6.877 Computational Evolutionary Biology
Lecture 4: Climb *every* mountain?

The forces of evolution, part II

Agenda:

– The interaction of evolutionary forces, II: mutation-selection balance

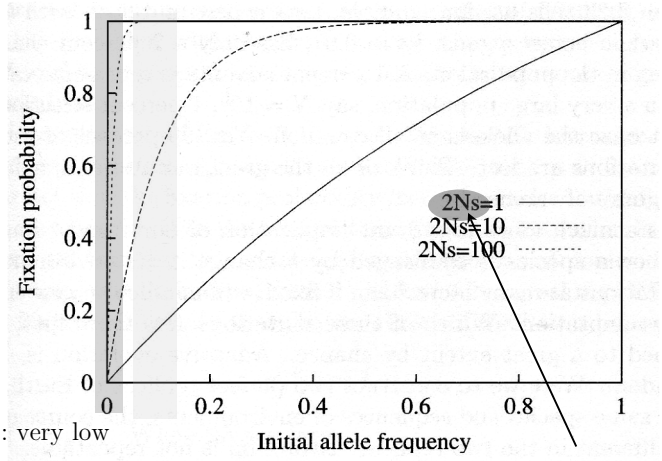
– Genetic drift, and genetic variation: how population size matters

– The interaction of mutation, drift, selection: when does one force prevail over another?

Climb every mountain? Some surprising results

- The power of selection: what is the fixation probability for a new mutation?
- If no selection, the pr of loss in a *single generation* is $1/e$ or 0.3679
- In particular: suppose new mutation has 1% selection advantage as heterozygote – this is a *huge* difference
- Yet this will have only a 2% chance of ultimate fixation, starting from 1 copy (in a *finite* population a Poisson # of offspring, mean $1+s/2$, the Pr of extinction in a *single generation* is $e^{-1}(1-s/2)$, e.g., 0.3642 for $s = 0.01$)
- Specifically, to be 99% certain a new mutation will fix, for $s = 0.001$, we need about 4605 allele copies (independent of population size N !!)
- Also very possible for a *deleterious* mutation to fix, if $2Ns$ is close to 1
- Why? Intuition: look at the shape of the selection curve – flat at the start, strongest at the middle
- To understand this, we'll have to dig into how variation changes from generation to generation, in finite populations

The fate of *selected* mutations



Regime 1: very low copy #

Regime 2: Frequency matters

$2Ns$ (compare to Nu factor)

But wrt selection: Don't make this mistake

BBC NEWS

Friday, 27 September, 2002, 11:51 GMT 12:51 UK

Blondes 'to die out in 200 years'

The last natural blondes will die out within 200 years, scientists believe.

A study by experts in Germany suggests people with blonde hair are an endangered species and will become extinct by 2202.

Researchers predict the last truly natural blonde will be born in Finland - the country with the highest proportion of blondes.

But they say too few people now carry the gene for blondes to last beyond the next two centuries.

The problem is that blonde hair is caused by a recessive gene.

Dyed rivals

The researchers also believe that so-called bottle blondes may be to blame for the demise of their natural rivals. They suggest that dyed-blondes are more attractive to men who choose them as partners over true blondes.

Image removed due to copyright restrictions.

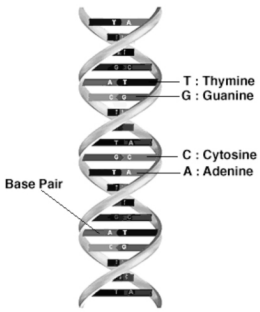
Scientists believe the last blondes will be in Finland

Image removed due to copyright restrictions.

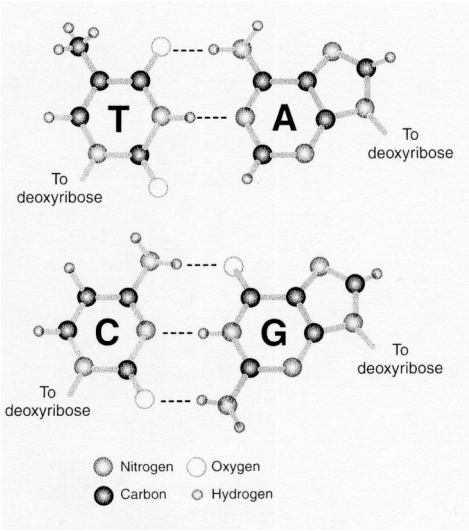
Bottle-blondes like Ann Widdecombe may be to blame

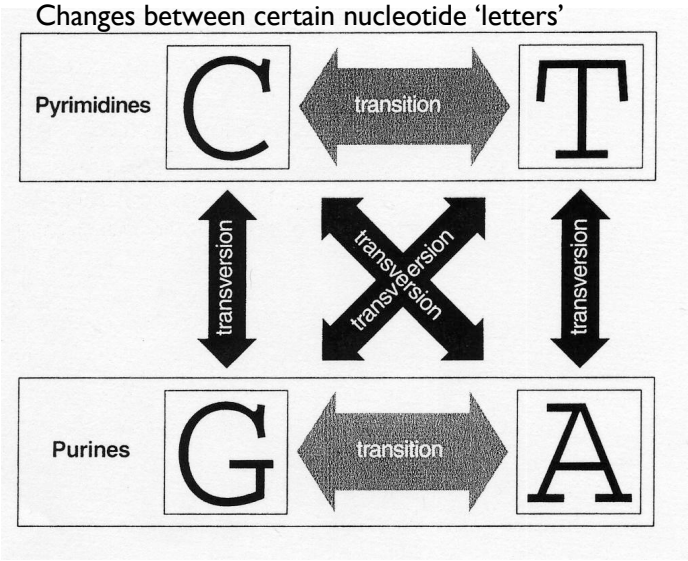
http://news.bbc.co.uk/1/hi/health/2284783.stm

