

Population Genetics of Selection

Jay Taylor

School of Mathematical and Statistical Sciences
Arizona State University

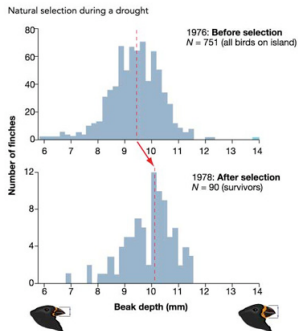
Evolution by Natural Selection

Darwin and Wallace (1859) observed that heritable traits that increase reproductive success will become more common in a population.

- **Variation within populations** - individuals have different traits (phenotypes).
 - height and weight are approximately normally distributed
 - variation for susceptibility to HIV-1 infection and progression to AIDS
- **Selection** - traits influence fecundity and survivorship (**fitness**).
 - larger body size may be beneficial in cold environments
 - height may influence mating success (sexual selection)
- **Heritability** - offspring are similar to their parents.
 - variation has both environmental and heritable components
 - differences in height are partly heritable, but are also influenced by childhood nutrition

Example: Beak size in the Medium Ground Finch (*Geospiza fortis*)

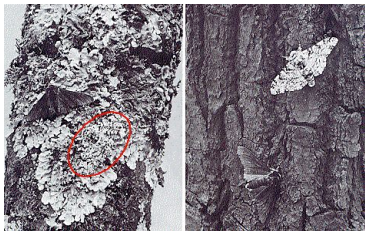
- Restricted to the Galapagos Islands.
- Forages mainly on seeds.
- Large seeds are handled more efficiently by birds with larger bills.
- Large seeds predominate following drought years (e.g., 1977).



Example: The rise and fall of the peppered moth (*Biston betularia*)

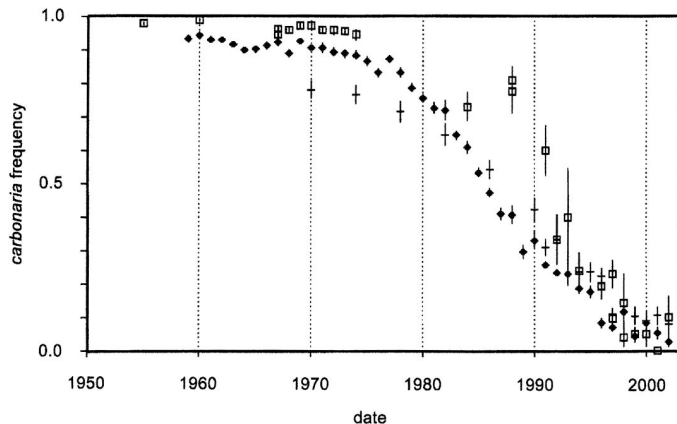
The peppered moth has two color morphs:

- white (wild-type)
- black (*carbonaria*)



- The black morph was rarely recorded at the beginning of the 19th century.
- *carbonaria* became common first in Lancashire/Yorkshire in the 1850's, then spread in urban areas throughout the UK.
- Similar increases of melanic forms occurred on the continent and in NA.
- Melanic forms appear to be favored in industrialized regions due to soot deposition on trees and declines in lichen cover.

With the decline in coal usage and the enactment of clean air legislation in the 1970's, the frequency of the melanic morph decreased in the UK:



source: Cook (2003)

The most serious weakness of Darwin's theory was his model of heredity, which was based on:

- **blending inheritance** - offspring traits are averages of parental traits.

This is problematic because it leads to a loss of variation.

Unknown to Darwin, Mendel (1859) proposed a particulate model of inheritance:

- Traits are determined by **genes**.
- Each gene can have finitely-many different types called **alleles**.
- Different alleles may produce different traits.
- Offspring are similar to their parents because they inherit their genes.

Mendel was essentially correct, but his work was largely ignored for 40 years.

A coherent theory explaining how natural selection could operate in the context of Mendelian genetics did not develop until the 1930's with the development of theoretical population genetics (Fisher, Wright, Haldane). This led to the **Modern Synthesis**:

- Genes are physical entities carried on chromosomes.
- Heritable variation is produced by mutation and recombination.
- Continuous variation can arise from the contribution of many loci of small effect.
- Selection causes changes in the frequencies of genotypes that in turn affect traits that influence fitness.
- Population genetics can explain both microevolutionary and macroevolutionary changes.

Population genetics focuses on understanding **evolution at the molecular level**: how does natural selection affect the dynamics of gene frequencies?

