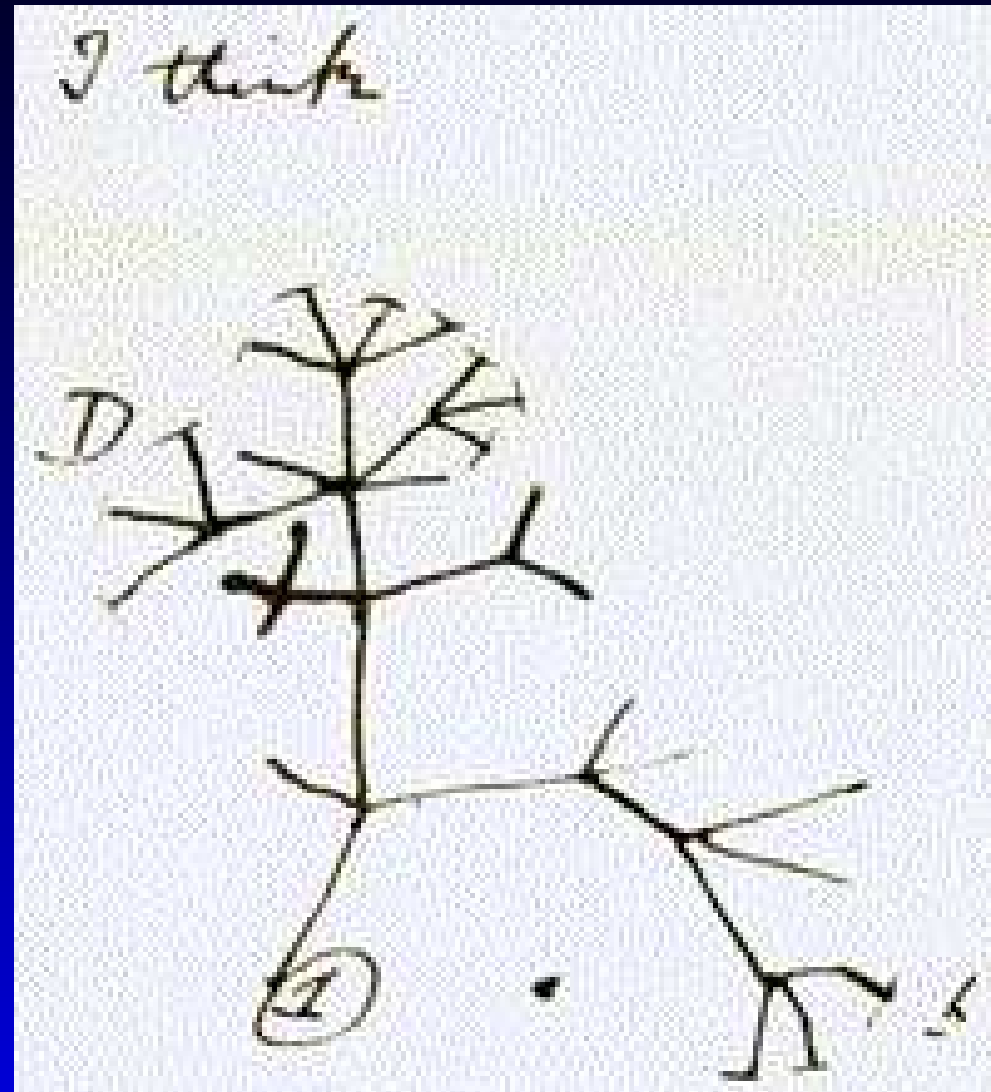


**Evolutionary trees, coalescents, and gene trees:**  
*can mathematicians find the woods?*

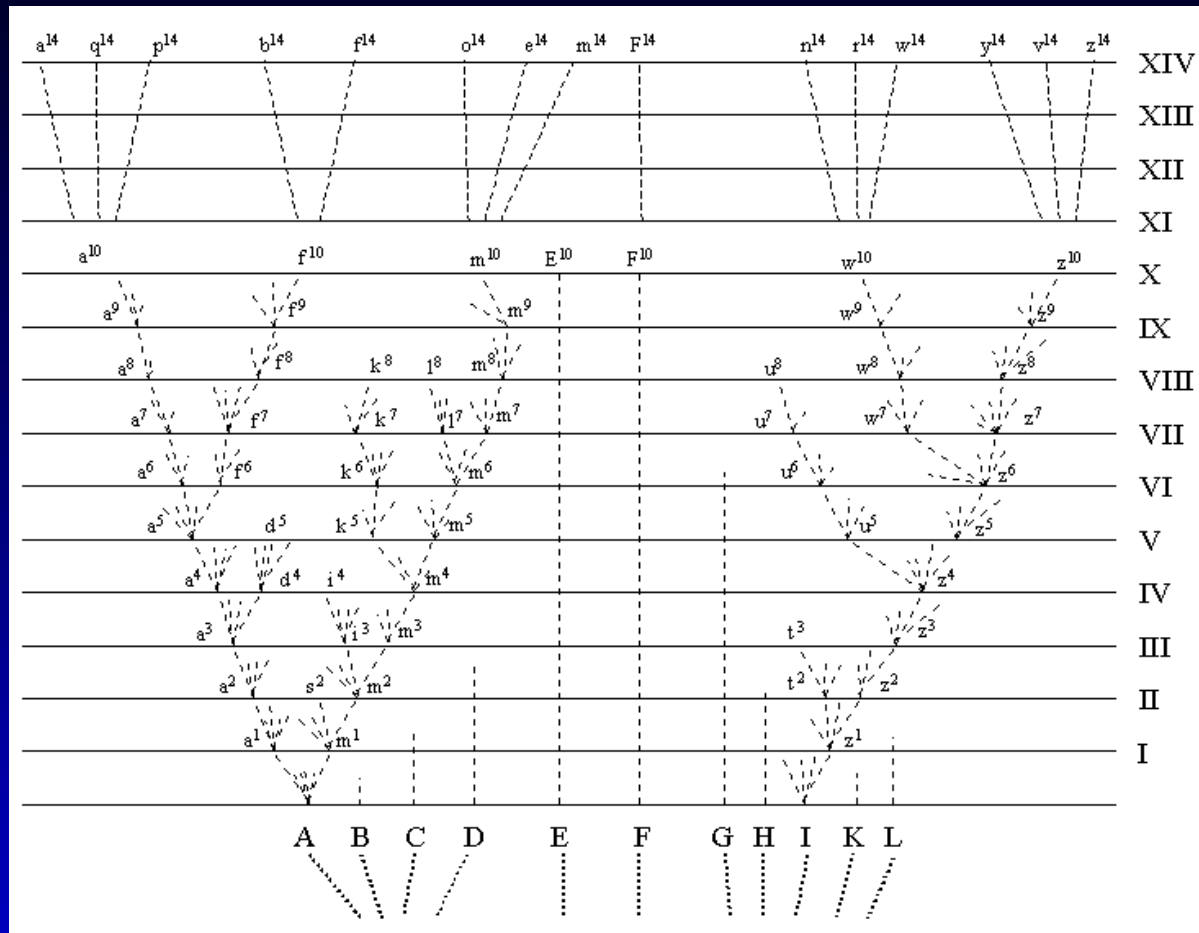
Joe Felsenstein

Department of Genome Sciences and Department of Biology

# Sketch in Darwin's notebook, 1842

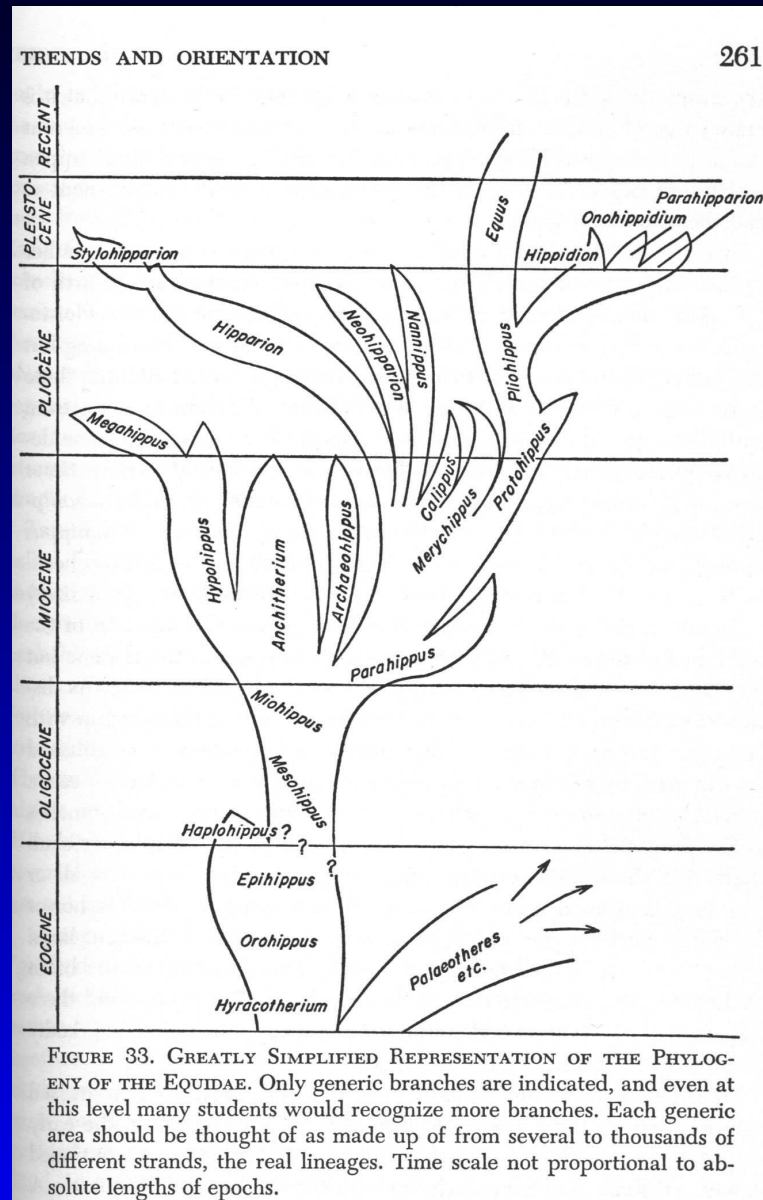


# The only figure in Darwin's book, 1859

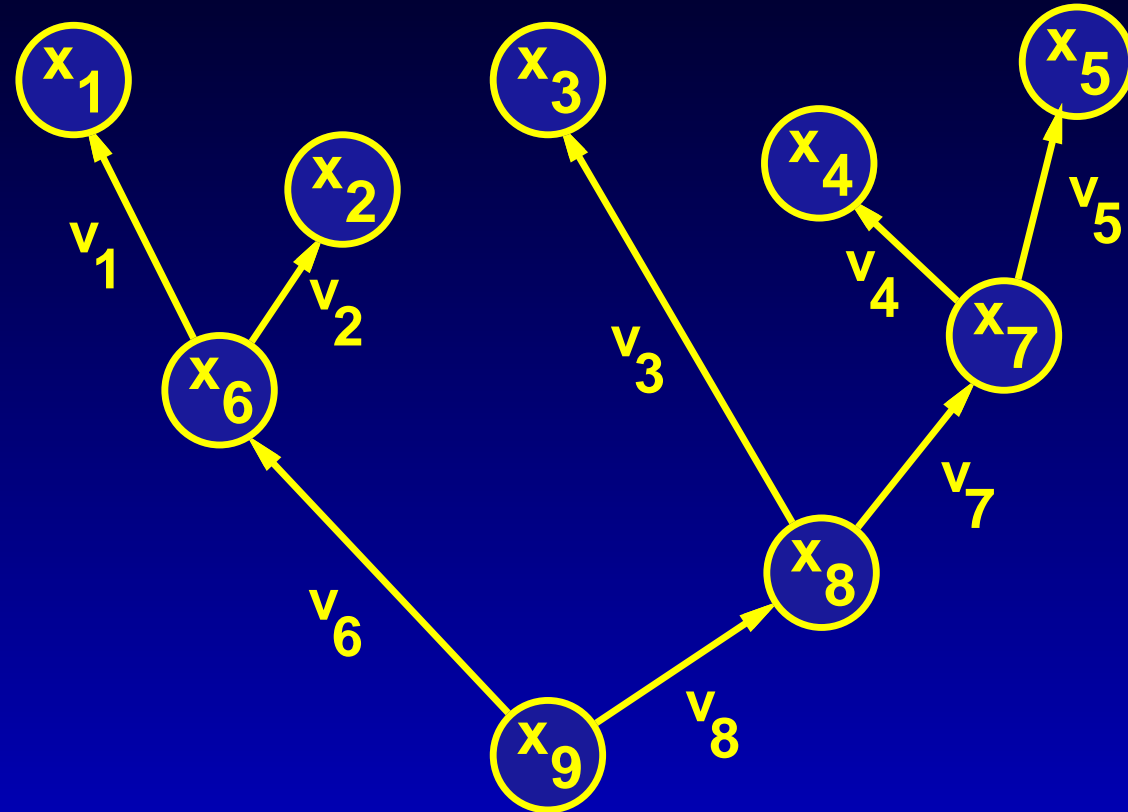




# Simpson's tree of horses, 1940's

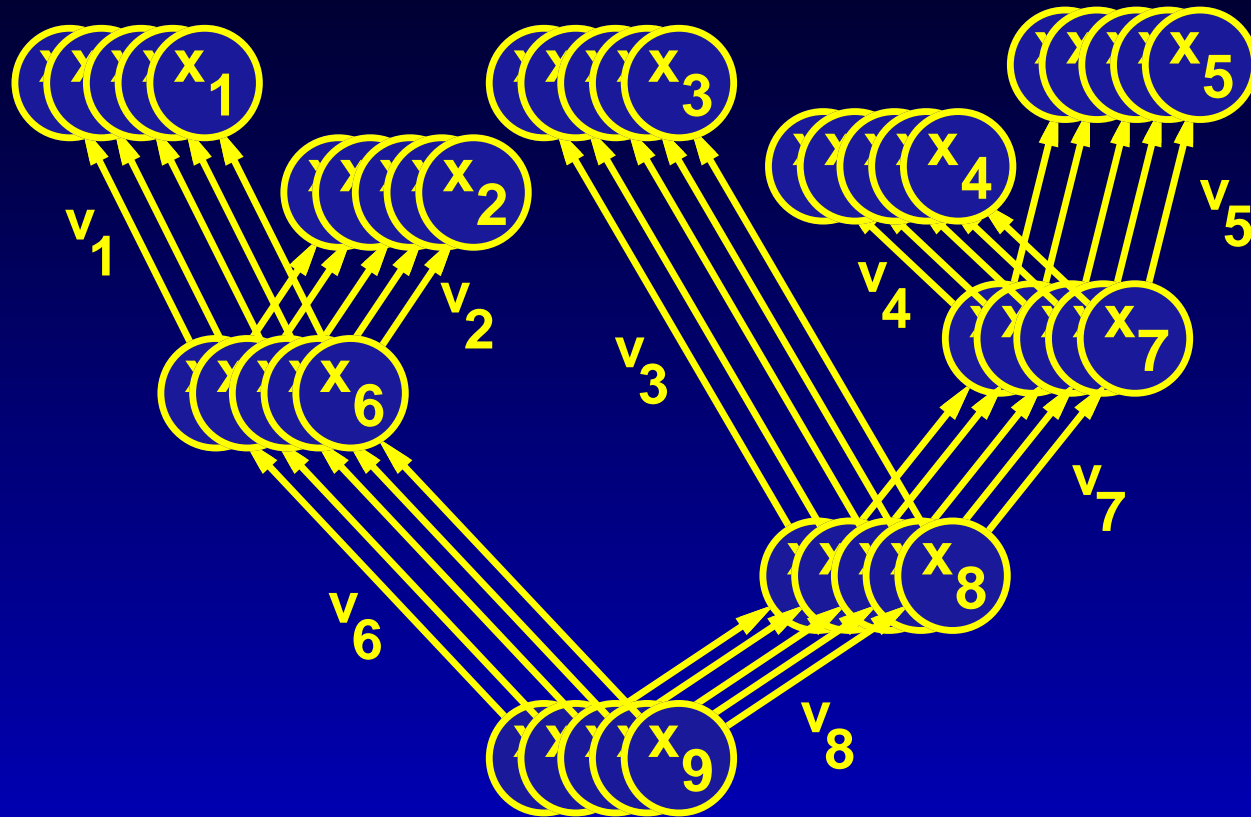


## Evolution of characters along trees



Characters are modeled (in simple cases) as changing by Markov processes along branches of the tree, according to graphical models like this.

## Evolution of characters along trees



With multiple characters (such as sites in DNA) sites are often assumed to change independently, so we have stacked graphical models.

