealogy, do the following:

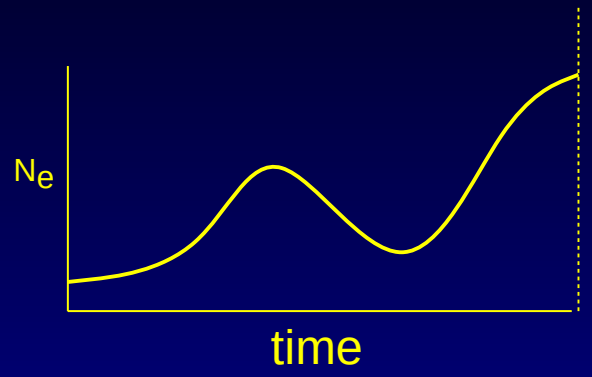
1. Start with k lineages
2. Draw an exponential time interval with mean $4N/(k(k-1))$ generations.
3. Combine two randomly chosen lineages.
4. Decrease k by 1.
5. If $k = 1$, then stop
6. Otherwise go back to step 2.

Random coalescent trees with 16 lineages

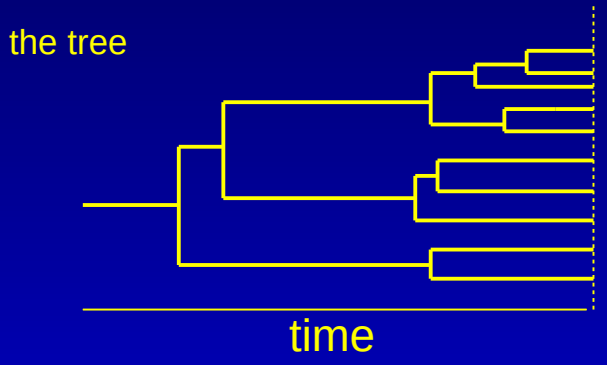


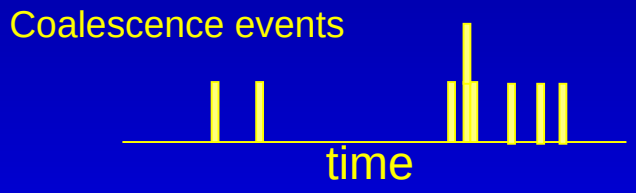
Coalescence is faster in small populations

Change of population size and coalescents



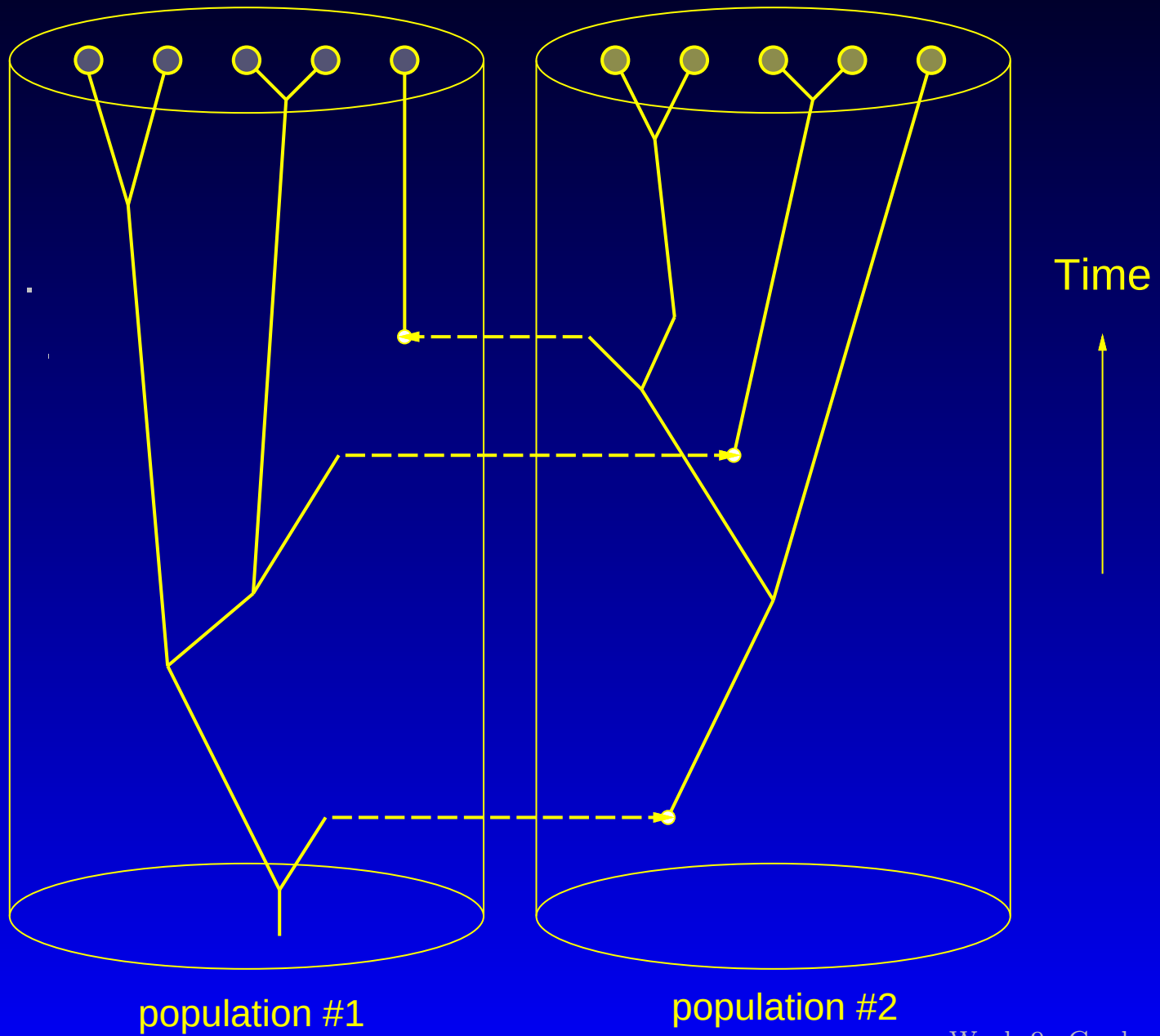
the changes in population size will produce waves of coalescence



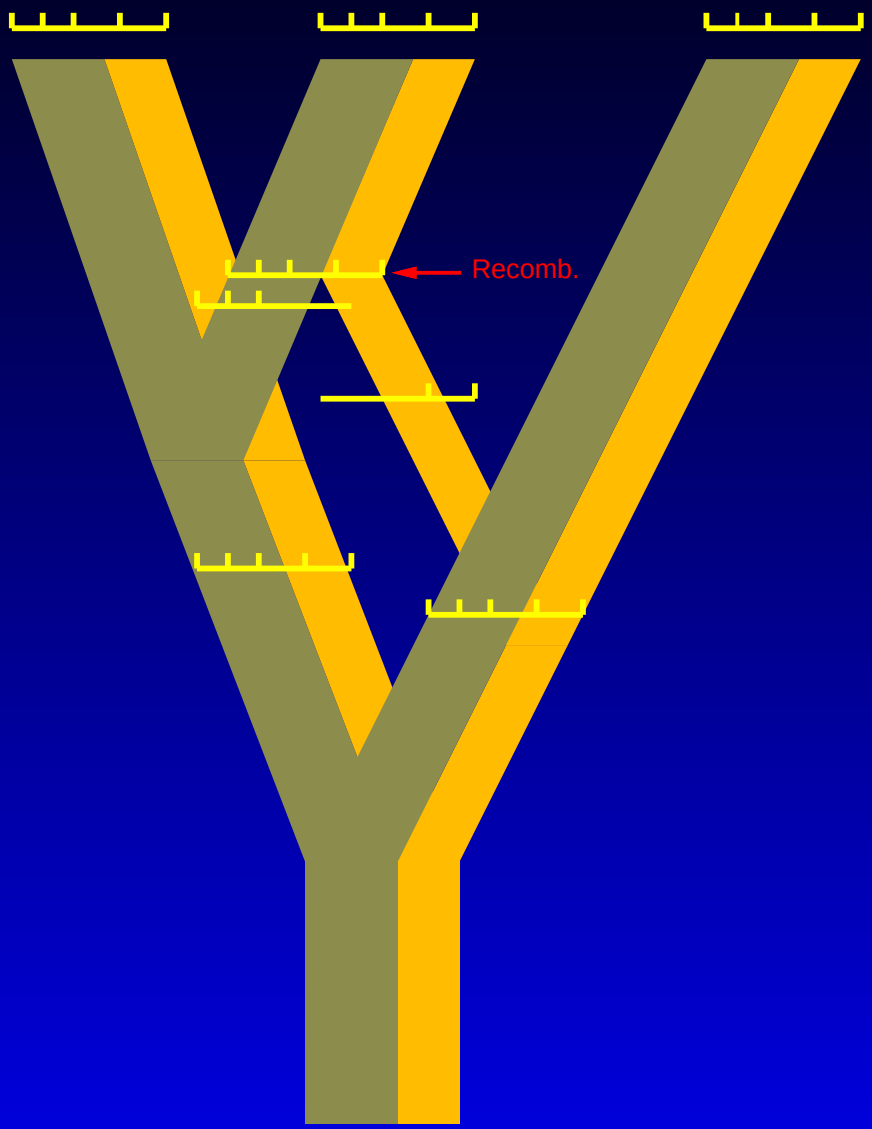


The parameters of the growth curve for N_e can be inferred by likelihood methods as they affect the prior probabilities of those trees that fit the data.