From DNA to mutatians



Nucleotide base pairs T(hymine) – A(denine)
C(ytosine) – Guanine





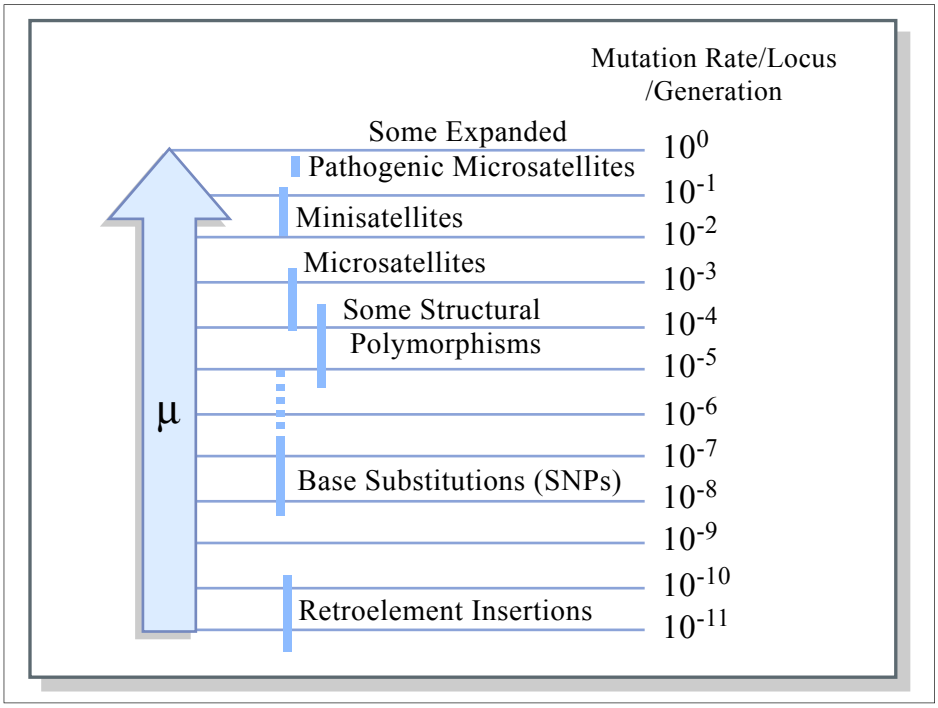
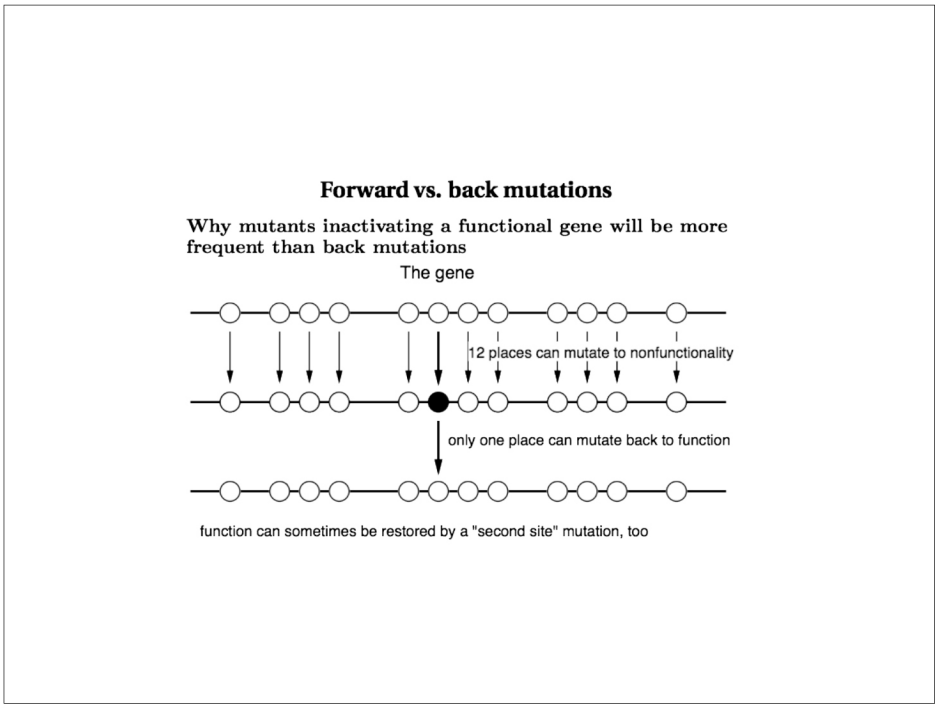


Figure by MIT OCW.



Coat color mutants in mice. From

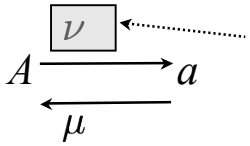
Schlager G. and M. M. Dickie. 1967. Spontaneous mutations and mutation rates in the house mouse. *Genetics* **57:** 319-330

Locus	Gametes tested	No. of Mutations	Rate
Nonagouti	67,395	3	4.4×10^{-6}
Brown	919,619	3	3.3×10^{-6}
Albino	150,391	5	33.2×10^{-6}
Dilute	839,447	10	11.9×10^{-6}
Leaden	243,444	4	16.4×10^{-6}
Total	2,220,376	25	11.2×10^{-6}

Estimation of mutation rate in a bacterial chemostat. Image removed due to copyright restrictions.

Mutation - the weak force

but...sets the *context*



10^{-7} per locus per generation

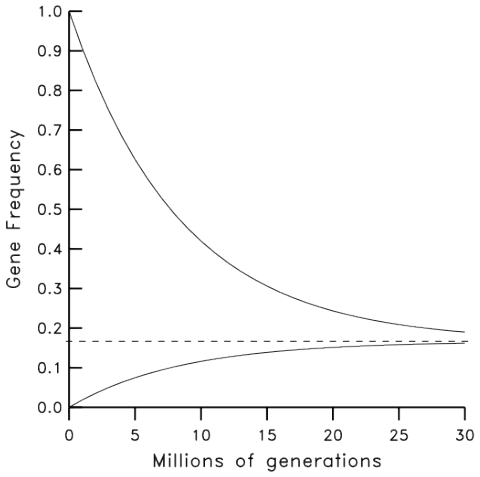
x= current frequency of A

$$x' = x(1-u) + (1-x)v.$$

$$\Delta x = x' - x = -ux + v(1-x).$$

$$x_e = \frac{v}{u+v}$$

Mutation critical for introducing new alleles but very slow at changing them



Approach of gene frequency to equilibrium in a two-allele case starting from fixation at either allele when $u = 5v$ with $u = 10^{-7}$. Note the large number of generations on the horizontal time scale.

Mutation-selection balance: an intuition

$$q_e = u/s$$

Rare mutant a has risk s being eliminated each generation
Each mutant remains avg of $1/s$ generations (coin toss until big D)
So, with this number of generations and rate u of producing a's per generation we have $q_e = u \times 1/s = u/s$

Mutation-selection balance: deleterious dominant allele, a

Assumptions: frequency of a is small (= 1–p = q)
no heterozygote selection effect (h=0)
q is small due to selection

Then:

$$\Delta_s p = \frac{pqs[ph+q(1-h)]}{1-2pqhs-q^2 s} \approx \boxed{qs}$$

$$\Delta_u p = -u$$

$$0 = \Delta_u p + \Delta_s p$$

$$\approx -u + qs; \text{ So}$$

$$\hat{q} = \frac{u}{s}$$

Exchange p and q (x= freq of q)

- Dominant disease

Genotype	AA	Aa	aa
Fitness	1	1 - s	1 - s

$E_s[\Delta x] \approx -sx(1-x)$ Selection

$E_u[\Delta x] = (1-x)u$ Mutation

$\Rightarrow \quad \tilde{x} = \frac{u}{s}$

By an equilibrium calculation. Huntington's disease. Dominant. Does not express itself until after age 40. $1/100,000$ of people of European ancestry have the gene. Reduction in fitness maybe 2%.

