# Tutorial on Theoretical Population Genetics

Joe Felsenstein

Department of Genome Sciences and Department of Biology
University of Washington, Seattle

replacing Lee Altenberg

who cannot be here this time, but has made helpful comments.

# History of theoretical population genetics

Darwin's work on natural selection was done in the absence of an understanding of how heredity worked. In the 1880's and 1890's there was increased interest in heredity. Francis Galton (1822-1911) and Karl Pearson (1857-1936) tried to construct statistical regression formulae that summarized inheritance empirically. This was not successful, though they made major contributions to the development of multivariate statistics.

Soon after the rediscovery of Mendel's work in 1900, geneticists were concerned with its implications for populations of organisms. One argument, about whether a dominant mutant, syndactyly, would necessarily increase and take over the human population, led to the work of G. H. Hardy on the Hardy-Weinberg law.

# History of theoretical population genetics, continued

R. A. Fisher

J. B. S. Haldane

Sewall Wright

Much of its structure was developed by three people: R. A. Fisher (1890-1962), Sewall Wright (1889-1988), and J. B. S. Haldane (1892-1962). Their work in the 1910’s to 1940’s was the basis for the “neodarwinian synthesis" which united genetics and evolution.

# Resources in theoretical population genetics

## Textbooks

Ewens, W. J. 2004. *Mathematical Population Genetics. I. Theoretical introduction. Second edition.* Springer-Verlag, New York. (Advanced-level text).

Rice, S. H. 2004. *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer Associates, Sunderland, Massachusetts. (not yet published)

Hartl, D. H. and A. J. Clark. 1997. *Principles of Population Genetics, 3rd edition.* Sinauer Associates, Sunderland, Massachusetts. (A standard introduction, emphasizing experimental more than theoretical results).

## Web text

Felsenstein, J. 2003. *Theoretical Evolutionary Genetics.* Free as a PDF at: `http://evolution.gs.washington.edu/pgbook/pgbook.html`

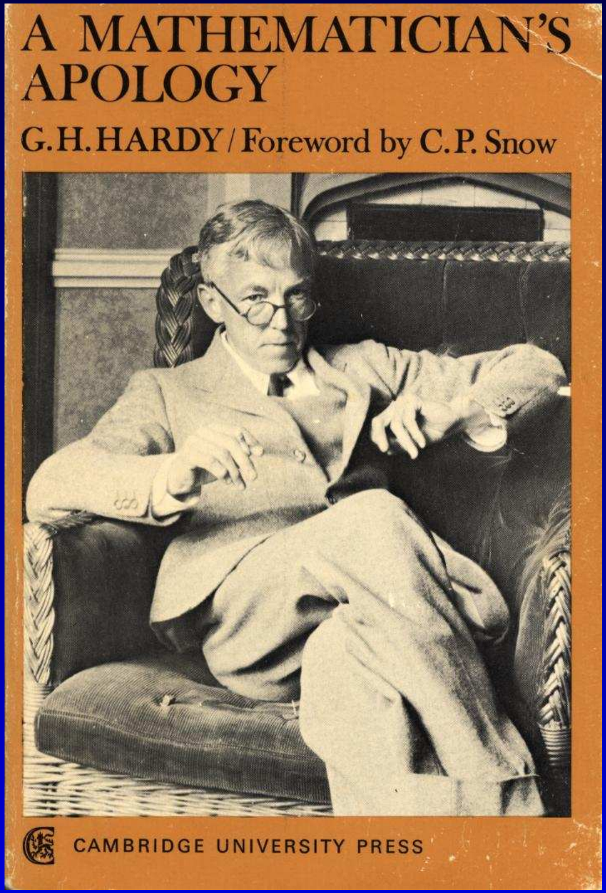
## Bibliography

Felsenstein, J. 1981. *A Bibliography of Theoretical Population Genetics.* Dowden, Hutchinsin, and Ross, Stroudsburg, Pennsylvania. Out of print as a book but the file of references can be downloaded from: `ftp://evolution.gs.washington.edu/bible/`

(7,982 references 1867-1981, the entirety of the theoretical literature to that date.)

# Random mating; Hardy-Weinberg proportions

This was established by Hardy (1908) and Weinberg (1908):


A MATHEMATICIAN'S APOLOGY
G.H.HARDY / Foreword by C.P. Snow
CAMBRIDGE UNIVERSITY PRESS

Godfrey Harold Hardy (1877-1947)

Wilhelm Weinberg (1862-1937)

## Hardy-Weinberg proportions

Given a population with a locus with two alleles and discrete generations and these assumptions:

- The population is infinitely large
- There is no mutation
- There is no immigration or emigration
- There are no differences among the genotypes in viability
- There are no differences among the genotypes in fertility
- There are no differences in frequencies of genotypes between females and males

then if we start with genotype frequencies $\mathsf{P}$, $\mathsf{Q}$, and $\mathsf{R}$,we can calculate the frequencies of the genotypes in the next generation as

$$\begin{array}{llcr} \mathsf{AA} & \mathsf{p}^2 & = & (\mathsf{P}+\frac{1}{2}\mathsf{Q})^2 \\ \mathsf{Aa} & 2\mathsf{pq} & = & 2\,(\mathsf{P}+\frac{1}{2}\mathsf{Q})\,(\frac{1}{2}\mathsf{Q}+\mathsf{R}) \\ \mathsf{aa} & \mathsf{q}^2 & = & (\frac{1}{2}\mathsf{Q}+\mathsf{R})^2 \end{array}$$

# Hardy-Weinberg proportions, continued

the quantities $P + \frac{1}{2}Q$ and $Q + \frac{1}{2}R$ are the “gene frequencies" (they are really the allele frequencies but we are stuck with the terminology). These, and not the individual genotype frequencies, determine the next generation’s genotype frequencies.