Migration can be taken into account



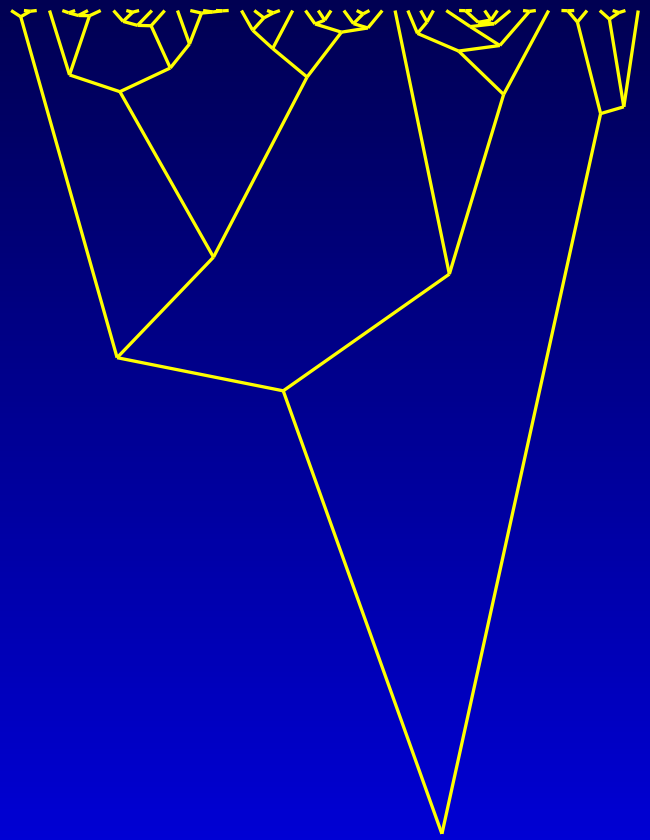
Recombination creates loops



Different markers have slightly different coalescent trees

If we have a sample of 50 copies

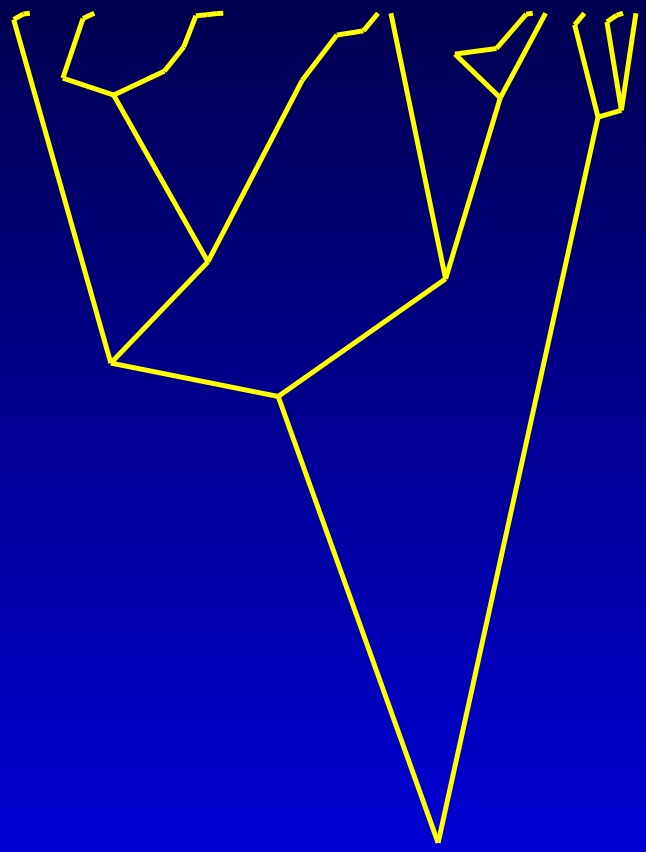
50−gene sample in a coalescent tree



The first 10 account for most of the branch length

10 genes sampled randomly out of a

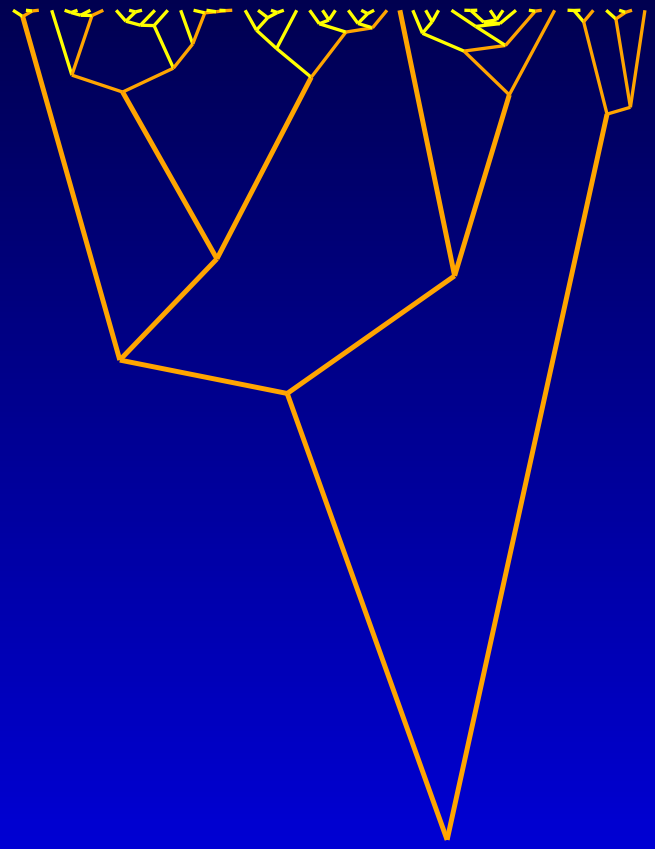
50−gene sample in a coalescent tree



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

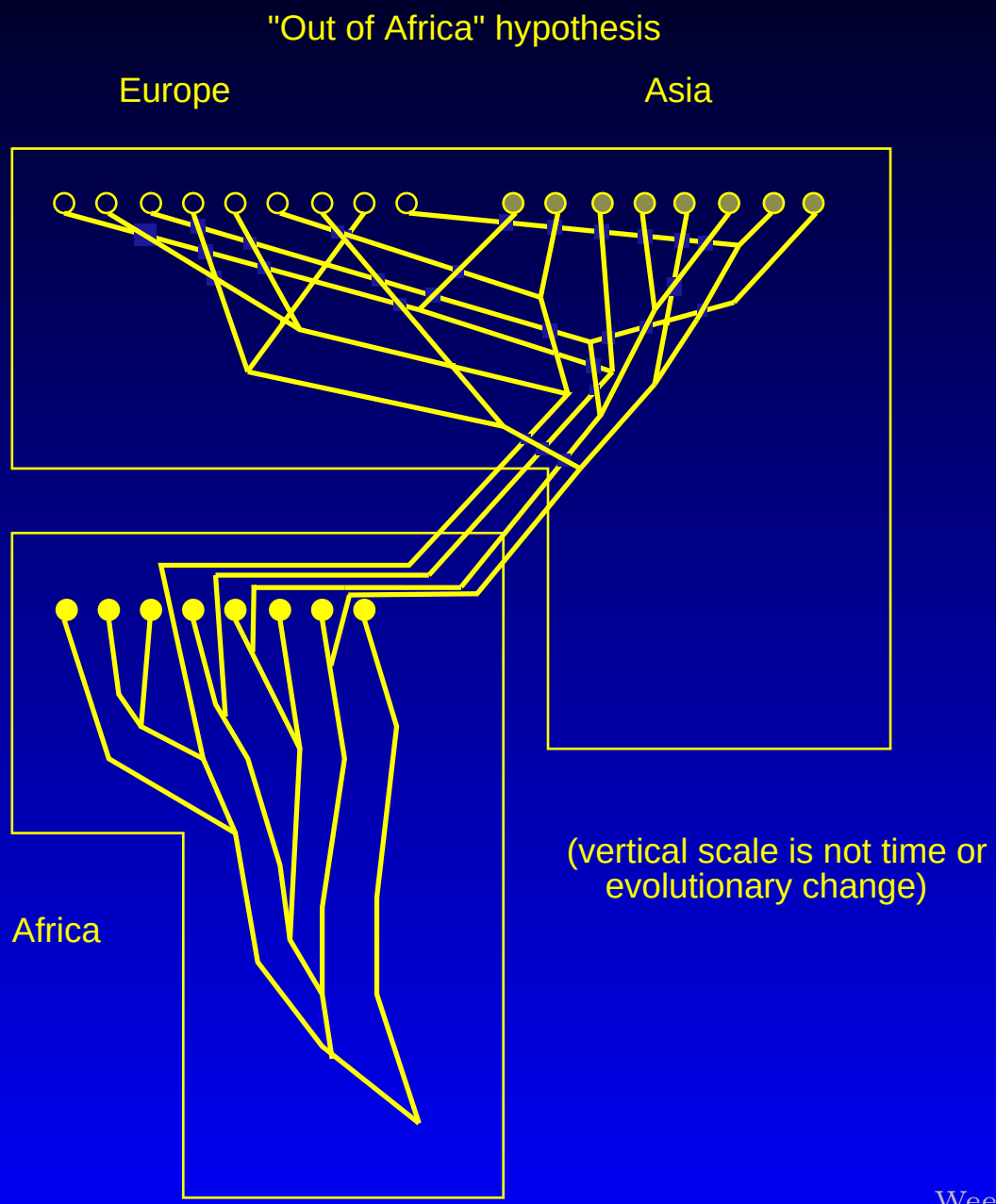
10 genes sampled randomly out of a

50−gene sample in a coalescent tree



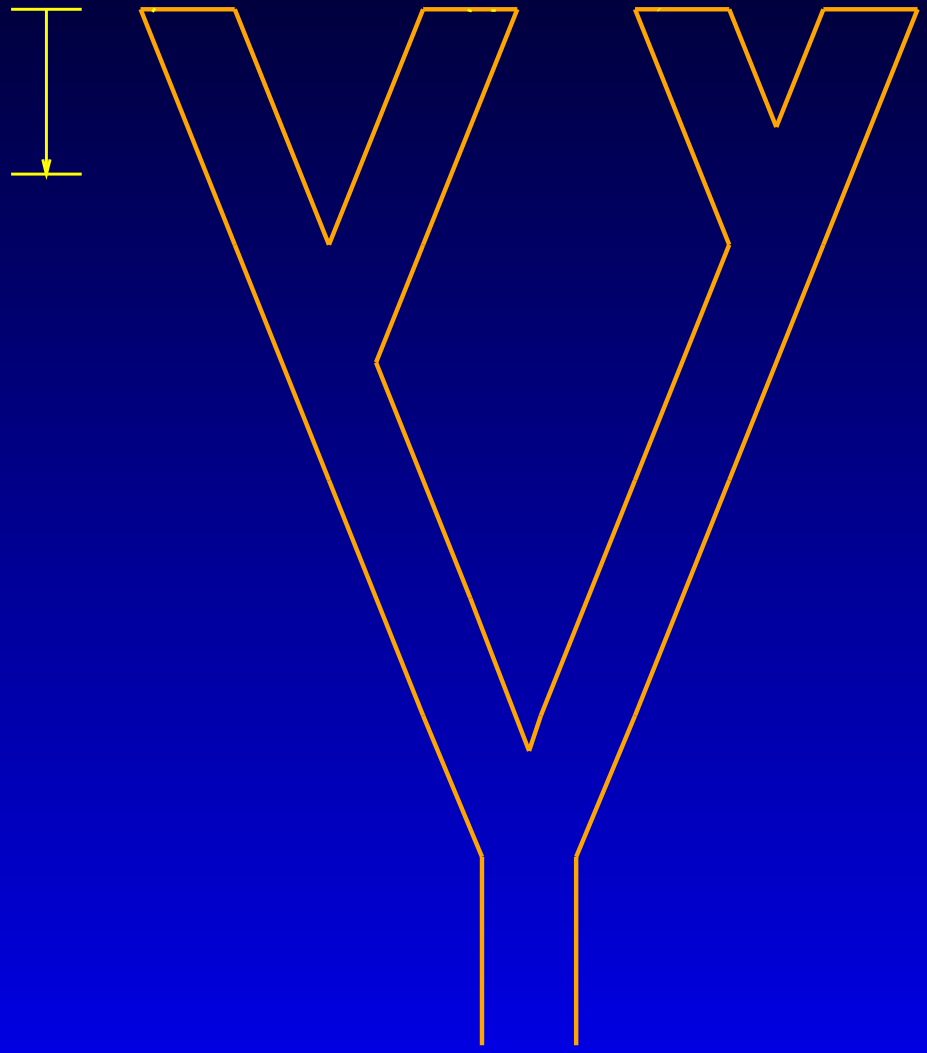
(orange lines are the 10−gene tree)

