Comparative method, coalescents, and the future

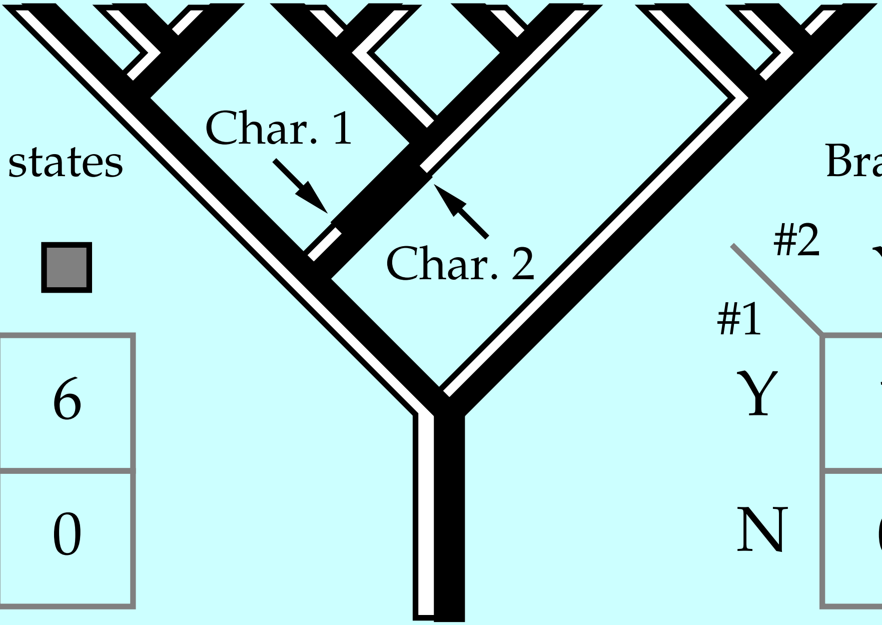
Joe Felsenstein

Depts. of Genome Sciences and of Biology, University of Washington

Correlation of states in a discrete-state model

Character 1: □ □ □ ■ ■ ■ ■ □ □ □

Character 2: ■ ■ ■ □ □ □ □ ■ ■ ■



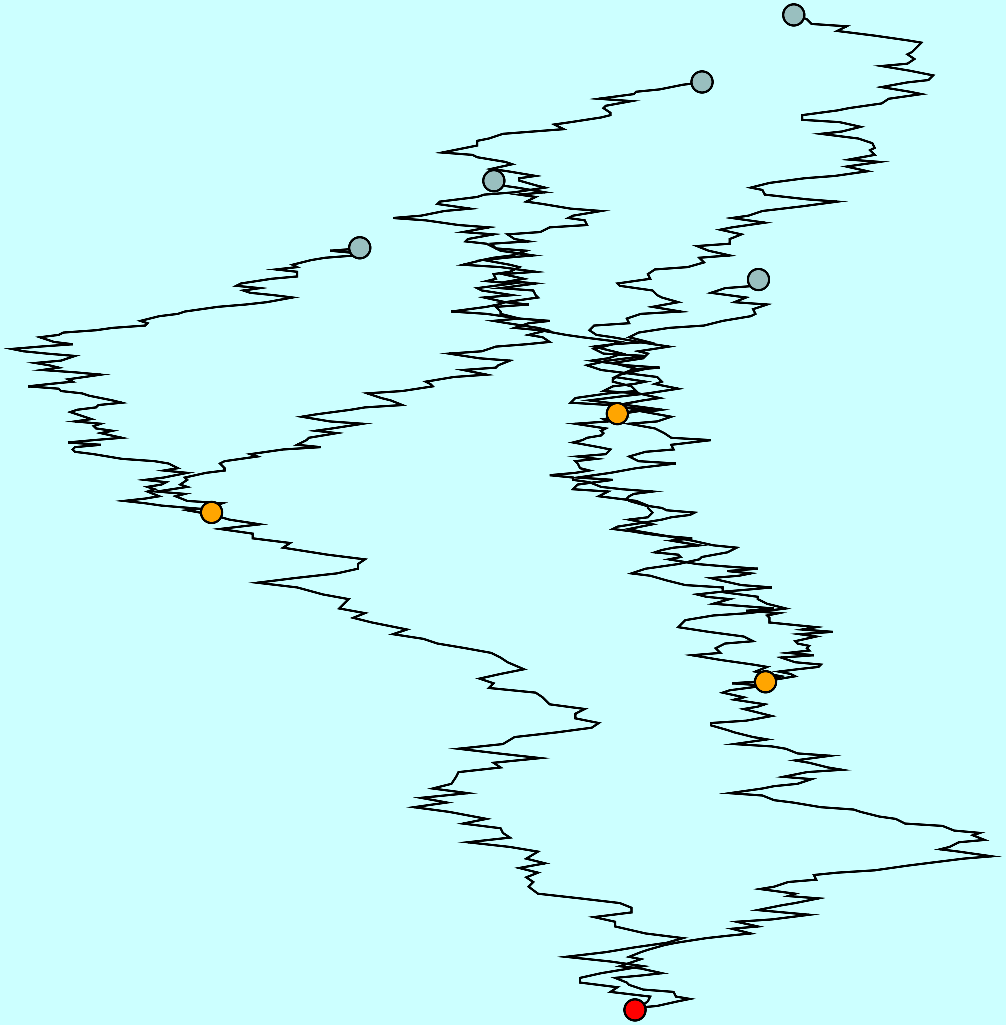
Species states

#2 #1	□	■
□	0	6
■	4	0

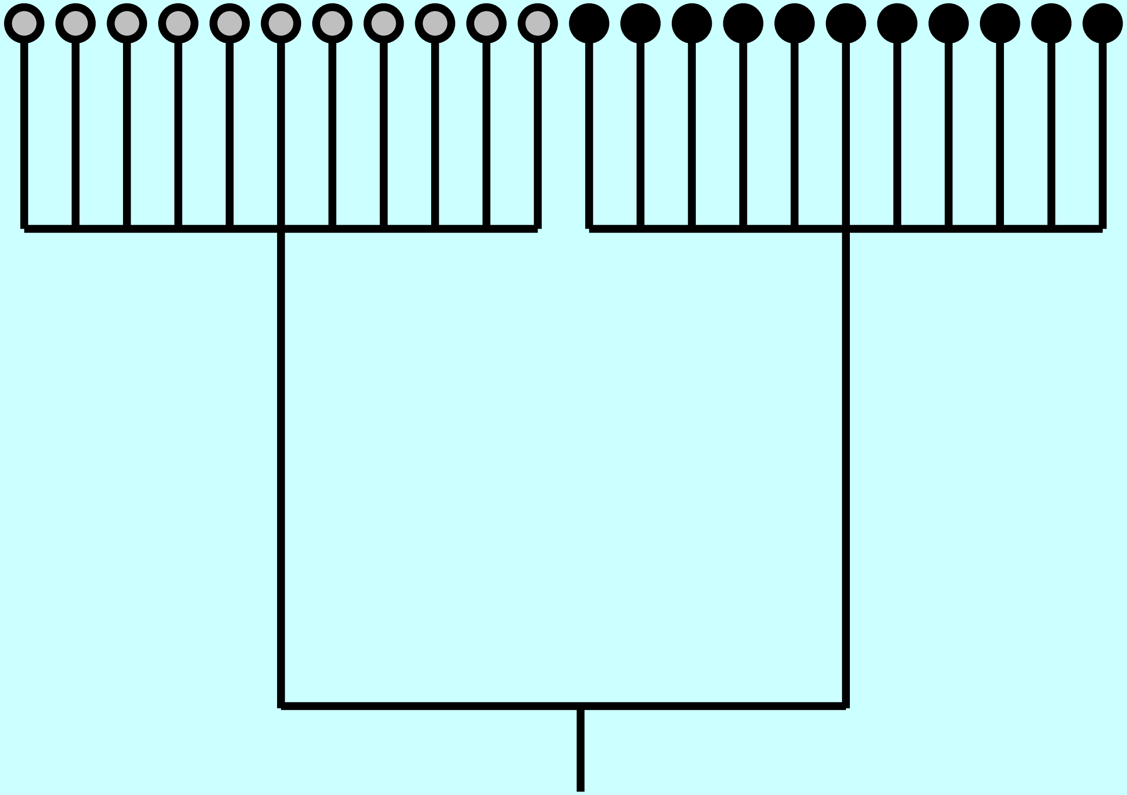
Branch changes

#2 #1	Y	N
Y	1	0