Essentially, the outcome of random mating under these conditions is that everybody contributes equally to an infinitely large “gamete pool" and the offspring are equivalent to draws of pairs of gametes from this.

Furthermore, the generation after that has gene frequencies:

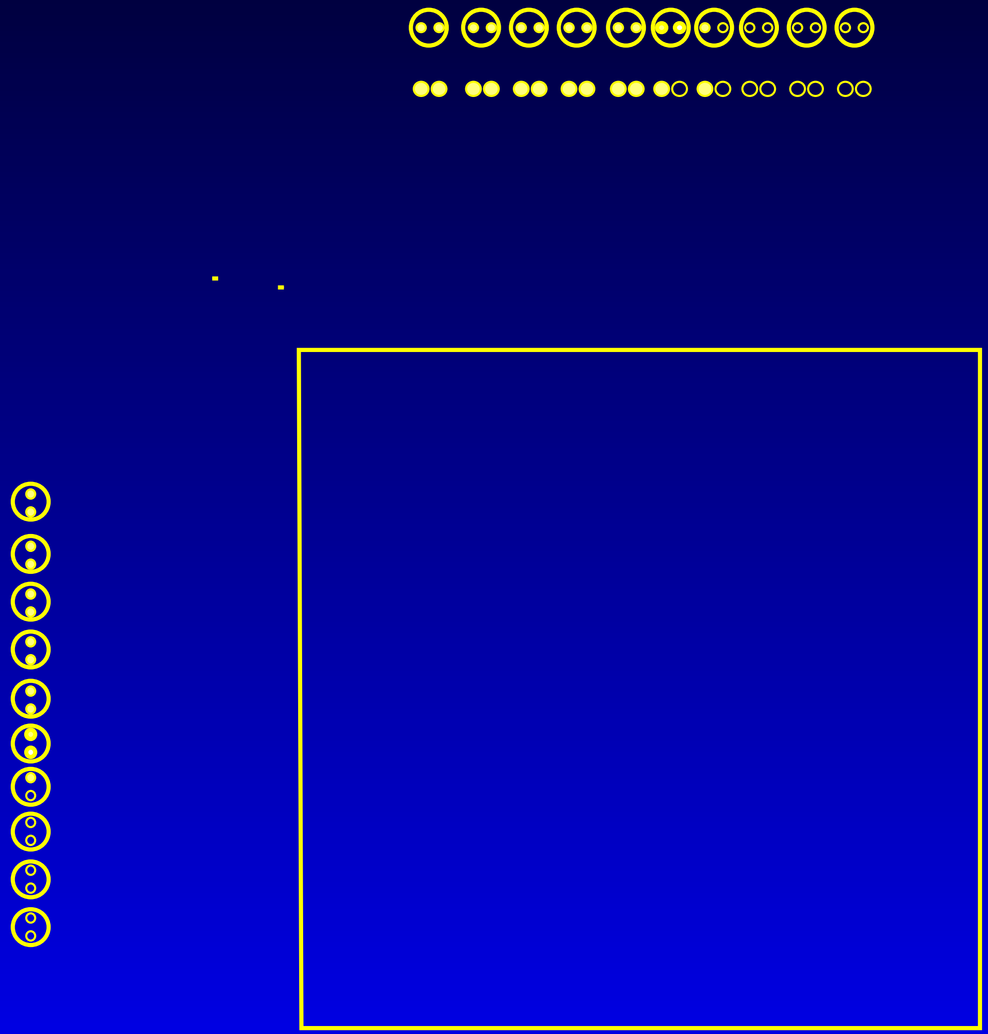
$$p' \;=\; P' + \frac{1}{2}Q' \;=\; p^2 + \frac{1}{2}(2pq) \;=\; p$$

So the gene frequency does not change from generation to generation, and the Hady-Weinberg proportions do not change after the first generation.