We want to be able to analyze human evolution



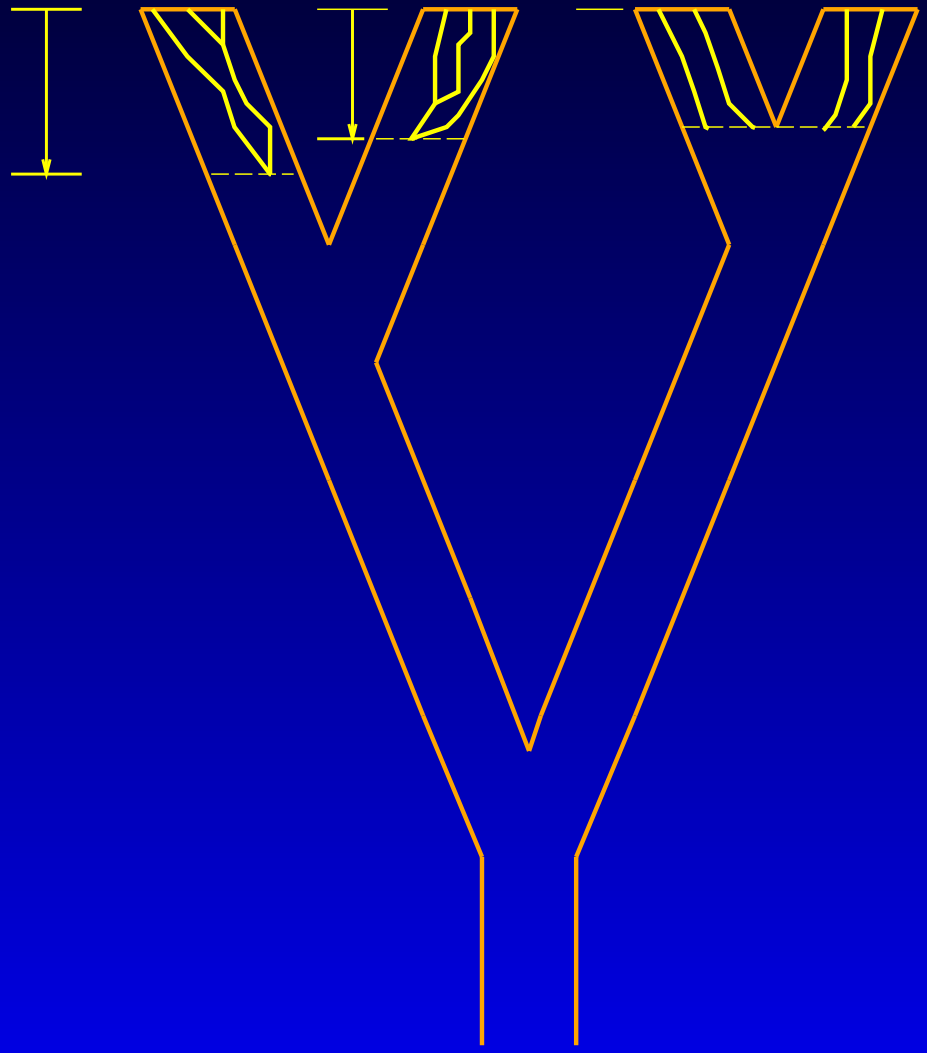
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



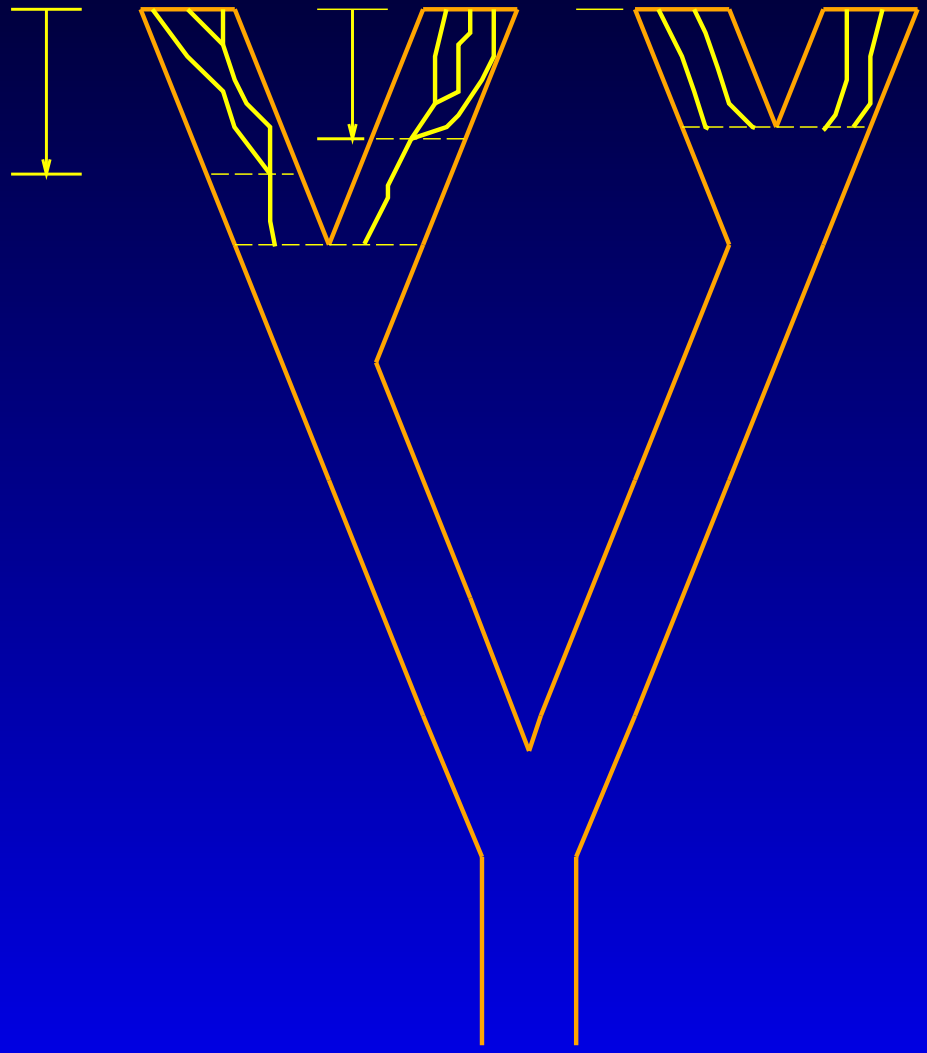
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



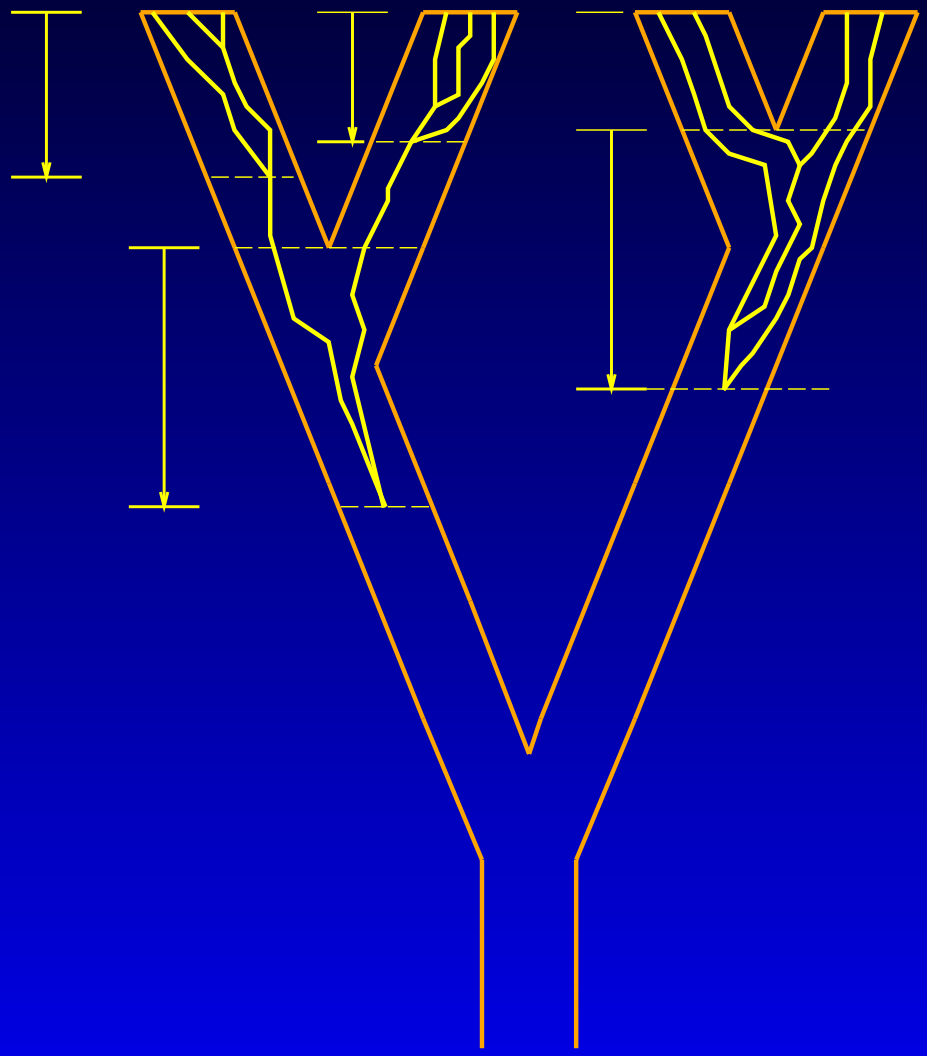
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



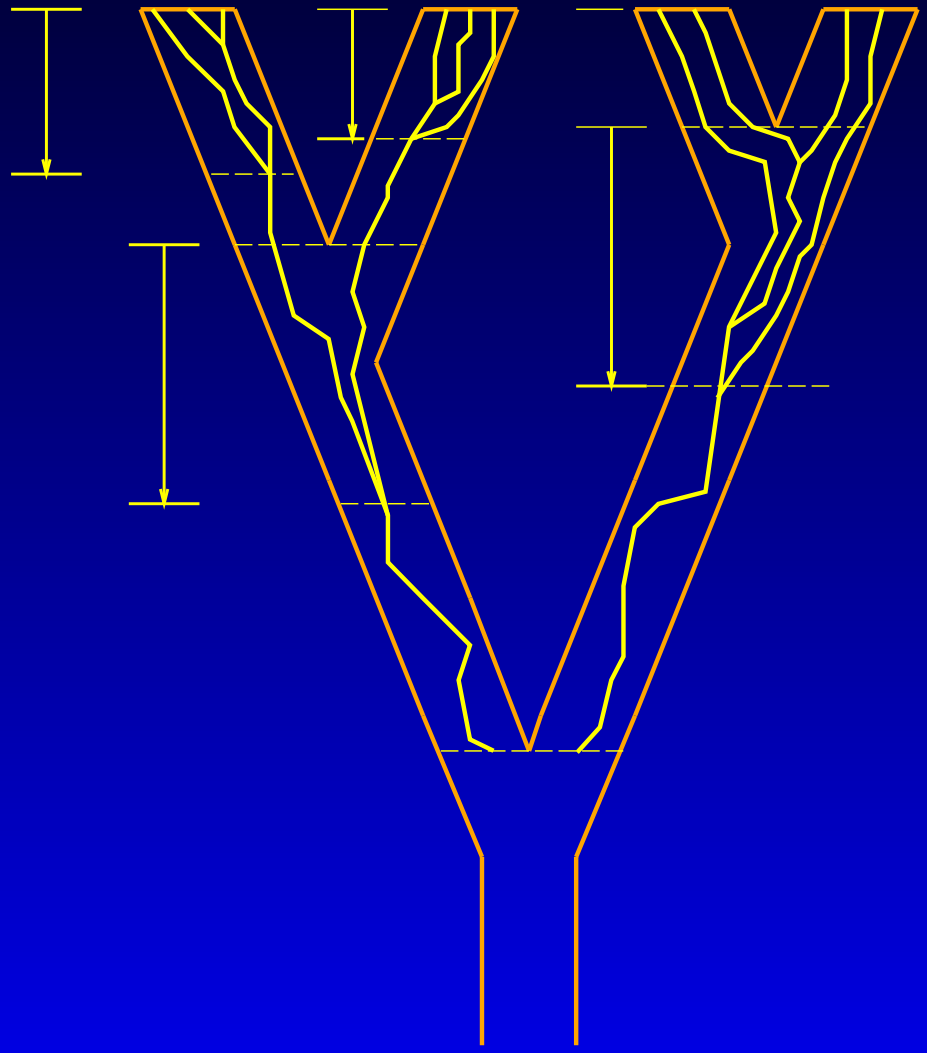
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree

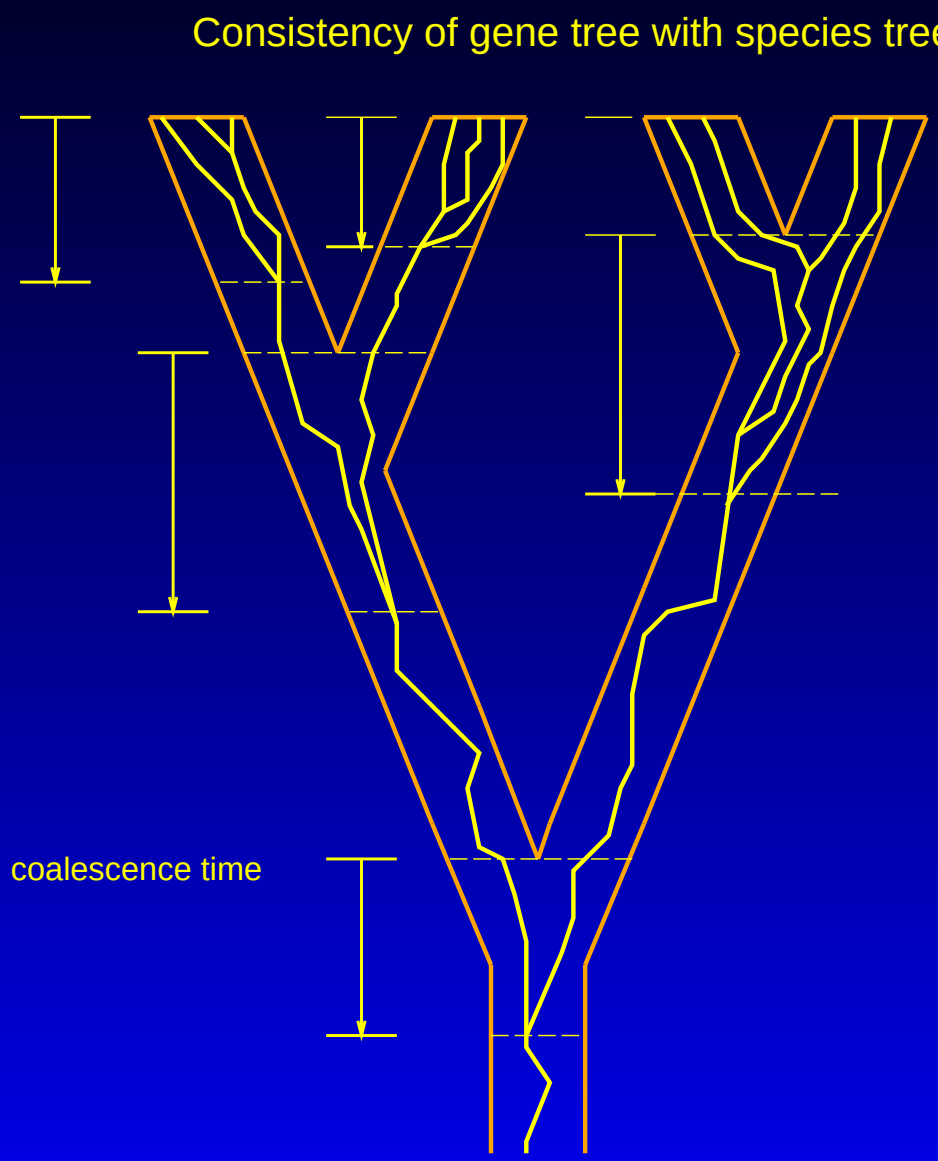


coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree

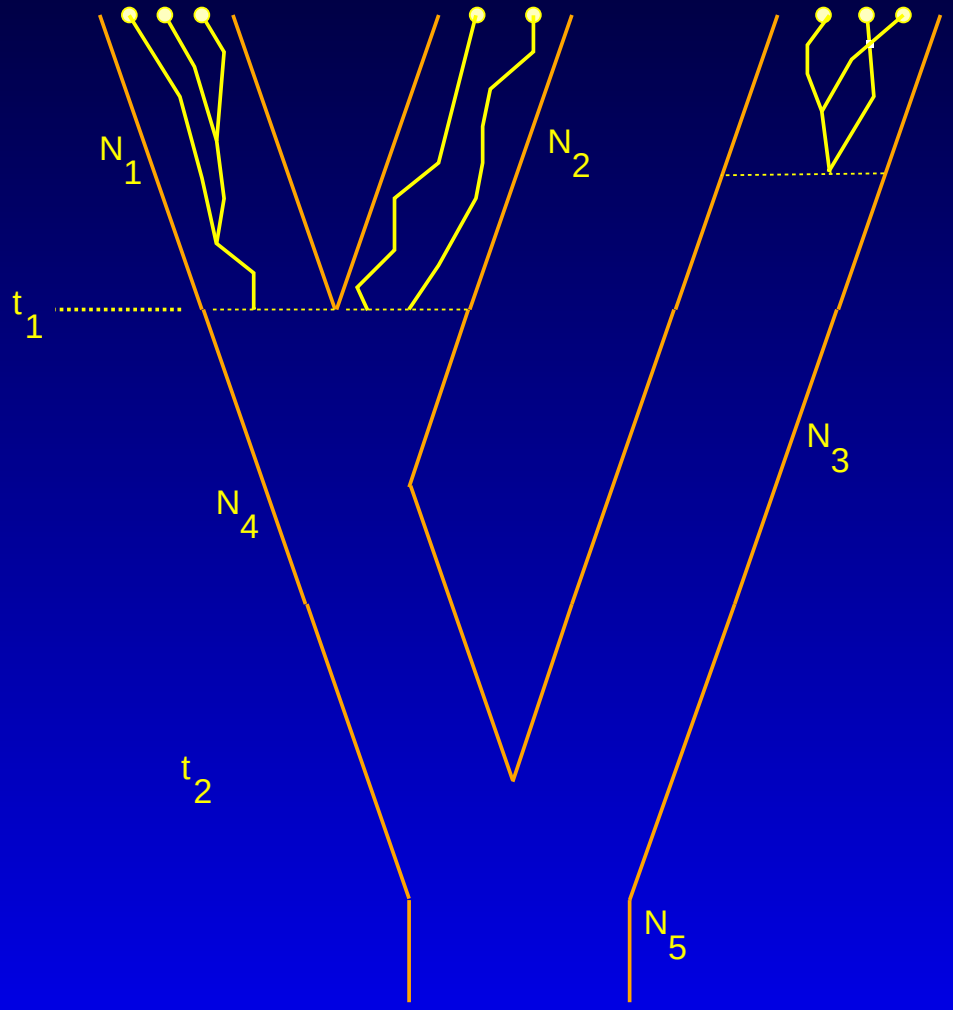


coalescent and “gene trees” versus species trees



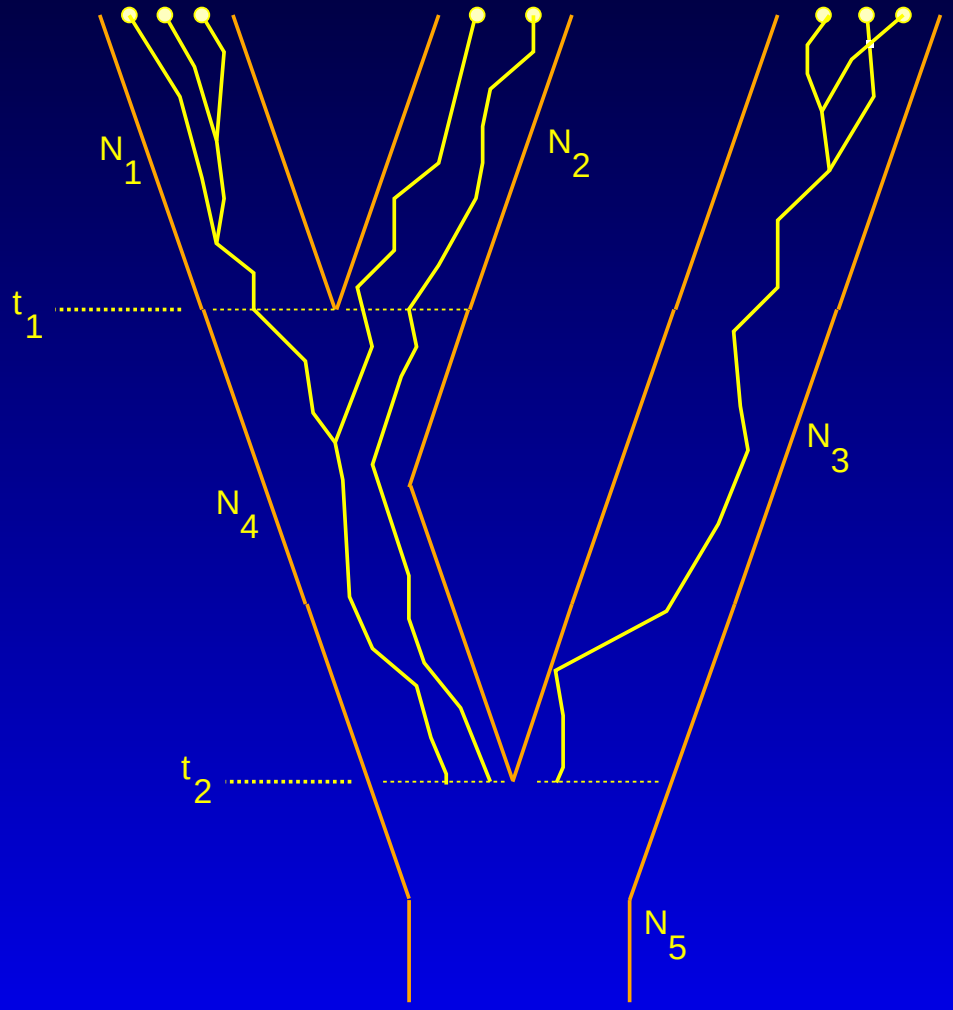
If the branch is more than N_e generations long ...