N	0	18

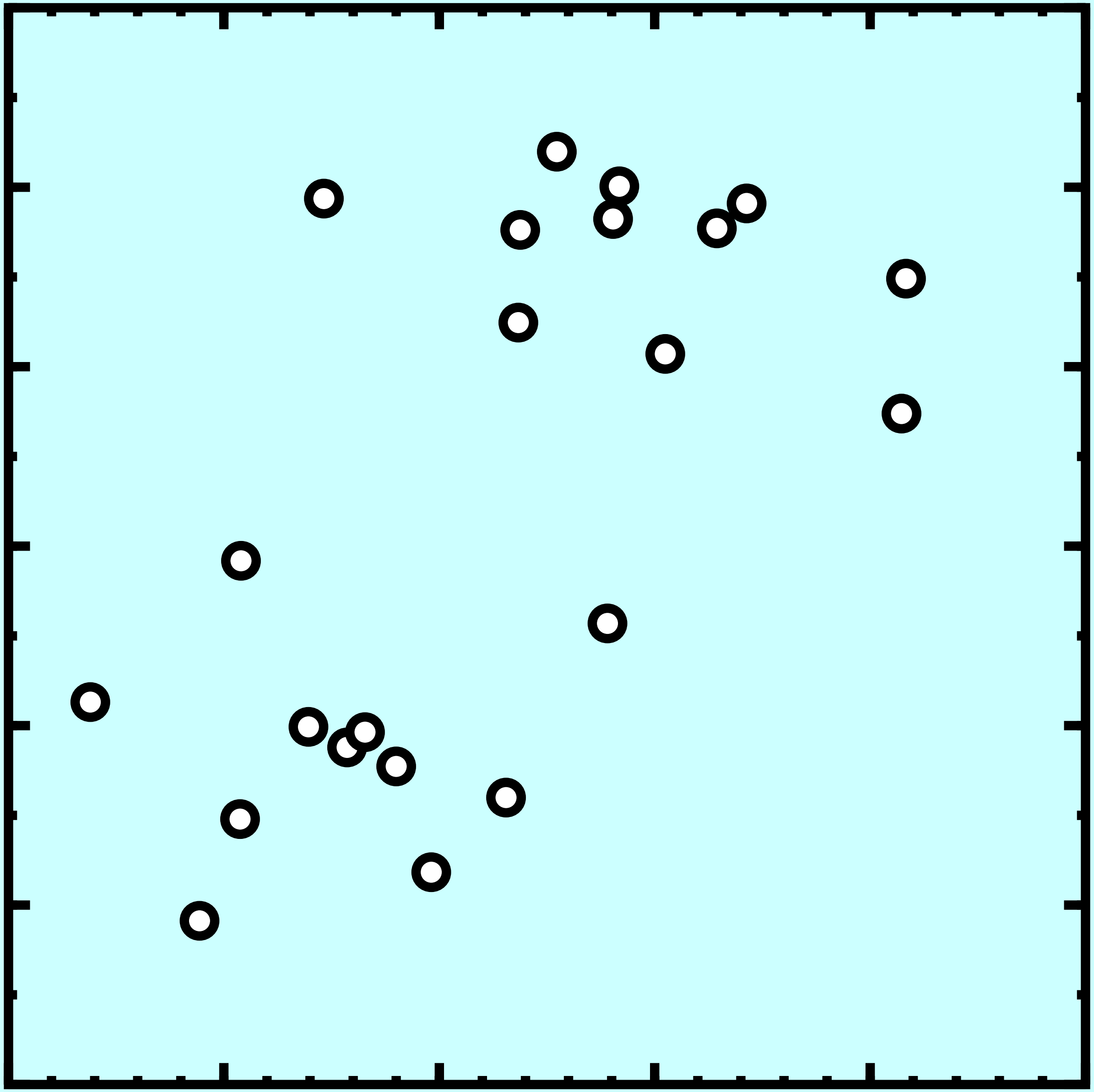
A simple model: Brownian motion



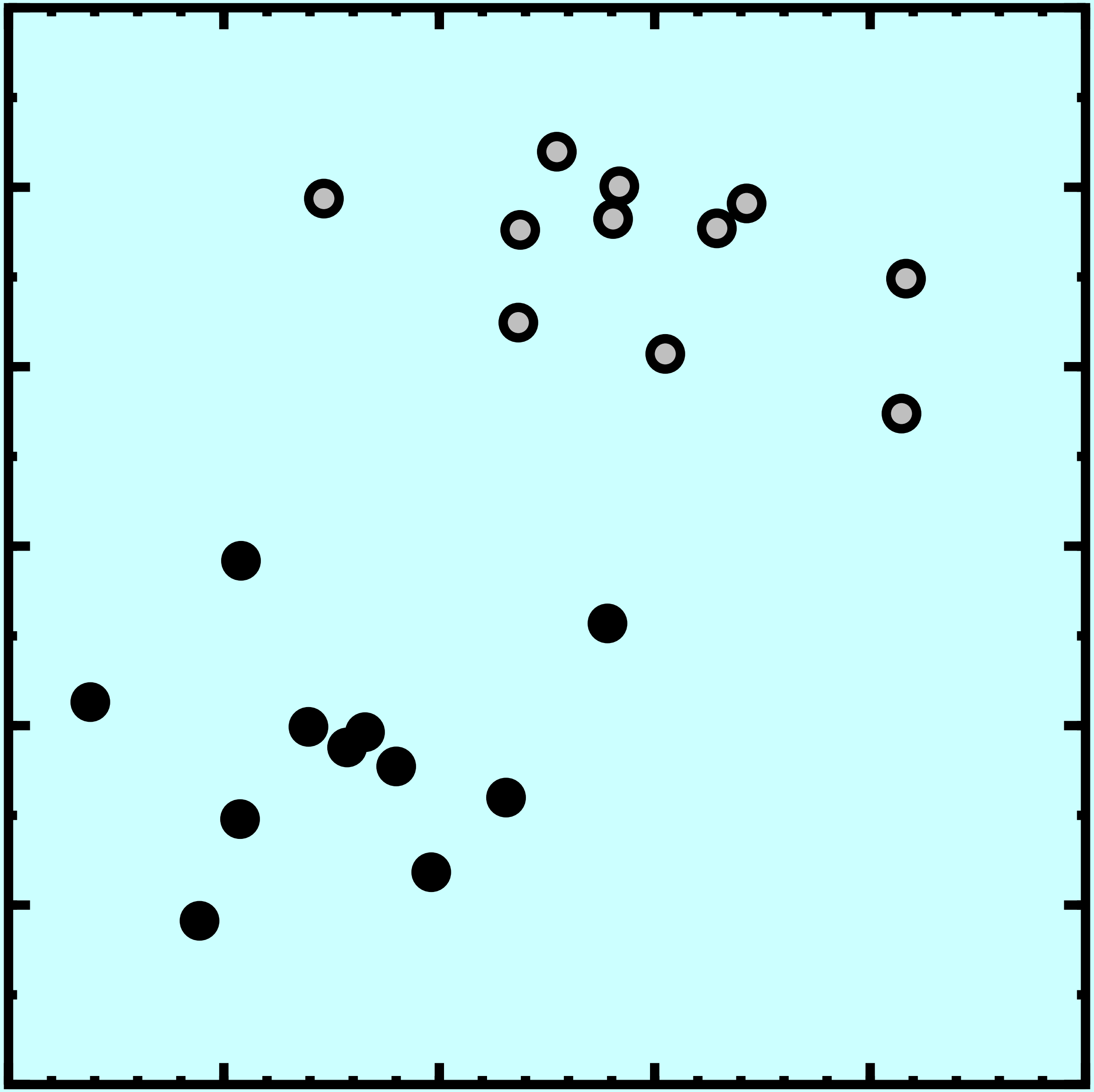
A simple case to show effects of phylogeny



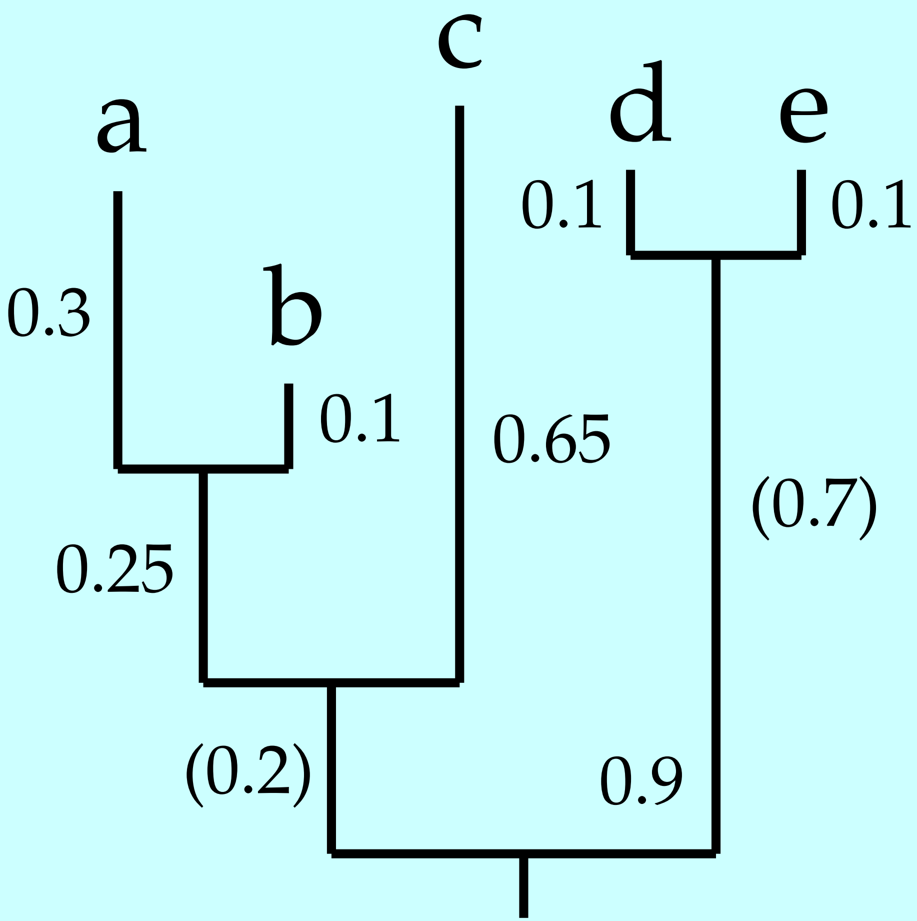
Two uncorrelated characters evolving on that tree



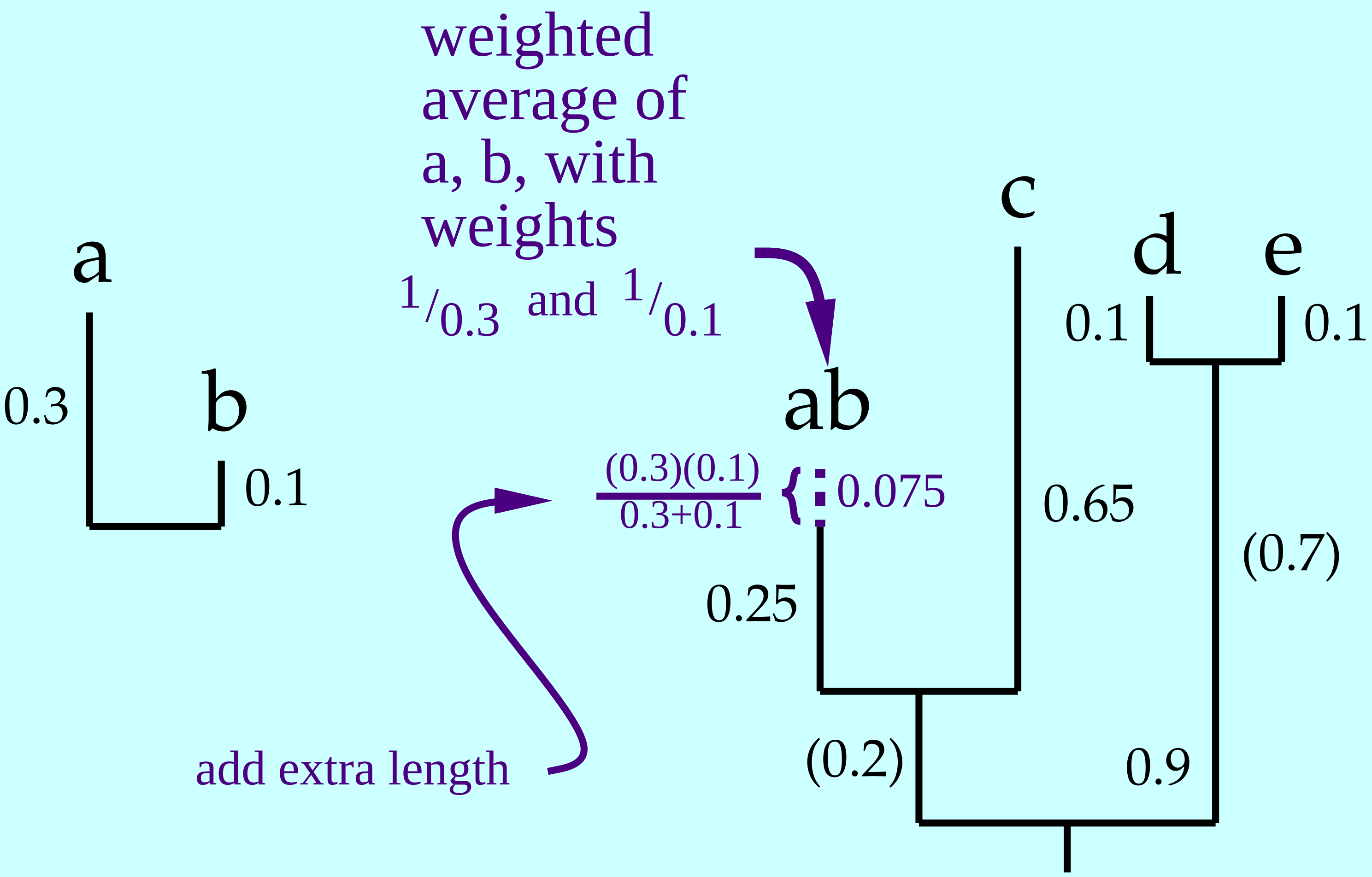
Identifying the two clades



A tree on which we are to observe two characters



This turns out to be statistically equivalent to ...