- If allele frequency is q, then $2q(1-q)$ of everyone are heterozygotes.
- 0.02 of these die. Each has half its copies the Huntington's allele.
- So as the frequency of people with the gene is $\simeq 1/100,000$, the fraction of all copies that are mutations that are eliminated is $0.00001 \times 1/2 \times 0.02 \simeq 10^{-7}$
- If we are at equilibrium between mutation and selection, this is also the fraction of copies that have a new mutation.

Similar calculations can be done with recessive alleles.

Selection-mutation equilibrium

What does this mean?
In almost every case where we can see selection operating on phenotype,
$s \gg u$ (hard to imagine $s < 10^{-6}$)
Exception: DNA and protein data
$u = 10^{-7}$ and $s = 10^{-3}$, then $q_e = 0.0001$
Note: at each gen, a fraction $u(1-q) = 0.9999 \times 10^{-7}$ mutate A to a
A fraction $uq = 10^{-11}$ mutate from a to A
(So back mutation safely ignored)

- Recessive disease

Genotype	AA	Aa	aa
Fitness	1	1	1 - s

$E_s[\Delta x] \approx -sx^2(1-x)$ Selection

$E_u[\Delta x] = (1-x)u$ Mutation

$\Rightarrow \quad \tilde{x} = \sqrt{\frac{u}{s}}$

$u = 10^{-6}$, $s = 2\%$

Dominant = 1 in 20,000
Recessive = 1 in 140

The selection-mutation equilibrium: recessive case

Diploid Newborns	*Selection* ⟶	Diploid Adults	*Meiosis* ⟶	Haploid Gametes	*Mutation* ⟶	Haploid Gametes	*Mating* ⟶	Diploid Newborns

AA	Aa	aa
1	1	$1-s.$

After selection (check this!):

$$p^* = \frac{p(p \times 1 + (1-p) \times 1)}{p^2 \times 1 + 2p(1-p) \times 1 + (1-p)^2 \times (1-s)}$$

$$p^* = \frac{p}{1 - s(1-p)^2}$$

After adding mutation:

$$p' = \frac{p(1-u)}{1 - s(1-p)^2}$$

Computing the equilibrium

$$p' = \frac{p(1-u)}{1 - s(1-p)^2}$$

$$1 - s(1-p_e)^2 = 1 - u$$

$$(1-p_e)^2 = u/s$$

$$q_e = 1 - p_e = \sqrt{u/s}.$$

For $u = 10^{-7}$ and $s = 10^{-3}$, $q_e = 0.01$

This is 100 times greater than the recessive case…Why?

Informal argument for recessive diploid

Key: must be *homozygous* to lose
from H-W: frequency of affected organisms the same:
$q_e^2 = u/s$
Pr sq_e of being eliminated in each gen
Average mutant persists $1/(sq_e)$ generations
Population has $1/sq_e0$ generations worth of mutants
Times u mutants per generation =
$q_e = u \times 1/sq_e$

What about the other forces?

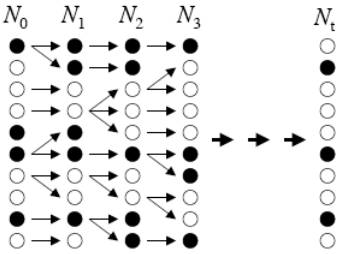
Genetic variability is lost in finite populations

Image removed due to copyright restrictions.

Buri (1956):

107 Drosophila populations, each started with 16 heterozygotes for a brown eye mutation (bw)

The Wright-Fisher model



We get a binomial tree that depends on frequency, p, and total population size, N.

→ Binomial sampling $\Pr\{j \mid i\} = \frac{2N!}{j!(2N-j)!}\left(\frac{i}{2N}\right)^{j}\left(1-\frac{i}{2N}\right)^{2N-j}$

What is the pr that a particular allele has at least 1 copy in the next generation?
Well, what is the pr of *not* picking an allele on *one* draw?
Ans: 1-(1/2N). There are 2N draws (why?). So, pr of *not* picking for this many draws is $[1-(1/2N)]^{2N} = e^{-1}$ for large N

Let's explore the consequences…

Binomial sampling already implies some results

Pr that generation t has i copies of an allele A_1, given $2N$ independent trials is:

For example, the probability that generation t has 10 copies of A, where pr(A)=11/20=0.55 in gene pool for generation t–1 is: 20!/10! 10! $(0.55)^{10}(0.45)^{10}$= 0.1593

Mean and variance of *frequencies* p (nb, not just the allele numbers)

Because this is a binomial draw with parameters p, $2N$, the mean of this distribution (the expected # of A_1 alleles drawn) is just $2Np$, i.e., mean <u>frequency</u> is p
The variance in allele # is: $2Np(1\text{-}p)$
So the variance in <u>allele frequency</u> is:

$$E[p'] = E[X]/2N = 2Np/2N = p$$

The variance of p <u>goes down</u> as the population size increases:

$$\mathrm{Var}[p'] = \mathrm{Var}[X]^2/4N^2 = 2Np(1-p)/4N^2 = p(1-p)/2N$$

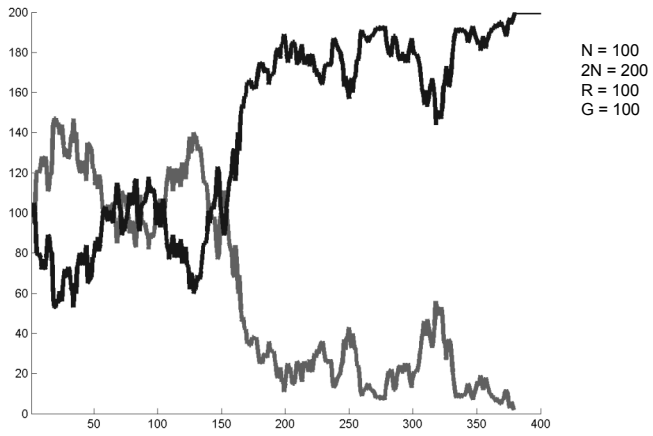
First consequence: new mutations, if neutral…

What is the probability that a particular allele has at least 1 copy in the next generation? In other words: that a brand-new mutation survives?