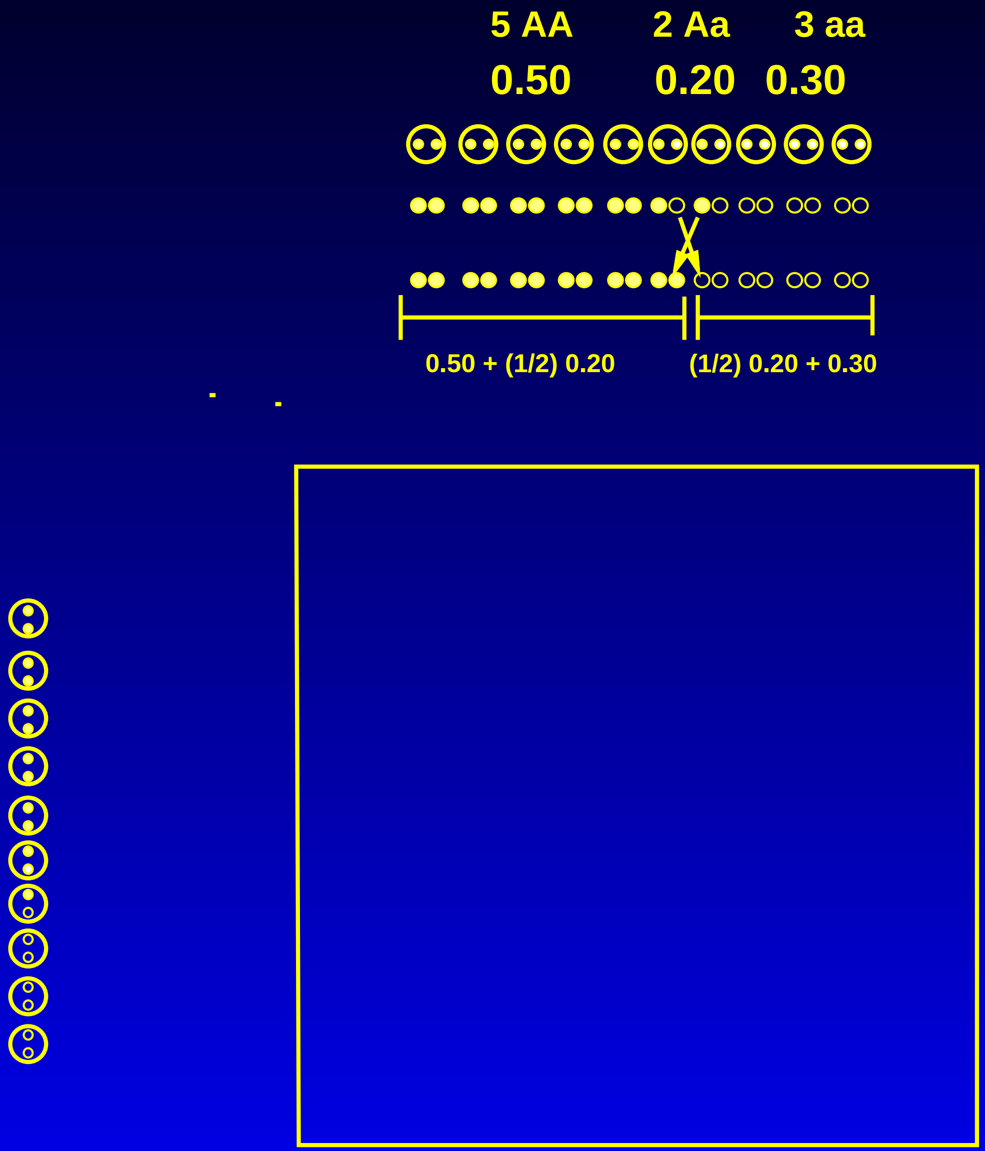
# A Hardy-Weinberg calculation

5 AA 2 Aa 3 aa

0.50 0.20 0.30



# A Hardy-Weinberg calculation



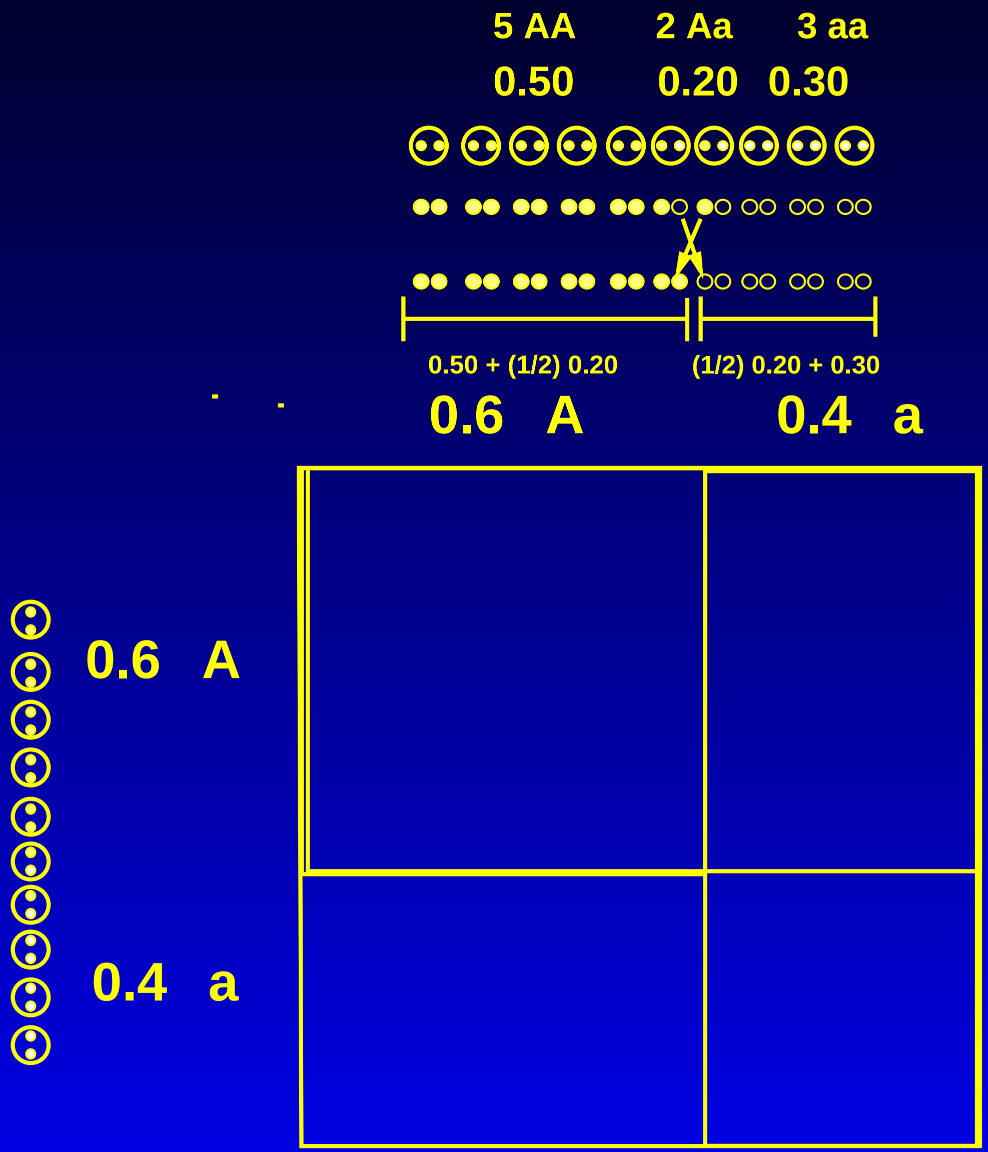
5 AA 2 Aa 3 aa

0.50 0.20 0.30

0.50 + (1/2) 0.20

(1/2) 0.20 + 0.30

# A Hardy-Weinberg calculation



5 AA 2 Aa 3 aa

0.50 0.20 0.30

0.50 + (1/2) 0.20

(1/2) 0.20 + 0.30

0.6 A

0.4 a

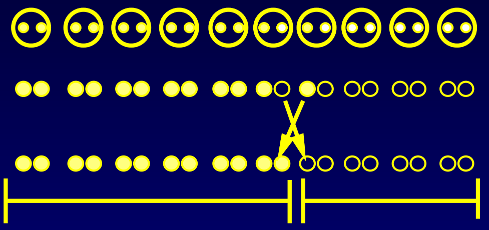
0.6 A

0.4 a

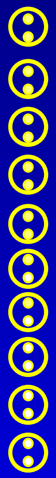
# A Hardy-Weinberg calculation

5 AA 2 Aa 3 aa

0.50 0.20 0.30



0.50 + (1/2) 0.20 (1/2) 0.20 + 0.30

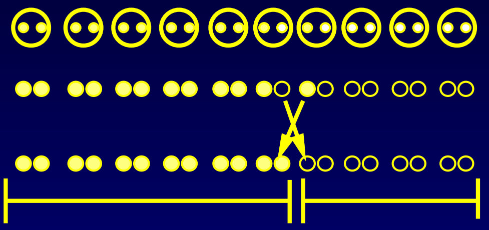


|  | 0.6 A | 0.4 a |
|---|---|---|
| 0.6 A | 0.36 AA | 0.24 Aa |
| 0.4 a | 0.24 Aa | 0.16 aa |

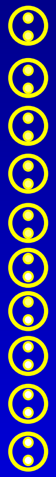
# A Hardy-Weinberg calculation

5 AA 2 Aa 3 aa

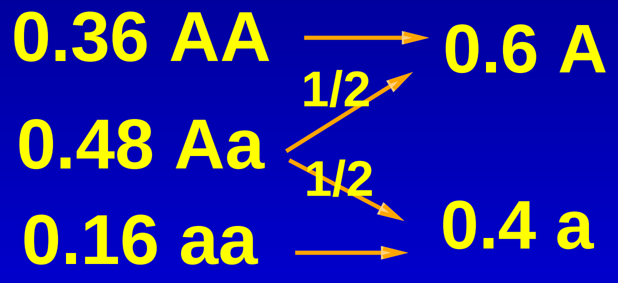
0.50 0.20 0.30



0.50 + (1/2) 0.20 (1/2) 0.20 + 0.30



|  | 0.6 A | 0.4 a |
|---|---|---|
| 0.6 A | 0.36 AA | 0.24 Aa |
| 0.4 a | 0.24 Aa | 0.16 aa |



Result:

0.36 AA → 0.6 A

0.48 Aa —1/2→ 0.6 A, —1/2→ 0.4 a

0.16 aa → 0.4 a

# The algebra of natural selection

(J. B. S. Haldane, 1924)