weighted average of a, b, with weights $^{1}/_{0.3}$ and $^{1}/_{0.1}$

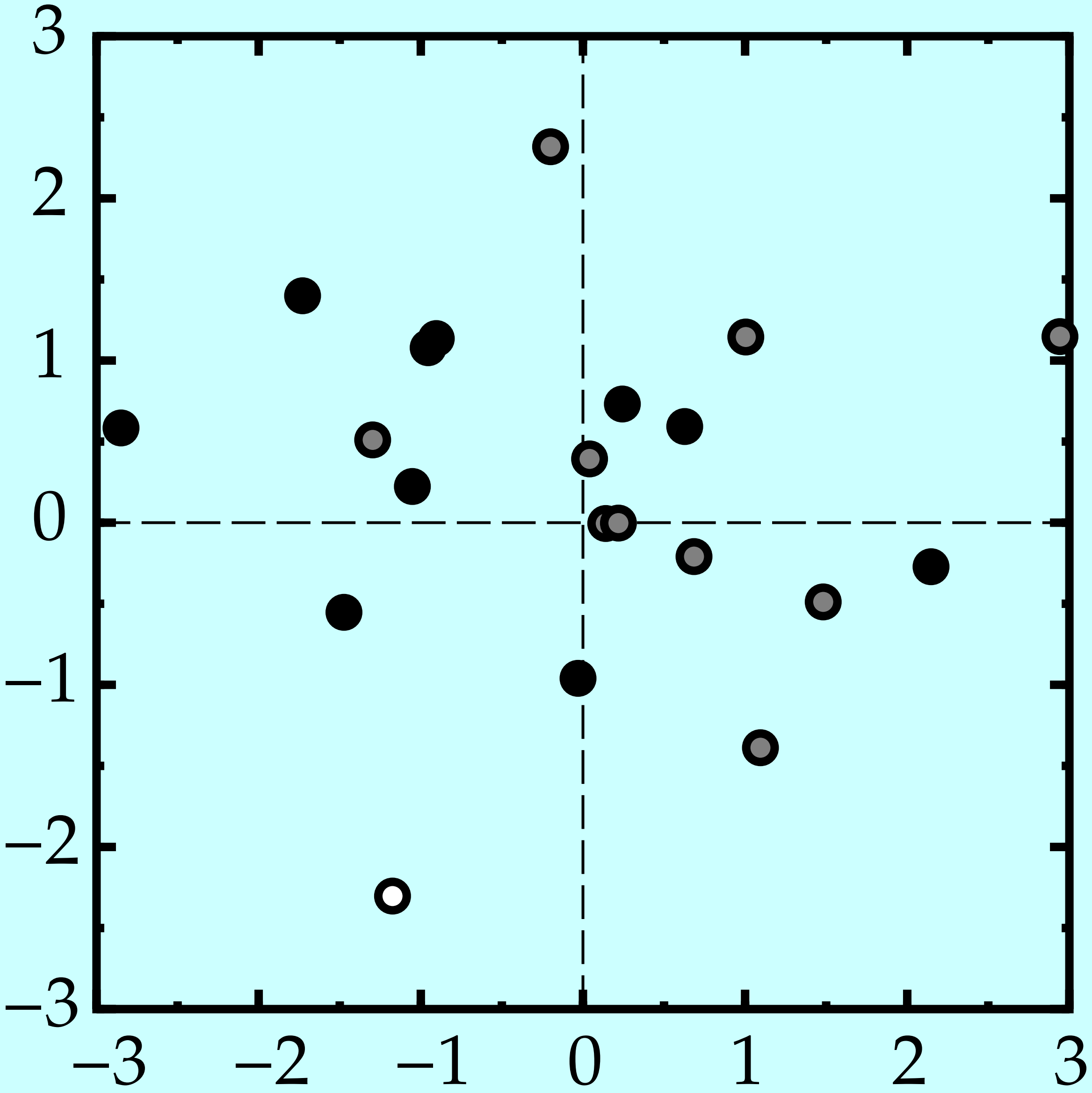
$\frac{(0.3)(0.1)}{0.3+0.1}$ { 0.075

add extra length

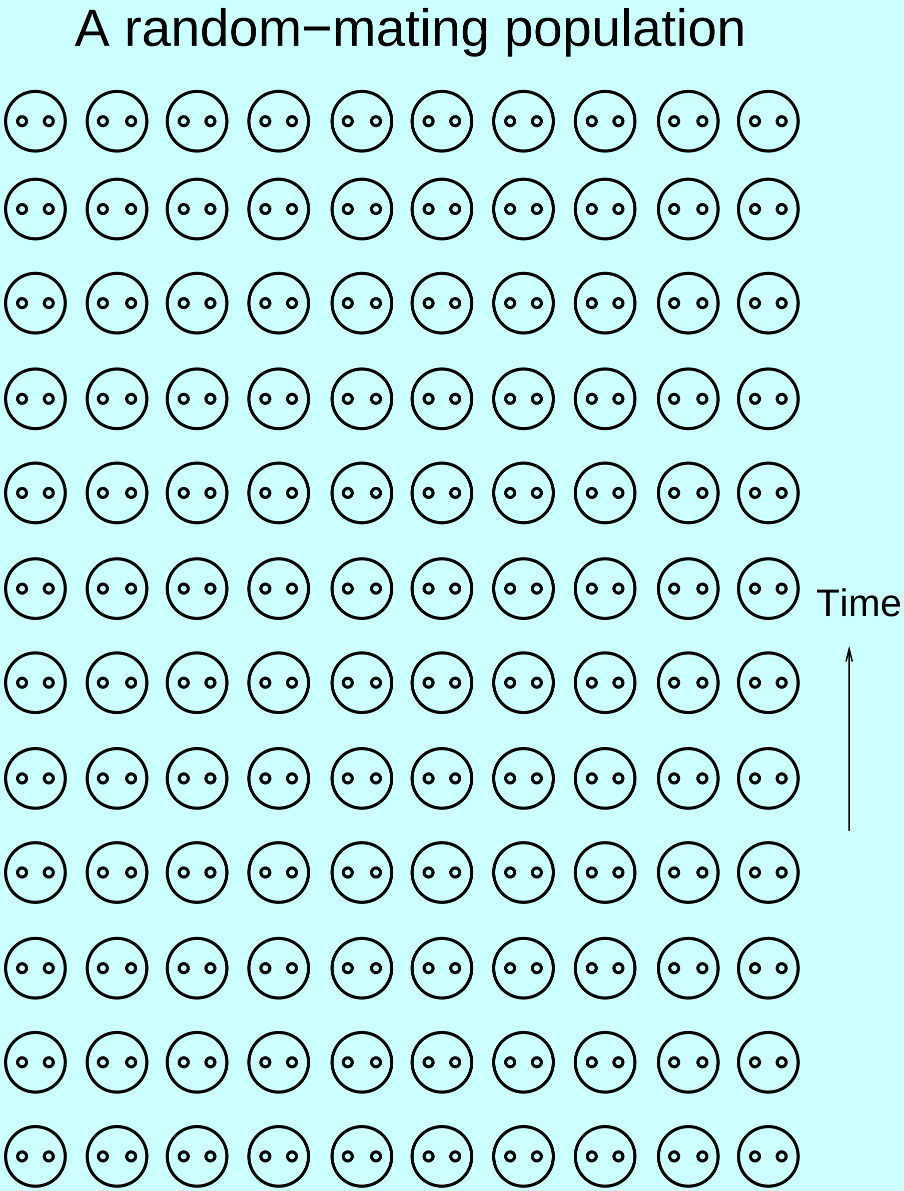
Contrasts on that tree

Contrast	Variance proportional to
$y_1 = x_a - x_b$	0.4
$y_2 = \frac{1}{4}\, x_a + \frac{3}{4}\, x_b - x_c$	0.975
$y_3 = x_d - x_e$	0.2
$y_4 = \frac{1}{6}\, x_a + \frac{1}{2}\, x_b + \frac{1}{3}\, x_c - \frac{1}{2}\, x_d - \frac{1}{2}\, x_e$	1.11666

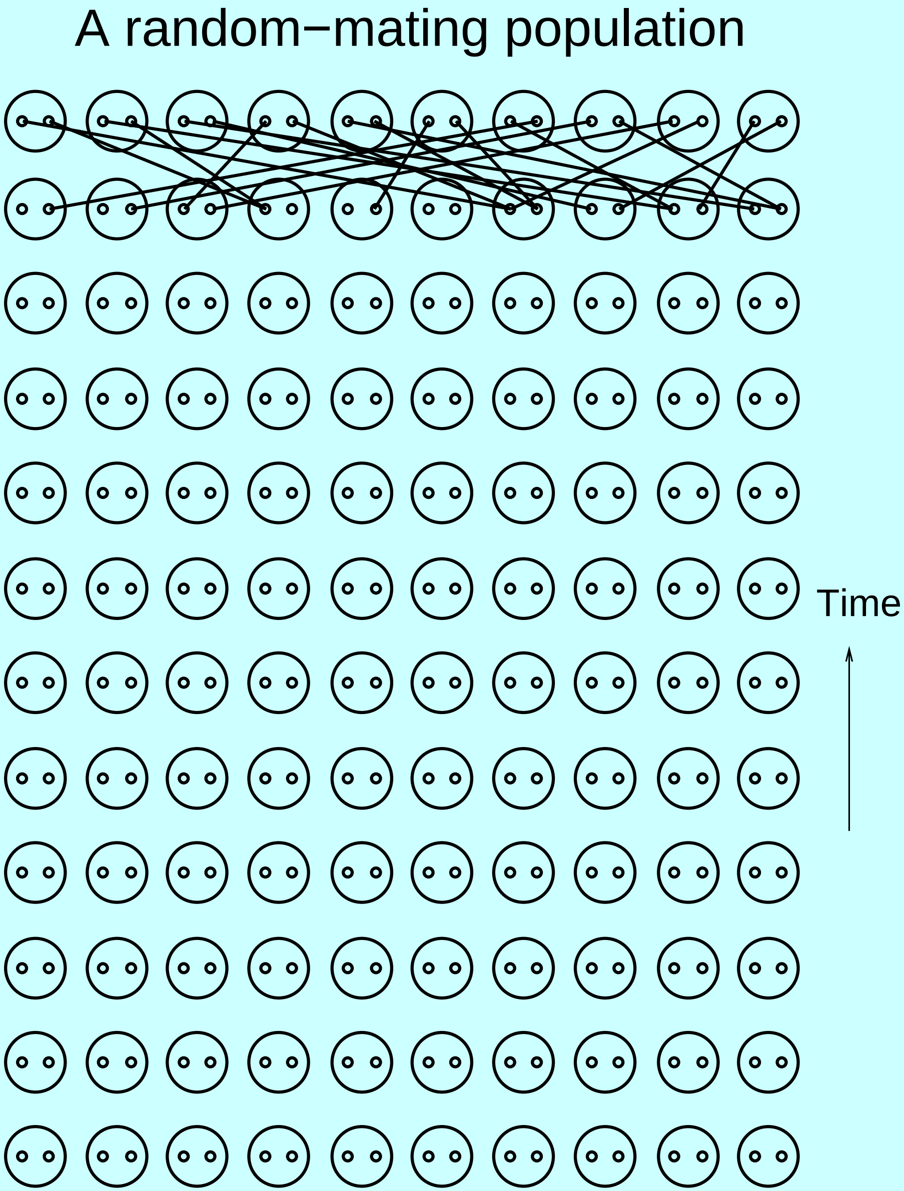
Plot standardized contrasts against each other



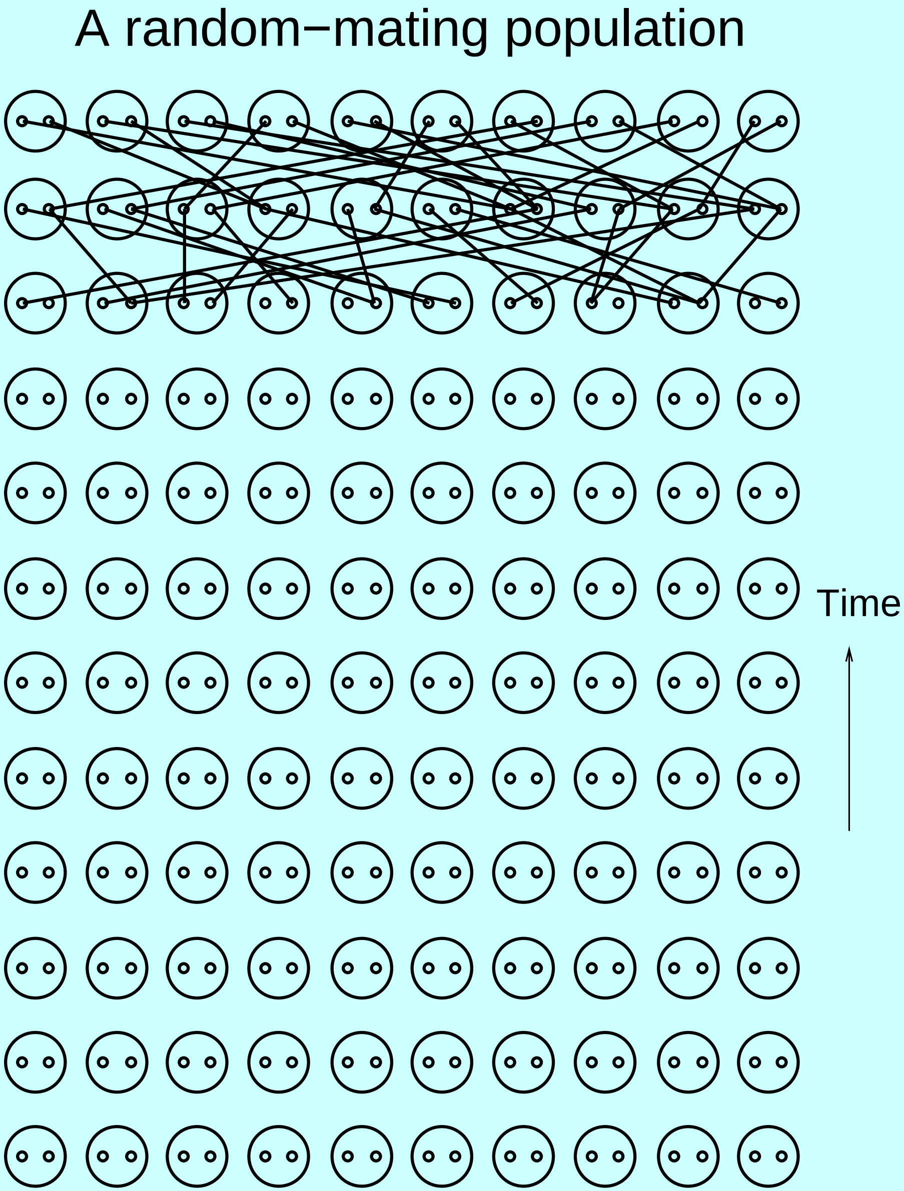
Gene copies in a population of 10 individuals