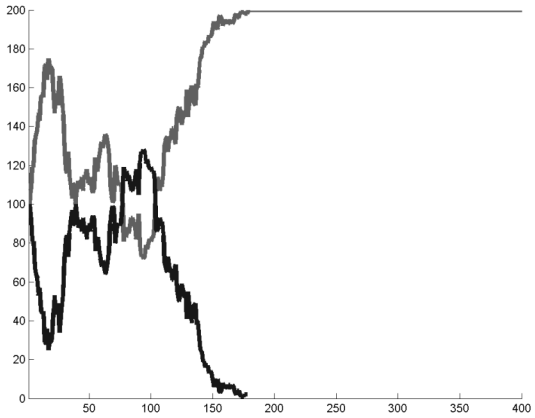
Well, what is the pr of *not* picking an allele on *one* draw?
Ans: 1-(1/2N). There are 2N draws (why?).
So, pr of *not* picking for this many draws is $[1\text{-}(1/2N)]^{2N} = e^{-1}$ for large N

So: probability of a new mutation being lost simply due to 'Mendelian bad luck' is $1/e$ or 0.3679

Why doesn't population size N matter?
Answer: it's irrelevant to the # of offspring produced initially by the new gene



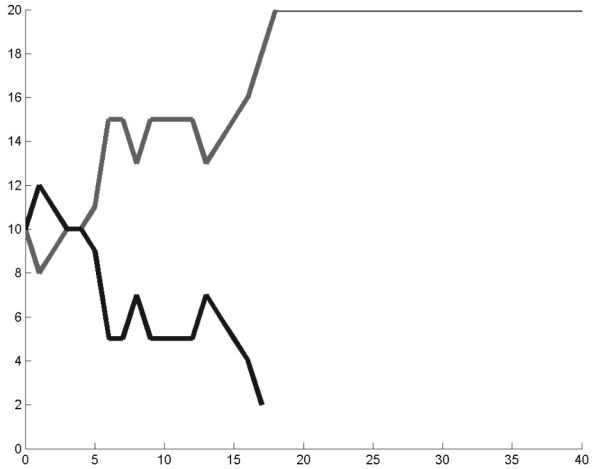
One allele always wins!
Survival of the fittest? Down with Darwin?