If we relax the assumption of no differential viability or fertility, we essentially multiply each genotype's frequency by its fitness (reflecting the weighting of it by its probability of survival and its relative participation in reproduction)

| Genotype frequency | Frequency before | Frequency after | Normalized to add to 1 |
|---|---|---|---|
| AA | $p^2$ | $p^2\ w_{AA}$ | $p^2\ w_{AA}/\bar{w}$ |
| Aa | $2pq$ | $2pq\ w_{Aa}$ | $2pq\ w_{Aa}/\bar{w}$ |
| aa | $q^2$ | $q^2\ w_{aa}$ | $q^2\ w_{aa}/\bar{w}$ |

where the mean fitness is $\bar{w} = p^2\ w_{AA} + 2pq\ w_{Aa} + q^2\ w_{aa}$

# The algebra of natural selection, continued

so that the offspring are the result of random mating at a gene frequency of

$$p' = \frac{p^2\, w_{AA} + \frac{1}{2}\, 2pq\, w_{Aa}}{p^2\, w_{AA} + 2pq\, w_{Aa} + q^2\, w_{aa}}$$

$$= p\, \frac{p\, w_{AA} + q\, w_{Aa}}{p^2\, w_{AA} + 2pq\, w_{Aa} + q^2\, w_{aa}}$$

$$= p\, \frac{\bar{w}_A}{\bar{w}}$$

# A numerical example of natural selection

| Genotypes: | AA | Aa | aa | |
|---|---|---|---|---|
| relative fitnesses: | 1 | 1 | 0.7 | (assume these are viabilities) |