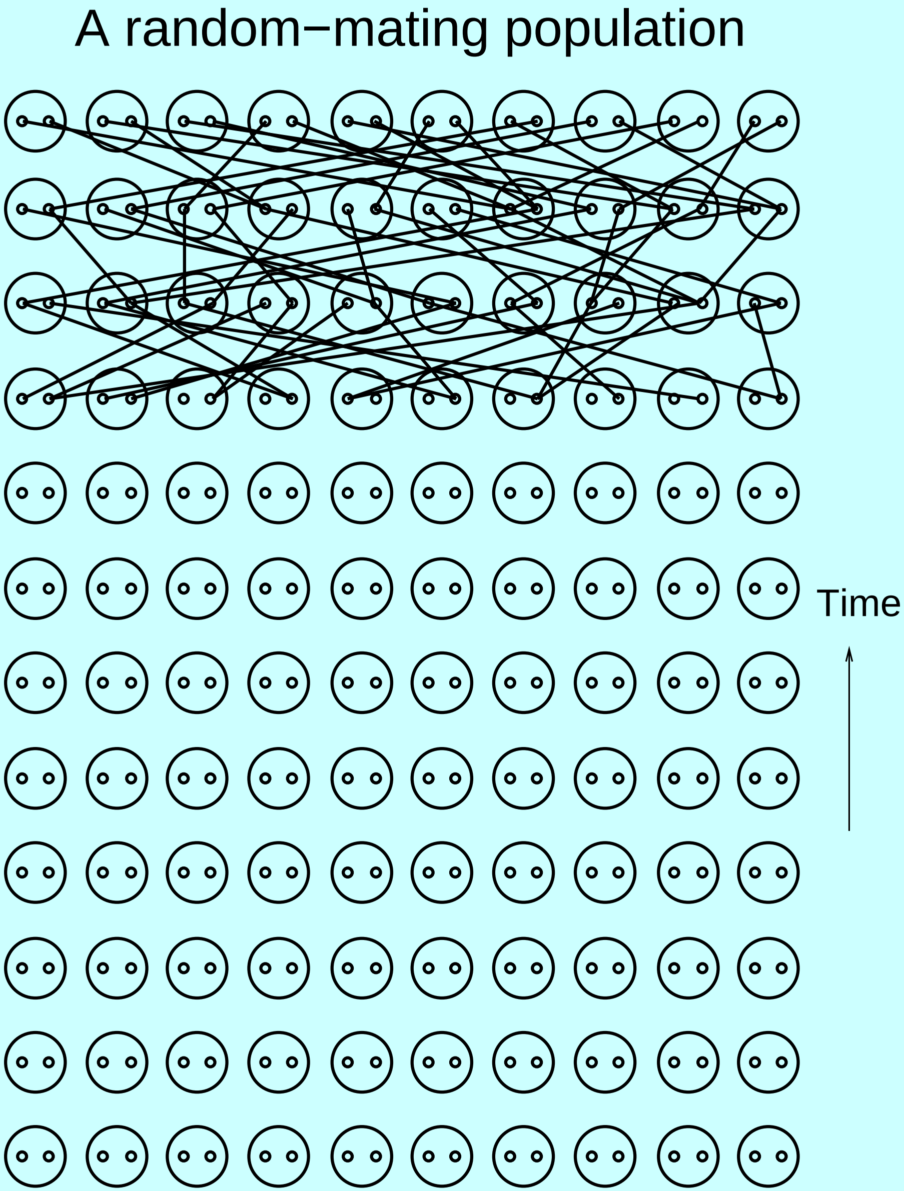
Going back one generation



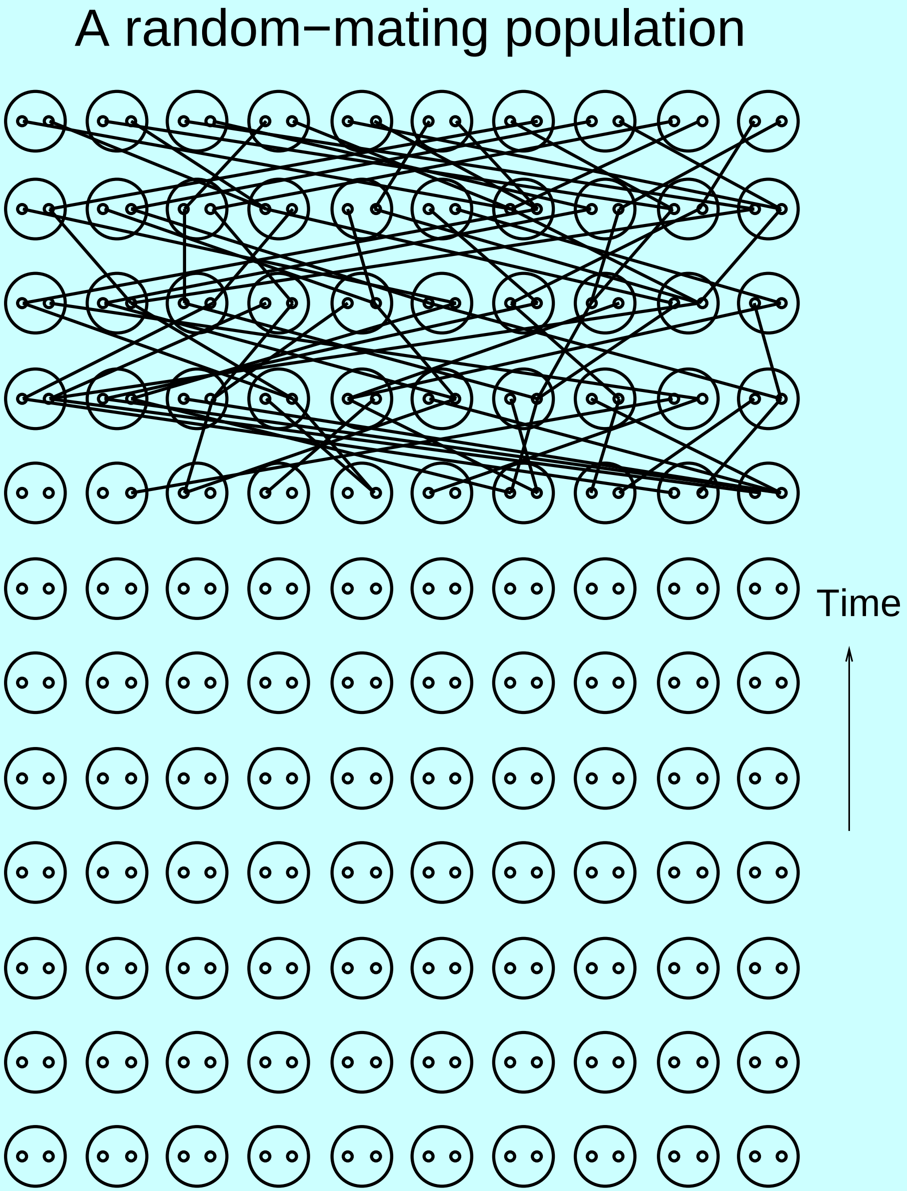
... and one more



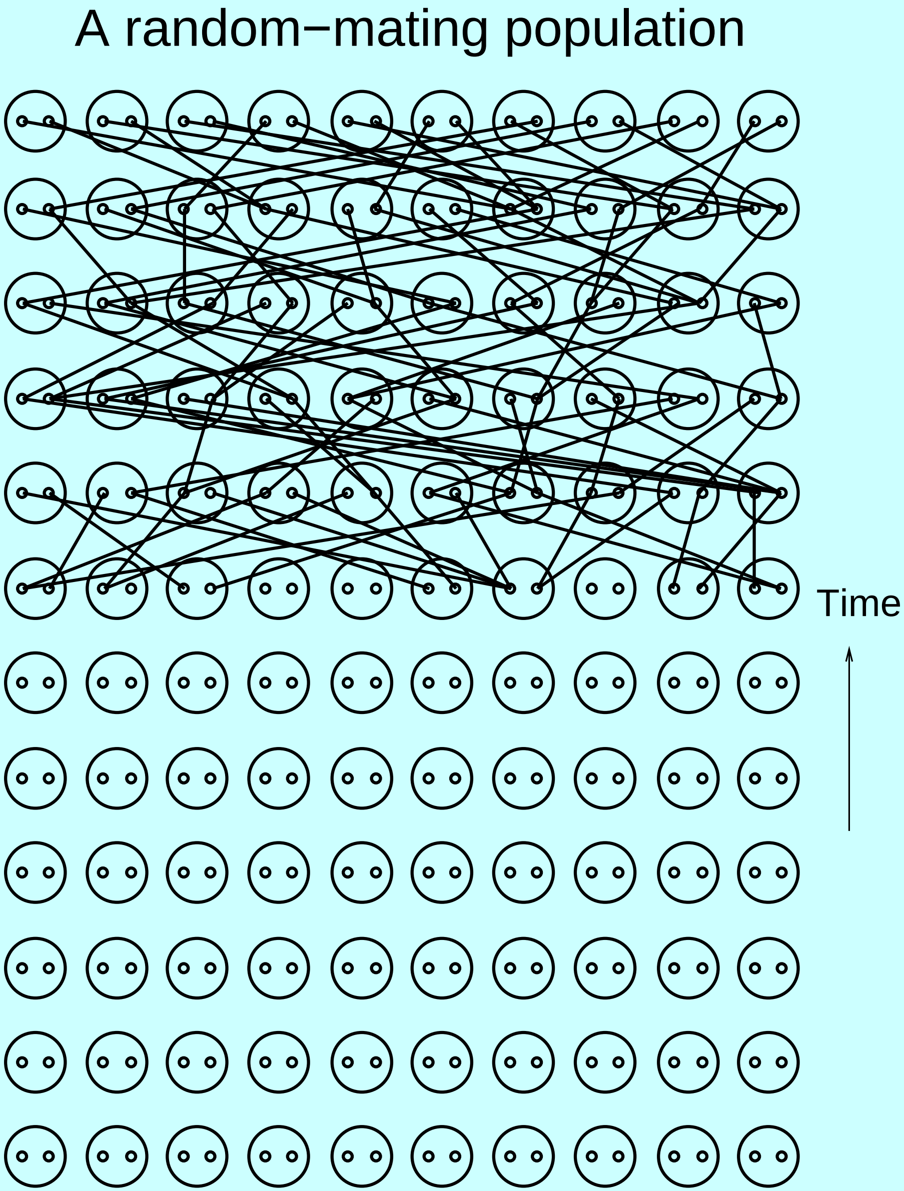
... and one more



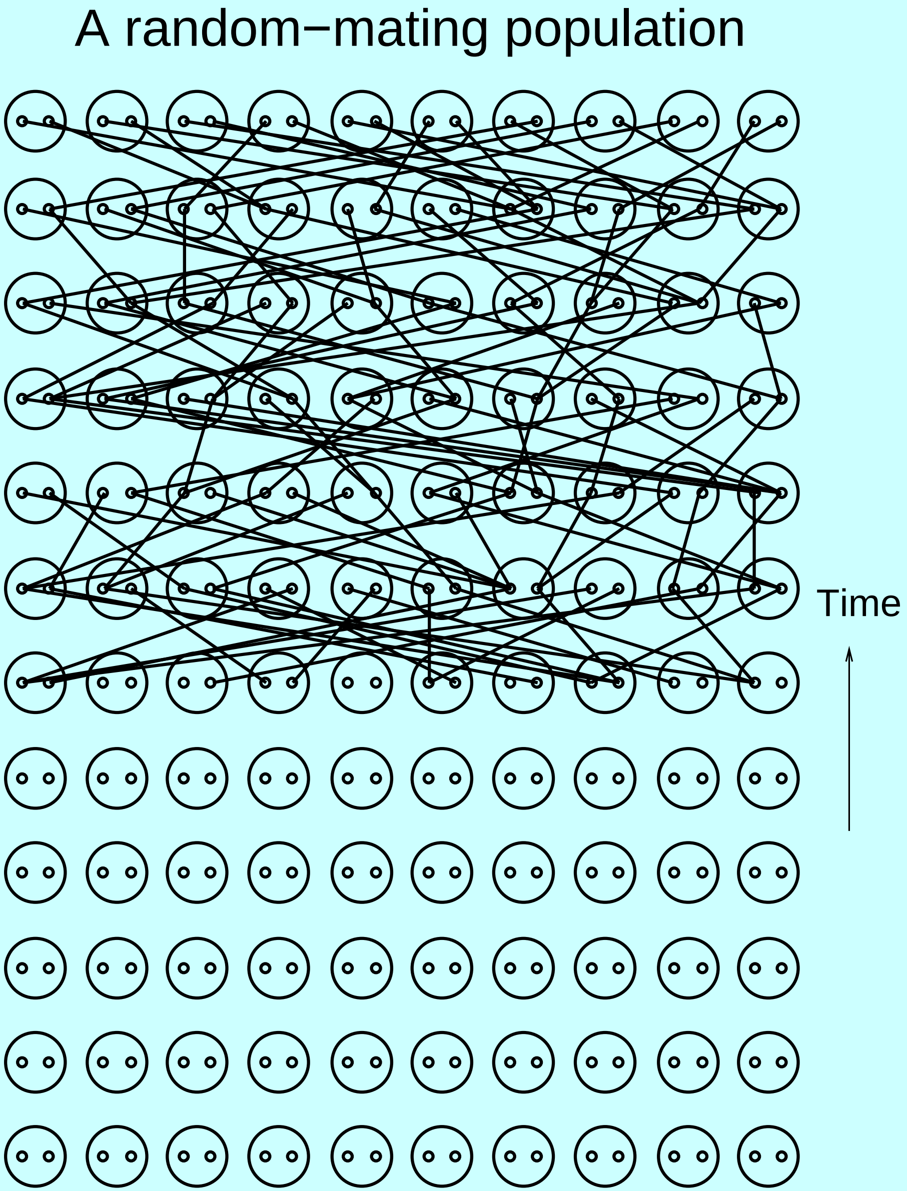
... and one more



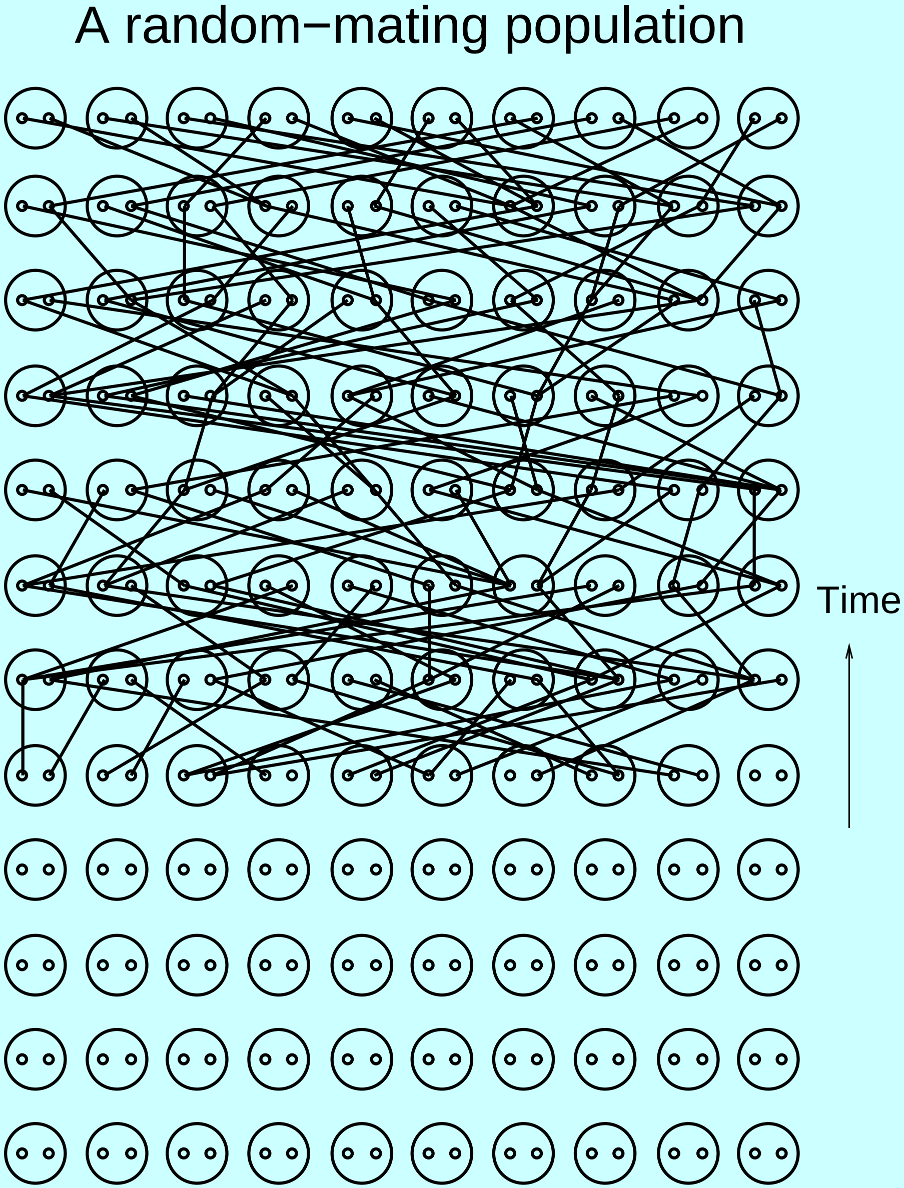
... and one more



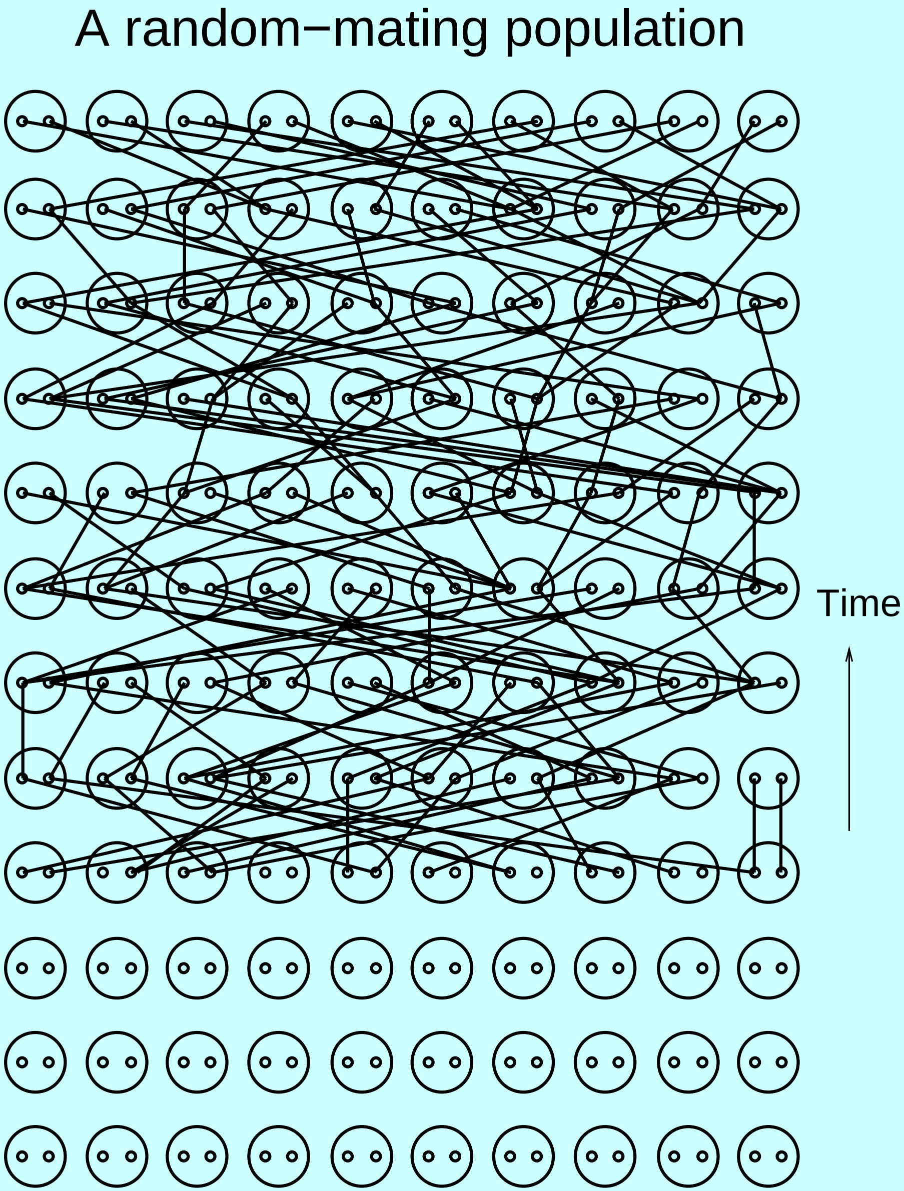