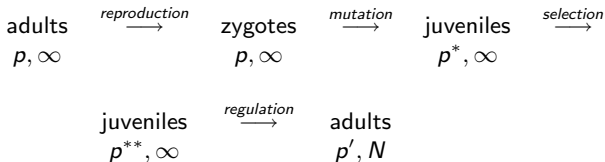
Some key questions about selection and adaptation are:

- What fraction of the genome is under selection?
- How frequently does selection lead to changes in the genome?
- How often are deleterious mutations fixed in a population?
- How do demography and life history influence the rate of adaptation?
- Does adaptation rely mainly on standing variation or on new mutations?
- Does adaptation occur through the fixation of many mutations of small effect or through the fixation of a few mutations of large effect?

A Wright-Fisher model with directional selection and mutation

Assumptions:

- non-overlapping generations (e.g., an annual plant)
- N haploid adults in each generation
- genotypes A_1, A_2 with frequencies p and $1 - p$
- mutation from A_i to A_j at rate μ_{ij}
- relative fitness of $A_1 : A_2$ is $1 + s : 1$
- population regulation by binomial sampling



To say that $A_1 : A_2$ have relative fitness $1 + s_1 : 1 + s_2$ means that, on average, each individual with genotype A_1 contributes $1 + s_1$ offspring to the next generation for every $1 + s_2$ offspring contributed by an individual with genotype A_2 .

In an infinite population subject only to selection, the frequency of A_1 changes from p to p' given by

$$p' = \frac{p(1 + s_1)}{p(1 + s_1) + (1 - p)(1 + s_2)} = \frac{p(1 + s_1)}{1 + ps_1 + (1 - p)s_2}.$$

- The denominator of this expression is the mean fitness of the population, weighted by the allele frequencies.
- The quantities s_1 and s_2 are called **selection coefficients**.

In the full model, the allele frequencies at the different stages of the life cycle are given by the following equations:

$$p^* = p(1 - \mu_{12}) + (1 - p)\mu_{21} \quad (\text{mutation})$$

$$p^{**} = \frac{p^*(1 + s)}{p^* \cdot (1 + s) + (1 - p^*) \cdot 1} \quad (\text{viability selection})$$

$$p' \sim \frac{1}{N} \cdot \text{Binomial}(N, p^{**}) \quad (\text{regulation}).$$

Remark: We say that selection is **directional** or **purifying** in this model because the same allele is always favored and tends to increase in frequency:

- A_1 is favored if $s > 0$
- A_2 is favored if $s < 0$

Recall that for the neutral W-F model, the variance of the change of allele frequencies over one generation is of order $\frac{1}{N}$:

$$\mathbb{E}_p \left[(p^N(1) - p)^2 \right] = \frac{1}{N} p(1-p) + O\left(\frac{1}{N^2}\right)$$

(assuming $\mu_{ij} = s = 0$).

This is also true under mutation and selection, provided that we choose the mutation rates and selection coefficient so that the expected change of p over over one generation is also of order $\frac{1}{N}$. This requirement motivates the following assumption:

$$\mu_{ij} \equiv \mu_{ij}^{(N)} = \frac{\theta_{ij}}{N} \quad \text{and} \quad s \equiv s^{(N)} = \frac{\sigma}{N}.$$

With these scalings, we have

$$\begin{aligned}\mathbb{E}_p \left[p^N(1) - p \right] &= \frac{1}{N} \left((1-p)\theta_{21} - p\theta_{12} + \sigma p(1-p) \right) + O \left(\frac{1}{N^2} \right) \\ \mathbb{E}_p \left[(p^N(1) - p)^2 \right] &= \frac{1}{N} p(1-p) + O \left(\frac{1}{N^2} \right) \\ \mathbb{E}_p \left[(p^N(1) - p)^e \right] &= O \left(\frac{1}{N^2} \right) \quad (e \geq 3).\end{aligned}$$

However, this shows that the process $(p^N(\lfloor Nt \rfloor) : t \geq 0)$ converges to a diffusion process with infinitesimal mean and variance coefficients

$$\begin{aligned}m(p) &= \theta_{21}(1-p) - \theta_{12}p + \sigma p(1-p) \\ v(p) &= p(1-p).\end{aligned}$$

Transition Semigroups and Infinitesimal Generators

Suppose that $X = (X_t : t \geq 0)$ is a continuous-time Markov process with values in \mathbb{R} . The **transition semigroup** of X is the family of operators $(T_t; t \geq 0)$ on the space of bounded continuous functions $f : (R) \rightarrow \mathbb{R}$ defined by

$$T_t f(x) = \mathbb{E}[f(X_t) | X_0 = x] \equiv \mathbb{E}_x[f(X_t)].$$

We say that this family is a semigroup because $T_0 = Id$ is the identity operator and because it satisfies the following property

$$T_{t+s} = T_t \circ T_s$$

for all $t, s \geq 0$. This is a consequence of the Markov property of X .