Gene tree and Species tree



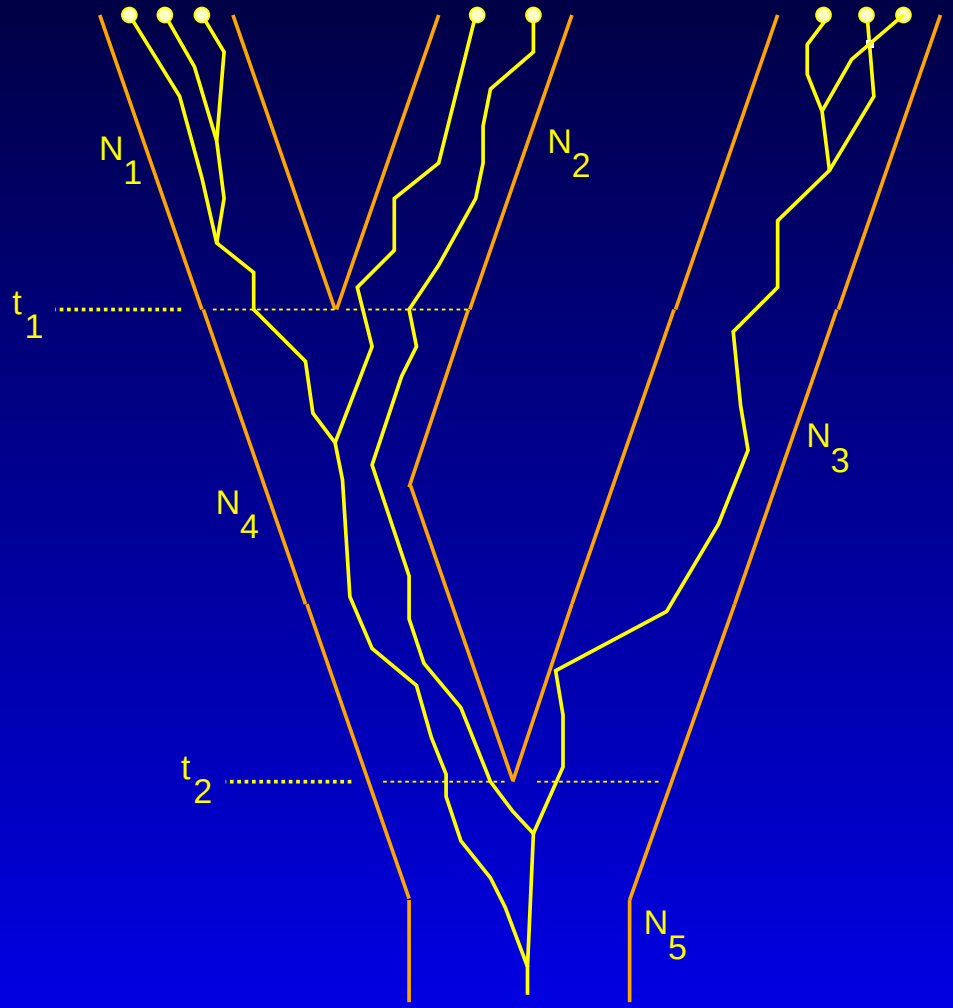
If the branch is more than N_e generations long ...

Gene tree and Species tree



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

Gene tree and Species tree

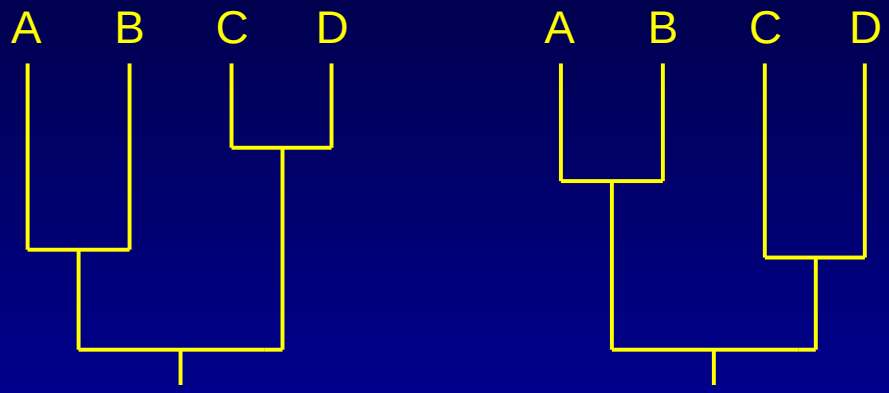


Labelled histories

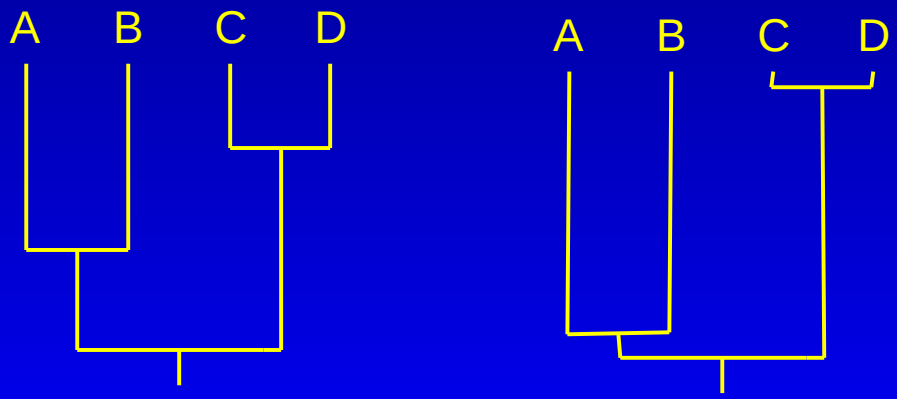
Labelled Histories (Edwards, 1970; Harding, 1971)

Trees that differ in the time−ordering of their nodes

These two are different:



These two are the same:



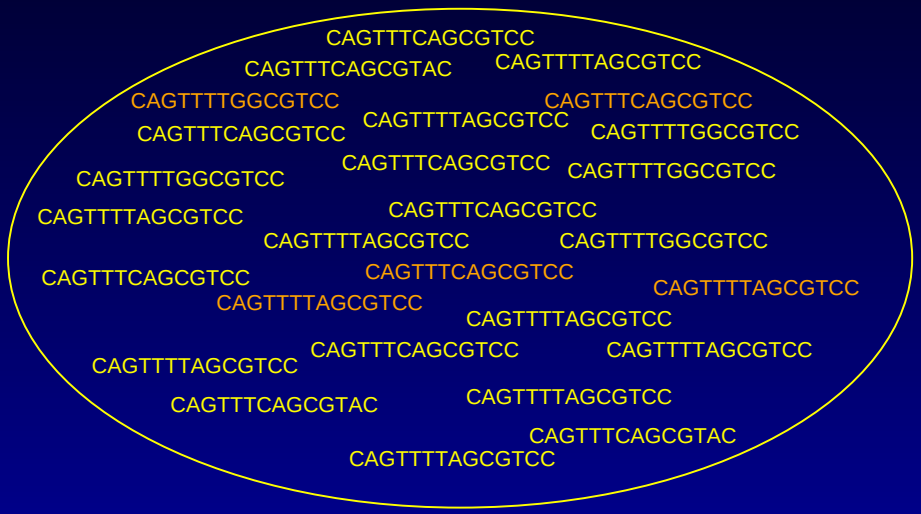
Inconsistency of estimation from concatenated gene sequences

Degnan and Rosenberg (2006) show that the most likely topology for a gene tree is not necessarily the tree that agrees with the phylogenetic tree.

For some phylogenetic shapes (e.g. imbalanced trees with short internal nodes) there exists (at least) one other tree shape that has a higher probability of agreeing with a gene tree.

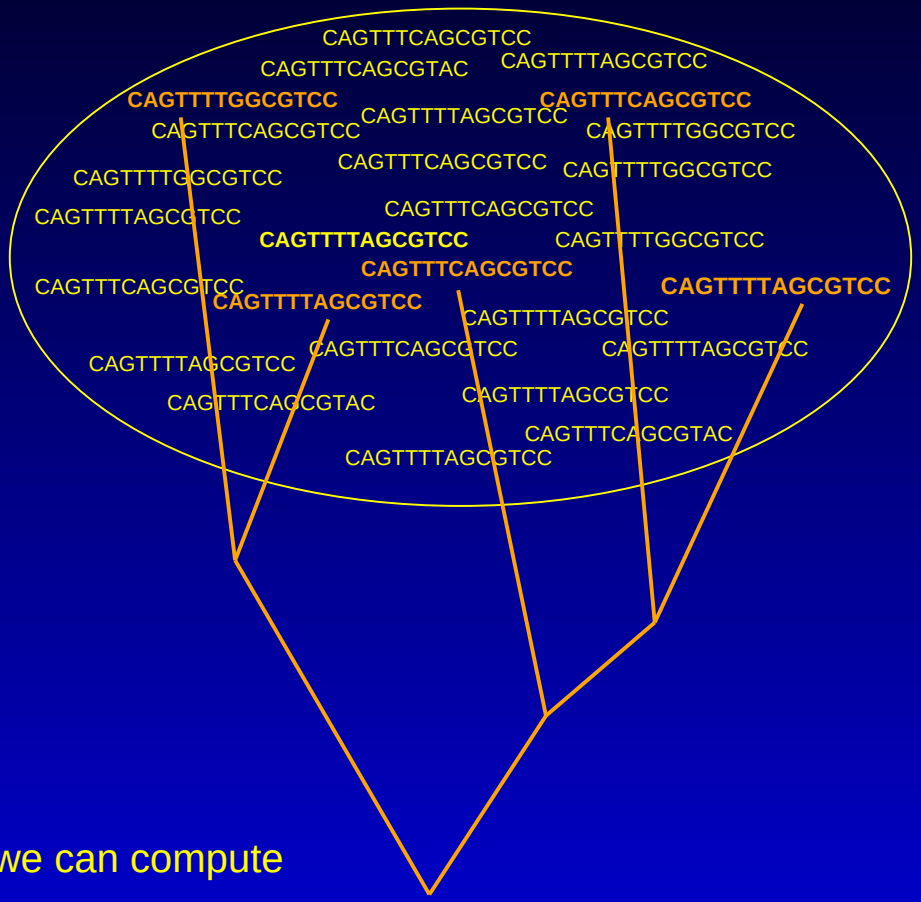
Argues for explicitly considering the coalescent process in phylogenetic inference.