... and one more



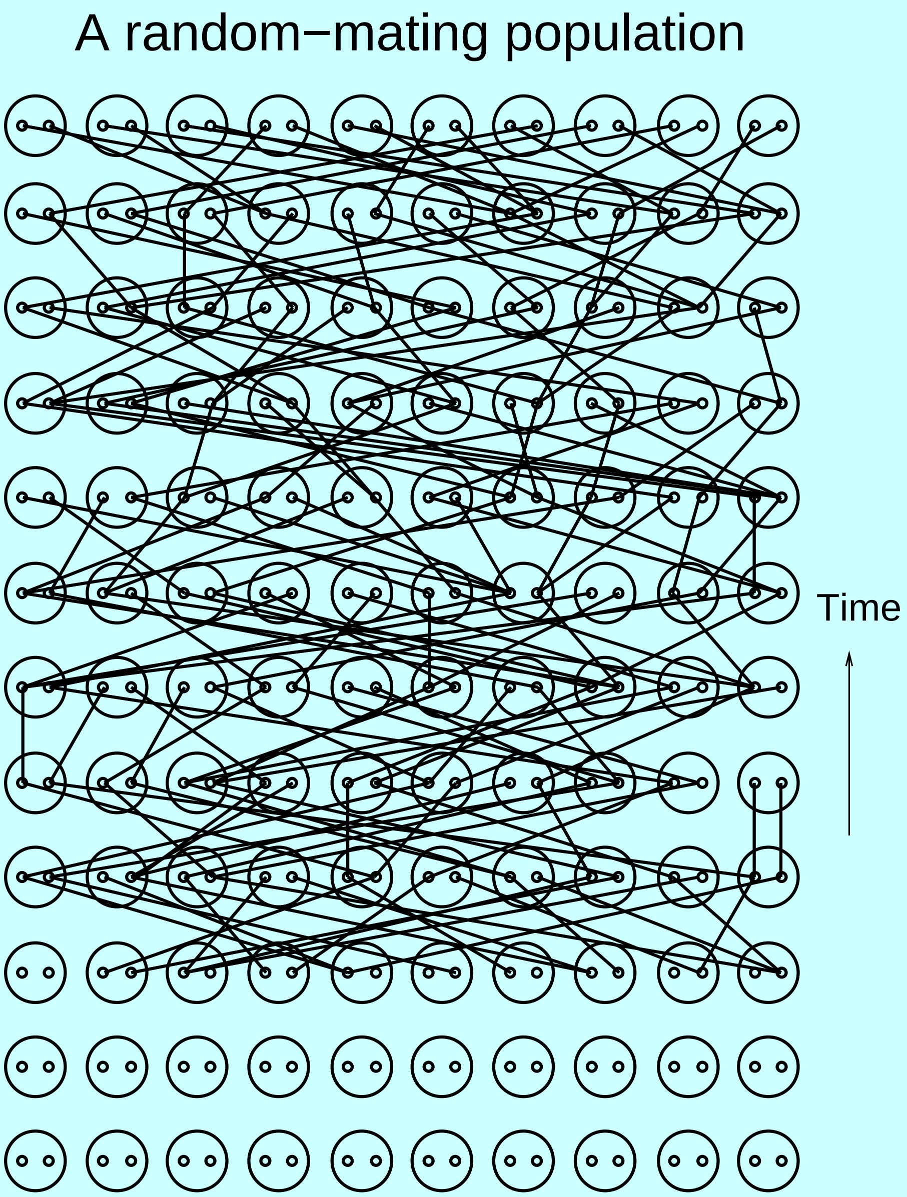
... and one more



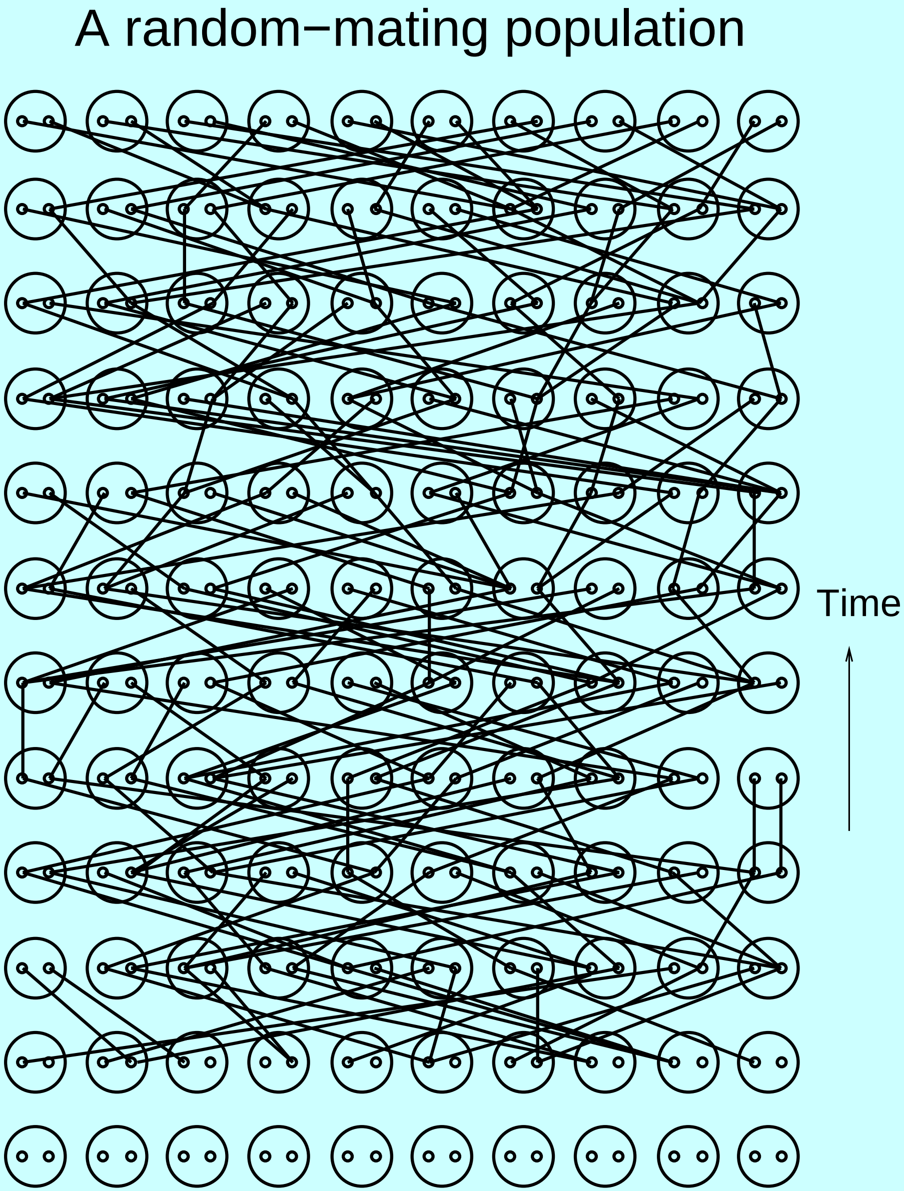
... and one more



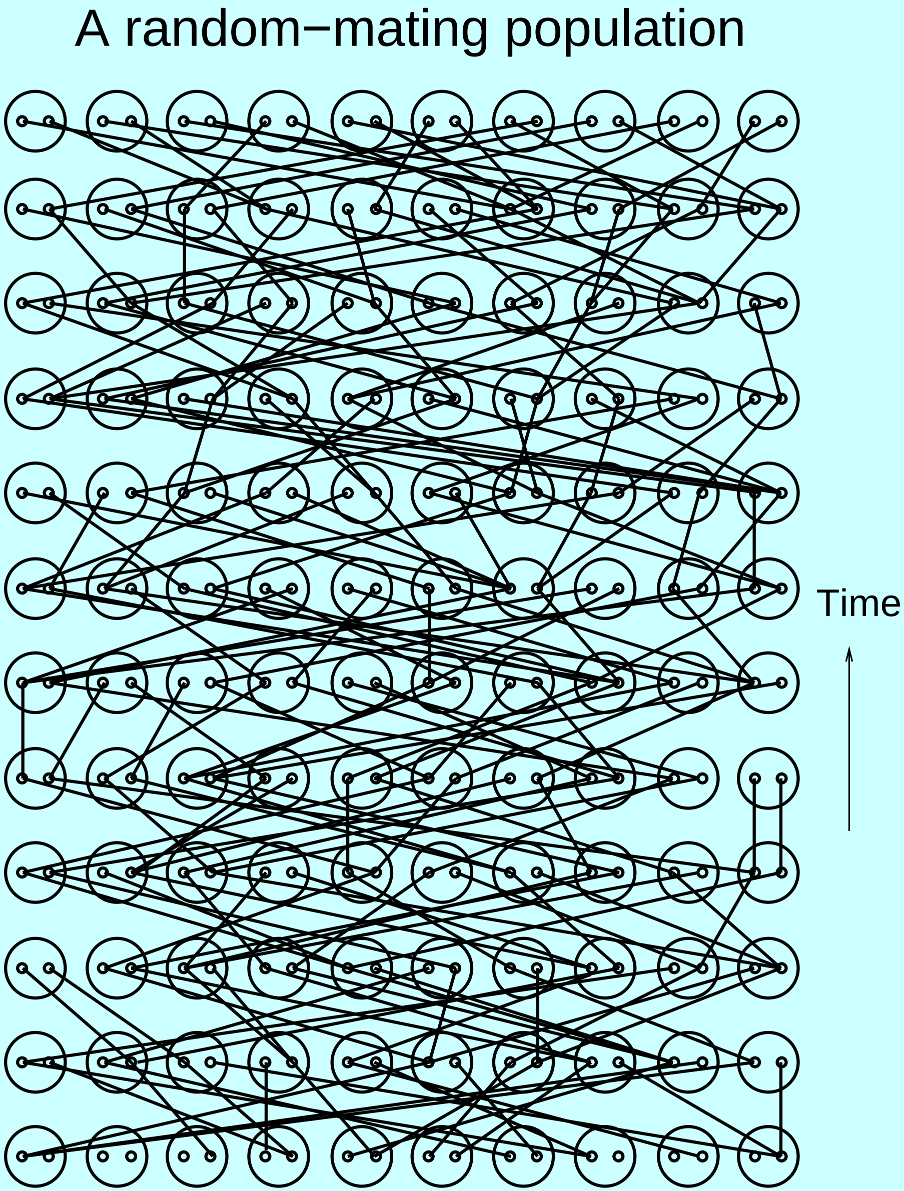
... and one more



... and one more

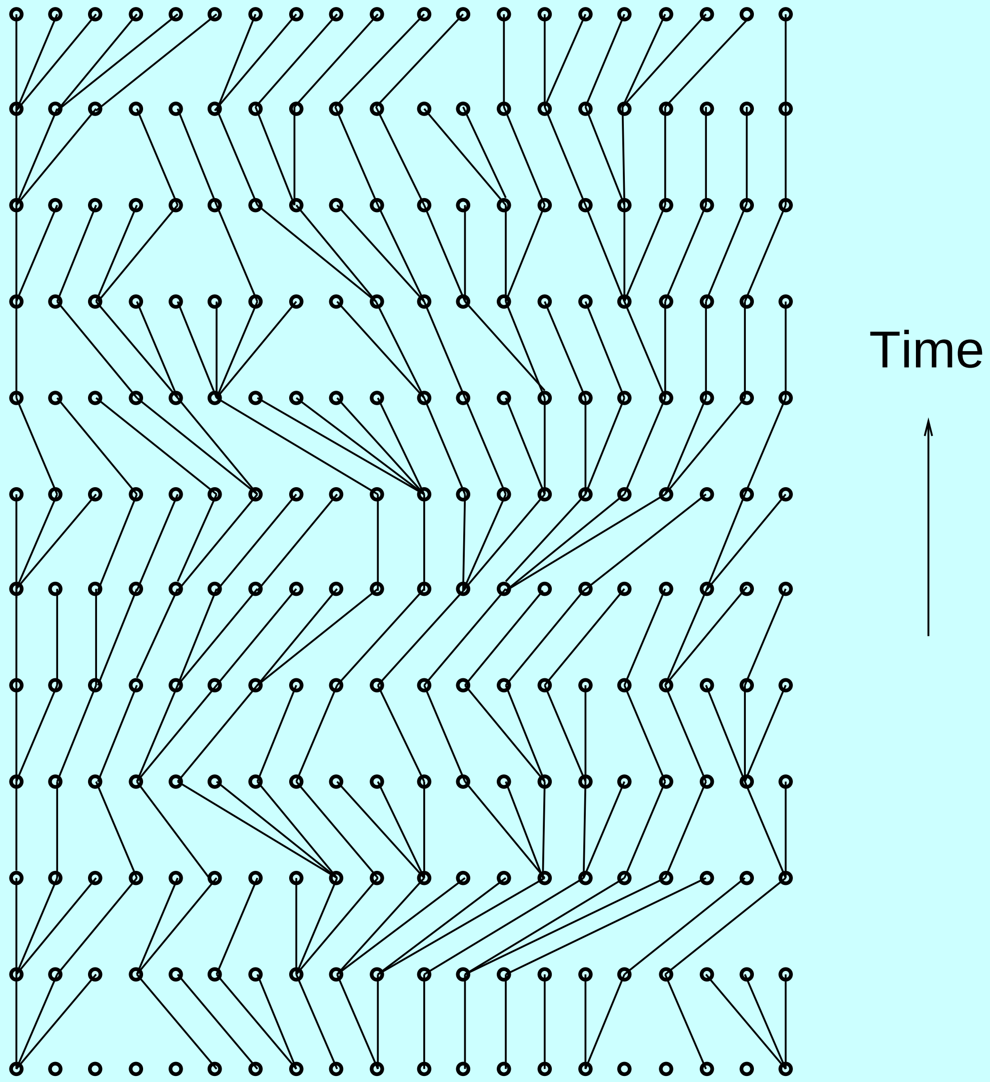


showing ancestry of gene copies



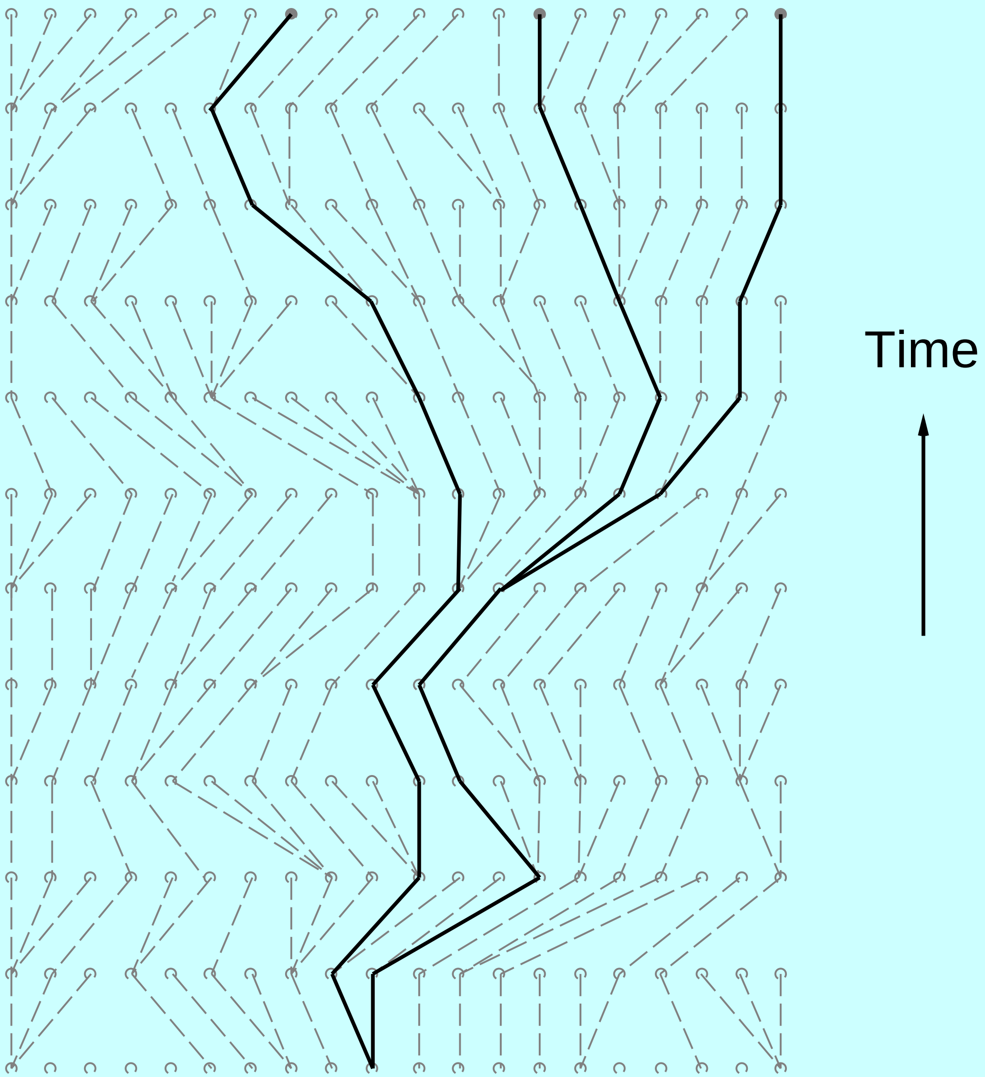