# A data example: mitochondrial D-loop sequences

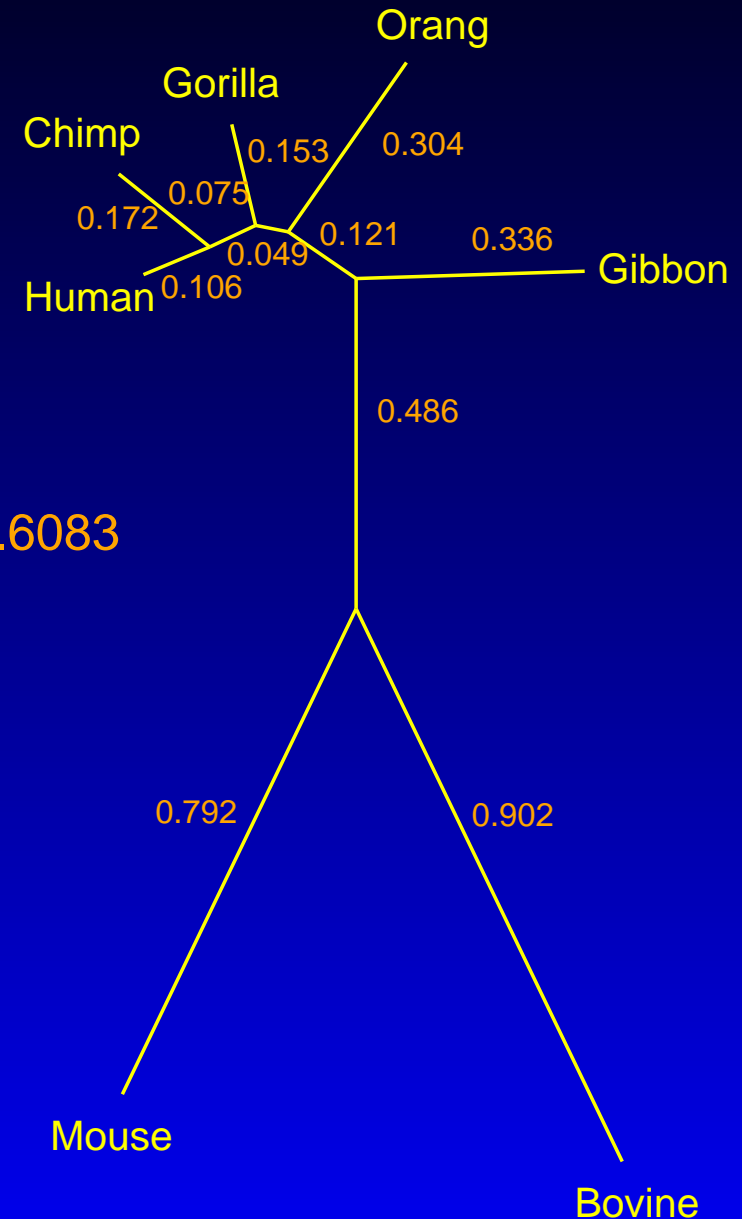
Bovine	CCAAACCTGT	CCCCACCATC	TAACACCAAC	CCACATATAC	AAGCTAAACC	AAAAATACCA
Mouse	CCAAAAAAC	ATCCAAACAC	CAACCCCAGC	CCTTACGCAA	TAGCCATACA	AAGAATATTA
Gibbon	CTATACCCAC	CCAACCTGAC	CTACACCAAT	CCCCACATAG	CACACAGACC	AACAACCTCC
Orang	CCCCACCCGT	CTACACCAGC	CAACACCAAC	CCCCACCTAC	TATACCAACC	AATAACCTCT
Gorilla	CCCCATTTAT	CCATAAAAC	CAACACCAAC	CCCCATCTAA	CACACAAACT	AATGACCCCC
Chimp	CCCCATCCAC	CCATACAAAC	CAACATTACC	CTCCATCCAA	TATACAAACT	AACAACCTCC
Human	CCCCACTCAC	CCATACAAAC	CAACACCACT	CTCCACCTAA	TATACAAATT	AATAACCTCC

TACTACTAAA	AACTCAAATT	AACTCTTTAA	TCTTTATACA	ACATTCCACC	AACCTATCCA
TACAACCATA	AATAAGACTA	ATCTATTTAA	ATAACCCATT	ACGATACAAA	ATCCCTTTCCG
CACCTTCCAT	ACCAAGCCCC	GACTTTACCG	CCAACGCACC	TCATCAAAAC	ATACCTACAA
CAACCCCTAA	ACCAAACACT	ATCCCCAAAA	CCAACACACT	CTACCAAAAT	ACACCCCCAA
CACCCTCAA	GCCAAACACC	AACCCTATAA	TCAATACGCC	TTATCAAAAC	ACACCCCCAA
CACTCTTCAG	ACCGAACACC	AATCTCACAA	CCAACACGCC	CCGTCAAAAC	ACCCCTTCAG
CACCTTCAGA	ACTGAACGCC	AATCTCATAA	CCAACACACC	CCATCAAAGC	ACCCCTCCAA

CACAAAAAAA	CTCATATTTA	TCTAAATACG	AACTTCACAC	AACCTTAACA	CATAAACATA
TCTAGATACA	AACCACAACA	CACAATTAAT	ACACACCACA	ATTACAATAC	TAAACTCCCA
CACAAACAAA	TGCCCCCCCA	CCCTCCTTCT	TCAAGCCCAC	TAGACCATCC	TACCTTCCTA
TTCACATCCG	CACACCCCCA	CCCCCCCTGC	CCACGTCCAT	CCCATCACCC	TCTCCTCCCA
CATAAACCCA	CGCACCCCCA	CCCCTTCCGC	CCATGCTCAC	CACATCATCT	CTCCCCTTCA
CACAAATTCA	TACACCCTTA	CCTTTCCTAC	CCACGTTCAC	CACATCATCC	CCCCCTCTCA
CACAAACCCG	CACACCTCCA	CCCCCCTCGT	CTACGCTTAC	CACGTTCATCC	CTCCCTCTCA

CCCAGCCCA	ACACCCTTCC	ACAAATCCTT	AATATACGCA	CCATAAATAA	CA
TCCCACCAA	TCACCCTCCA	TCAAATCCAC	AAATTACACA	ACCATTAACC	CA
GCACGCCAAG	CTCTCTACCA	TCAAACGCAC	AACTTACACA	TACAGAACCA	CA

# which gives the ML tree



$$\ln L = -1405.6083$$

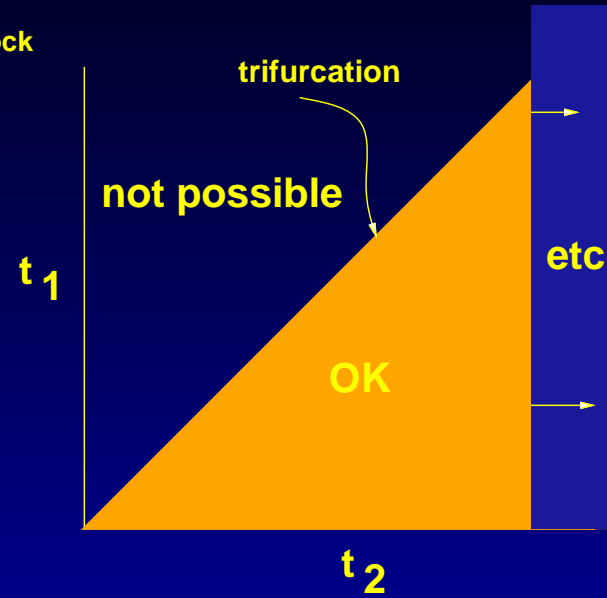
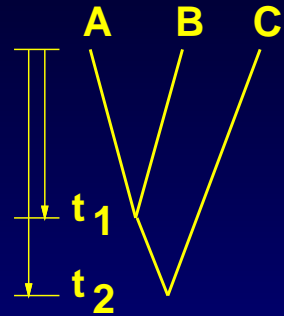
Maximum likelihood tree for the Hasegawa 232-site mitochondrial D-loop data set, with Ts/Tn set to 2, analyzed with maximum likelihood (DNAML)

# How many trees? Rooted, bifurcating, tips labelled

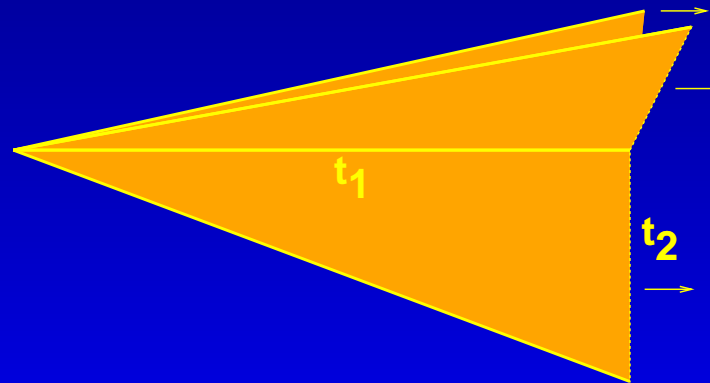
species	number of trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
30	$4.9518 \times 10^{38}$
40	$1.00005 \times 10^{57}$

# Tree space in a small case

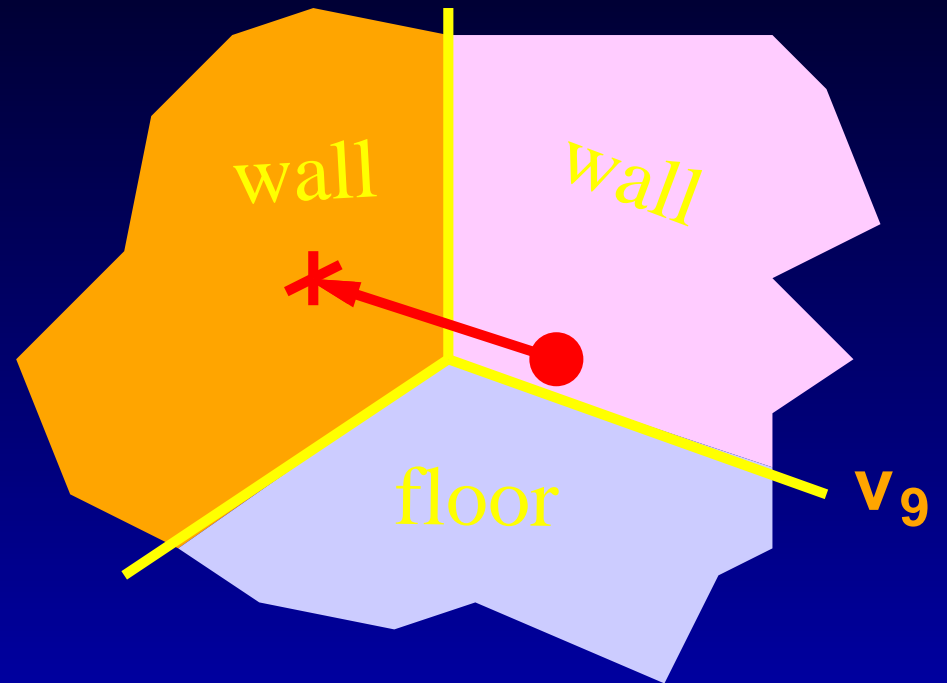
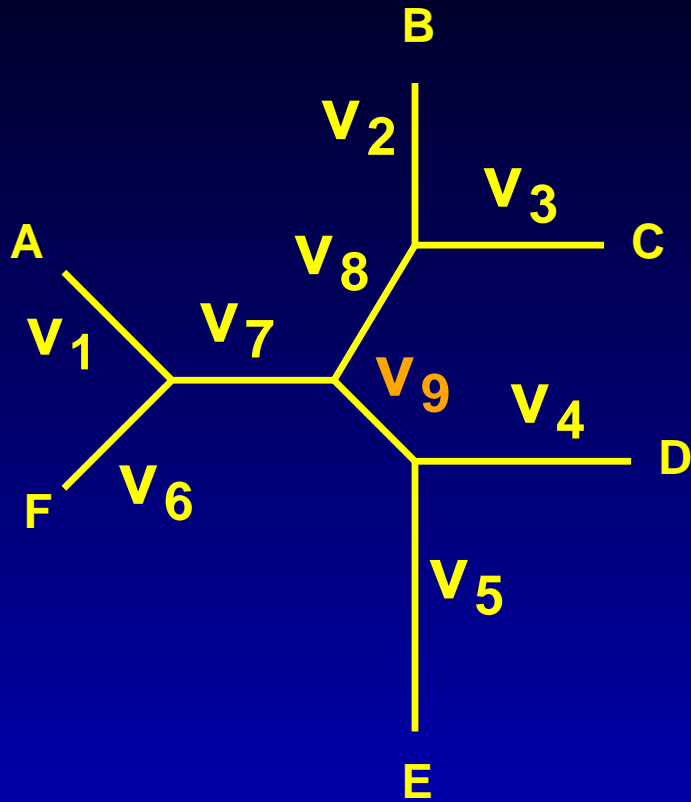
an example: three species with a clock



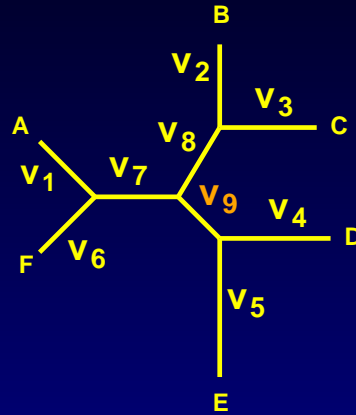
when we consider all three possible topologies, the space looks like:



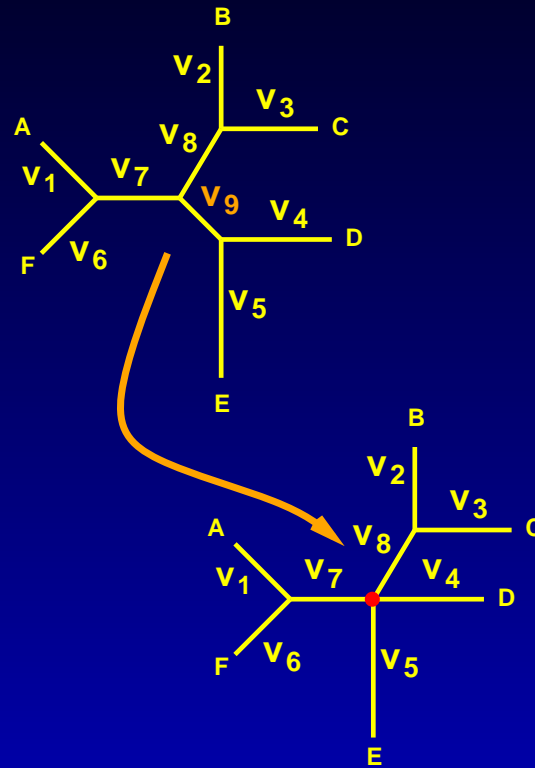
# Through the looking glass



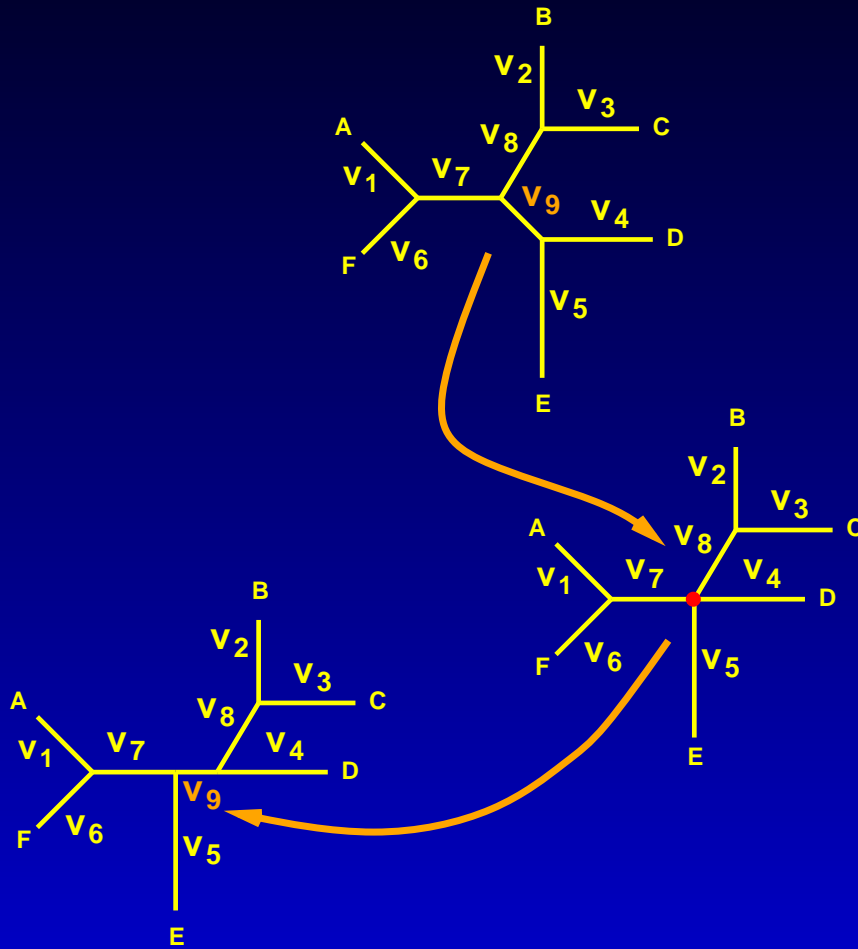
# Nearest-neighbor interchanges



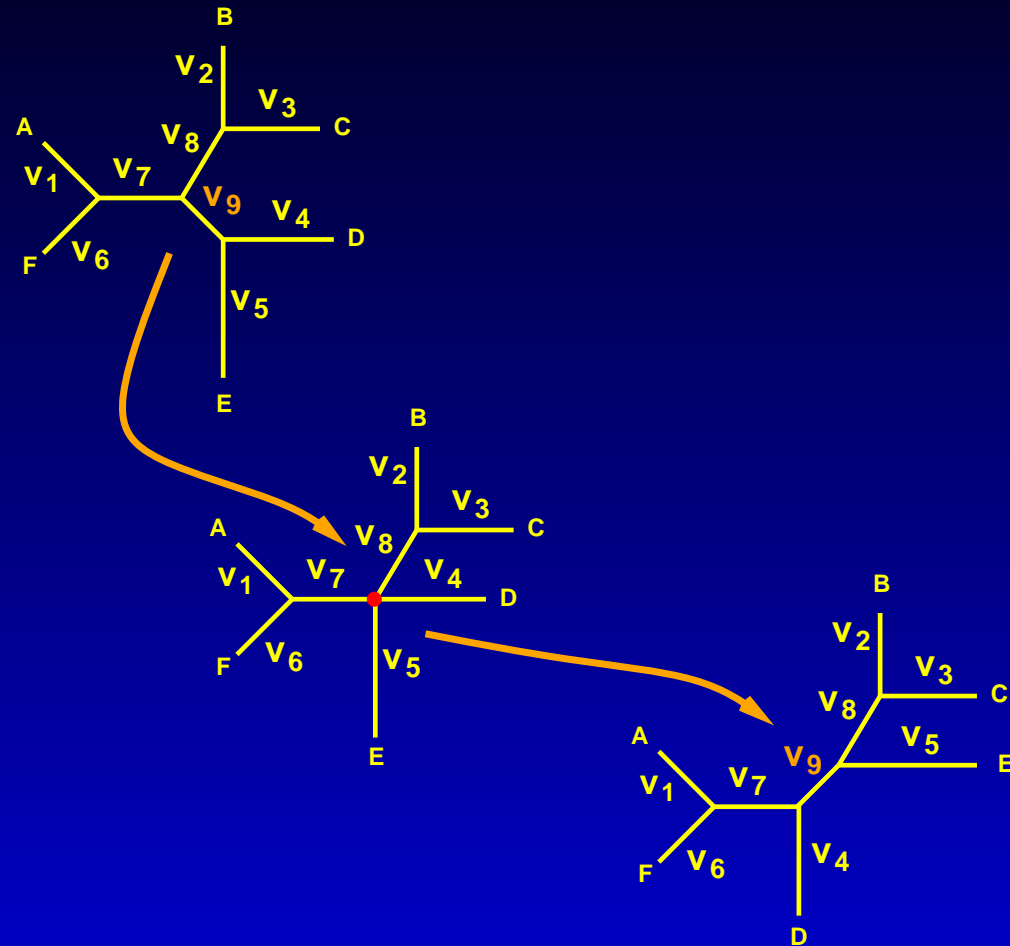
# Nearest-neighbor interchanges



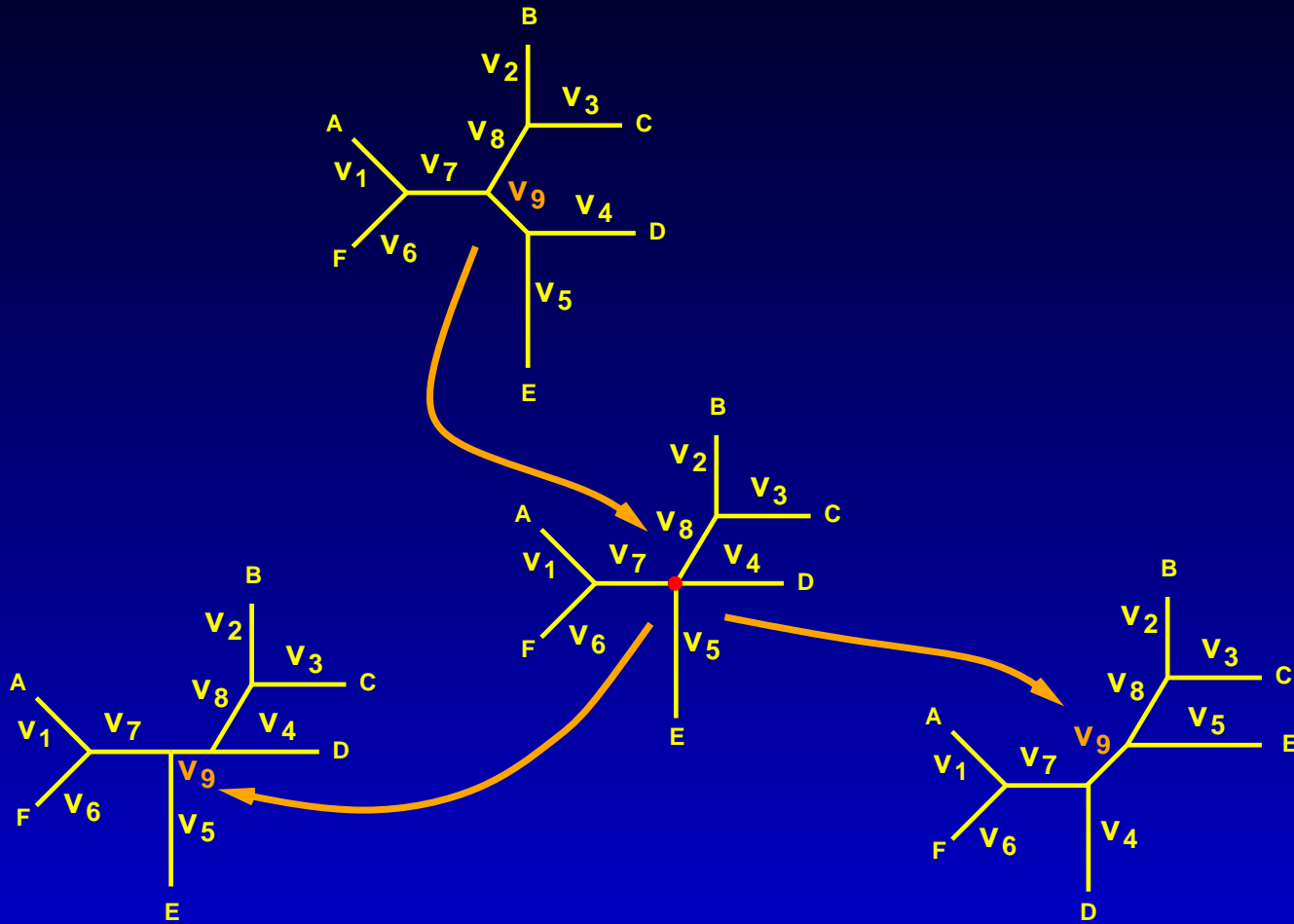
# Nearest-neighbor interchanges



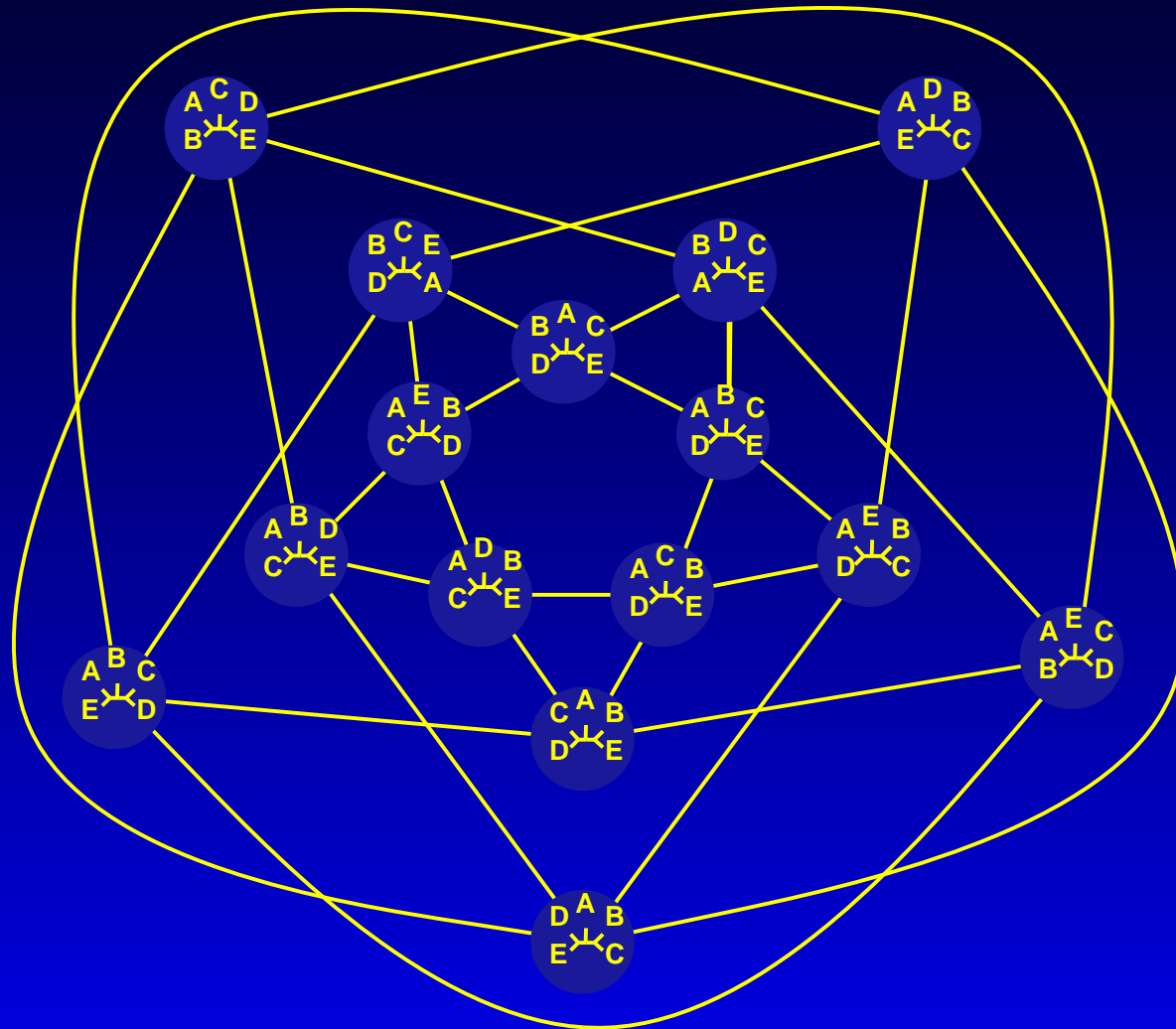
# Nearest-neighbor interchanges



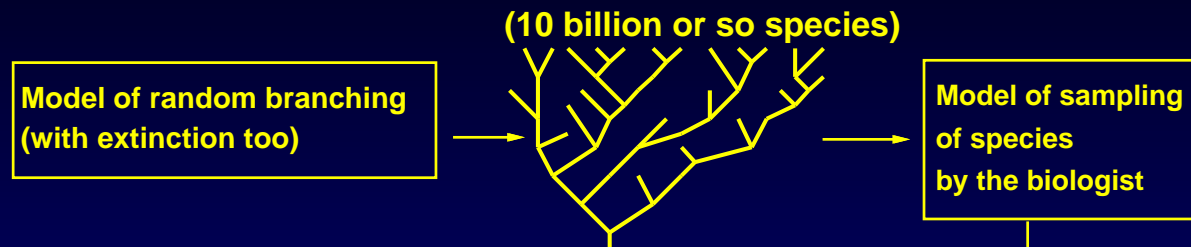
# Nearest-neighbor interchanges



# The Schoenberg graph



# Where does the true tree come from?



## Models, Trees, and Data



model of evolution of characters along the tree

Aligned DNA (or protein) sequences

other sequences too

A	ggtca	aactt	gagtc	aacat	...
B	ggtcc	aagtt	gagtc	gacat	...
C	ggtca	aactt	gagtc	aacat	...
D	ggtcc	aactt	gactc	aatat	...
E	gctca	aagtt	gactc	ttcat	...

## Birth-and-death processes as priors?

- If they have parameters  $(\lambda, \mu)$ , what are the priors on their values?

## Birth-and-death processes as priors?

- If they have parameters  $(\lambda, \mu)$ , what are the priors on their values?
- How do we determine how long the BD process runs?

## Birth-and-death processes as priors?

- If they have parameters  $(\lambda, \mu)$ , what are the priors on their values?
- How do we determine how long the BD process runs?
- If the systematist selects species, how do we characterize that process?

## Birth-and-death processes as priors?

- If they have parameters  $(\lambda, \mu)$ , what are the priors on their values?
- How do we determine how long the BD process runs?
- If the systematist selects species, how do we characterize that process?
- How do we sample  $N$  species out of a huge tree-of-life? It's not so easy

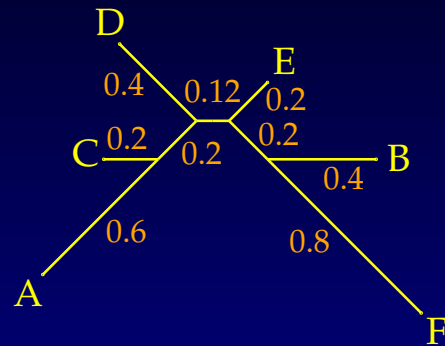
## What has mathematics done for us?

**The analogy between recursion and trees** Discovered many times, basis for dynamic programming algorithms for computing likelihoods and other similar quantities.

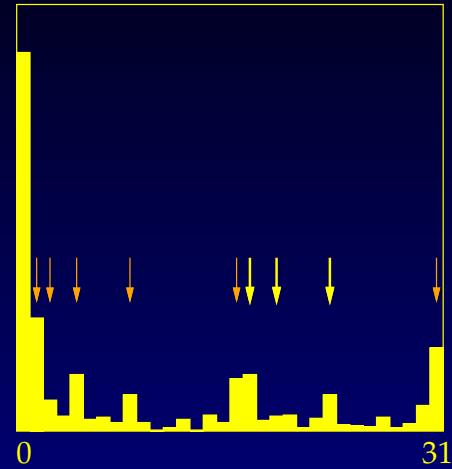
**Hadamard transform methods** Hendy and Penny (1989) discovered that a Hadamard transform applied to a list of frequencies of occurrences of different patterns of character values (at the tips of the tree) yield support for different splits in the tree.

**Invariants** Cavender and also Lake (both in 1987) initiated the exploration, not of “tree space” but of polynomial relationships between expected frequencies of occurrence of patterns of character values. Related to the practice of “algebraic statistics”.

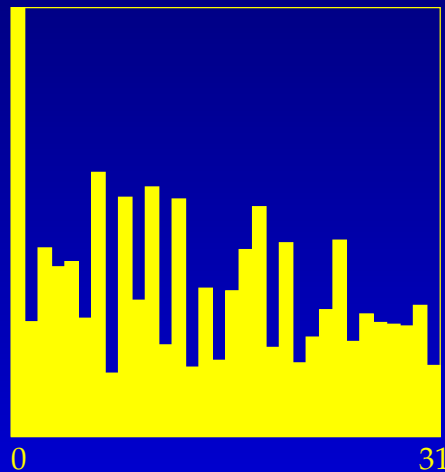
# The Hadamard Conjugation



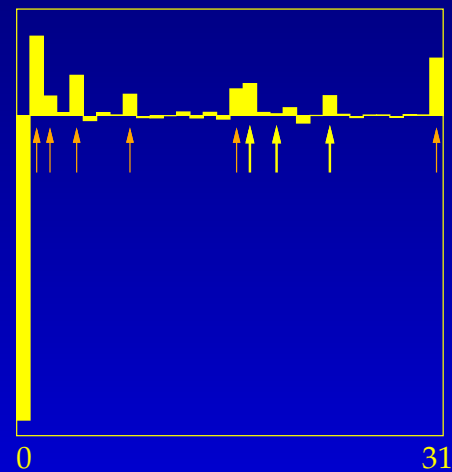
Tree



Partitions



Hadamard Transform



Conjugate Spectrum

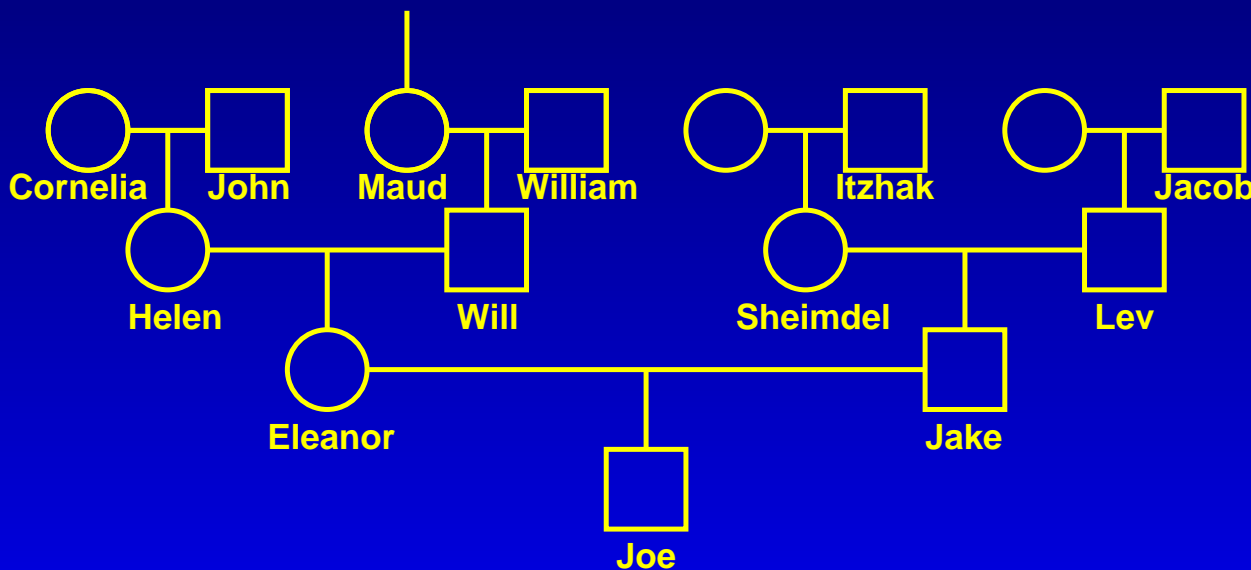
# My ancestor?