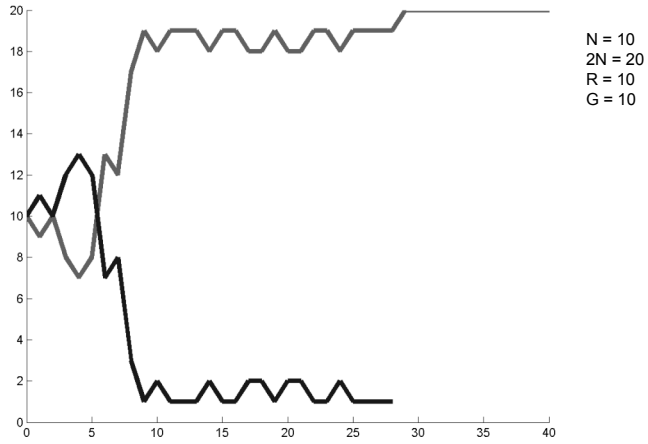
N = 100
2N = 200
R = 100
G = 100

Is this always so?
Let's try changing N and initial allele frequencies

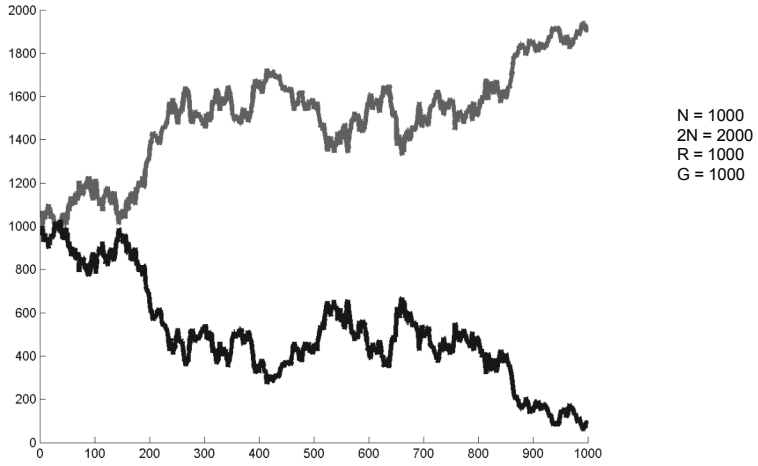
Reduce N to 10

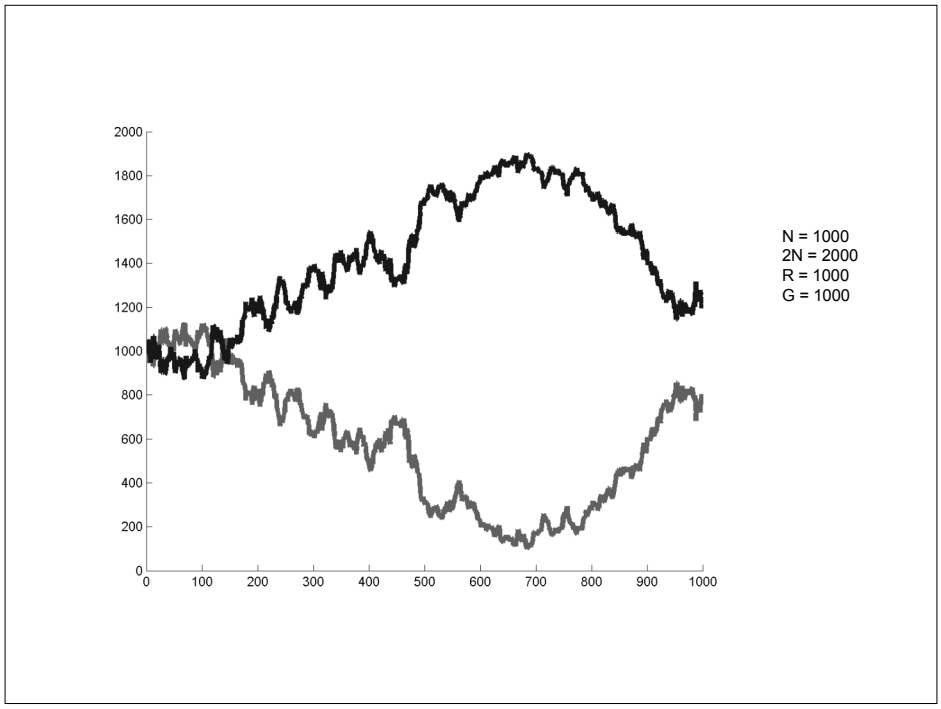


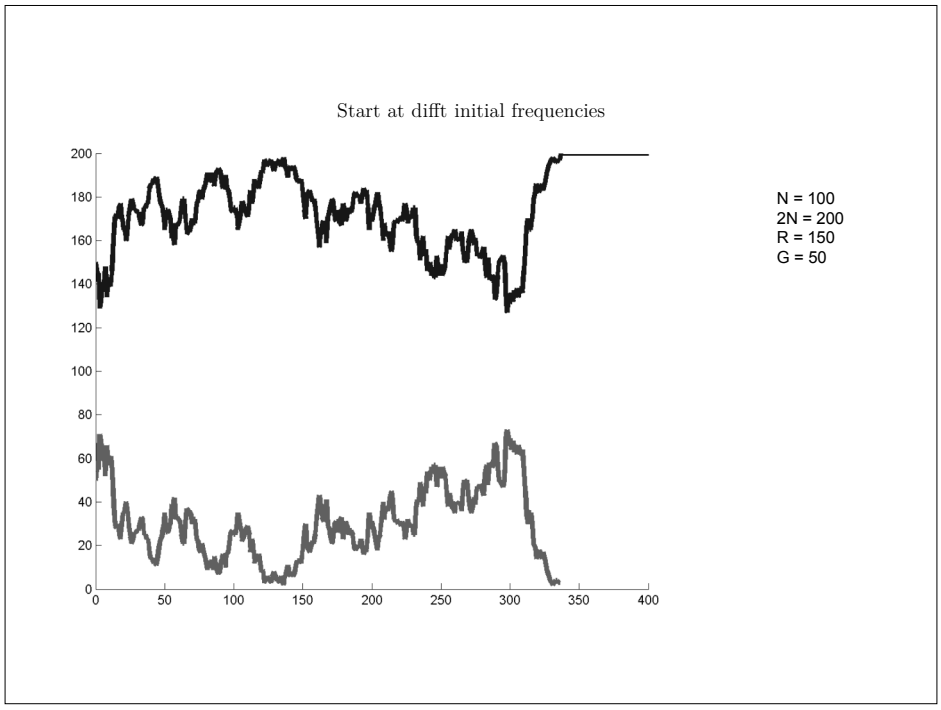
N = 10
2N = 20
R = 10
G = 10

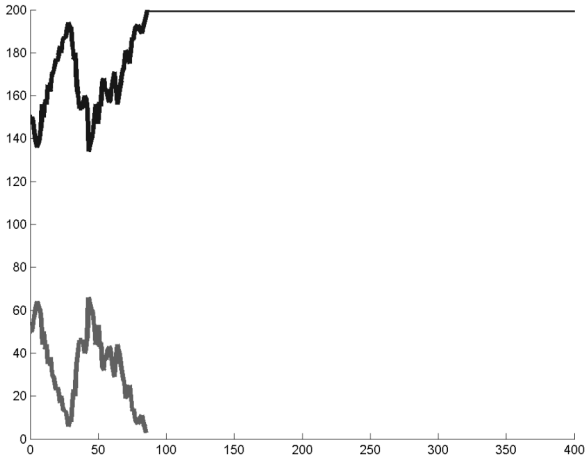


Boost N to 1000

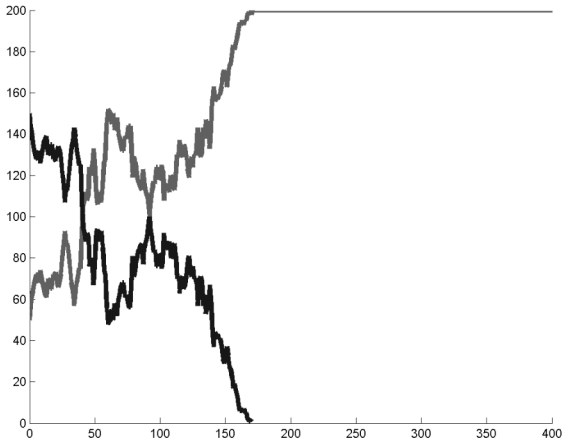



2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
100
200
300
400
500
600
700
800
900
1000
N = 1000
2N = 2000
R = 1000
G = 1000


Start at difft initial frequencies
200
180
160
140
120
100
80
60
40
20
0
0
50
100
150
200
250
300
350
400
N = 100
2N = 200
R = 150
G = 50



N= 100
2N = 200
R = 150
G = 50

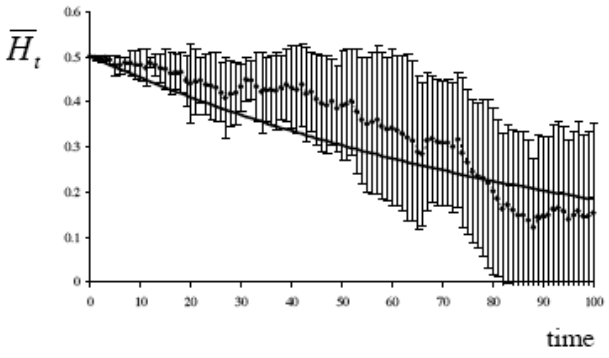


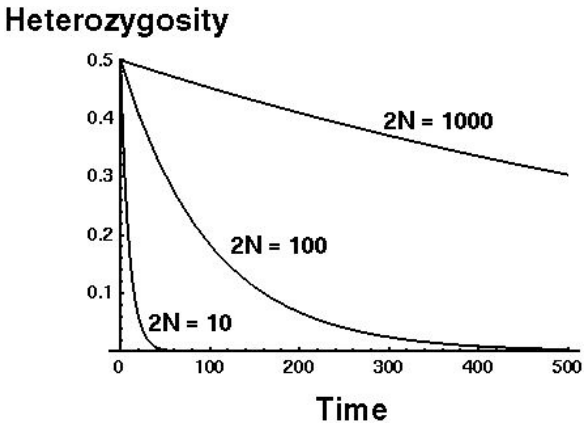
N= 100
2N = 200
R = 150
G = 50

What are the general rules?

- Higher population size = alleles stick around longer.
- Less susceptibility to "random walk"
- Probability of winning seems related to initial frequencies.
- At 50/50 initial allele frequency, 50% chance of either allele winning.
- Hypothesis: probability of winning is proportional to initial allele frequency. (Proof follows)
- Hypothesis: One allele must always win.

Drift & the inevitable decay of heterozygosity (variation), H_t

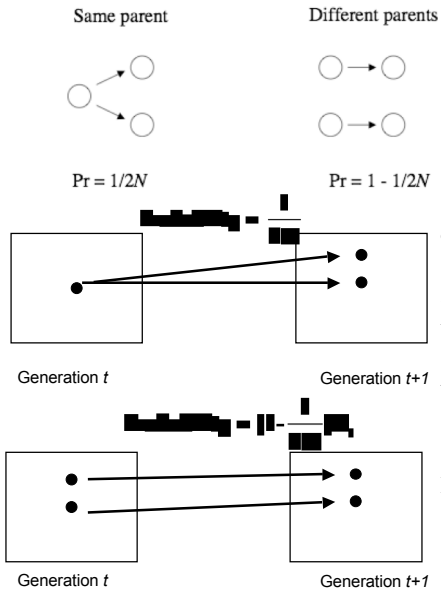




A mathematical analysis of drift: the decay of heterozygosity (loss of variation)

- Define H_t= probability in generation t that 2 alleles picked at random are **different** from one another ('heterozygous'); homozygosity, G_t as 1-H_t ('identical by descent')
- Now develop a recurrence relation for H_t

Recurrence relation for G, H



G_t= *Identical* in one of 2 possible ways: (1) clone from same ancestor; or (2) the two ancestors were *different*, **but** they were *identical by descent* back then; *add* these 2 probabilities

Recurrence relation for G_t, H_t

$G_t = \Pr\{\text{identical by descent}\}$

$G_1 = \frac{1}{2N} + (1-\frac{1}{2N})G_0$

$G_2 = \frac{1}{2N} + (1-\frac{1}{2N})G_1$

$H_t = \Pr\{\text{different by descent}\}$
$= 1 - G_t$

$H_1 = 1 - G_1$
$= (1-\frac{1}{2N})(1-G_0)$
$= (1-\frac{1}{2N})H_0$

$H_2 = 1 - G_2$
$= (1-\frac{1}{2N})^2 H_0$

This has <u>important</u> implications for allele fixation: eventually, one allele *always* wins, just as we said…and…we can now figure out the *pr* of fixation (assuming no selection – we will factor that back in …)

What is the half-life of H?