The genealogy of gene copies is a tree

Genealogy of gene copies, after reordering the copies

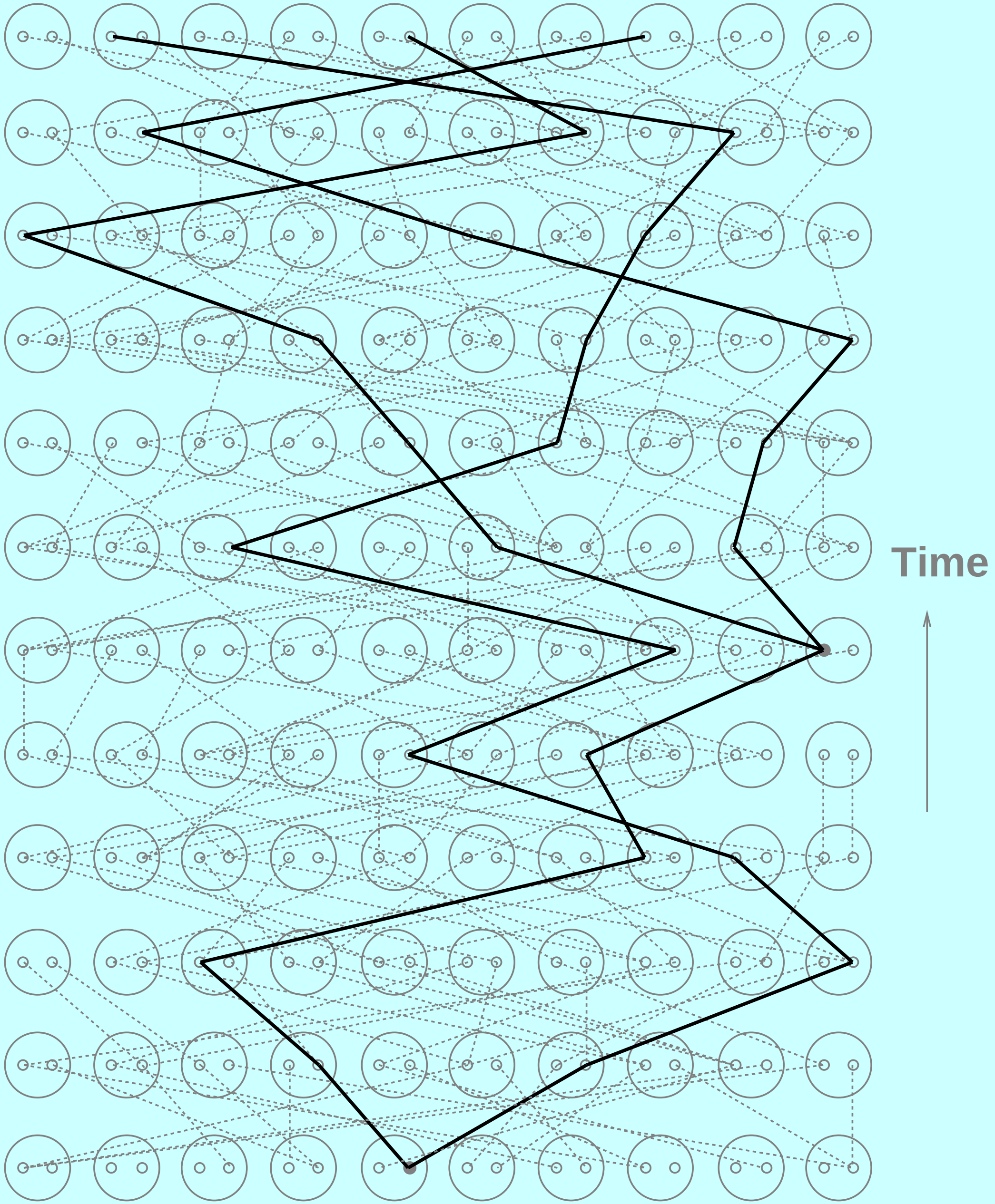


Ancestry of a sample of 3 copies

Genealogy of a small sample of genes from the population



Here is that tree of 3 copies in the pedigree



Kingman’s coalescent

Coalescent trees of gene copies within species (Kingman, 1982)