It is also useful to consider the **infinitesimal generator** of X , which is defined by

$$Gf(x) = \lim_{t \rightarrow 0} \frac{T_t f(x) - f(x)}{t}$$

for any f such that the limit exists for all x .

For a one-dimensional diffusion process with infinitesimal mean and variance coefficients $m(x)$ and $v(x)$, the generator has the form

$$Gf(x) = \frac{1}{2}v(x)f''(x) + m(x)f'(x)$$

for any function f for which the derivatives f' , f'' exist and are bounded.

Fixation Probabilities

Suppose that both mutation rates $\theta_{12} = \theta_{21} = 0$. Then the ultimate fate of the allele A_1 is to either be lost from or fixed in the population. Let

$$\tau = \inf\{t \geq 0 : p(t) = 0 \text{ or } 1\}$$

be the time when this event occurs and define

$$u(p) = \mathbb{P}_p\{p(\tau) = 1\}$$

to be the fixation probability of A_1 given that its initial frequency is p .

Question: We know that if A_1 and A_2 are neutral, then $u(p) = p$. How does selection alter this fixation probability?

Using the Markov property of the diffusion process, it can be shown that

$$u(p) = T_t u(p) = \mathbb{E}_p[u(p(t))].$$

This implies that

$$\begin{aligned} Gu(p) &= \lim_{t \rightarrow 0} \frac{T_t u(p) - u(p)}{t} \\ &= \lim_{t \rightarrow 0} \frac{u(p) - u(p)}{t} \\ &= 0 \end{aligned}$$

subject to the boundary conditions

$$u(0) = 0 \quad \text{and} \quad u(1) = 1.$$

Remark: It can also be shown that the process $(u(p(t)) : t \geq 0)$ is a martingale, in which case the optional sampling theorem can be used to deduce that $u(p)$ is the fixation probability of A_1 .

For the W-F diffusion with selection, we need to solve the equation

$$Gu(p) = \frac{1}{2}p(1-p)u''(p) + \sigma p(1-p)u'(p) = 0,$$

i.e., $u''(p) + 2\sigma u'(p) = 0,$

with $u(0) = 0$ and $u(1) = 1$.

The solution can be found by integrating, leading to the following expression for the fixation probability of a selected allele:

$$u(p) = \frac{1 - e^{-2\sigma p}}{1 - e^{-2\sigma}} = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}.$$

The most important case is when a single copy of a new allele is introduced into a population, either by mutation or immigration. Then the initial frequency is $p = 1/N$ and the fixation probability of the new allele is

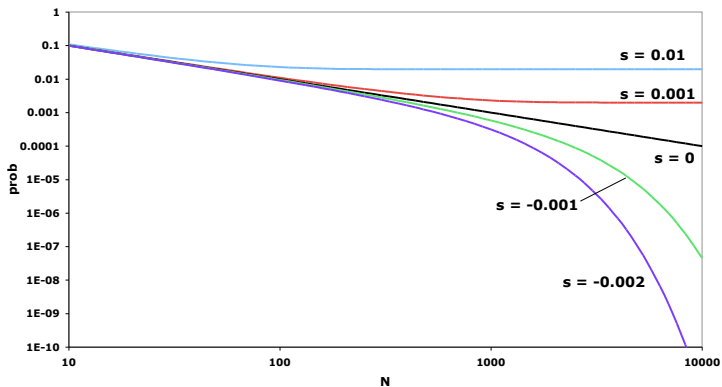
$$u\left(\frac{1}{N}\right) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \approx \begin{cases} 2s & \text{if } N^{-1} \ll s \ll 1 \\ 2|s|e^{-2N|s|} & \text{if } -1 \ll s \ll -N^{-1} \end{cases}$$

In particular, this shows that

- Novel beneficial mutations are likely to be lost from a population;
- Deleterious mutations can be fixed, but only if $N|s|$ is not too large;
- Selection is dominated by genetic drift when $|s| < \frac{1}{N}$.

Key result: Selection is more effective in larger populations.

Fixation Probabilities of New Mutants



Selective Constraints and Divergence

One prediction of this theory is that sites that are under **purifying selection** should diverge more slowly than neutrally evolving sites.