**Charles the Great  
(Charlemagne)**



**born  
747**

**about 44 more generations**



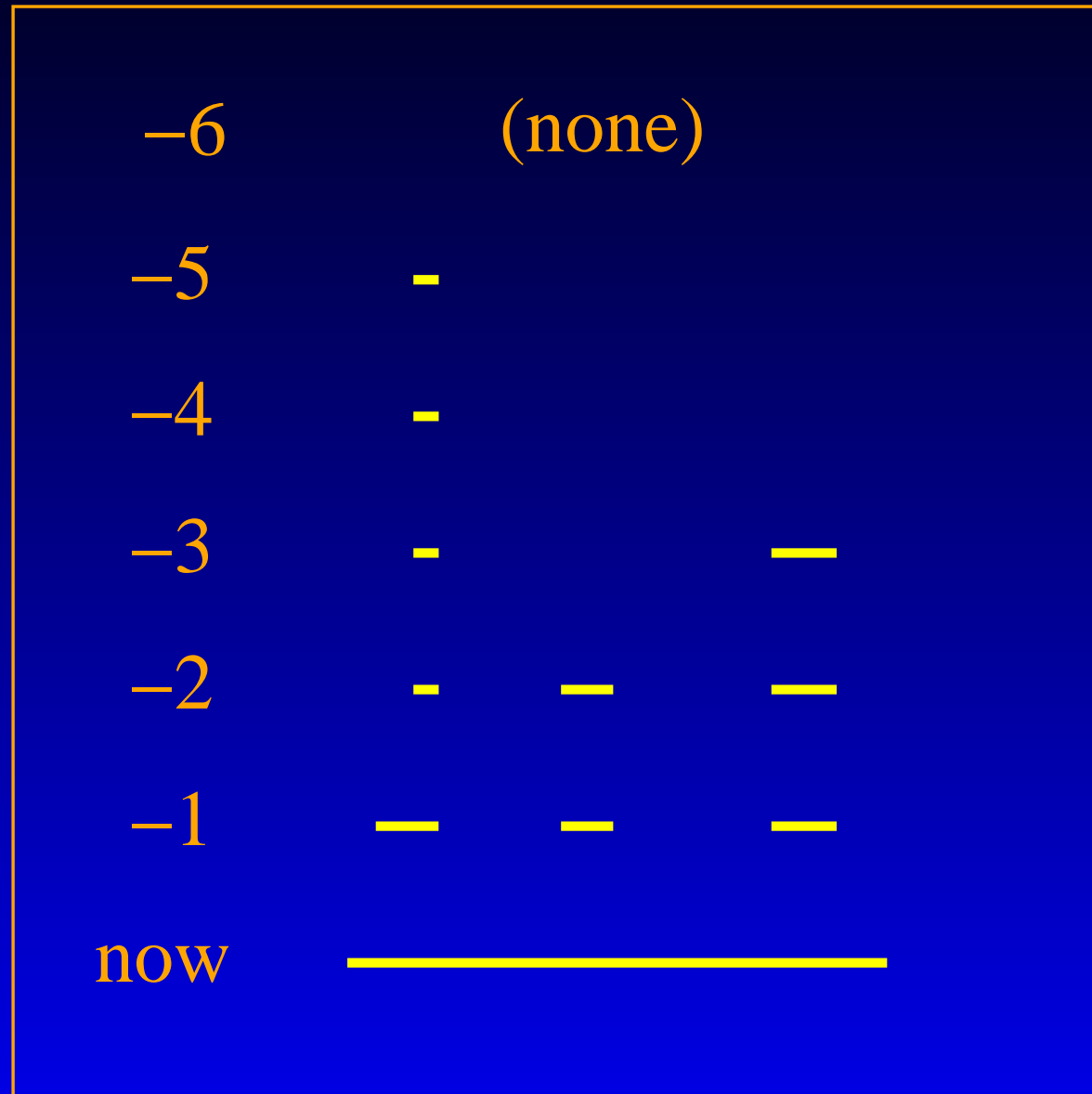
**1850s**

**1880s**

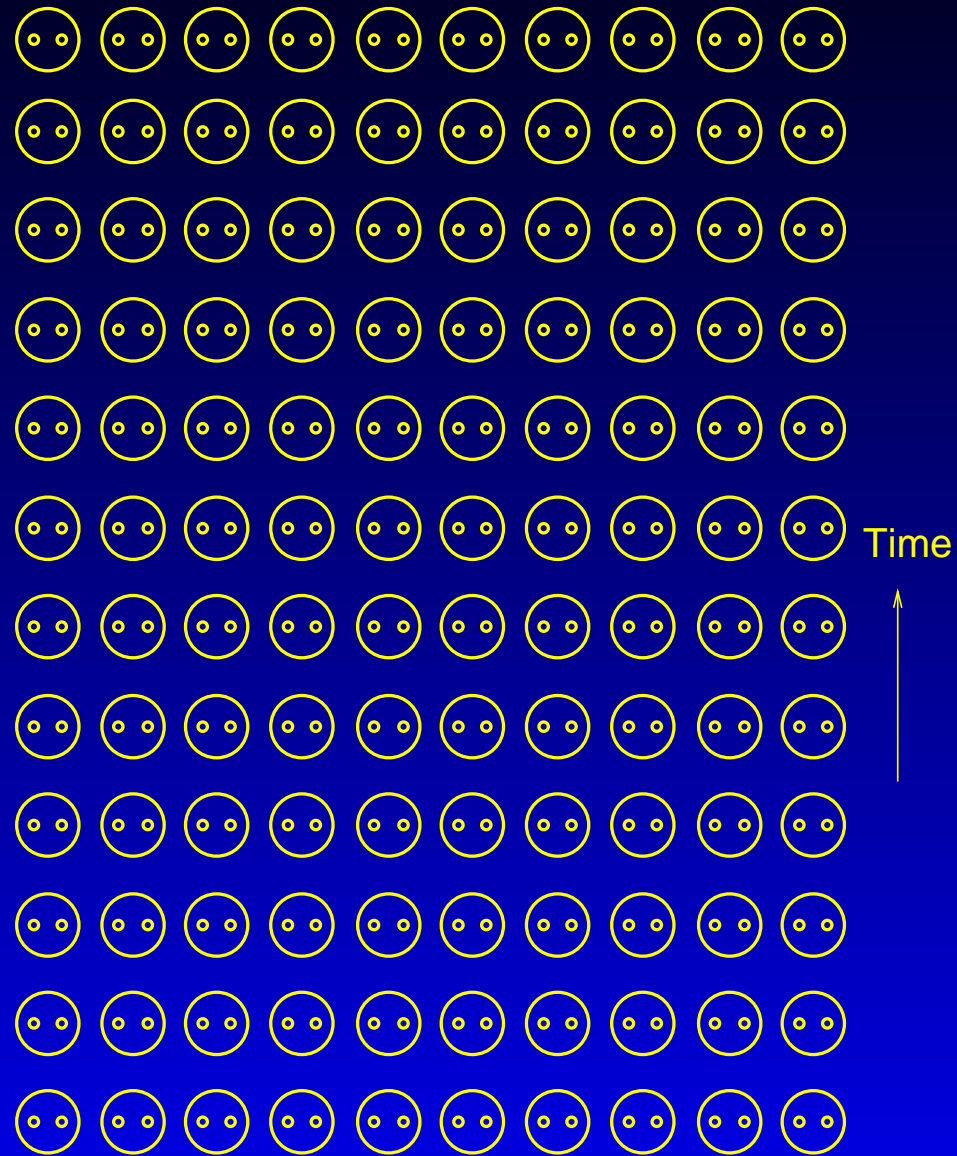
**1910s**

**1942**

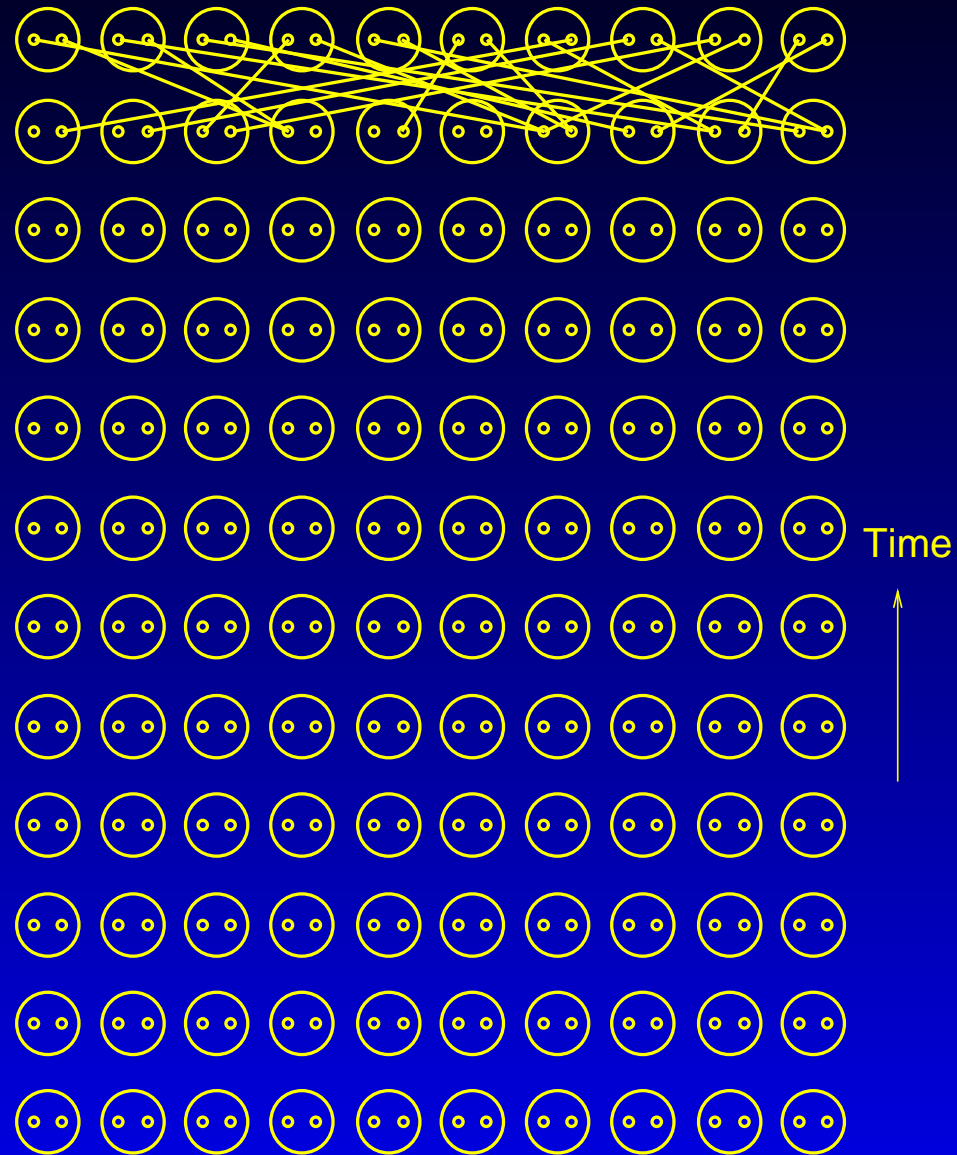
# Chromosome 1, back up one lineage



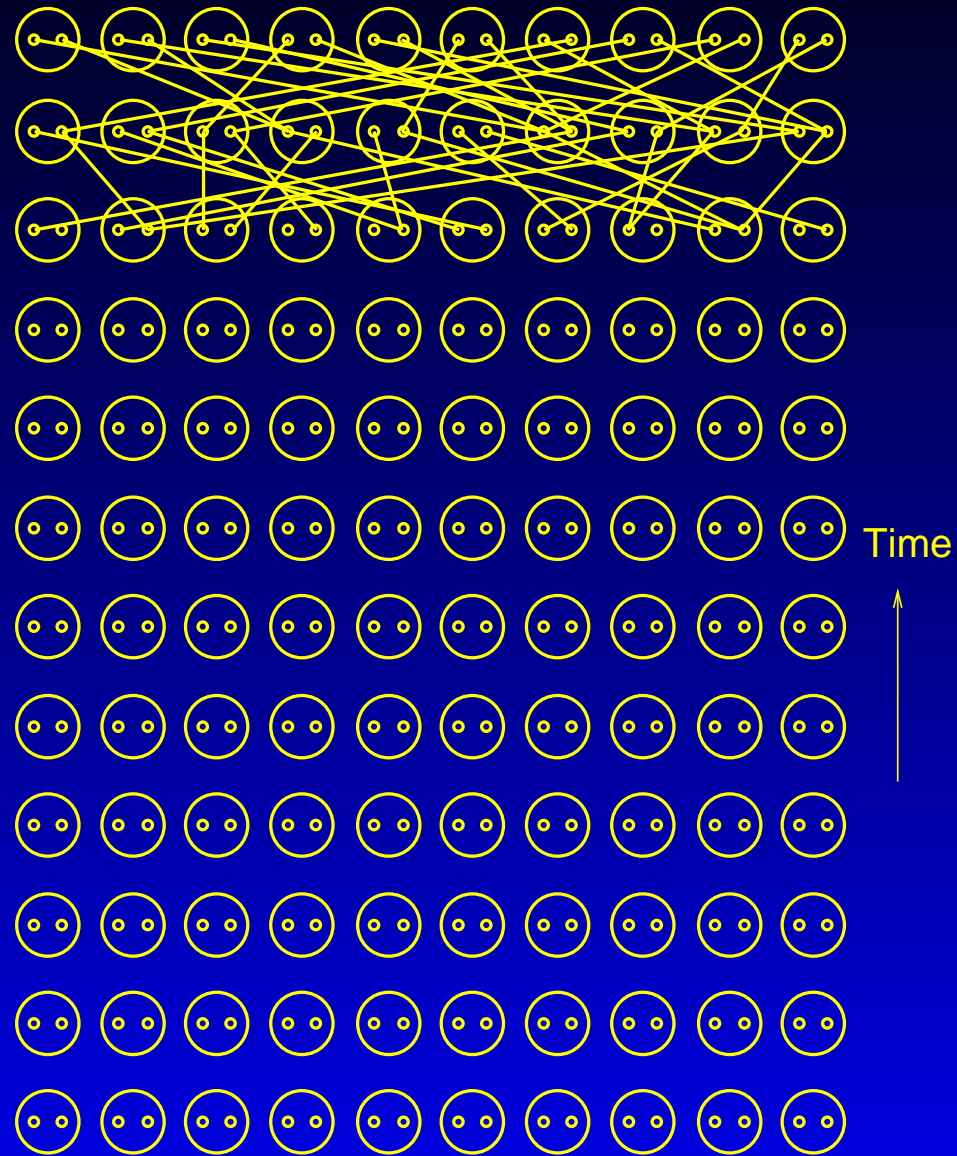
# Coalescent genealogy for one gene



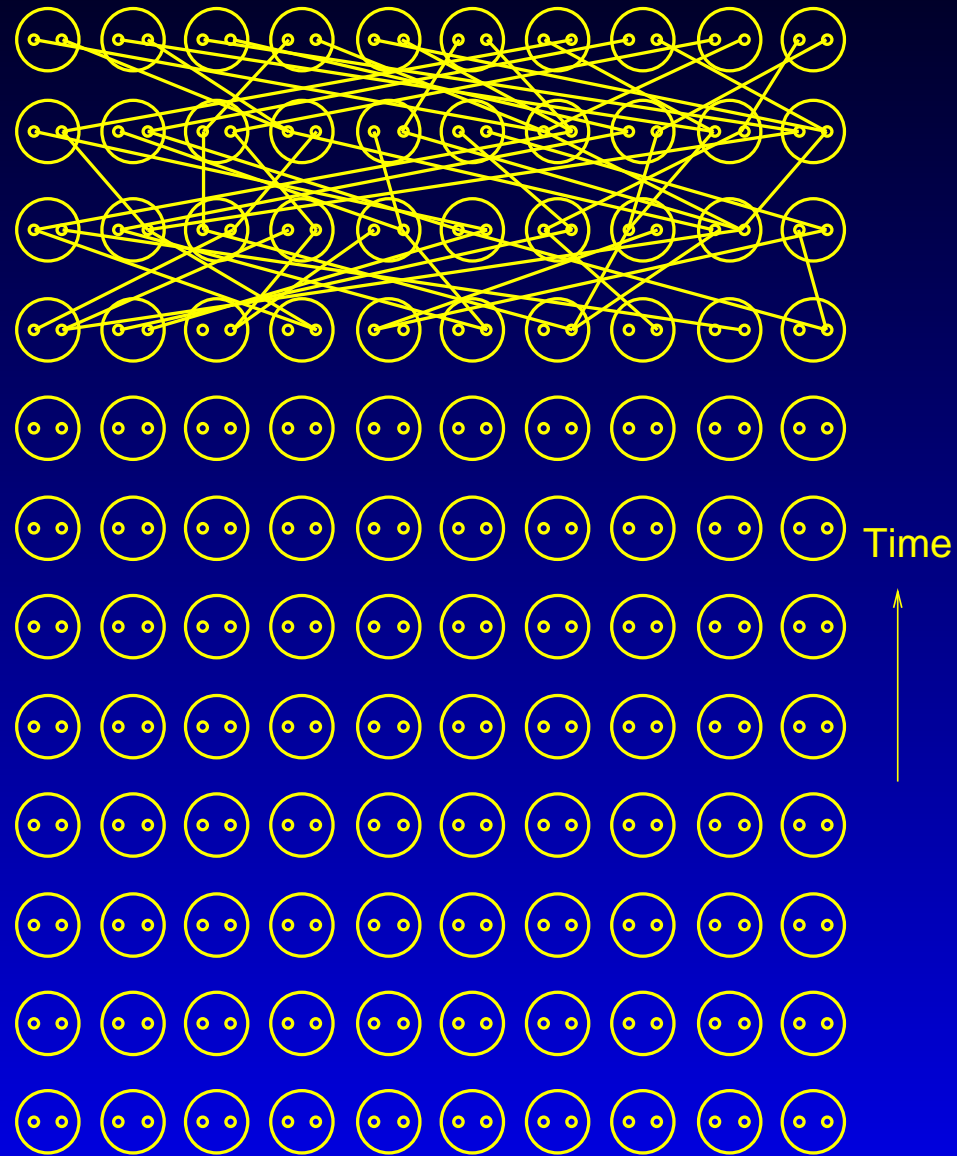
# Coalescent genealogy for one gene



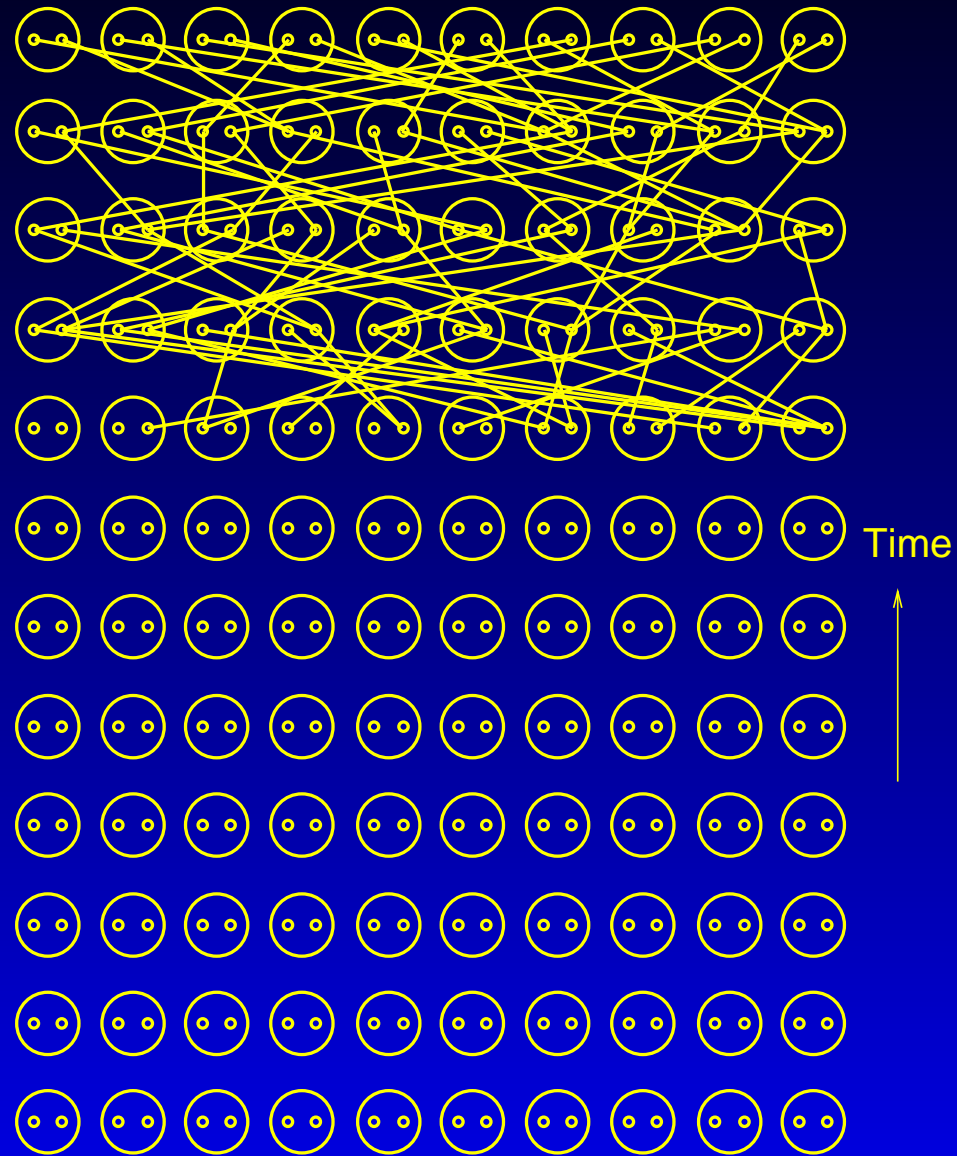