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

Collision is faster the smaller the effective population size

Average time for

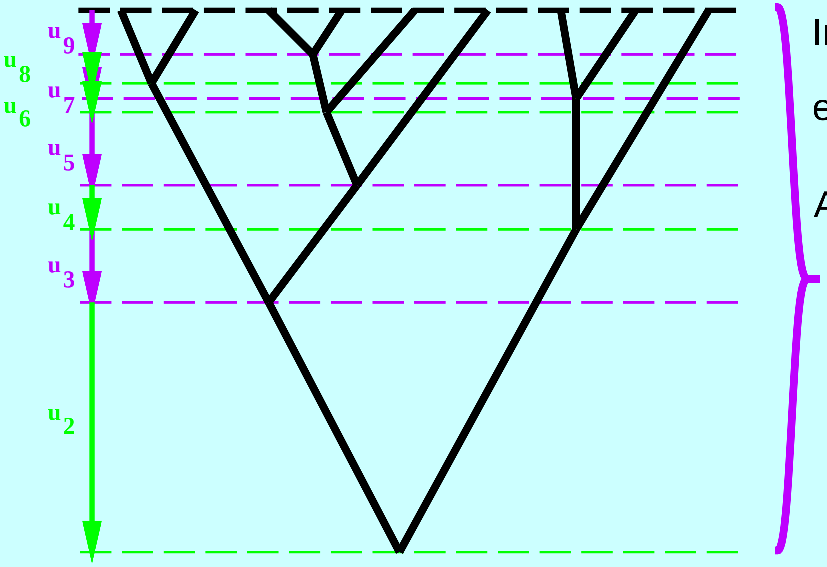
k copies to coalesce to

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

Average time for

two copies to coalesce

= 2N generations



In a diploid population of

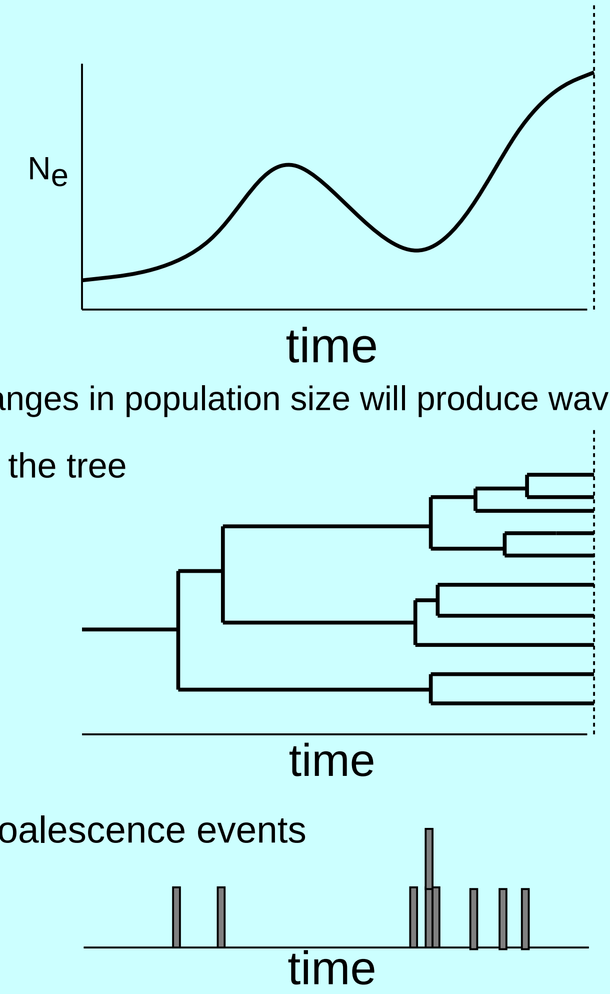
effective population size N,

Average time for n copies to coalesce

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

Coalescence is faster in small populations

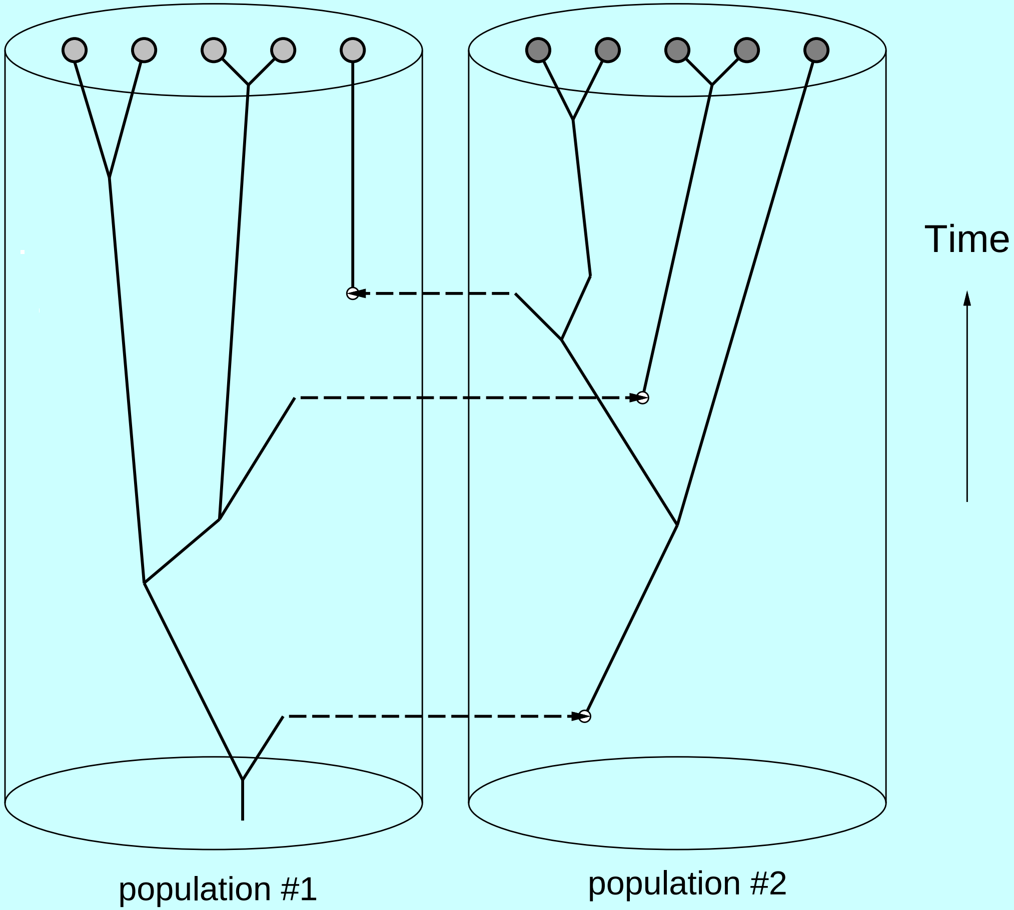
Change of population size and coalescents



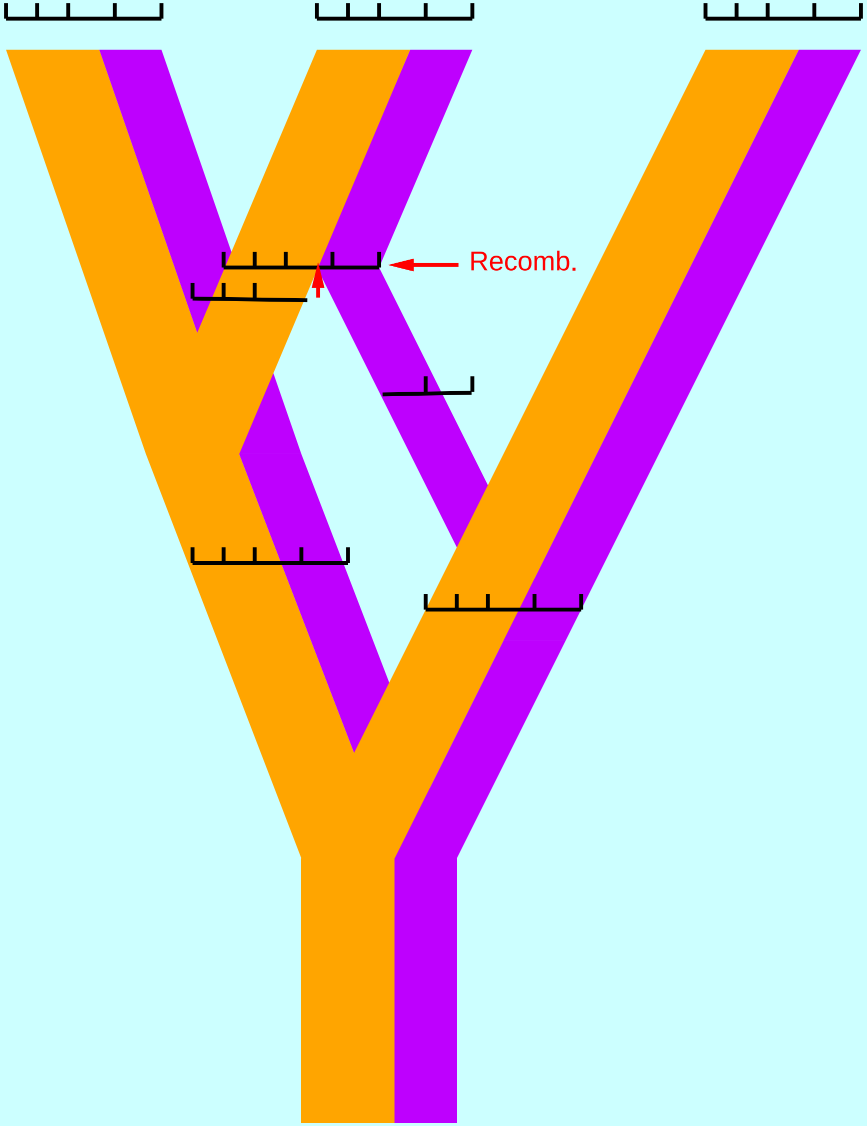
the changes in population size will produce waves of coalescence