$H_0/2 = H_0(1\text{-}1/2N)^t$ – cancel H_0 from both sides, take natural logs, solve for t

$t = 2N \ln 2$ (using $\ln(1+x)$ approx x)

$N = 10^6$, $t =$ 1.38e6 generations

Important part: this says something about the time-scale of drift – it's roughly the population size

Time scale & interaction of forces

Drift: $2N$ generations
HW: 1 or 2 generations
So: these 'forces' don't interact w/ each other…

Important: after $2N$ generations, all variation is gone - this is how far back we can 'see' - everybody derived from this single allele

Fixation probability of an allele is *proportional* to its initial frequency

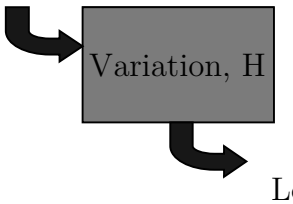
All variation is ultimately lost, so eventually 1 allele is ancestor of *all* alleles
There are $2N$ alleles
So the chance that any one of them is ancestor of all is $1/2N$

If there are i initial copies, the fixation chance is $i/2N$

(Simple argument because all alleles are equivalent – there is no natural selection)

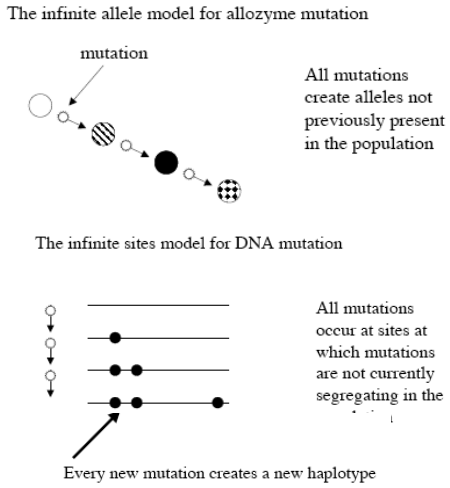
Adding mutations – the mutation-drift balance

Mutation gain $2Nu$



Loss at rate $1/(2N)$

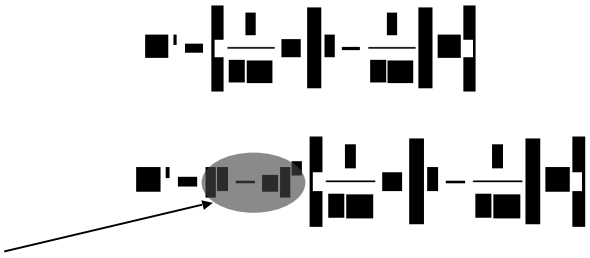
Modeling mutations - 2 ways



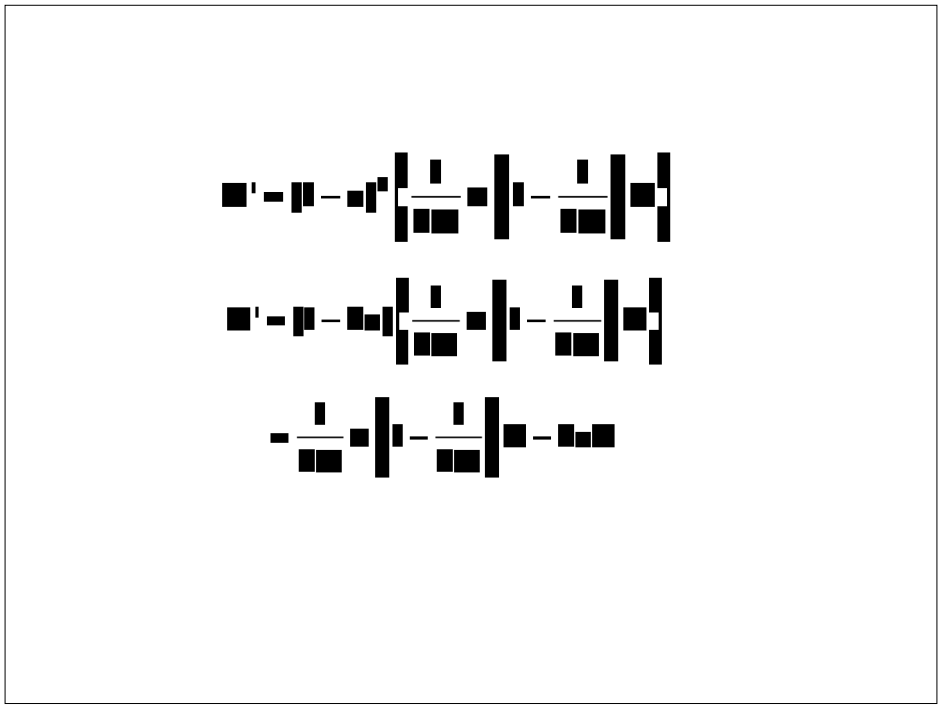
Modeling the balance

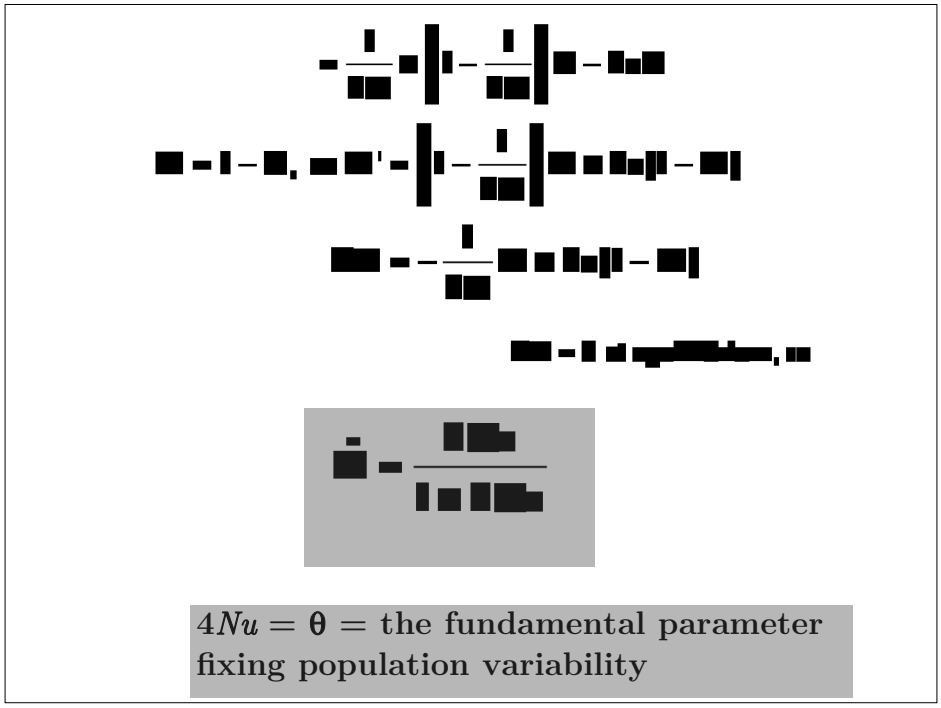
Assume N is large, compared to u

Take our existing formula for G and factor in mutation rate u (which *reduces* G, increases H):