# Coalescent genealogy for one gene



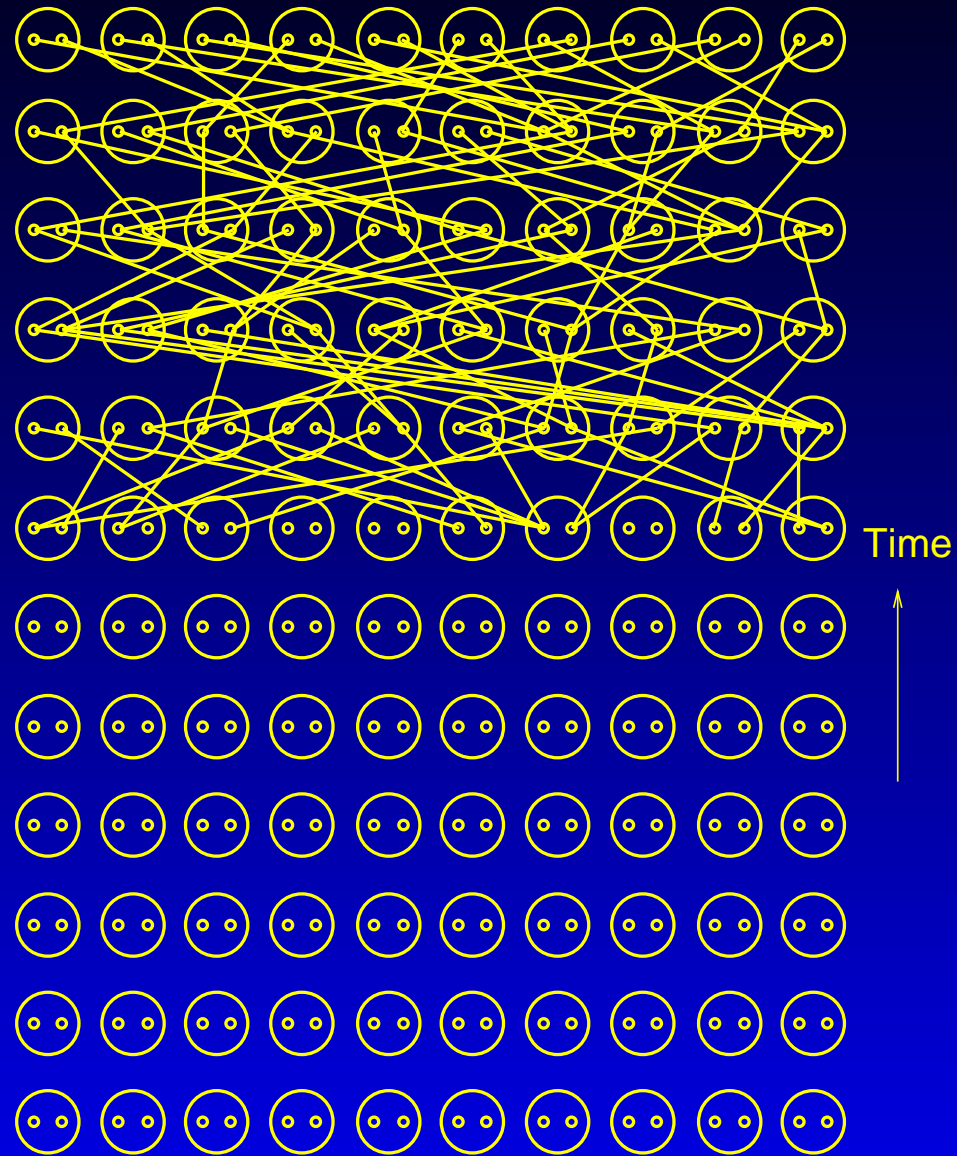
# Coalescent genealogy for one gene



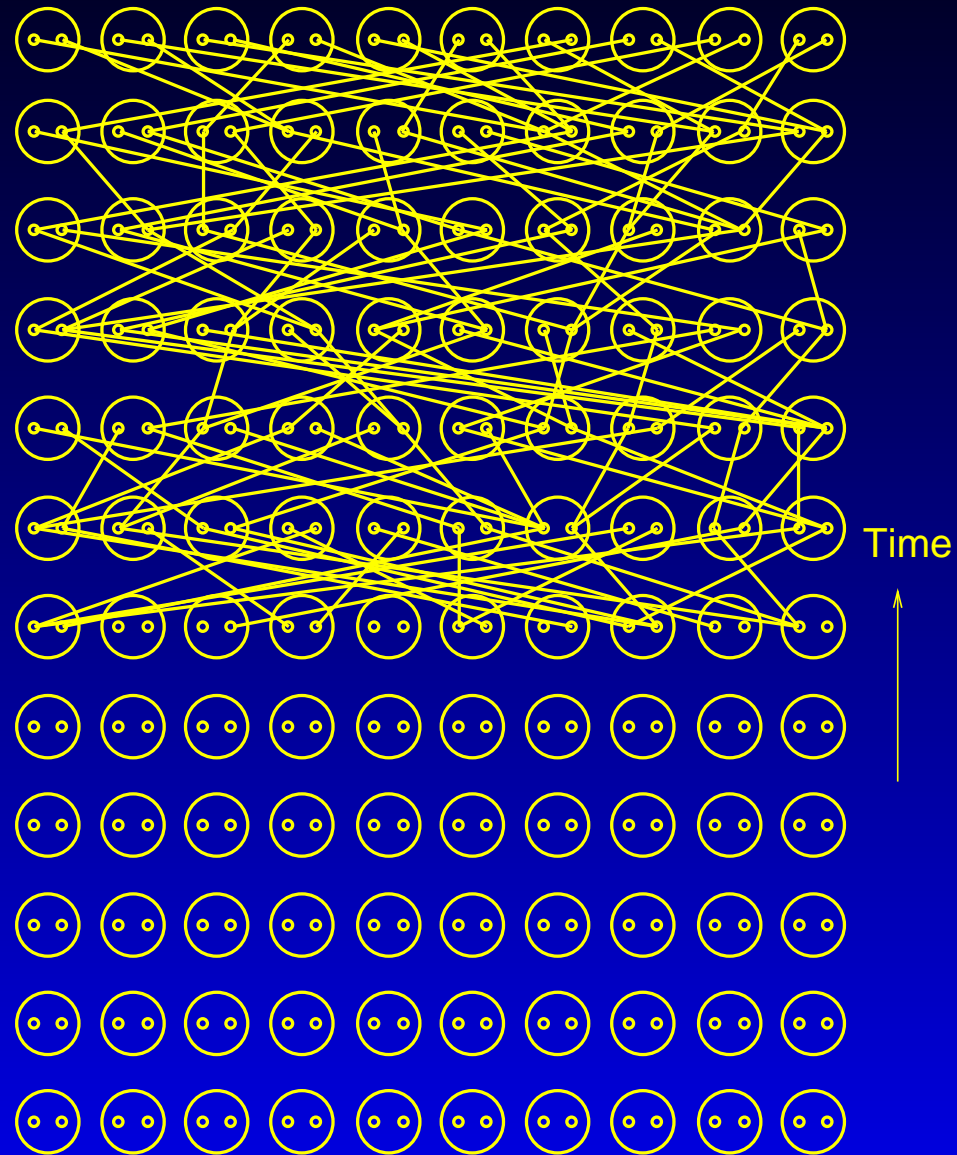
# Coalescent genealogy for one gene



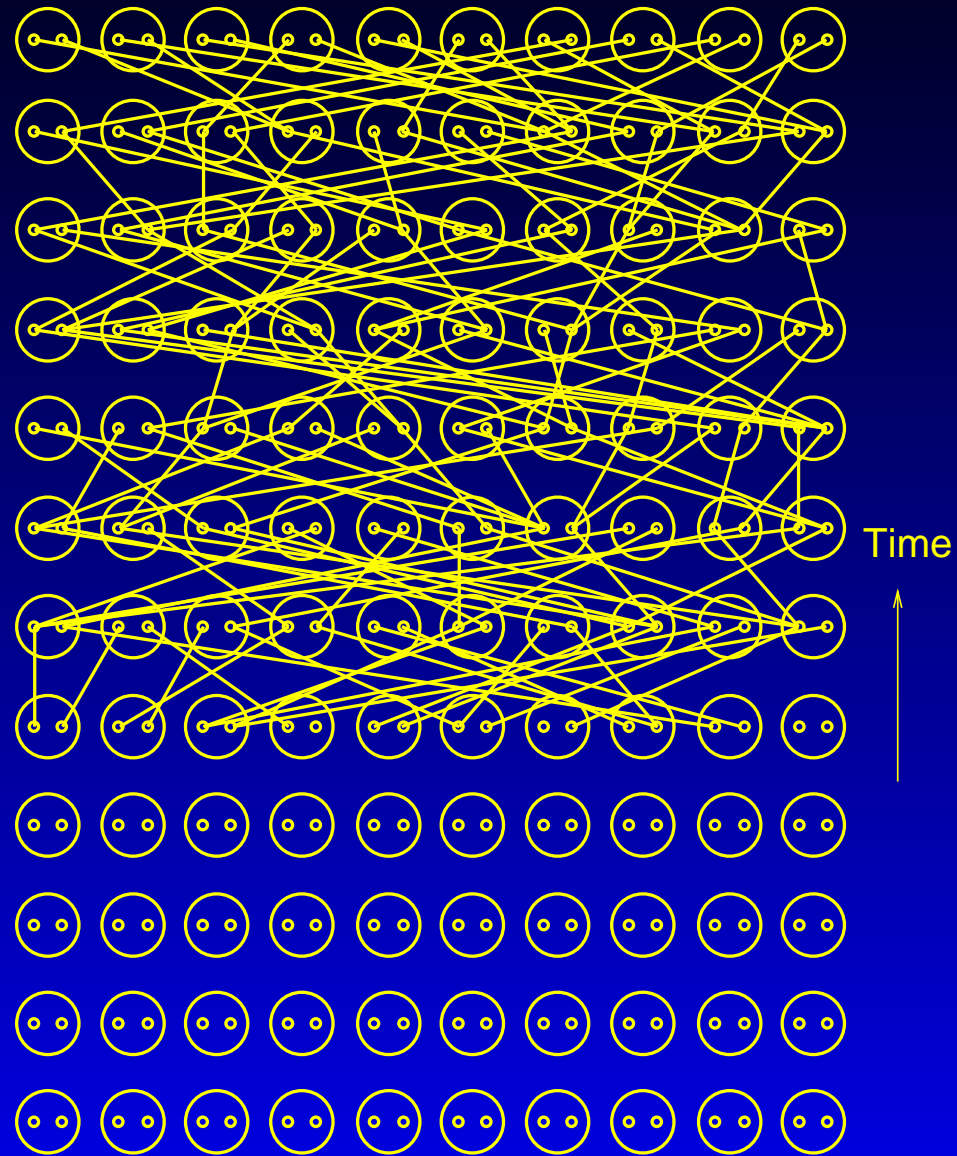
# Coalescent genealogy for one gene



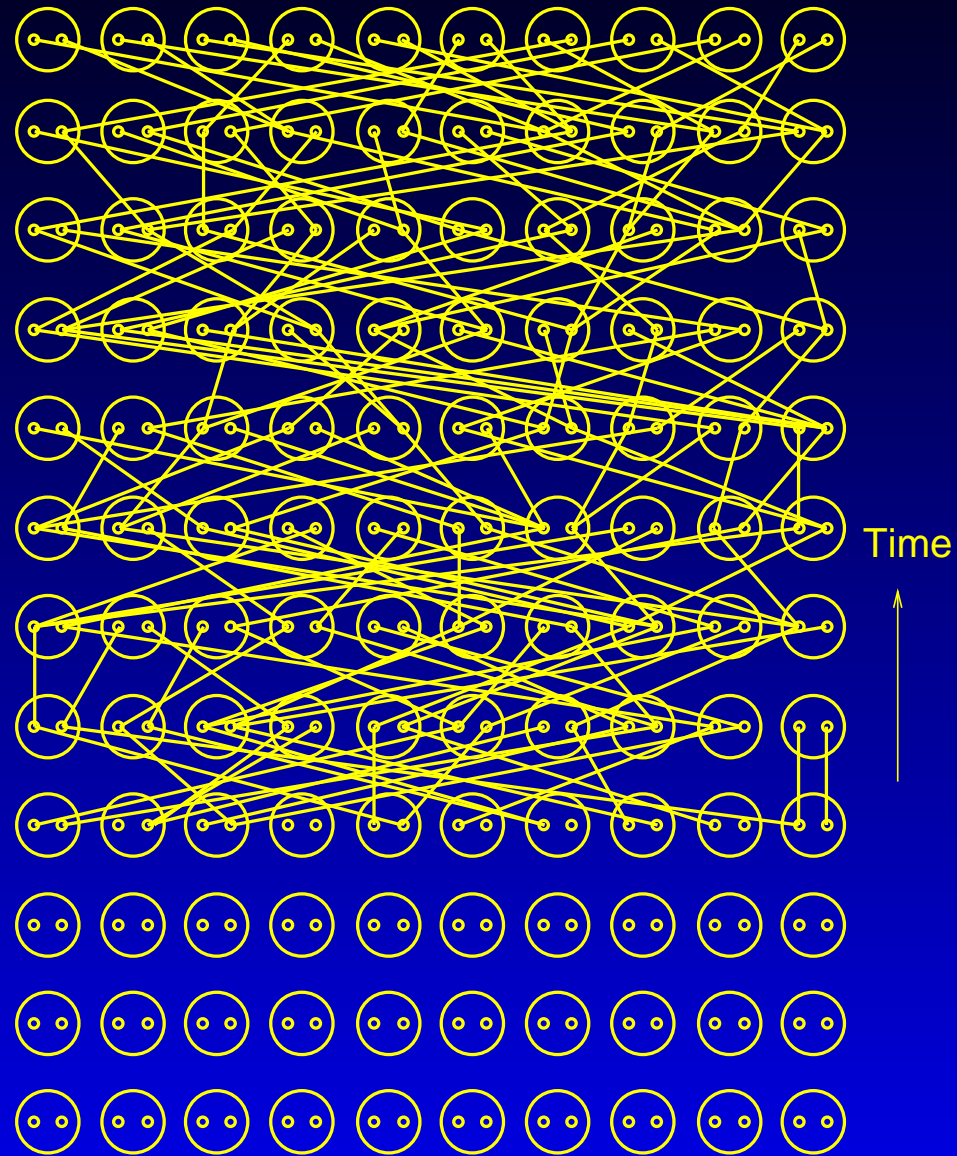
# Coalescent genealogy for one gene



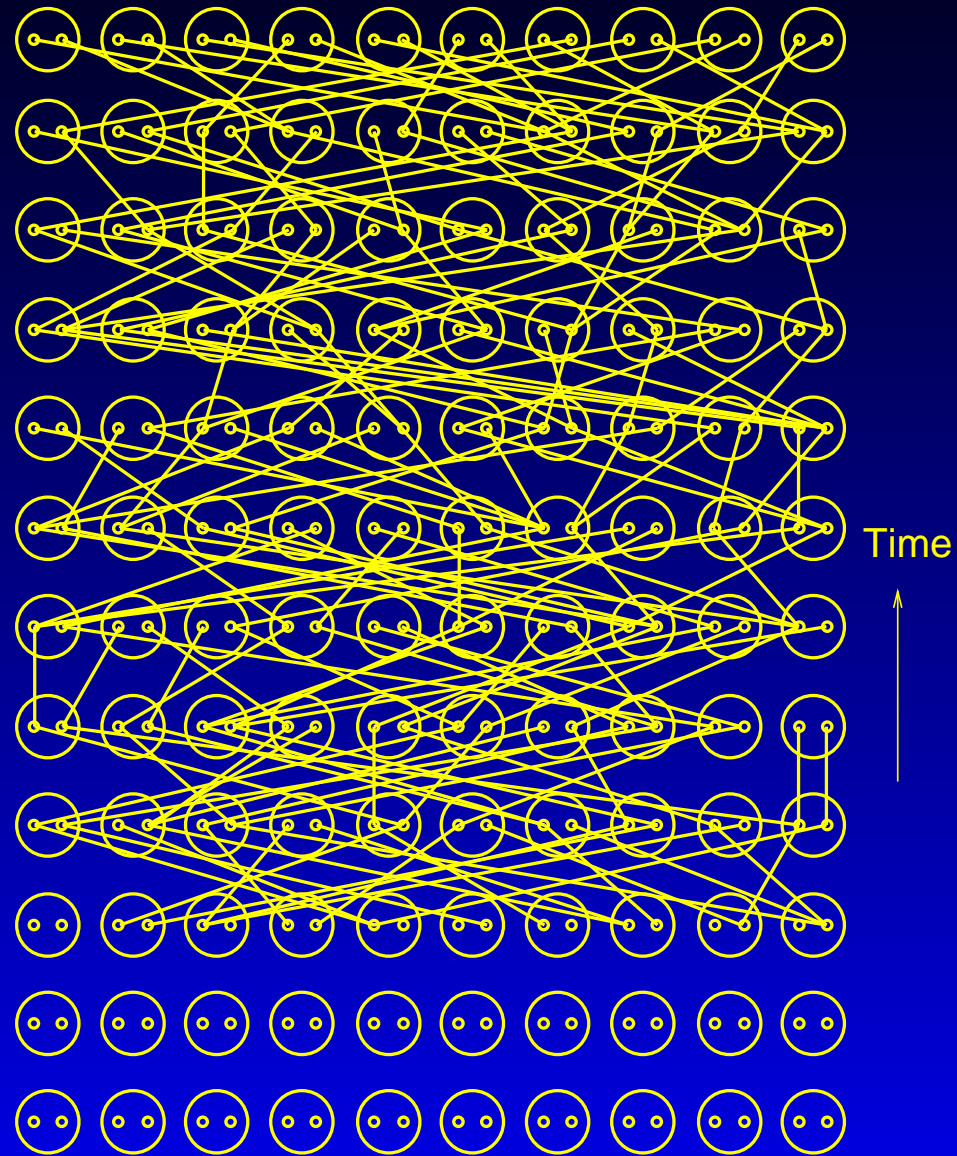
# Coalescent genealogy for one gene



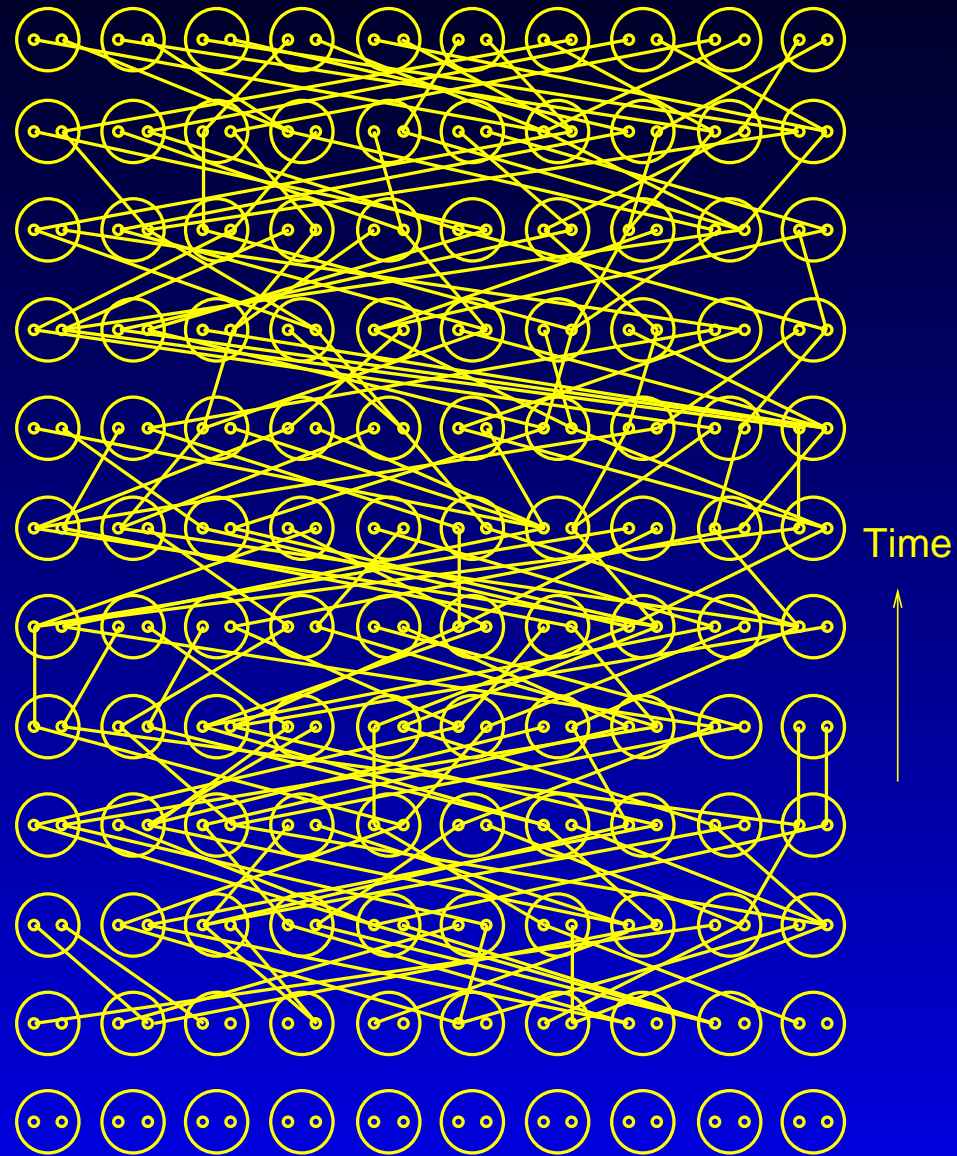