Initial gene frequency of A = 0.2

Initial genotype frequencies (from Hardy−Weinberg)

| | | | |
|---|---|---|---|
| (newborns) | 0.04 | 0.32 | 0.64 |

# A numerical example of natural selection

| Genotypes: | AA | Aa | aa | |
|---|---|---|---|---|
| relative fitnesses: | 1 | 1 | 0.7 | (assume these are viabilities) |

Initial gene frequency of A = 0.2

Initial genotype frequencies (from Hardy−Weinberg)

| | | | |
|---|---|---|---|
| (newborns) | 0.04 | 0.32 | 0.64 |
| | x 1 | x 1 | x 0.7 |

Survivors (these are relative viabilities)

0.04 + 0.32 + 0.448 = Total: 0.808

# A numerical example of natural selection

| Genotypes: | AA | Aa | aa |
|---|---|---|---|
| relative fitnesses: | 1 | 1 | 0.7 (assume these are viabilities) |

Initial gene frequency of A = 0.2

Initial genotype frequencies (from Hardy−Weinberg)

| | | | |
|---|---|---|---|
| (newborns) | 0.04 | 0.32 | 0.64 |
| | x 1 | x 1 | x 0.7 |

Survivors (these are relative viabilities)

0.04 + 0.32 + 0.448 = Total: 0.808

genotype frequencies among the survivors: (divide by the total)

| | | |
|---|---|---|
| 0.0495 | 0.396 | 0.554 |

# A numerical example of natural selection

| Genotypes: | AA | Aa | aa |
|---|---|---|---|
| relative fitnesses: | 1 | 1 | 0.7 (assume these are viabilities) |

Initial gene frequency of A = 0.2

Initial genotype frequencies (from Hardy−Weinberg)

|  |  |  |  |
|---|---|---|---|
| (newborns) | 0.04 | 0.32 | 0.64 |
|  | x 1 | x 1 | x 0.7 |

Survivors (these are relative viabilities)

0.04 + 0.32 + 0.448 = Total: 0.808

genotype frequencies among the survivors: (divide by the total)

0.0495 0.396 0.554

gene frequency

A: 0.0495 + 0.5 x 0.396 = 0.2475

a: 0.554 + 0.5 x 0.396 = 0.7525

# A numerical example of natural selection

| Genotypes: | AA | Aa | aa |
|---|---|---|---|
| relative fitnesses: | 1 | 1 | 0.7 (assume these are viabilities) |

Initial gene frequency of A = 0.2

Initial genotype frequencies (from Hardy−Weinberg)

| | | | |
|---|---|---|---|
| (newborns) | 0.04 | 0.32 | 0.64 |
| | x 1 | x 1 | x 0.7 |

Survivors (these are relative viabilities)

0.04 + 0.32 + 0.448 = Total: 0.808

genotype frequencies among the survivors: (divide by the total)

0.0495 0.396 0.554

gene frequency

A: 0.0495 + 0.5 x 0.396 = 0.2475

a: 0.554 + 0.5 x 0.396 = 0.7525

genotype frequencies: (among newborns)

0.0613 0.3725 0.5663

## Mutation

(R. A. Fisher, 1922, J. B. S. Haldane, 1926-1927)

If there is no natural selection, but now we allow mutations between alleles: say at these rates:

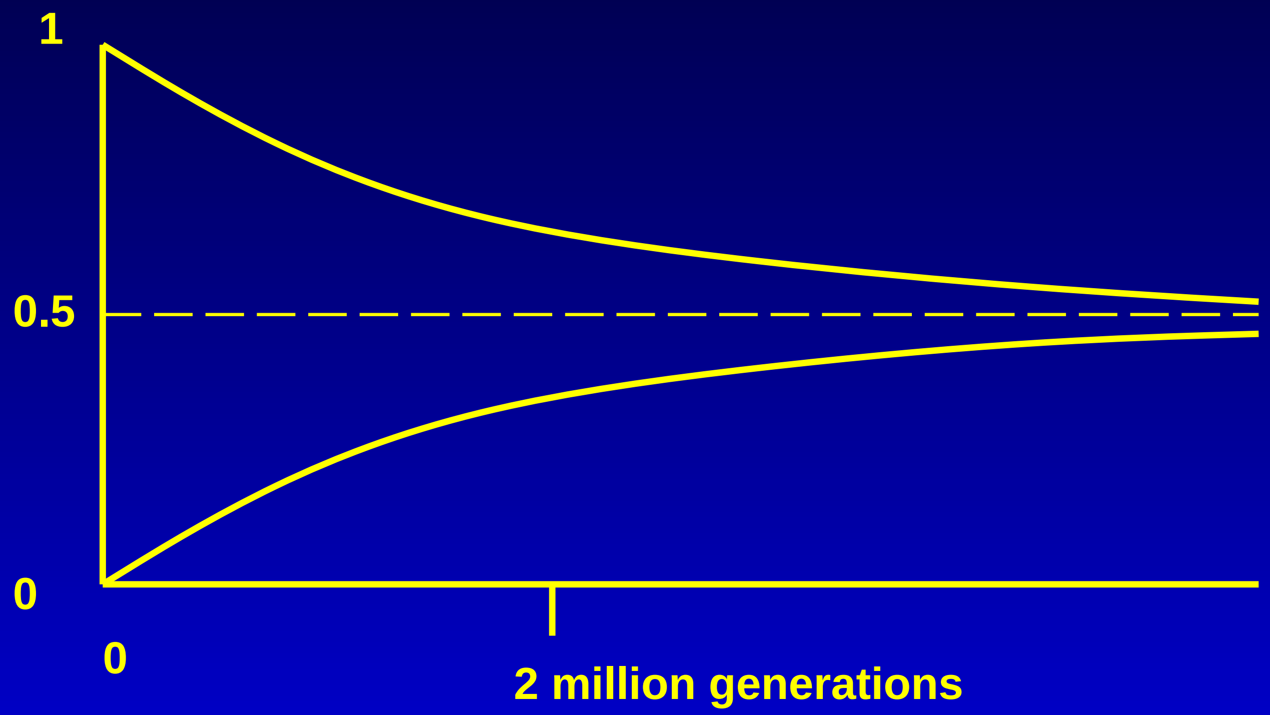
$$\mathsf{A} \quad \underset{\mathsf{v}}{\overset{\mathsf{u}}{\rightleftarrows}} \quad \mathsf{a}$$