Pr that we did <u>not</u> mutate (both alleles)





$4Nu = \theta$ = the fundamental parameter fixing population variability

Analysis…implications

- $H_{eq} = 4Nu/(1+4Nu)$
- Let Nu be large compared to 1. Then the population is almost always heterozygous. (Mutations occur before drift can remove)
- Let Nu be very small compared to 1. Then the population has little variation. (Drift removes variation before a new mutation occurs)
- If $1/u \ll N$, time scale of mutation is much less than drift, so population will have many unique alleles; if $N \ll 1/u$, then time scale of drift is shorter, population will be devoid of variation

Examples

Example: HIV virus.

$\mu = 10^{-5}$ per nucleotide and N = 10^7-10^8 infected cells in a host. This means almost every nucleotide is variable in the population.

Example: Human

$\mu = 10^{-8}$ per nucleotide and N = 10^3-10^5 (?)
A typical nucleotide shows almost no variation in the population.

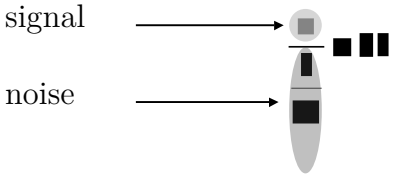
$\mu = 10^{-5}$ per gene. A typical gene will have few variants in a population.

$\mu > 1$ per genome. Every genome is essentially unique.

The forces of evolution…

$$E[H] = \frac{4N_e u}{1 + 4N_e u}$$

Goal: understand relation between forces: u, $1/N$

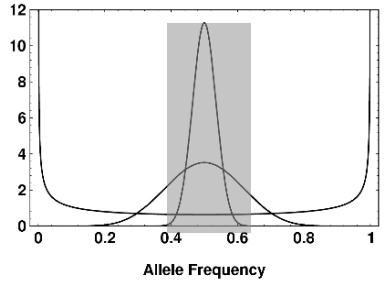


Mutation vs. drift: the key number is $4N\mu$ vs. 1

$N\mu > 1$, diversity *increases*
heterozygosity maintained around 0.5

Gain heterozygosity → variance stays high

Population "large" wrt genetic drift



“Follow the variation”

Heterozygosity= [redacted] $4Nu = \theta$

Homozygosity (identity)= $1 - H =$ G= $1/\theta$

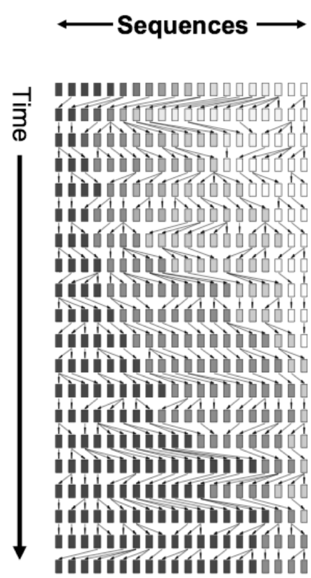
These are the key measures of how ‘variant’ two genes (loci), sequences, etc. are

What can we learn about their distributions?

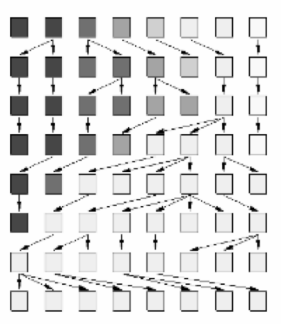
How can we estimate them from data?

How can we use them to test hypotheses about evolution?

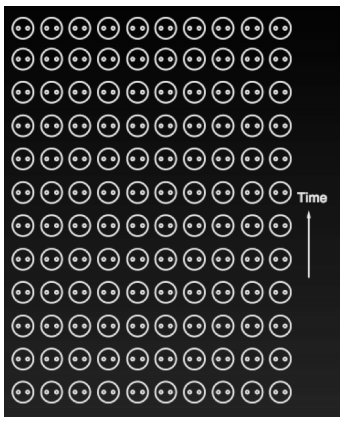
Loss of ancestral lineages: why lineages ‘coalesce’

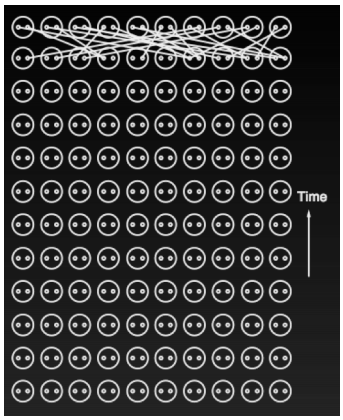


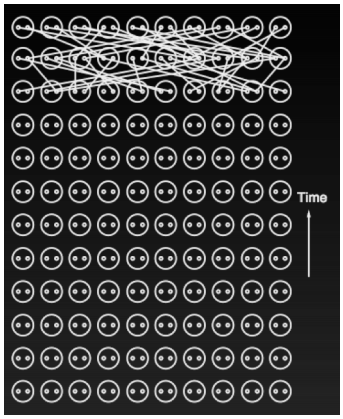
Eventually, only *one* copy of an allele will survive
(assuming no selection, migration in, etc.)

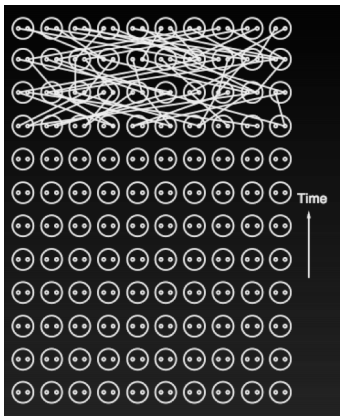


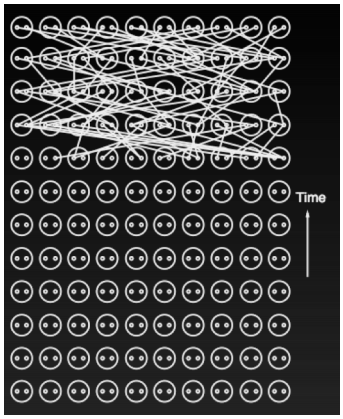
Wright-Fisher random mating… large population