# Coalescent genealogy for one gene



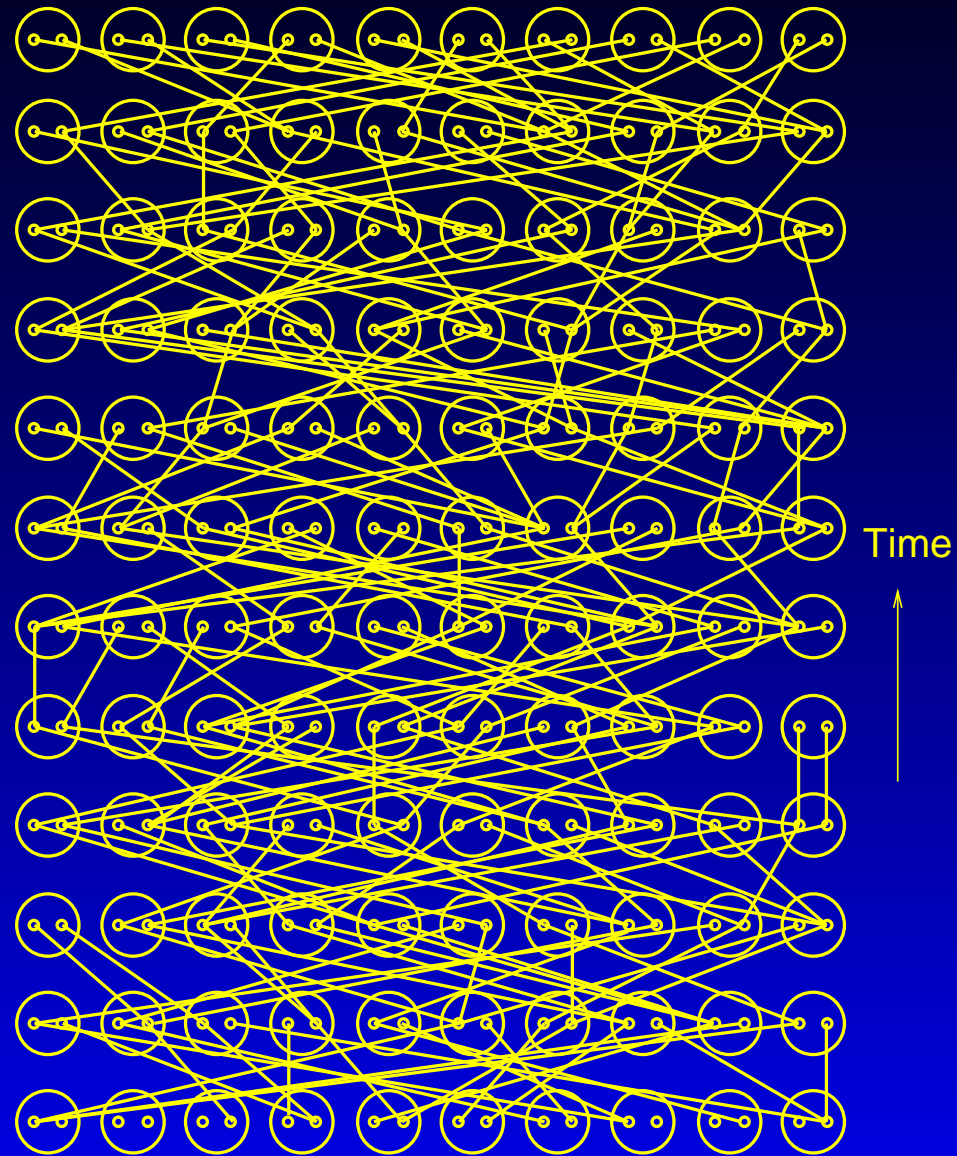
# Coalescent genealogy for one gene



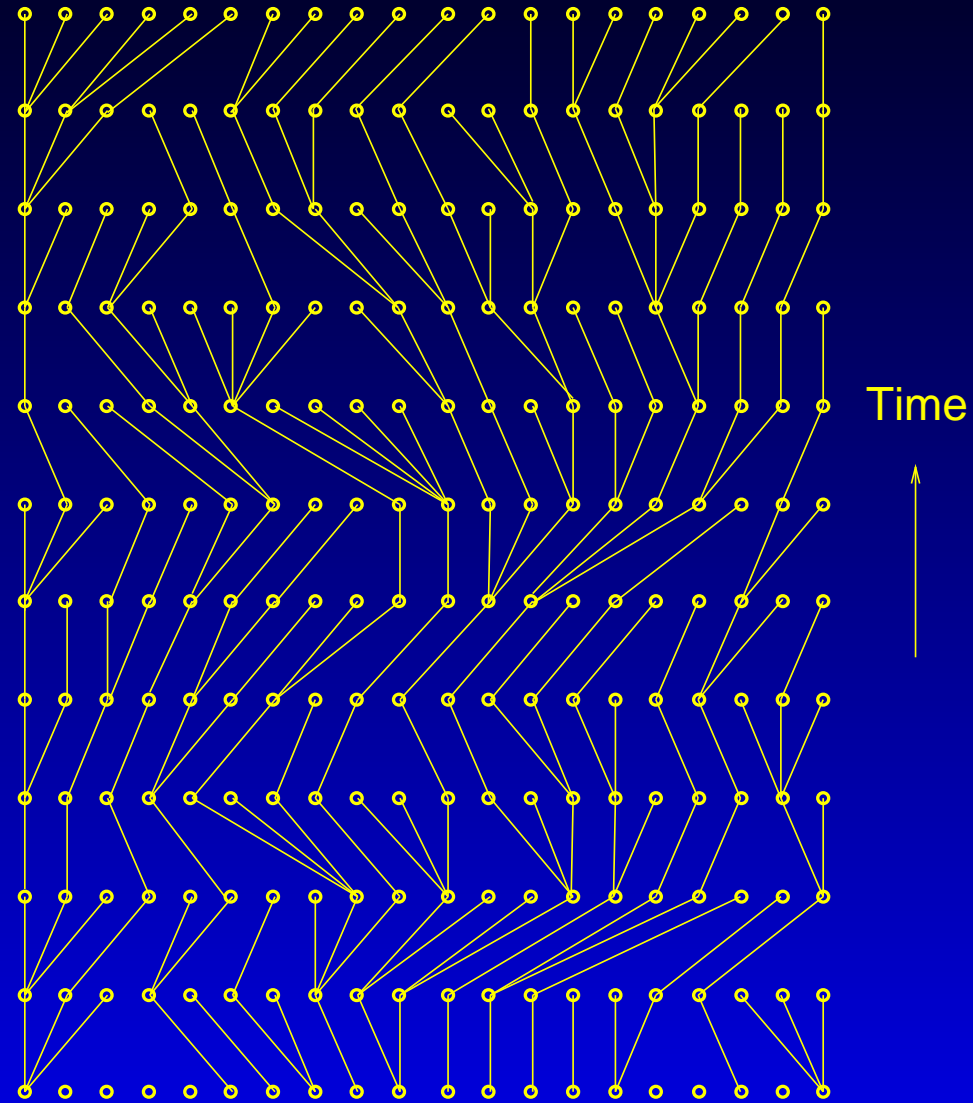
# Coalescent genealogy for one gene



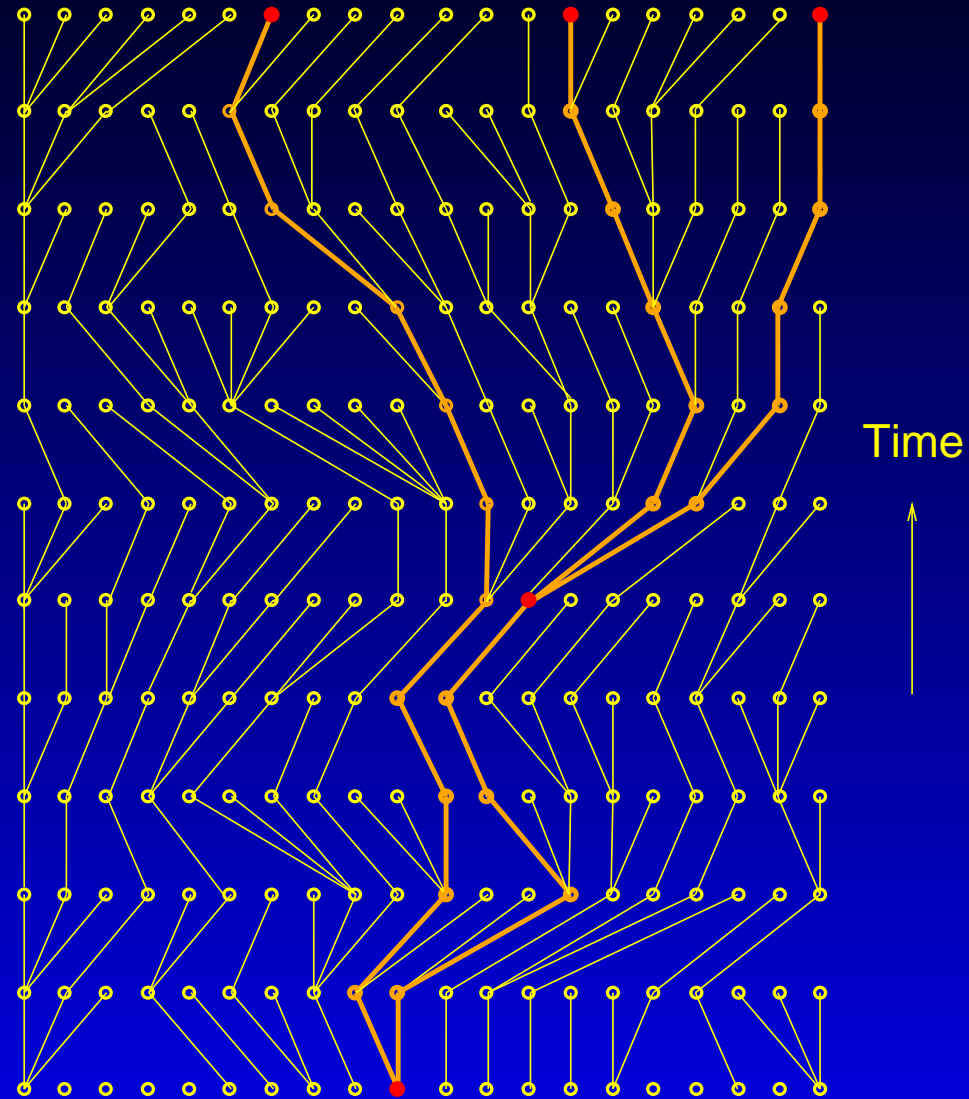
# Coalescent genealogy for one gene



# Untangling the crossed lines ...



# Genealogy of a sample of 3 copies



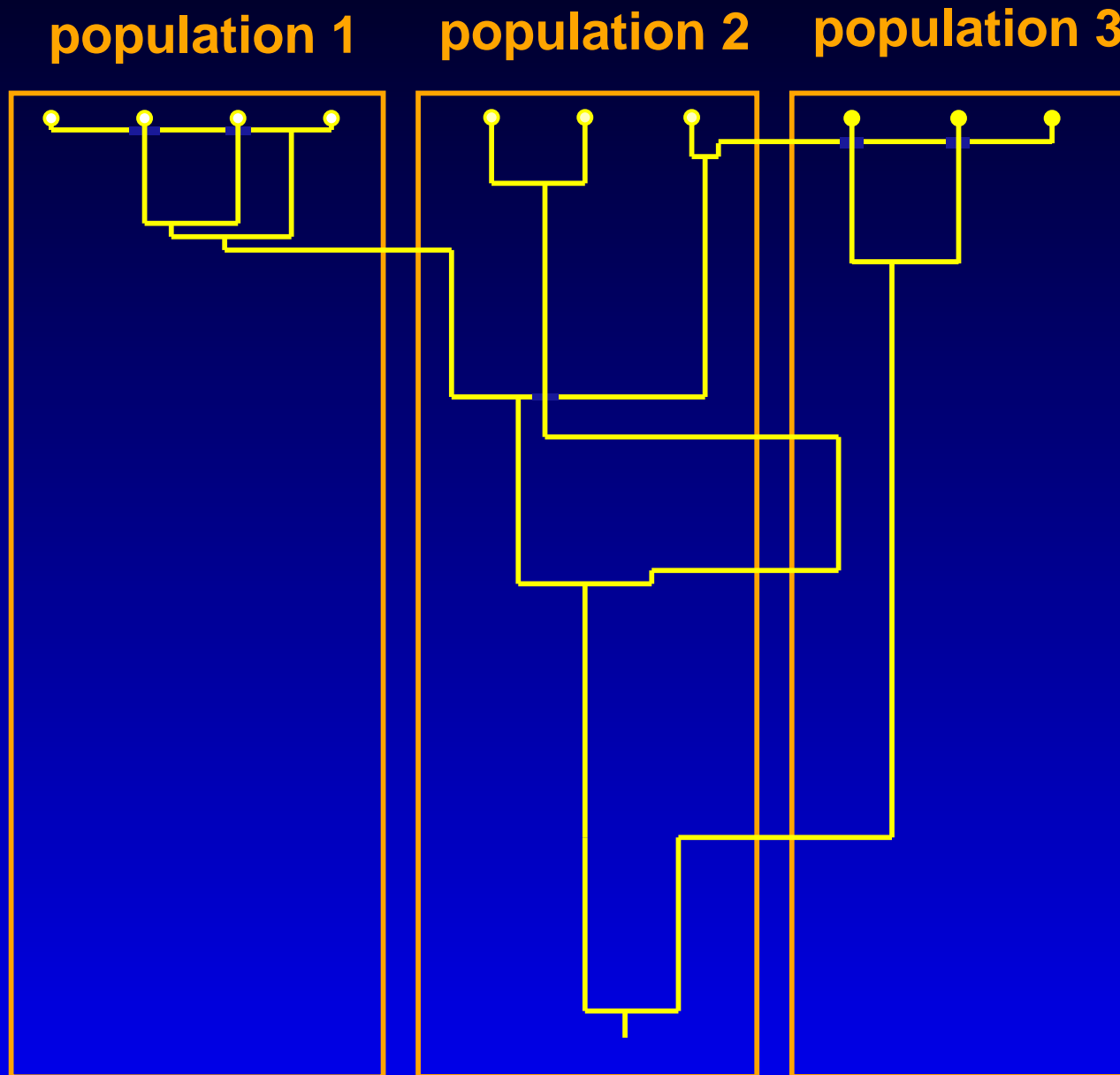
## J. F. C. Kingman's (1982) "coalescent"