Then the equation for change of gene frequency of A is

$$p' \;=\; p\,(1-u) + (1-p)\,v$$

# Mutation as an evolutionary force

If we have two alleles A and a, and mutation rate from A to a is $10^{-6}$ and mutation rate back is the same,



Mutation is critical in introducing new alleles but is very slow in changing their frequencies

# Mutation versus selection

What will happen if we have mutation from $A$ to $a$ at rate $u$, but natural selection against the homozygote $aa$, with fitness $1 - s$? If selection acts first, then mutation, we get equations:

$$\begin{aligned} p^* &= \frac{p}{1-(1-p)^2 s} \\ p' &= p^*(1-u) \end{aligned}$$

so that over the whole generation, the net result is

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

which yields when we reach an equilbrium where $p' = p$

$$p = 1 - \sqrt{\frac{u}{s}}$$

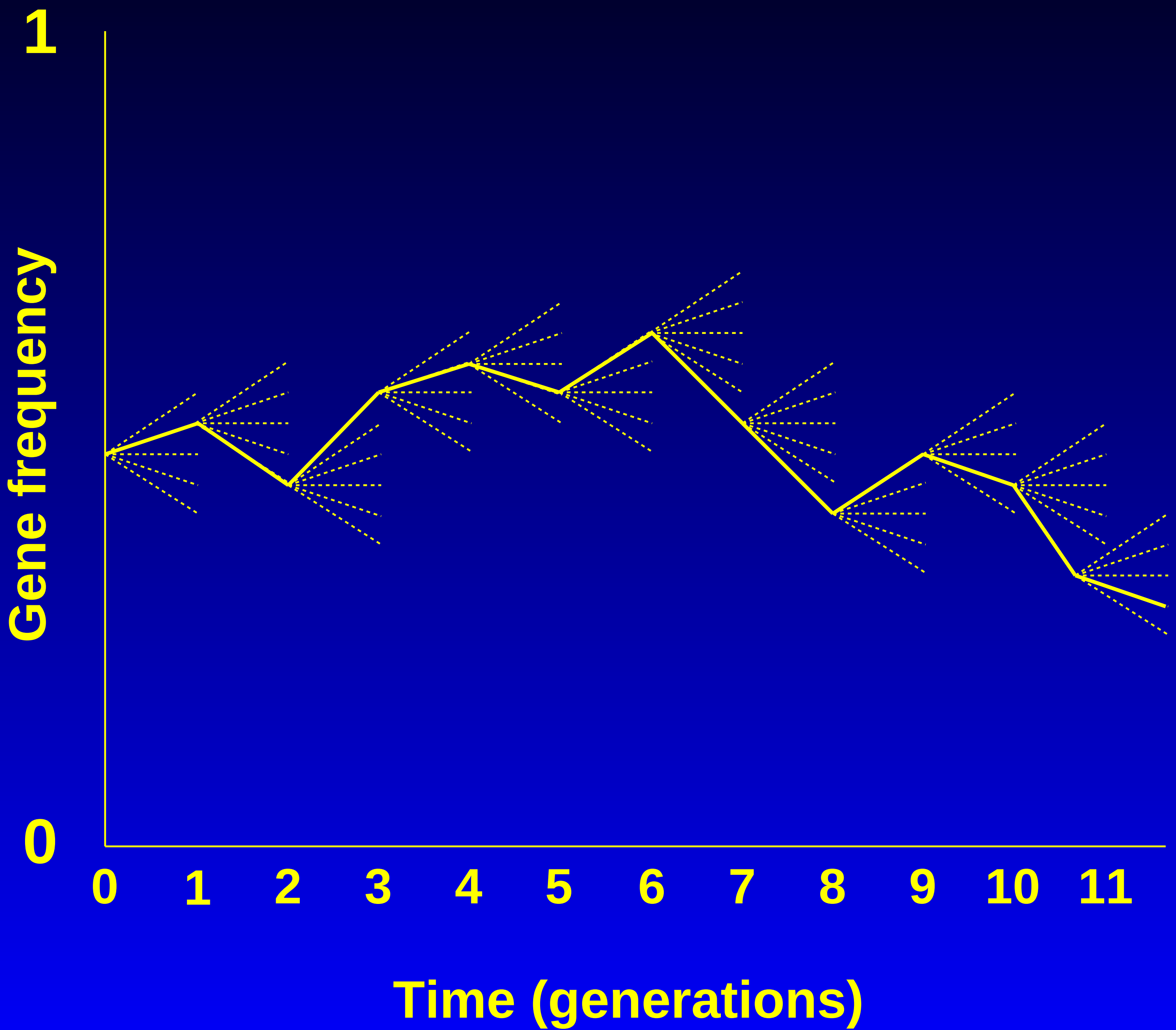
## A new result?