The **degeneracy** of the genetic code illustrates this effect.

- Amino acids are encoded by triplets of DNA bases called **codons**.
- There are $64 = 4^3$ different codons, but only 20 amino acids.
- On average, there are 3 different codons per amino acid.

It follows that there are two kinds of mutations in coding DNA:

- (i) A **non-synonymous** mutation is one that changes an amino acid.
- (ii) A **synonymous** substitution is one that changes only the DNA sequence.

		SECOND POSITION					
		T	C	A	G		
FIRST POSITION	T	TTT (F)	TCT (S)	TAT (Y)	TGT (C)	T	
		TTC (F)	TCC (S)	TAC (Y)	TGC (C)	C	
		TTA (L)	TCA (S)	TAA STOP	TGA STOP	A	
		TTG (L)	TCG (S)	TAG STOP	TGG (W)	G	
	C	CTT (L)	CCT (P)	CAT (H)	CGT (R)	T	
		CTC (L)	CCC (P)	CAC (H)	CGC (R)	C	
		CTA (L)	CCA (P)	CAA (Q)	CGA (R)	A	
		CTG (L)	CCG (P)	CAG (Q)	CGG (R)	G	
	A	ATT (I)	ACT (T)	AAT (N)	AGT (S)	T	
		ATC (I)	ACC (T)	AAC (N)	AGC (S)	C	
		ATA (I)	ACA (T)	AAA (K)	AGA (R)	A	
		ATG (M)	ACG (T)	AAG (K)	AGG (R)	G	
	G	GTT (V)	GCT (A)	GAT (D)	GGT (G)	T	
		GTC (V)	GCC (A)	GAC (D)	GGC (G)	C	
		GTA (V)	GCA (A)	GAA (E)	GGA (G)	A	
		GTG (V)	GCG (A)	GAG (E)	GGG (G)	G	

- The mutation $TTT \rightarrow TTC$ is synonymous.
- The mutation $TTT \rightarrow TTA$ is non-synonymous because the amino acid changes from phenylalanine (F) to Leucine (L).

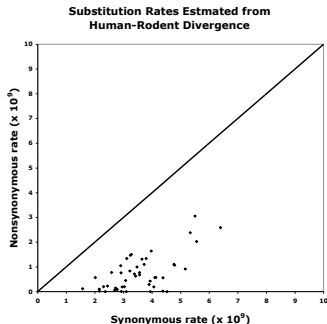
Hypothesis: In general, it is thought that non-synonymous mutations are more likely to be deleterious than synonymous mutations because they can change protein structure and function.

Prediction: If true, then synonymous substitution rates should be higher than non-synonymous substitution rates.

This is, in fact, what is observed:

	syn (yr^{-1})	non-syn (yr^{-1})	ratio (syn/non-syn)
influenza A	13.1×10^{-3}	3.5×10^{-3}	3.8
HIV-1	9.7×10^{-3}	1.7×10^{-3}	5.7
Hepatitis B	4.6×10^{-5}	1.5×10^{-5}	3.1
<i>Drosophila</i>	15.6×10^{-9}	1.9×10^{-9}	8.2
human-rodent	3.51×10^{-9}	0.74×10^{-9}	4.7

The following plot shows synonymous and nonsynonymous substitution rates estimated from comparisons of human and rodent genes. In every case the non-synonymous substitution rate is less than the synonymous substitution rate.



Average substitution rates:

- Synonymous: 3.51
- Non-synonymous: 0.74

Source: Li (1997)

Stationary Distributions, Ergodicity and Polymorphism

- If $\theta_{12}, \theta_{21} > 0$, then the boundaries $p = 0, 1$ are no longer absorbing states:

$$m(0) = \theta_{21} > 0 \quad \text{and} \quad m(1) = -\theta_{12} < 0.$$

- Instead, **recurrent mutation** between A_1 and A_2 continually introduces new variation into the population.
- In turn, this variation is eroded both by genetic drift and selection.

Question: What can we say about the long-term distribution of allele frequencies in a population subject to mutation, drift and selection?

One way to address this question is to study the stationary distribution of the diffusion approximation.

Definition: We say that a distribution $\pi(dx)$ is a **stationary distribution** for a Markov process $X = (X_t; t \geq 0)$ if whenever the initial distribution of X_0 is $\pi(dx)$, then the marginal distribution of X_t is $\pi(dx)$ for every $t \geq 0$.