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

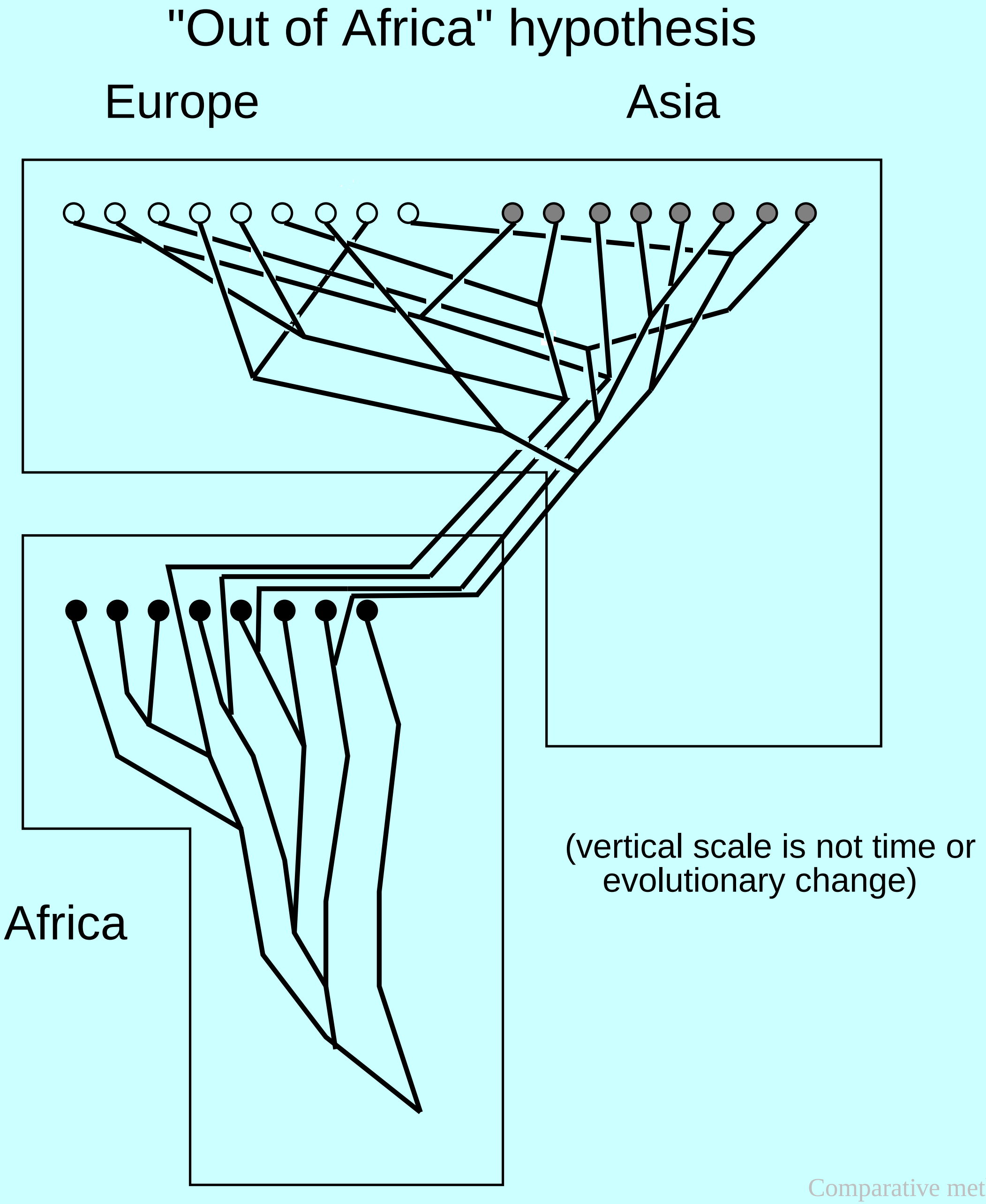


Recombination creates loops



Different markers have slightly different coalescent trees

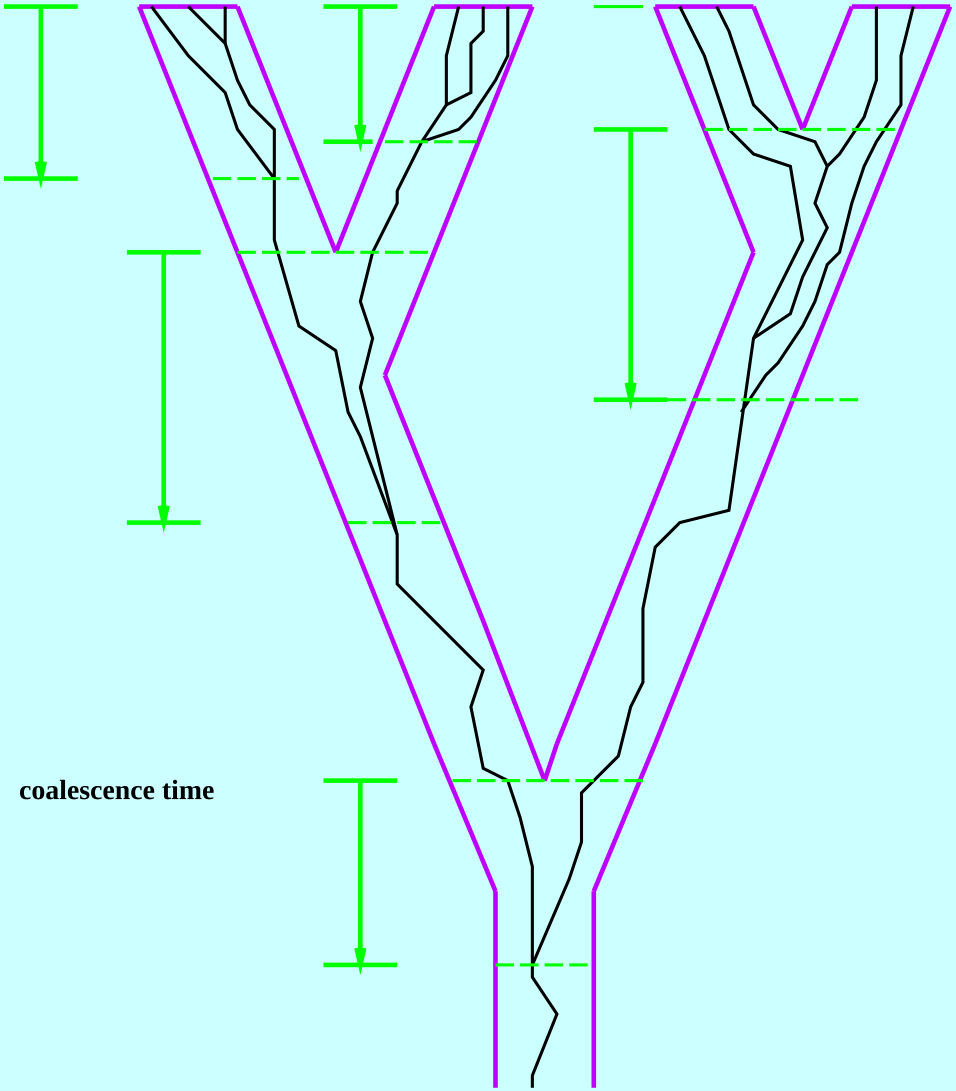
We want to be able to analyze human evolution



(vertical scale is not time or evolutionary change)

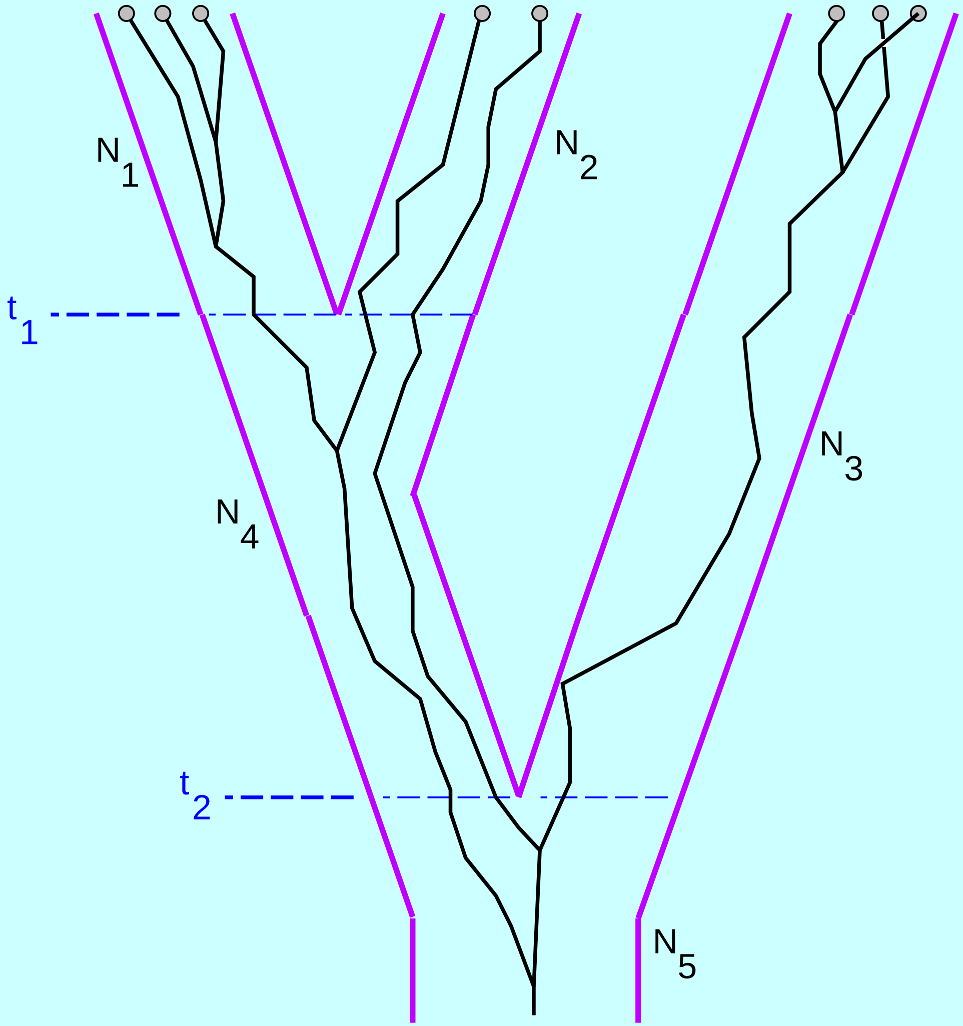
coalescent and “gene trees” versus species trees

Consistency of gene tree with species tree



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

Gene tree and Species tree



What to do with coalescents?

- They are poorly estimated (often only a modest number of sites is available for each tree).

What to do with coalescents?

- They are poorly estimated (often only a modest number of sites is available for each tree).
- Our interest is *not* in the coalescent tree itself, it is in the population and genetic parameters (population size, mutation rate, migration rate, population growth rate, rate of recombination).

What to do with coalescents?

- They are poorly estimated (often only a modest number of sites is available for each tree).
- Our interest is *not* in the coalescent tree itself, it is in the population and genetic parameters (population size, mutation rate, migration rate, population growth rate, rate of recombination).
- So we want to sum up likelihoods over our uncertainty about the tree, or do the equivalent in Bayesian terms.

What to do with coalescents?

- They are poorly estimated (often only a modest number of sites is available for each tree).
- Our interest is *not* in the coalescent tree itself, it is in the population and genetic parameters (population size, mutation rate, migration rate, population growth rate, rate of recombination).
- So we want to sum up likelihoods over our uncertainty about the tree, or do the equivalent in Bayesian terms.
- Got that? Our objective is *not* to “get the tree”! We don’t end up with a tree!

What to do with coalescents?

- They are poorly estimated (often only a modest number of sites is available for each tree).
- Our interest is *not* in the coalescent tree itself, it is in the population and genetic parameters (population size, mutation rate, migration rate, population growth rate, rate of recombination).
- So we want to sum up likelihoods over our uncertainty about the tree, or do the equivalent in Bayesian terms.
- Got that? Our objective is *not* to “get the tree”! We don’t end up with a tree!
- This can be done by Markov Chain Monte Carlo (MCMC) methods, in programs such as LAMARC, BEAST, MIGRATE, IMa or BEST (there are others too).

What to do with coalescents?

- They are poorly estimated (often only a modest number of sites is available for each tree).
- Our interest is *not* in the coalescent tree itself, it is in the population and genetic parameters (population size, mutation rate, migration rate, population growth rate, rate of recombination).
- So we want to sum up likelihoods over our uncertainty about the tree, or do the equivalent in Bayesian terms.
- Got that? Our objective is *not* to “get the tree”! We don’t end up with a tree!
- This can be done by Markov Chain Monte Carlo (MCMC) methods, in programs such as LAMARC, BEAST, MIGRATE, IMa or BEST (there are others too).
- ... and more approximately by Approximate Bayesian Computation (ABC) methods. Faster but not necessarily as efficient statistically.

Topics for the future ...

- Use of many loci

Topics for the future ...

- Use of many loci
- Use of SNP data on a large scale (if relevant)

Topics for the future ...

- Use of many loci
- Use of SNP data on a large scale (if relevant)
- Use of whole-genome sequences (in the longer run)

Topics for the future ...

- Use of many loci
- Use of SNP data on a large scale (if relevant)
- Use of whole-genome sequences (in the longer run)
- Integration of between-species and between-population studies with multiple loci across multiple species. IMPORTANT: If you are within a species, not all loci will have the same tree (we have just explained why, in the discussion of recombination). So you ought to consider coalescents that differ between loci, between SNPs and *not* just infer "the tree". (Also, please do *not* do phylogenies of individuals).

Topics for the future ...

- Use of many loci
- Use of SNP data on a large scale (if relevant)
- Use of whole-genome sequences (in the longer run)
- Integration of between-species and between-population studies with multiple loci across multiple species. IMPORTANT: If you are within a species, not all loci will have the same tree (we have just explained why, in the discussion of recombination). So you ought to consider coalescents that differ between loci, between SNPs and *not* just infer “the tree”. (Also, please do *not* do phylogenies of individuals).
- Integration of between-species and between-population studies with QTL mapping

Topics for the future ...

- Use of many loci
- Use of SNP data on a large scale (if relevant)
- Use of whole-genome sequences (in the longer run)
- Integration of between-species and between-population studies with multiple loci across multiple species. IMPORTANT: If you are within a species, not all loci will have the same tree (we have just explained why, in the discussion of recombination). So you ought to consider coalescents that differ between loci, between SNPs and *not* just infer "the tree". (Also, please do *not* do phylogenies of individuals).
- Integration of between-species and between-population studies with QTL mapping
- Integration of between-species and between-population studies with morphological characters.

References

Comparative methods

Felsenstein, J. 1985. Phylogenies and the comparative method. *American Naturalist* **125:** 1-15. [The contrasts method]

Harvey, P. H. and M. D. Pagel. 1991. *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford. [Reviews early work by me, Mark Ridley and the authors on comparative methods]

Pagel, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London, Series B* **255:** 37-45. [Method for two-state discrete characters]

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts. [Especially chapter 25 which covers comparative methods]

Felsenstein, J. 2012. A comparative method for both discrete and continuous characters using the threshold model. *American Naturalist* **179:** 145-156. [Using Sewall Wright's 1934 "threshold model" to get a comparative method that can handle both discrete and continuous characaters]

The coalescent

Griffiths, R. C. and S. Tavaré. 1994a. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Socety of London, Series B (Biological Sciences)* **344:** 403-10. [The pioneering sampling method]

(continued)

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1995. Effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140:** 1421-1430. [Our MCMC coalescent likelihood method]

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

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

Nielsen, R. and M. Slatkin. 2013. An Introduction to Population Genetics. Theory and Applications. Sinauer Associates, Sunderland, Massachusetts. Population genetics textbook with more coverage of coalescents than usual.

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts. [Especially chapter 27 which covers MCMC likelihood approaches (but explanation of logic of Griffiths/Tavaré method is wrong)]

Felsenstein, J. 2007. Trees of genes in populations. pp. 3-29 in *Reconstructing Evolution. New Mathematical and Computational Advances*, pp. 3-27 in by O. Gascuel and M. Steel. Oxford University Press, Oxford. [Review of coalescents including MCMC, for a somewhat mathematical audience]