The preceding result with the square-root dependence on mutation rate, is due to Haldane (1929).

It is in all known population genetics texts.

I am told that it was rediscovered in the genetic algorithms / evolutionary computation literature in the 1990’s.

# Genetic drift



## Mathematics of genetic drift

A model of genetic drift is the Wright-Fisher model (Wright, 1932; Fisher, 1930)

In it, each individual produces an infinite and equal number of gametes, these go into a gamete pool, and the next generation is a draw of 2N gametes.

In effect, the gene frequency in the next generation is 2N draws (with replacement) from the gene frequency in the present generation.

The cumulative nature of genetic drift is that it starts each generation from the previous generation’s gene frequency. The original gene frequency for that population is forgotten. Genetic drift is thus a Markov Process, with no tendency of multiple generations to average out and bring the gene frequency back to the original gene frequency.

# Mathematics of genetic drift, continued

This gives the binomial transition probability, when $p = i/(2N)$

$$P_{ij} \;=\; \binom{2N}{j} p^j (1-p)^{2N-j} \;=\; \frac{(2N)!}{j!(2N-j)!} \left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j}$$

There has been a lot of work by Wright and others on the rate at which alleles fix or are lost and genetic variability is lost (in your terminology, “convergence"), using second moments arguments.

Basically it takes $4N$ generations for most genetic variability to be lost from a population.

## Diffusion equations

Later work (1945 on) used A. N. Kolmogorov's diffusion equations. Diffusion equations allow the incorporation of other evolutionary forces (work of Wright in 1932, Motoo Kimura, Geoff Watterson, and Warren Ewens).

They approximate (extremely accurately) the change of gene frequencies by taking the limit as the population size $N \to \infty$ and the deterministic parameters for selection, migration, etc. tend to 0, such that $Ns$, $Nu$, etc. are constant. At the same time the time scale is measured in units of $N$ generations.

Then one can write partial differential equations for the change of the distribution of gene frequencies over time, and also similar equations for the probability of fixation.

# Diffusion equations, continued

For example, with selection with fitnesses $1 : 1 - s$ in a haploid population of size $N$, the probability of ultimate fixation of allele $A$ is $U(p)$ when initial gene frequency is $p$, where

$$0 \;=\; Ns\, x(1-x)\,\frac{\partial U(x)}{\partial x} + \frac{1}{2}\,x(1-x)\,\frac{\partial^2 U(x)}{\partial x^2}$$

for which, with boundary conditions $U(0) = 0$ and $U(1) = 1$, we get the solution

$$U(p) \;=\; \frac{1 \;-\; e^{-2Ns\,p}}{1 \;-\; e^{-2Ns}}$$

## Diffusion equations

One upshot is that we can judge which evolutionary forces will have a greater impact by comparing $1/(4N)$, $s$, and $u$.

Thus if $N = 10^5$ while selection coefficients like $s = 0.0001$ and mutation rates $u = 10^{-8}$, we can say roughly that selection will have a large effect in the face of genetic drift, while mutation will be unable to keep variability in the population.

These calculations are relevant to controversies from 1932 on as to whether Sewall Wright's “Shifting Balance Theory" is valid. Wright suggested that adaptation would happen most readily if the species were split into local populations, each of moderate size, so that different populations could find different solutions to evolutionary problems, and then migration between populations could spread the best solutions.

## More than one locus

If there are more than one locus, we have many possible genotypes. With two loci, each with two alleles ($\mathsf{A}$, and $\mathsf{a}$, $\mathsf{B}$ and $\mathsf{b}$) there are 4 different possible gametes (or haploid genotypes – “haplotypes"), $\mathsf{AB}$, $\mathsf{Ab}$, $\mathsf{aB}$ and $\mathsf{ab}$.

If there is random mating forming a new generation, the same logic as one locus shows that a Hardy-Weinberg calculation works: we can generate all the genotype frequencies from the frequencies of these four haplotypes.

In the absence of selection, mutation, migration or genetic drift, Robbins (1918) showed that if a haplotype frequency is $\mathsf{P_{AB}}$, and if there is recombination with recombination frequency $\mathsf{r}$ between these two loci, since $(1 - \mathsf{r})$ of the haplotypes come through the generation unchanged:

$$\mathsf{P'_{AB}} \;=\; (1 - \mathsf{r})\,\mathsf{P_{AB}} + \mathsf{r}\,\mathsf{p_A}\mathsf{p_B}$$

and note that the individual gene frequencies $\mathsf{p_A}$ and $\mathsf{p_B}$ stay the same over generations.