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

$\mathsf{N_e}$ effective population size

μ mutation rate per site

and assuming $\mathsf{G'}$ stands for a coalescent genealogy and D for the sequences,

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

Rescaling the branch lengths

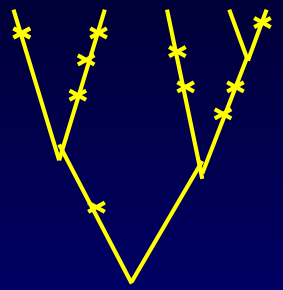
Rescaling branch lengths of G' so that branches are given in expected mutations per site, $G = \mu G'$, we get (if we let $\Theta = 4N_e\mu$)

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

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

The variability comes from two sources

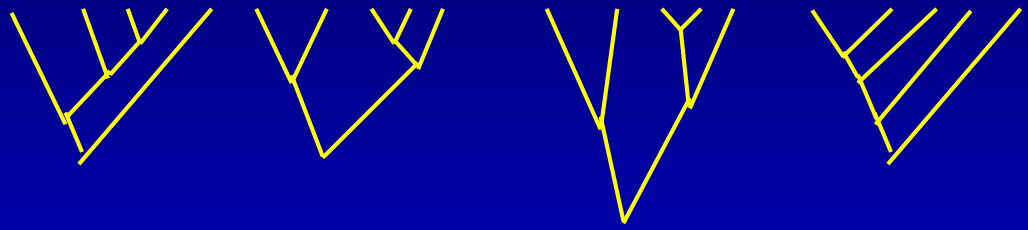
(1) Randomness of mutation



affected by the mutation rate u

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

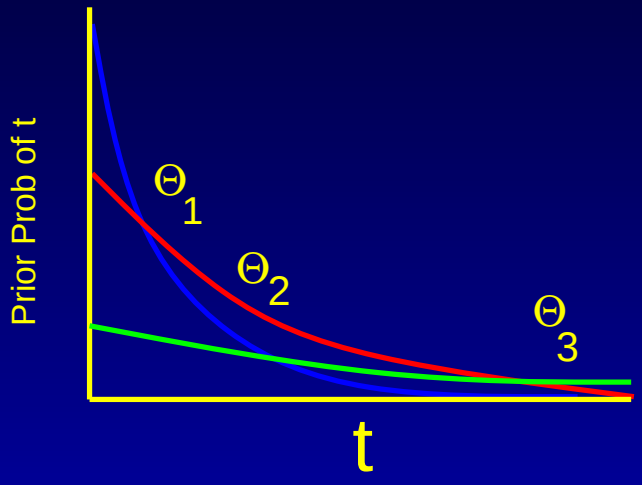
(i) more gene copies, or

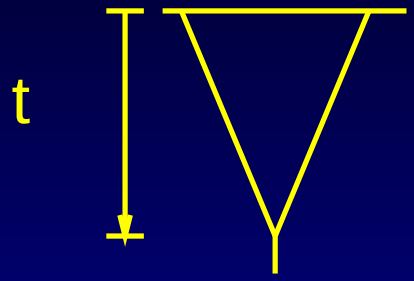
(ii) more loci

We can compute the likelihood by averaging over coalescents

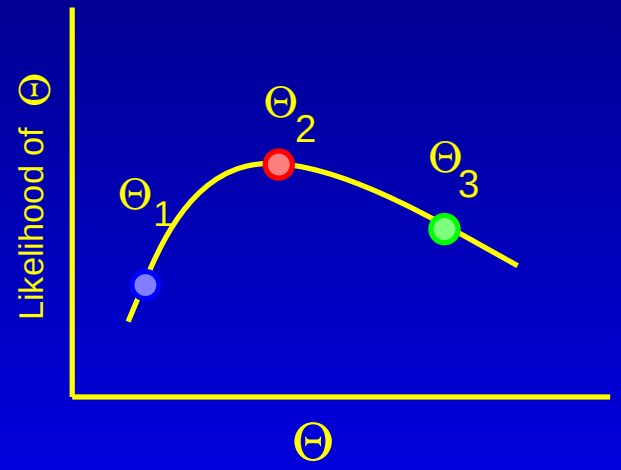
The likelihood calculation in a sample of two gene copies

The product of the prior on t,

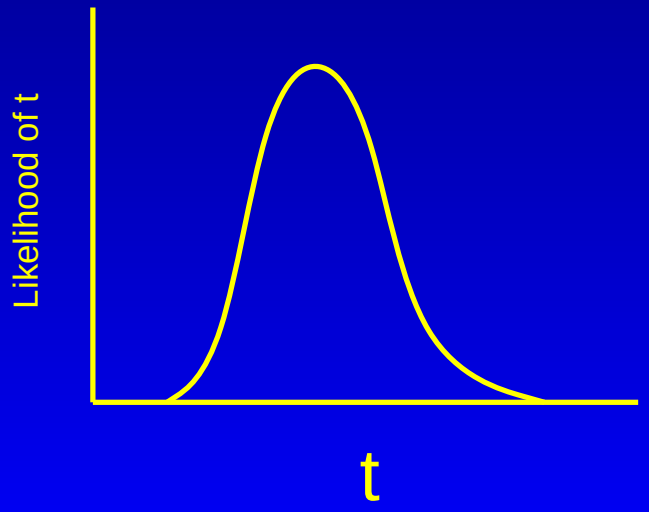




when integrated over all possible t's, gives the likelihood for the underlying parameter Θ

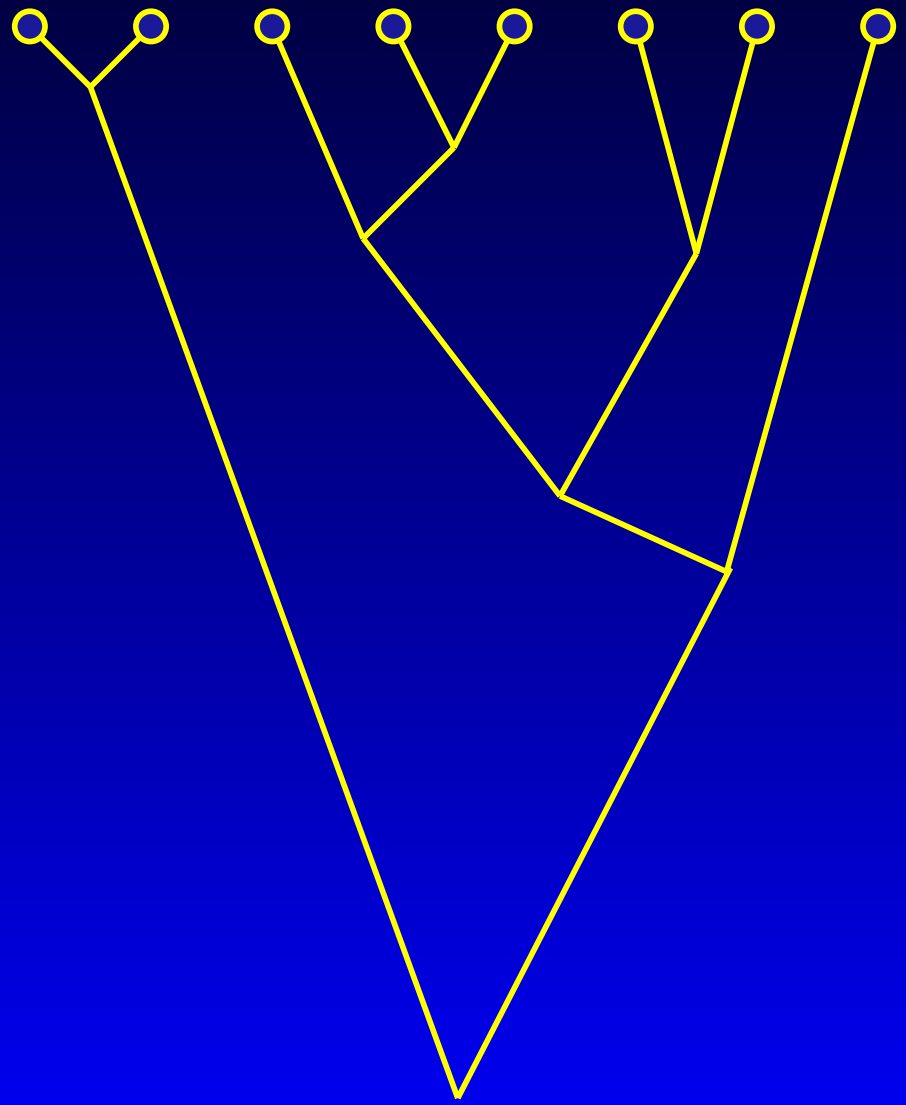


times the likelihood of that t from the data,



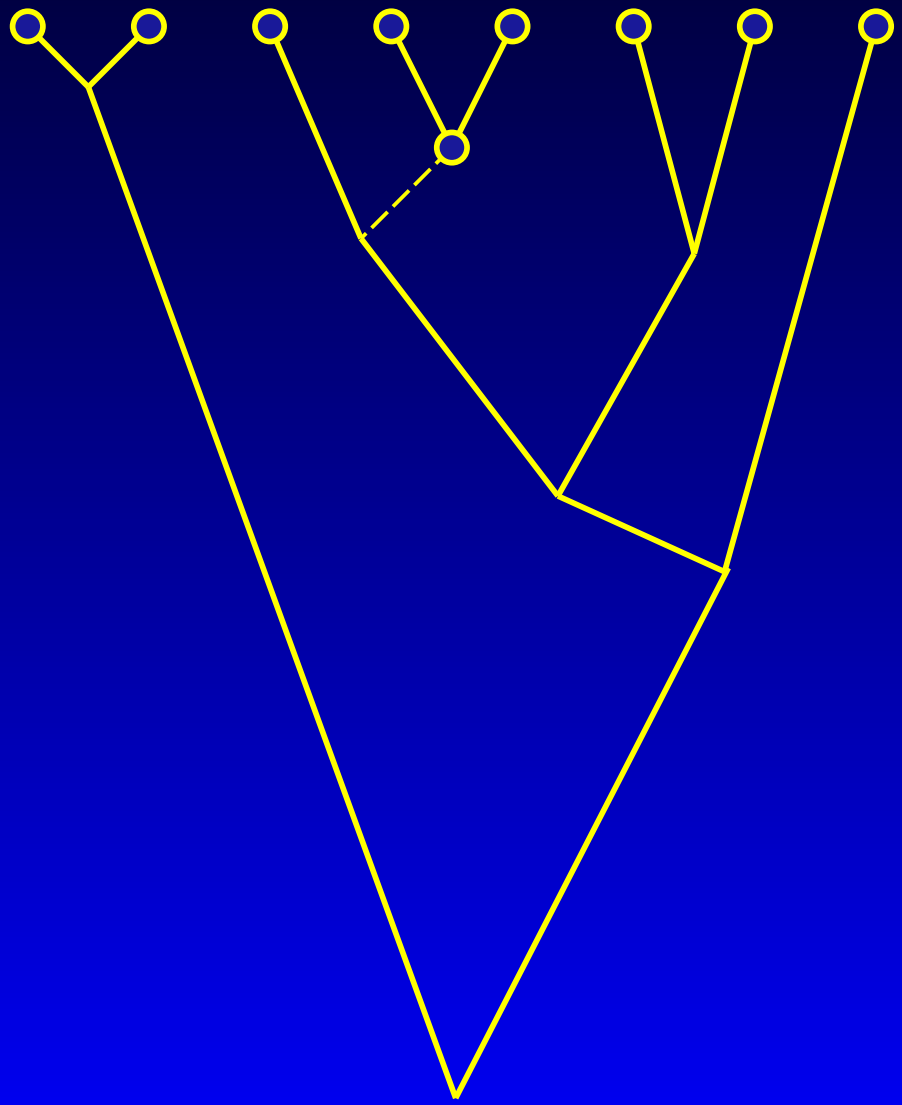
Rearrangement to sample points in tree space