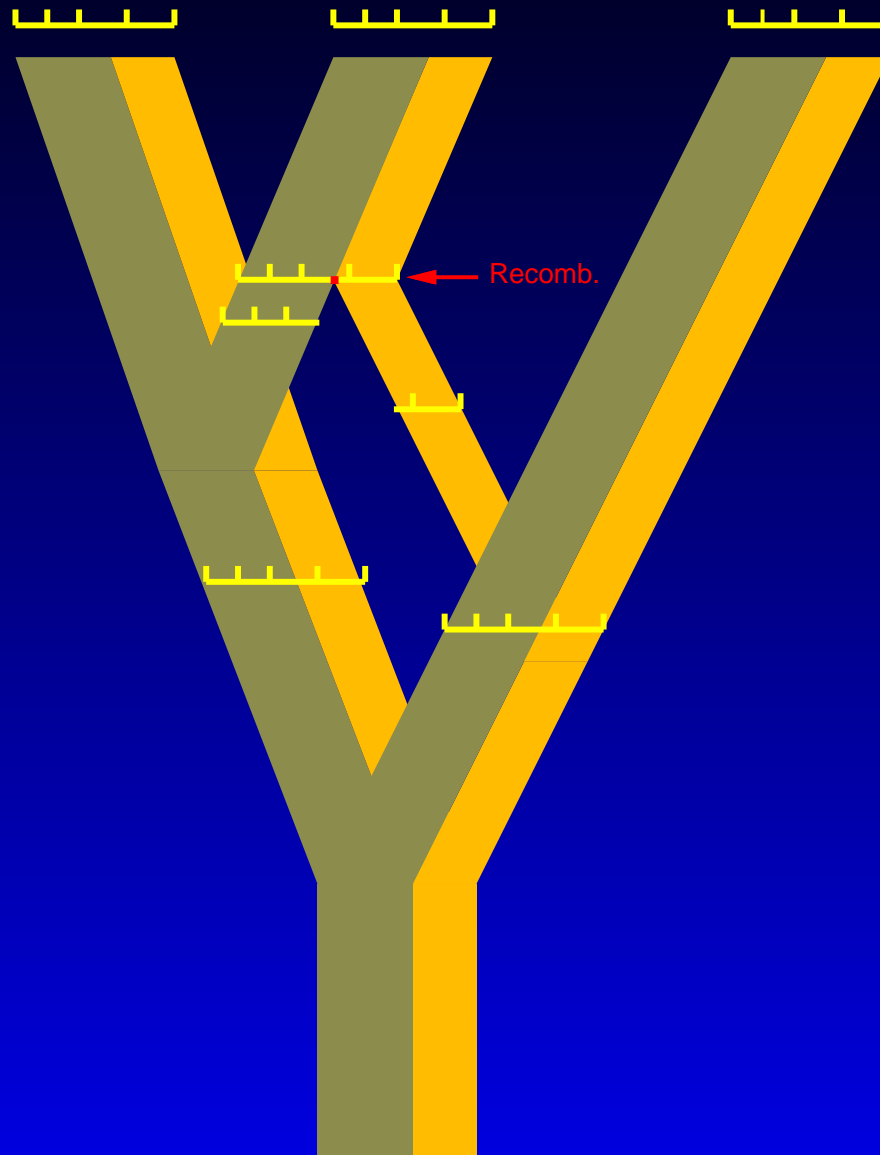
1. start with  $n$  tips
2. go back an amount of time  
drawn from Exponential  $\left(\frac{4N}{n(n-1)}\right)$
3. join a random pair of the  $n$
4.  $n \leftarrow n - 1$
5. if  $n = 1$  stop, else go to step 2.

This excellently approximates the distribution of genealogies which arise from samples from a standard (Wright-Fisher) population genetics model with a population size of  $N$ , provided  $n^2 \ll N$

# A coalescent with migration among populations

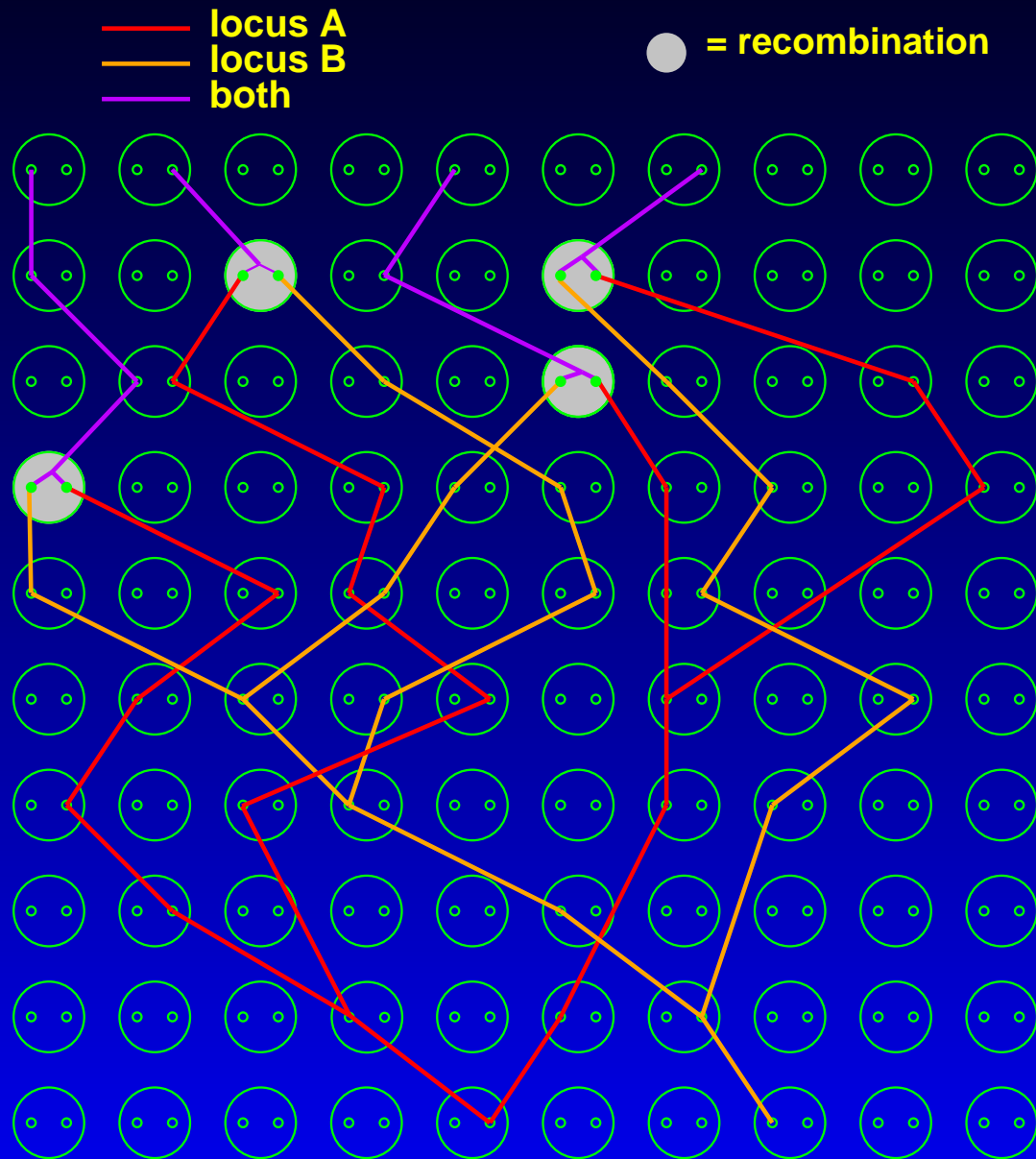


# A coalescent with recombination

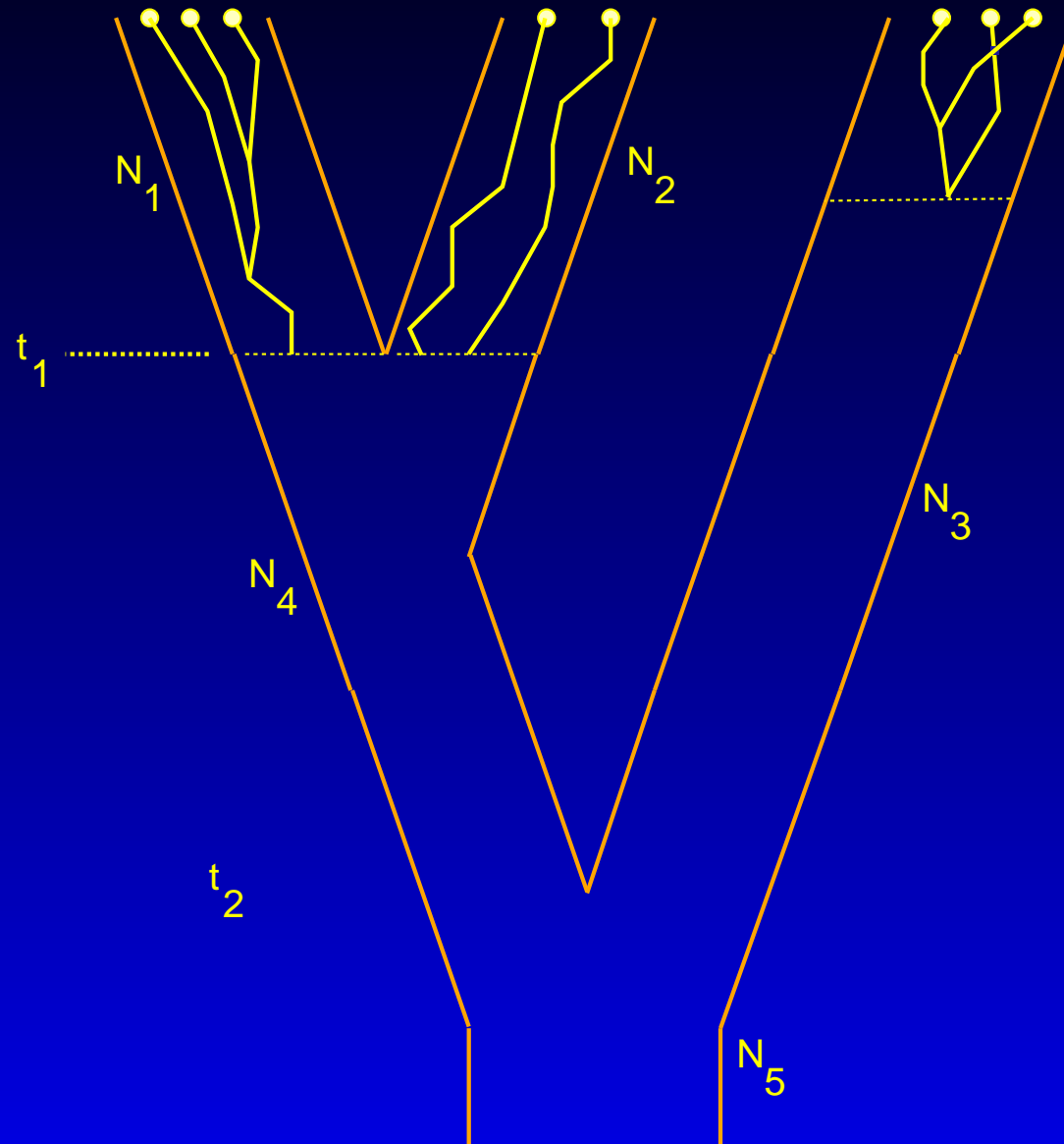


Different markers have slightly different coalescent trees

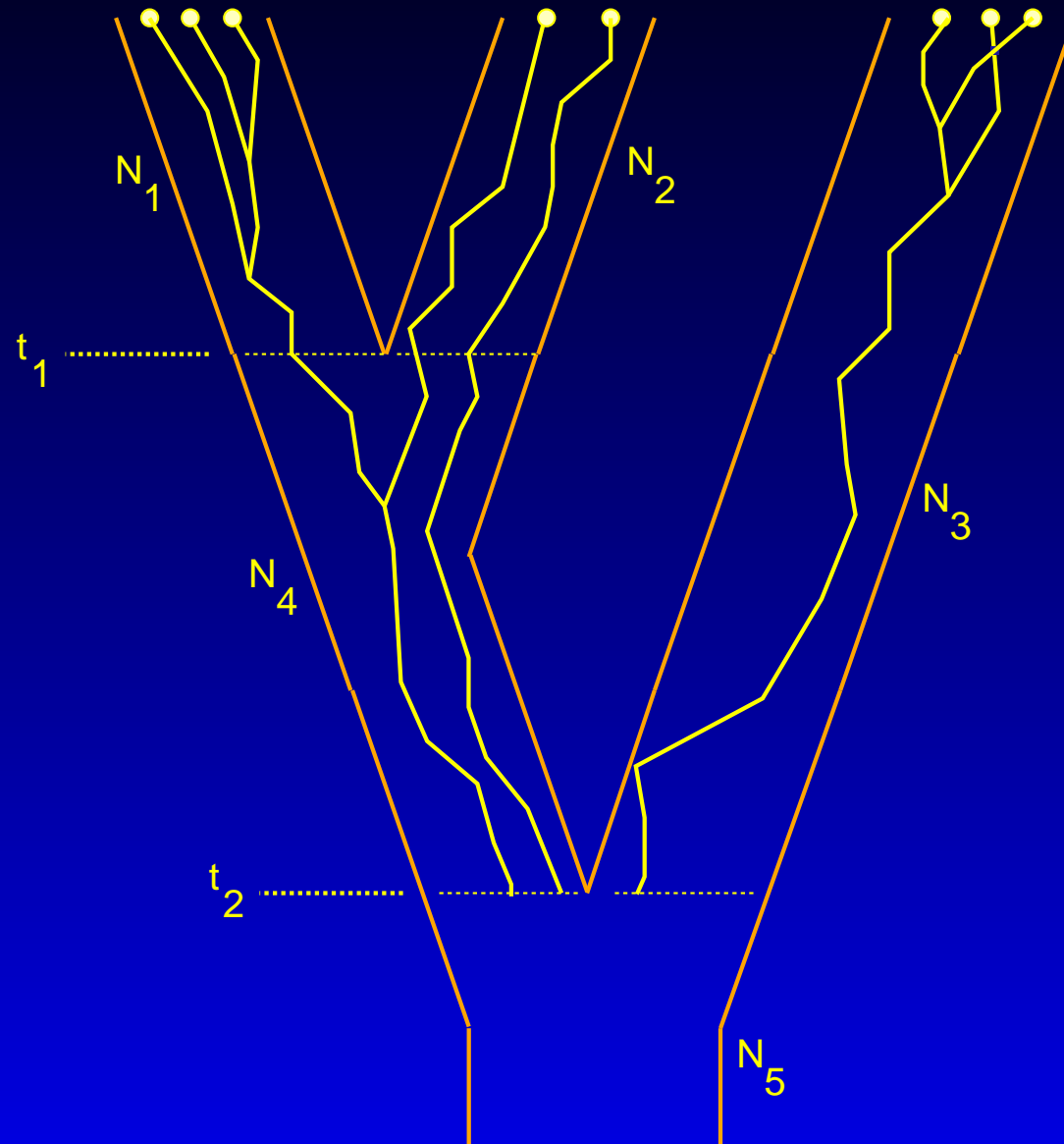
# Coalescents for two genes



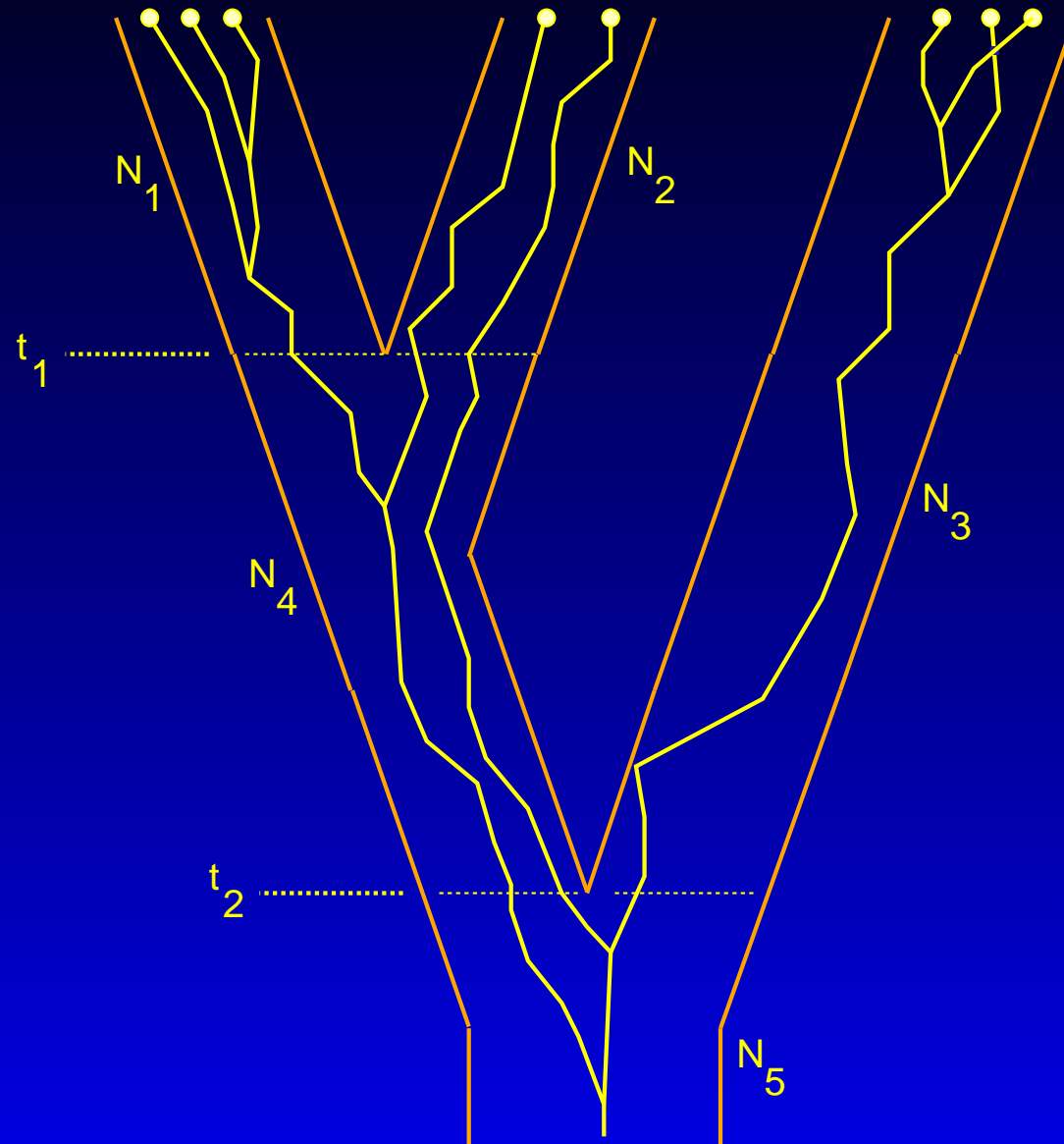
# Species trees and trees of gene copies