# Linkage disequilibrium

It follows that if we make a measure (Robbins, 1918) which is the difference between the haplotype frequency and its random-combination prediction $p_A \, p_B$ ,

$$D \;=\; P_{AB} \;-\; p_A \, p_B$$

it is easy to show that

$$D' \;=\; (1 - r) \, D$$

This shows that as time goes by, any initial departure from these random-combination proportions (called “linkage equilibrium") disappears, although only a fraction $r$ of this “linkage disequilibrum $D$ is lost each generation.

# Linkage disequilibrium – a numerical example

Suppose we have 2 loci each with 2 alleles. The haplotype frequencies are:

|  | B | b | Total |
|---|---|---|---|
| A | 0.49 | 0.01 | 0.5 |
| a | 0.02 | 0.48 | 0.5 |
| Total | 0.51 | 0.49 | 1.0 |

The linkage disequilibrium $D = 0.49 - (0.5)(0.51) = 0.235$ and it will ultimately decline to $0$. The haplotype frequency will ultimately reach its linkage equilibrium value of $(0.5)(0.51) = 0.255$.

If the loci have recombination fraction 0.5 (as when they are on different chromosomes) the half-life of $D$ will be 1 generation, but if they are close enough on the same chromosome to have $r = 0.001$, the disequilibrium will have a half-life of 693 generations.

## Linkage and selection

Natural selection can be one of the forces creating linkage disequilibrium. There is then a balance between the creation by selection and its destruction by recombination.

Lewontin and Kojima (1960) worked out the equations of change of haplotype frequencies with selection and recombination. There were many papers on this in the 1960's.
If the organisms are haploid, like yeast (up to now we have had them be diploid, like us), application of selection, followed by recombination, can rather easily be shown to give:

$$P'_{AB} \;=\; P_{AB}\frac{w_{AB}}{\bar{w}} - r\left(\frac{P_{AB}w_{AB}}{\bar{w}}\,\frac{P_{ab}w_{ab}}{\bar{w}} - \frac{P_{Ab}w_{Ab}}{\bar{w}}\,\frac{P_{aB}w_{aB}}{\bar{w}}\right)$$

with three similar equations for the other three haplotypes.

# Linkage and selection

A classic example by Crow and Kimura (1965) imagines two loci with two alleles and fitnesses (in a haploid organism):

| | |
|---|---|
| $AB$ | $1+s$ |
| $Ab$ | $1-t$ |
| $aB$ | $1-t$ |
| $ab$ | $1$ |

This is the dilemma in which, starting with $ab$ and adding one capital-letter alleles makes things worse, but changing both alleles makes things better.

Will a $AB$ haplotype increase when rare in a population that is mostly haplotype $ab$? In an infinite population, it can; they prove that the condition is the recombination fraction $r$ be less than $s$.

## The Royal Road?

The preceding fitness scheme is related to the Royal Road scheme for fitnesses, at times popular in the Genetic Algorithms / Evolutionary Computation literature. If we set $t = 0$ we get

| | |
|---|---|
| AB | $1 + s$ |
| Ab | $1$ |
| aB | $1$ |
| ab | $1$ |

As in the previous slide, progress in this case is impeded by recombination. Although in an infinite population AB always takes over ultimately, it goes more slowly the more recombination, because AB haplotypes can't take over by themselves without generating Ab and aB haplotypes as well. (Eshel and Feldman, 1970; Felsenstein, 1965).

The same phenomenon is seen with more than 2 loci, but with even more effect of recombination.

## Genetic simulation

Use of computer simulation of evolution started with the original and very eccentric Nils Aall Barricelli in 1954 (he was independently wealthy and had attached himself to the Institute of Advanced Study).

A more noticed origin was the papers by Alex Fraser in Australian Journal of Biological Sciences in 1957. He simulated selection on multilocus quantitative characters with recombination.

By the early 1960s there were multiple papers on computer simulation of evolution by population geneticists. Two books appeared, Fraser and Burnell’s “Computer Models in Genetics" (1970) and John Crosby’s “Computer Simulation in Genetics" (1973).

What work on genetic algorithms added was the application of these methods to nonbiological optimization problems, though this is also what Barricelli did.

Many of the early papers have been reprinted by David Fogel in his book, “Evolutionary Computation: The Fossil Record" (IEEE Press, 1998).

# And what of the Schema Theorem?

The previous (Crow/Kimura) result looks a lot like John Holland's "schema theorem". If there are more loci, the schema theorem has two problems:

- The relative fitness of a schema will not remain constant as it becomes associated with alleles at other loci
- The condition is too conservative; it allows for loss of the schema by recombination but not for assembly of it by recombination

A critique of it will be found in:

Altenberg, L. 1995. The Schema theorem and Price's theorem, pp. 23-49 in *Foundations of Genetic Algorithms 3*, ed. L. D. Whitley and M. D. Vose. Morgan Kaufmann, San Francisco.

## Why does recombination work so well, then?

If recombination tends to break up favorable haplotypes, why does it help in search for solutions?

Hill and Robertson, 1966, found an effect that explains this. When there is selection and recombination in a finite population, genetic drift creates random linkage disequibria. Some are helpful (good alleles associated with each other), some harmful (good alleles competing with each other).

Interestingly, the net effect is not zero. The net effect is to slow down the progress of selection. Recombination, by breaking down these disequibria, allows selection to be more effective.

This Hill-Robertson effect includes phenomena such as “hitch-hiking" and “Muller’s ratchet".

## How it was done

This projection produced

- 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.