A conditional coalescent rearrangement strategy



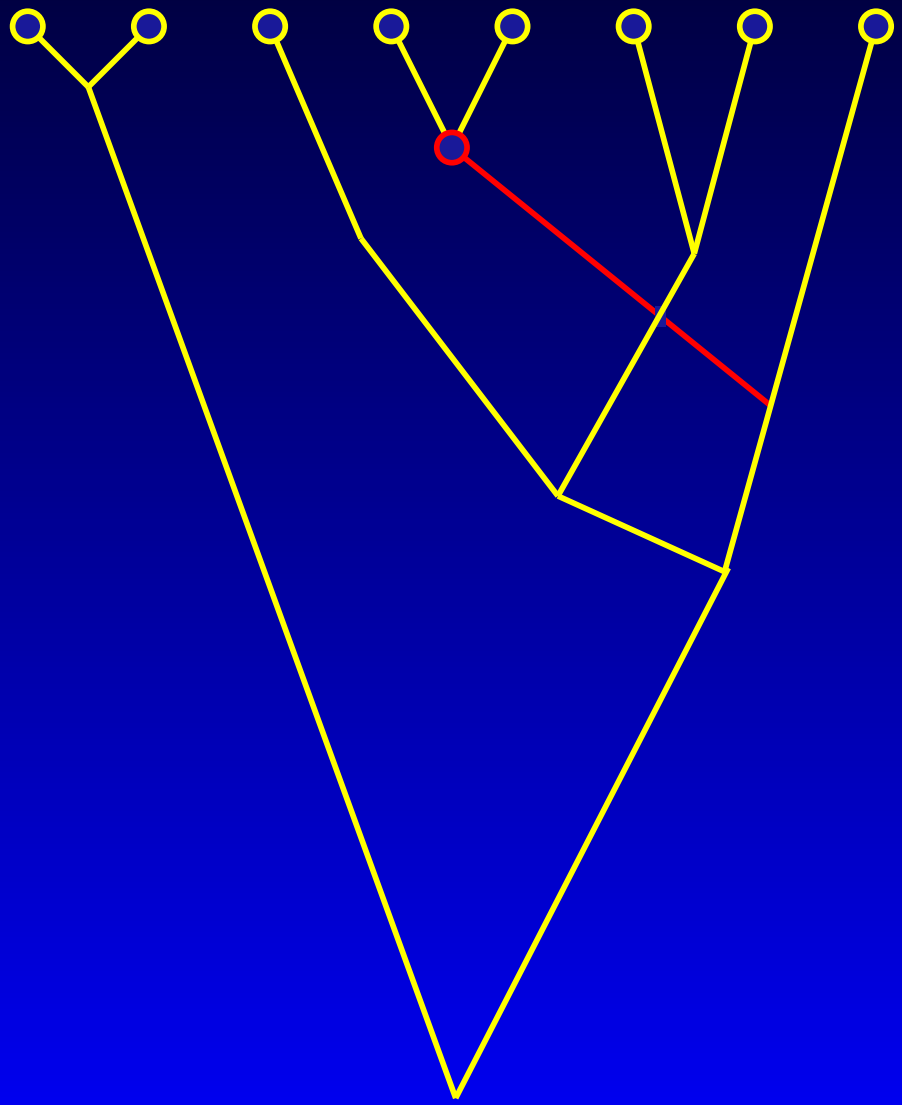
Dissolving a branch and regrowing it backwards

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



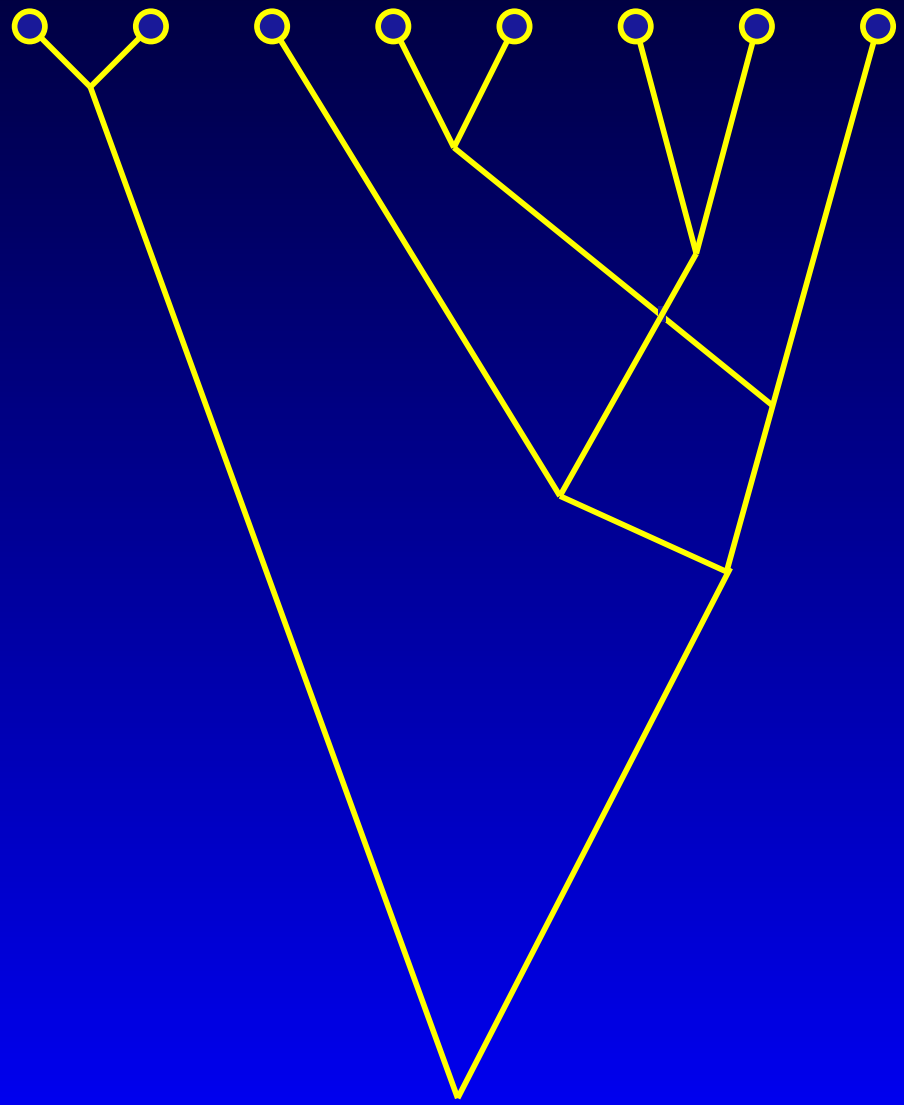
We allow it coalesce with the other branches

Then allow this node to re−coalesce with the tree



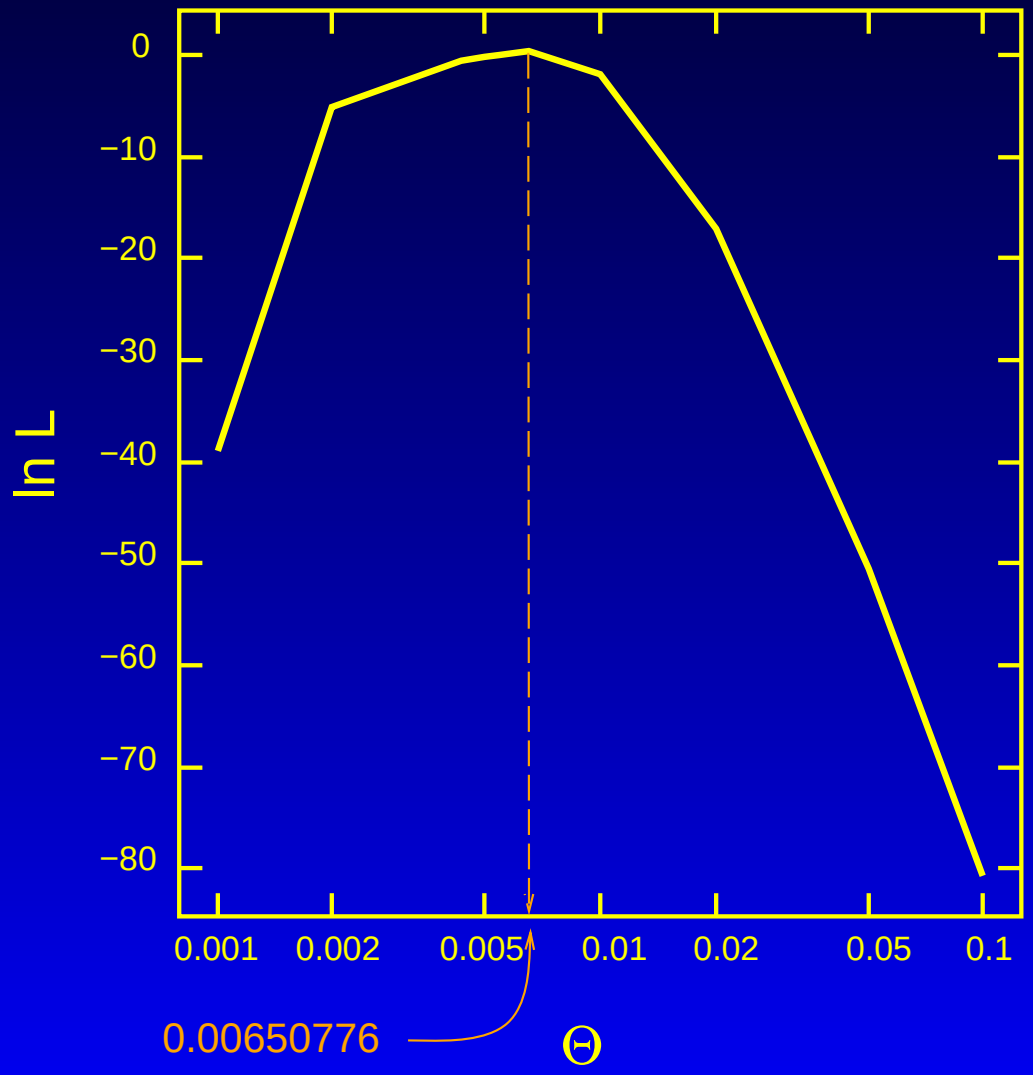
and this gives anothern coalescent

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, 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.
- **migrate** by Peter Beerli. Likelihood inference with multiple populations and migration rates.
- **IM** and **IMa** by Rasmus Nielsen and Jody Hey. Two populations allowing both historical splitting and migration after that.