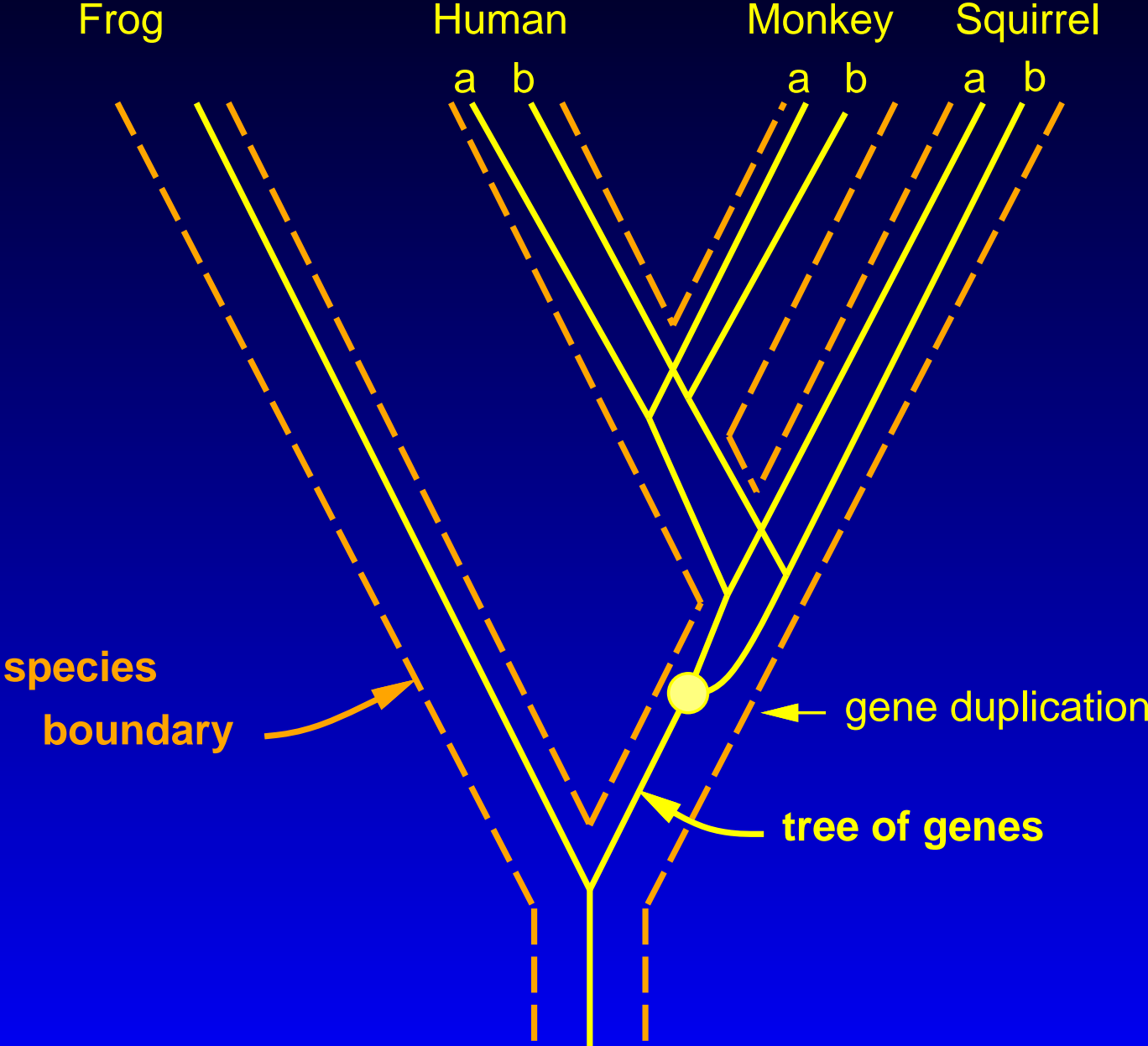
# Species trees and trees of gene copies



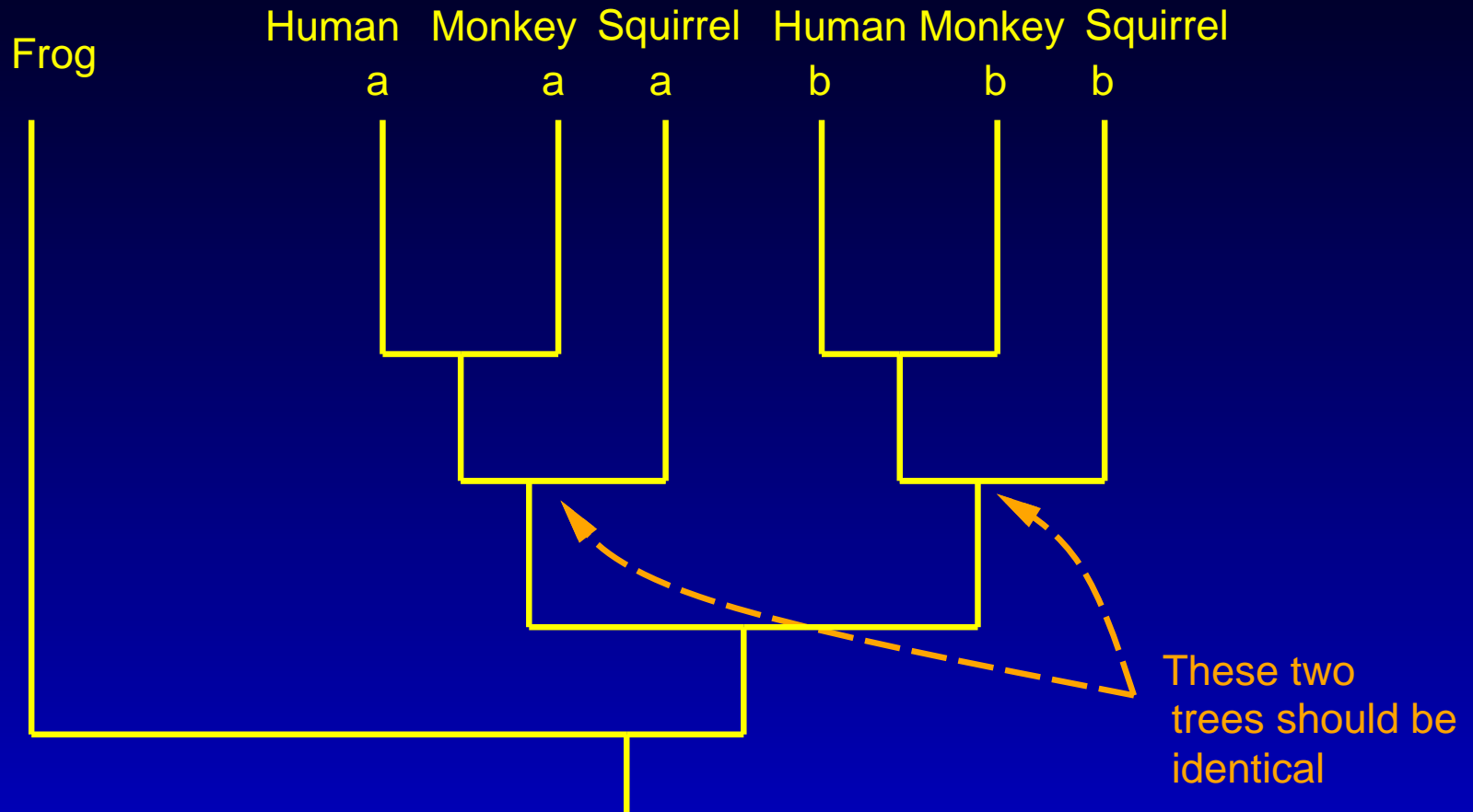
# Species trees and trees of gene copies



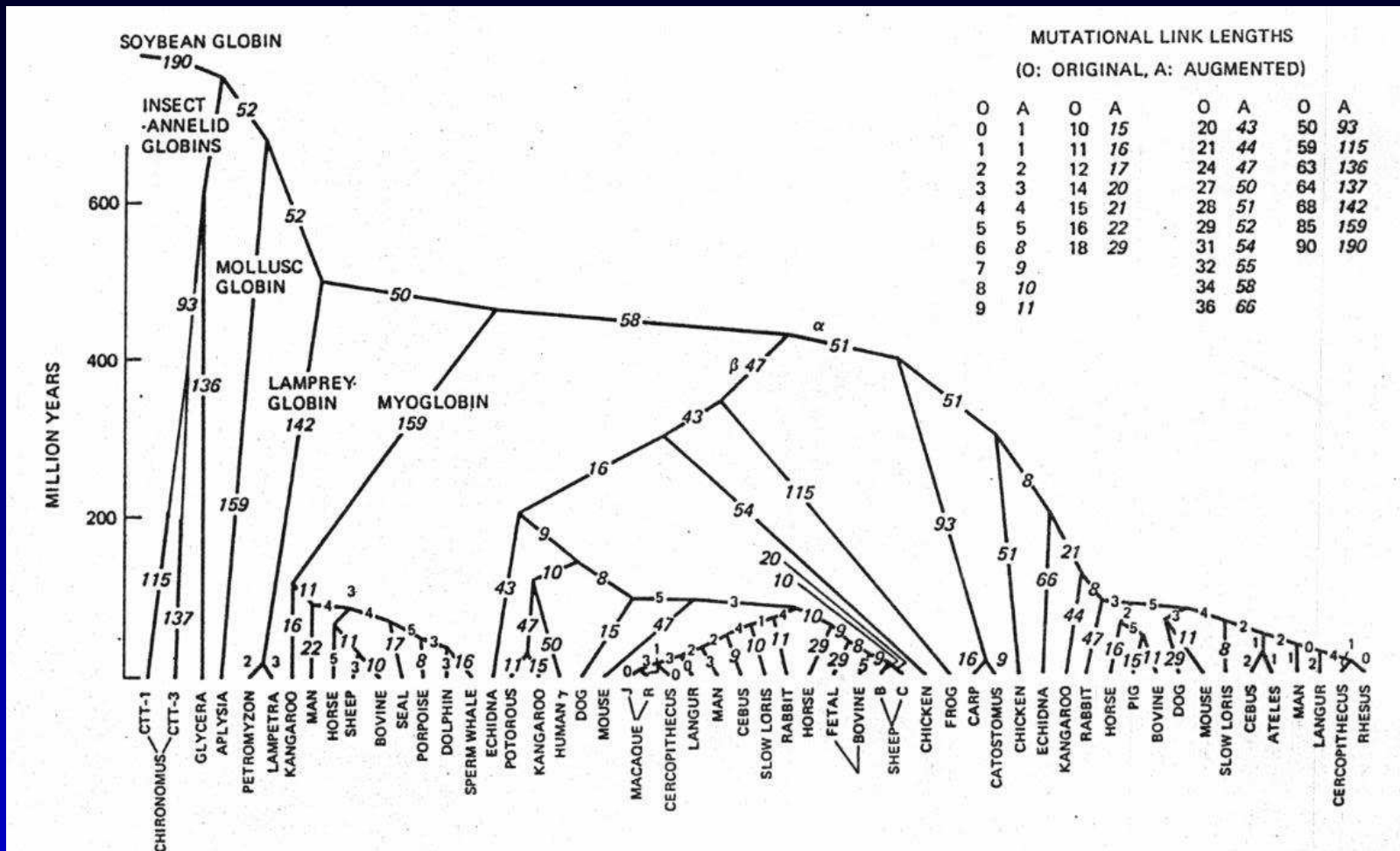
# A gene duplication in a phylogeny



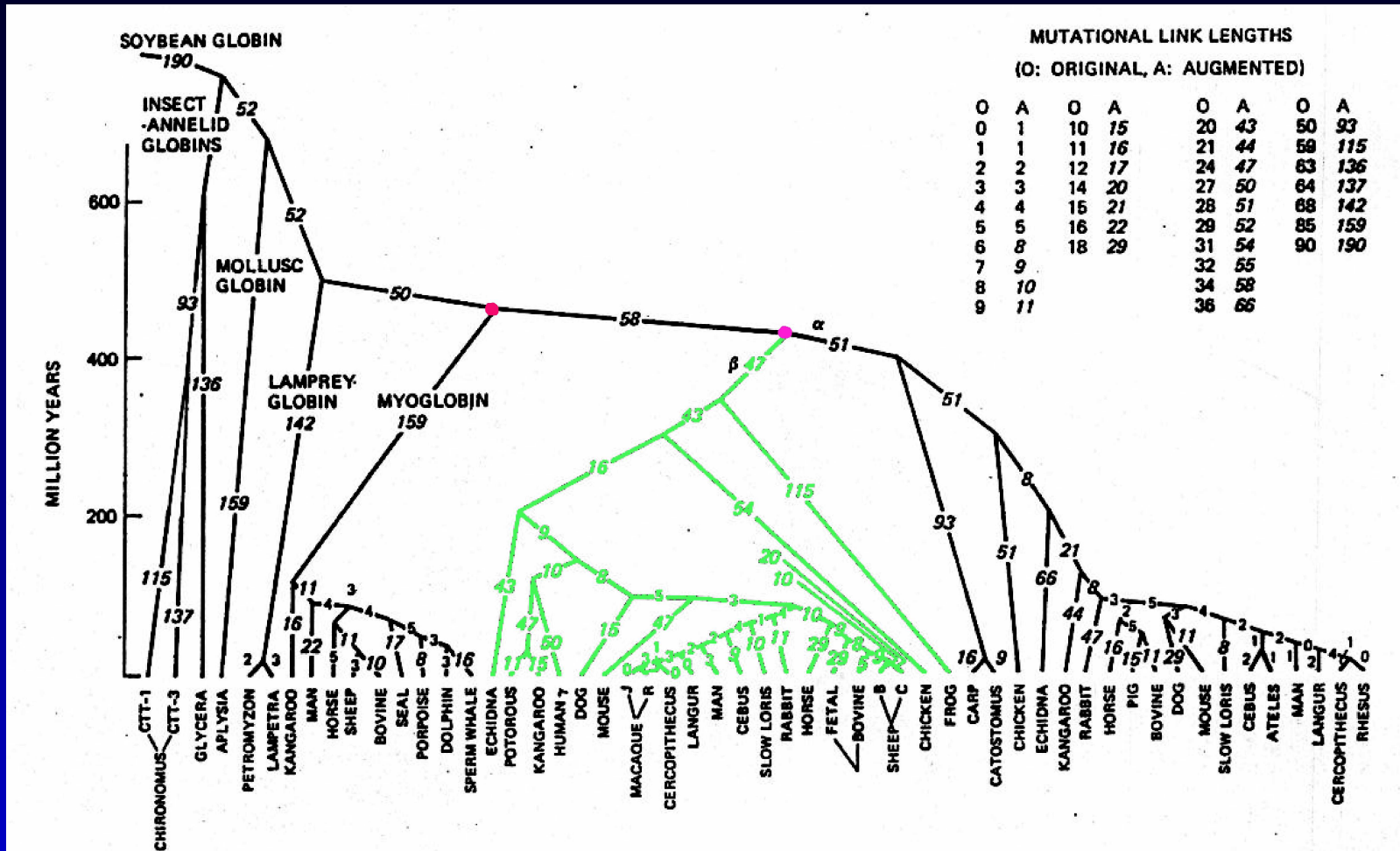
# If we just examine the tree of genes



# A tree of hemoglobins (Morris Goodman, 1975)



# Highlighting the duplication events



## References

- Billera, L. J., S. P. Holmes, and K. Vogtmann. 2001. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics* **27**: 733–767.
- Chang, J. T. 1999. Recent common ancestors of all present-day individuals. *Advances in Applied Probability* **31**(4): 1002-1026.
- Edwards, A. W. F. and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. pp. 67-76 in *Phenetic and Phylogenetic Classification*, ed. V. H. Heywood and J. McNeill. Systematics Association Publication No. 6. Systematics Association, London. [**First paper on numerical methods for estimating phylogenies (from gene frequencies)**]
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts. [**Book you and all your friends must rush out and buy**]
- Semple, C. and M. A. Steel. *Phylogenetics*. Oxford Lecture Series in Mathematics and Its Applications. Oxford University Press, Oxford.

## How it was done

This projection produced as a PDF and viewed using the Full Screen mode (in the View menu) of Adobe Acrobat Reader:  
I made my PDF using LaTeX (though Adobe Acrobat is another possibility):

- using the prosper style in LaTeX,
- using Latex to make a .dvi file,
- using dvips to turn this into a Postscript file,
- using ps2pdf to make it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.