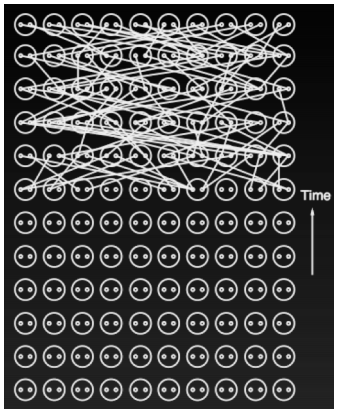


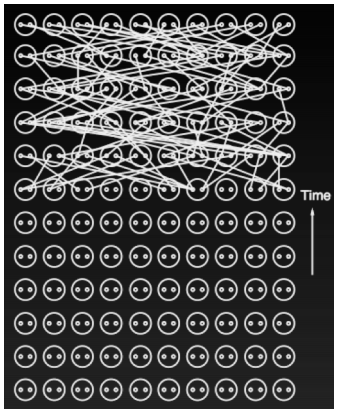

Time

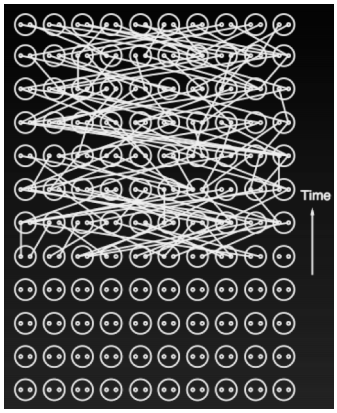

Time

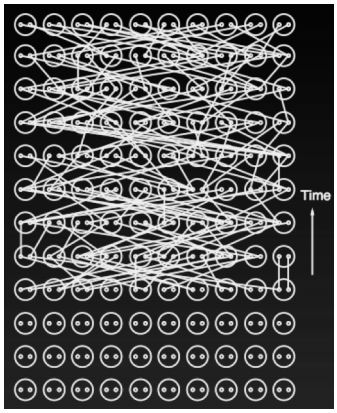

Time

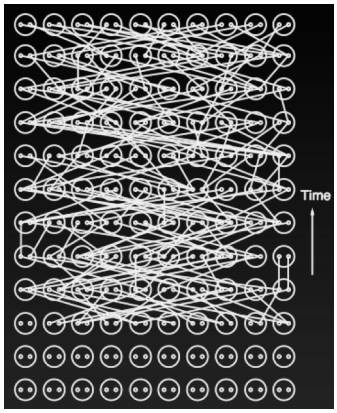

Time

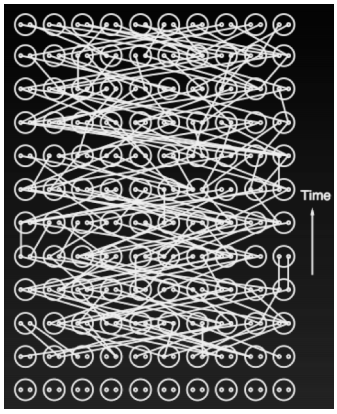

Time

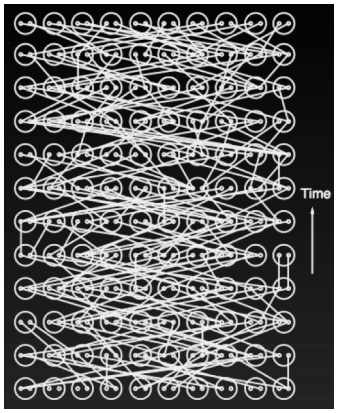

Time

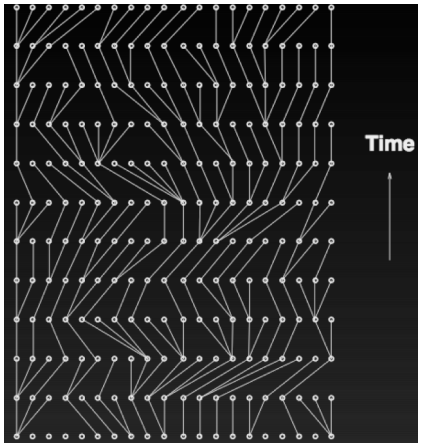

Time


Time

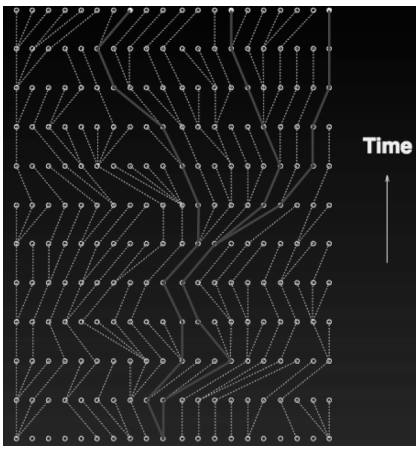

Time


Time

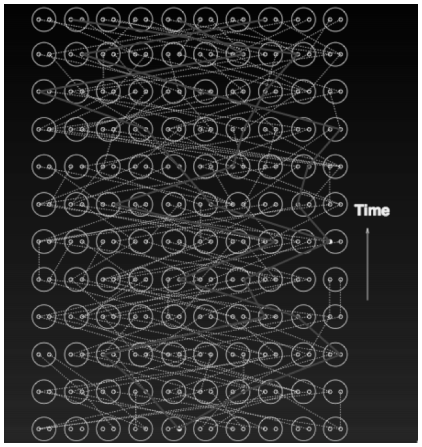

Time


Time

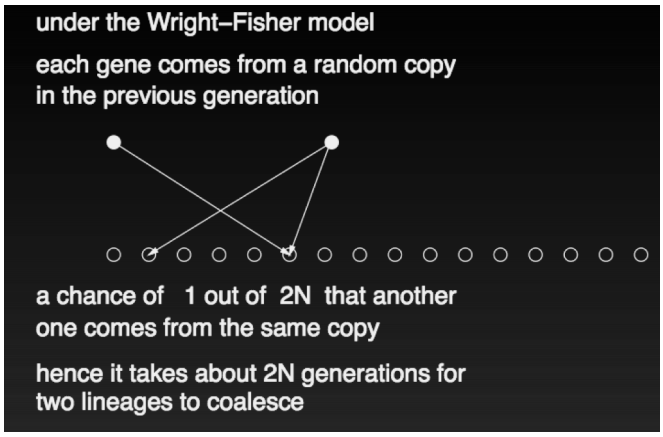
Genealogy of a sample of gene copies



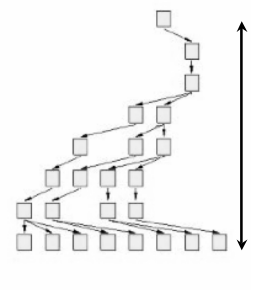
Ancestry of a sample in the population pedigree



Why lineages coalesce



In other words…



On average, depth $2N$ before collapse to 1 ancestor

We'll <u>prove</u> this next time – see ch. 3 of Rice book