“Skyline” and “Skyride” plots in BEAST

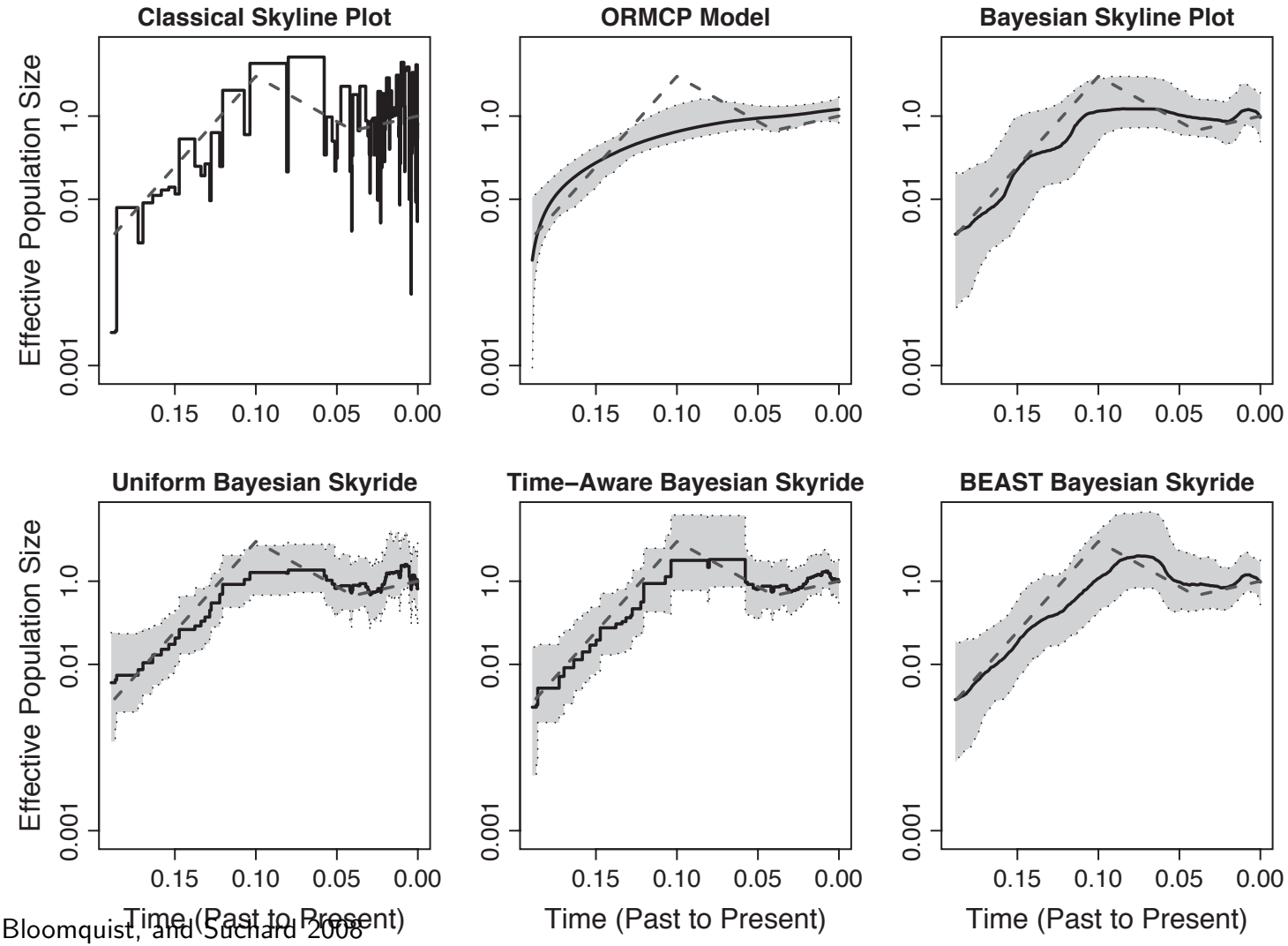


Figure from Minin, Bloomquist, and Suchard 2008

BEST Liu and Pearl (2007); Edwards et al. (2007)

- X – sequence data
- G – a genealogy (gene tree – with branch lengths)
- S – a species tree
- $\boldsymbol{\theta}$ – demographic parameters
- $\boldsymbol{\Lambda}$ – parameters of molecular sequence evolution

$$
\begin{aligned}
\Pr(S, \boldsymbol{\theta}|X) &= \frac{\Pr(S, \boldsymbol{\theta}) \Pr(X|S, \boldsymbol{\theta})}{\Pr(X)} \\
&= \Pr(S) \Pr(\boldsymbol{\theta}) \int \Pr(X|G) \Pr(G|S, \boldsymbol{\theta}) dG \\
&\propto \Pr(S) \Pr(\boldsymbol{\theta}) \int \left[\int \Pr(X|G, \boldsymbol{\Lambda}) \Pr(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \right] \Pr(G|S, \boldsymbol{\theta}) dG
\end{aligned}
$$

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
3. Use "importance weights" to correct the sample for the fact that an approximate prior was used

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
   (a) Use a tweaked version of MrBayes to sample N sets of gene trees, $\boldsymbol{G}$, from
   $$\overset{\dagger}{\Pr}(\boldsymbol{G}|X) = \frac{\Pr^{\dagger}(\boldsymbol{G})\Pr(X|\boldsymbol{G})}{\Pr^{\dagger}(X)}$$
   (b) $Pr^{\dagger}(\boldsymbol{G})$ is an approximate prior on gene trees from using a "maximal" species tree.
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
3. Use "importance weights" to correct the sample for the fact that an approximate prior was used

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
   (a) From each set of gene trees (G_j for $1 \leq j \leq N$) generate k species trees using coalescent theory:

$$\Pr(S_i|\boldsymbol{G}_j) = \frac{\Pr(S_i)\Pr(\boldsymbol{G}_j|S_i)}{\Pr(\boldsymbol{G}_j)}$$

3. Use "importance weights" to correct the sample for the fact that an approximate prior was used

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
3. Use "importance weights" to correct the sample for the fact that an approximate prior was used
   (a) Estimate $\widehat{\Pr}(\boldsymbol{G}_j)$ by using the harmonic mean estimator from the MCMC in step 2.
   (b) Compute a normalization factor

$$\beta = \sum_{j=1}^{N} \frac{\widehat{\Pr}(\boldsymbol{G}_j)}{\Pr(\boldsymbol{G}_j)}$$

   (c) Reweight all sampled species trees by

$$\frac{\widehat{\Pr}(\boldsymbol{G}_j)}{\Pr(\boldsymbol{G}_j)}\beta$$

BEST – conclusions

1. very expensive computationally (long MrBayes runs are needed)
2. should correctly deal with the variability in gene tree caused by the coalescent process.

*BEST

Similar model to BEST, but *much* more efficient implementation.

Both will be very sensitive to migration, but they represent the state-of-the-art for estimating species trees from gene trees.

Gene tree in a species tree w/ variable population size

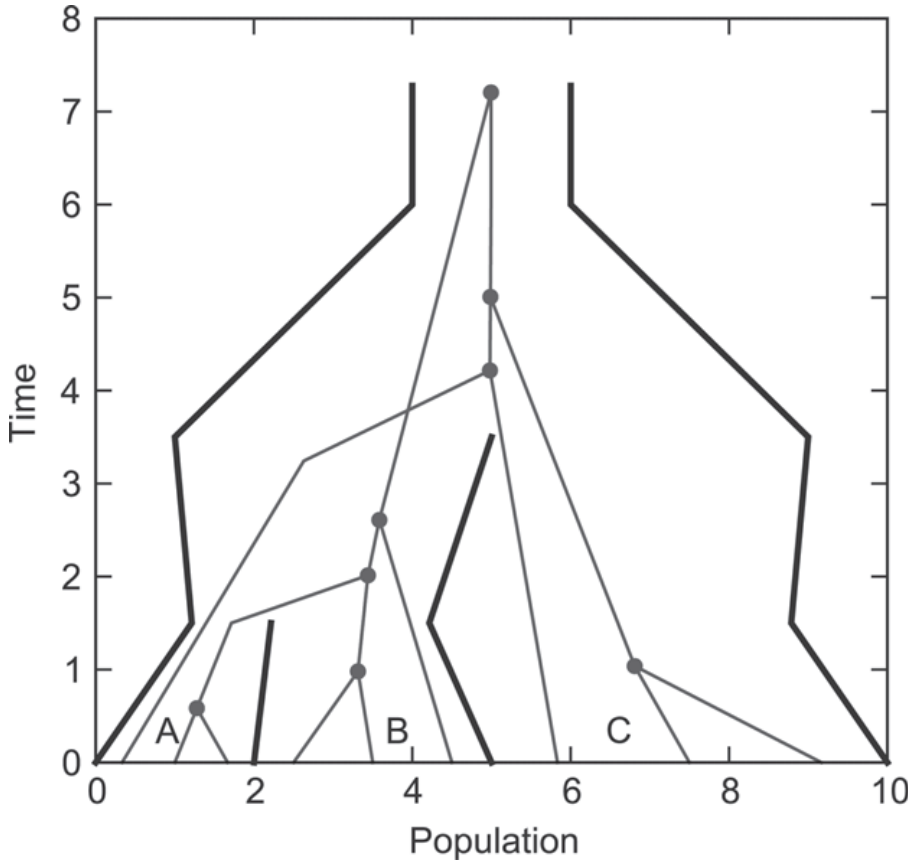


Figure from Heled and Drummond 2010

Multiple gene tree in a species tree w/ variable population size

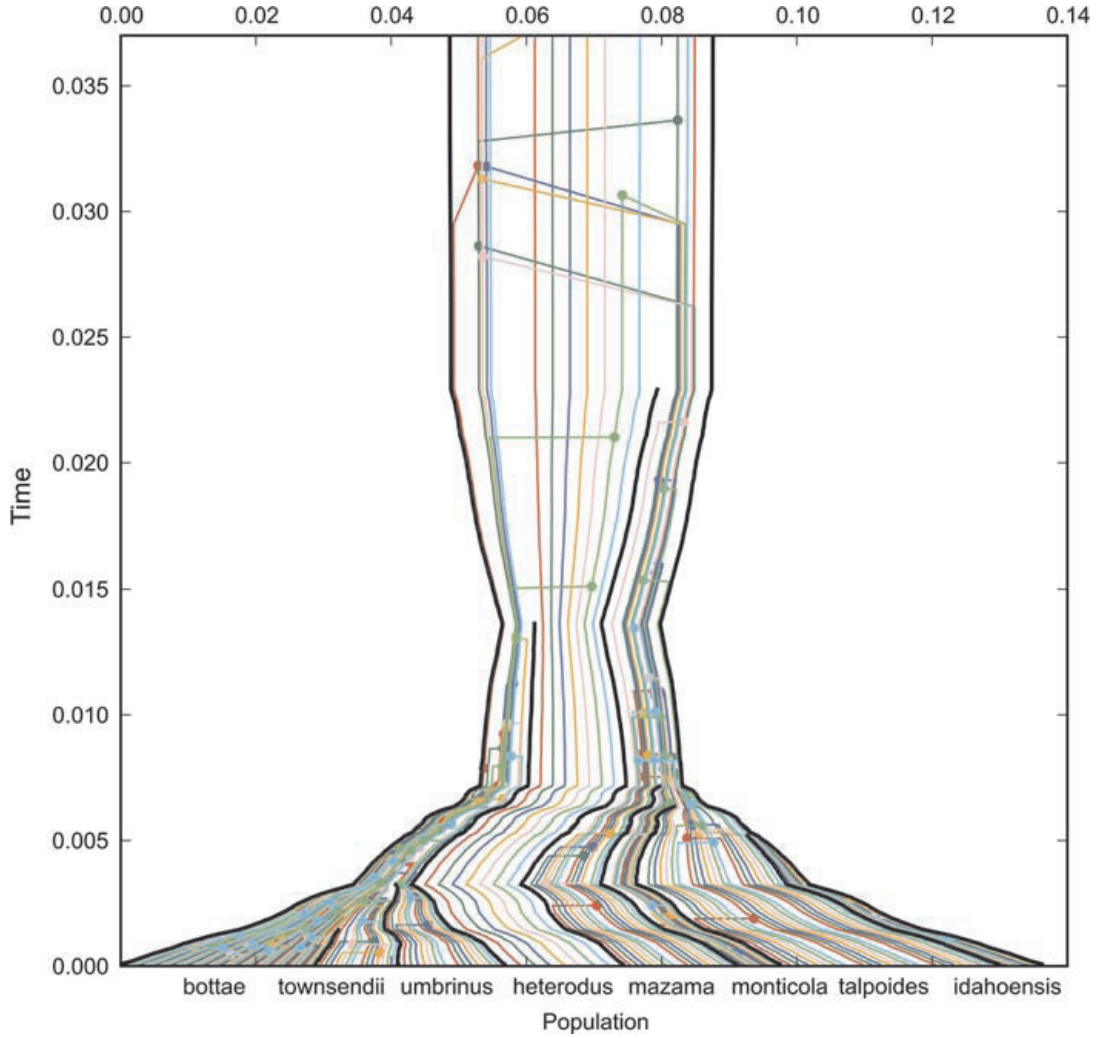


Figure from Heled and Drummond 2010

References

Degnan, J. and Rosenberg, N. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet*, 2(5).

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941.

Liu, L. and Pearl, D. K. (2007). Species trees from gene trees: reconstruction Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3):504–514.