In many cases, it can be shown that such a process is **ergodic**, meaning that for any initial value x the marginal distributions tend to the stationary distribution:

$$P(t; x, dy) \xrightarrow{w} \pi(dy),$$

as $t \rightarrow \infty$. This means that for any bounded continuous function f ,

$$\lim_{t \rightarrow \infty} \mathbb{E}_x[f(X_t)] = \int f(x)\pi(dx).$$

Thus the stationary distribution of an ergodic process tells us something about the typical long-term behavior of that process.

Remark: It can be shown that the Wright-Fisher diffusion with reciprocal mutation is ergodic and has a unique stationary distribution with density $\pi(p)$. We can identify this density using the following procedure.

We first note that because the marginal distributions of a stationary process are constant, we have

$$\begin{aligned} \int_0^1 T_t f(p) \pi(p) dp &= \mathbb{E}_\pi [f(p(t))] \\ &= \mathbb{E}_\pi [f(p(0))] \\ &= \int_0^1 f(p) \pi(p) dp \end{aligned}$$

for all $t \geq 0$.

Assuming that we can interchange the integration and the limit, this implies that

$$\begin{aligned}
 \int_0^1 Gf(p)\pi(p)dp &= \int_0^1 \lim_{t \rightarrow 0} \frac{T_t f(p) - f(p)}{t} \pi(p) dp \\
 &= \lim_{t \rightarrow 0} \frac{1}{t} \int_0^1 (T_t f(p) - f(p)) \pi(p) dp \\
 &= \lim_{t \rightarrow 0} \left\{ \int_0^1 T_t f(p) \pi(p) dp - \int_0^1 f(p) \pi(p) dp \right\} \\
 &= 0
 \end{aligned}$$

for any function f for which Gf is defined. In particular, this holds when f vanishes in a neighborhood of $p = 0$ and $p = 1$, in which case integration by parts gives

$$\int_0^1 f(p) \left\{ \frac{1}{2} (v(p)\pi(p))'' - (m(p)\pi(p))' \right\} dp = 0.$$

However, it can be shown that if this previous identity holds for all such f , then the density $\pi(p)$ must be a solution to the differential equation

$$\frac{1}{2}(v(p)\pi(p))'' - (m(p)\pi(p))' = 0.$$

This equation can be integrated twice to give

$$\pi(p) = \frac{1}{C} \frac{1}{v(p)} \exp\left(2 \int_c^p \frac{m(q)}{v(q)} dq\right),$$

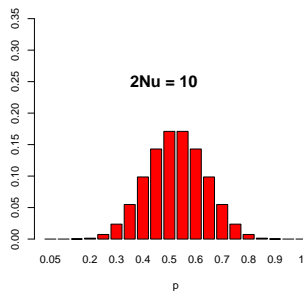
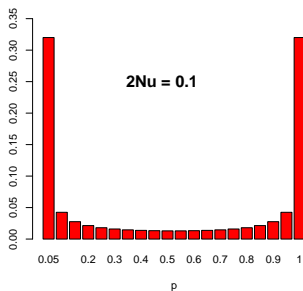
where the normalizing constant $C < \infty$ must be chosen (if possible) so that

$$\int_0^1 \pi(p) dp = 1.$$

Neutral Variation: The stationary distribution of the neutral Wright-Fisher diffusion with mutation is just the Beta distribution with parameters $2\theta_1$ and $2\theta_2$, which has density

$$\begin{aligned}
 \pi(p) &= \frac{1}{C} \frac{1}{2p(1-p)} \exp \left\{ \int_c^p \frac{(2\theta_1(1-q) - 2\theta_2q)}{q(1-q)} dq \right\} \\
 &= \frac{1}{C} \frac{1}{p(1-p)} \exp \left\{ 2\theta_1 \ln(p) + 2\theta_2 \ln(1-p) \right\} \\
 &= \frac{1}{\beta(2\theta_1, 2\theta_2)} p^{2\theta_1-1} (1-p)^{2\theta_2-1} \\
 &= \frac{1}{\beta(2N\mu_1, 2N\mu_2)} p^{2N\mu_1-1} (1-p)^{2N\mu_2-1}.
 \end{aligned}$$

The neutral stationary distribution reflects the competing effects of genetic drift, which eliminates variation, and mutation, which generates variation.



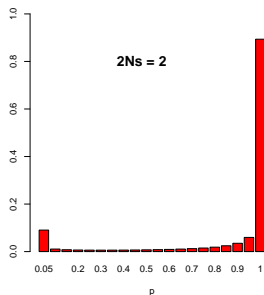
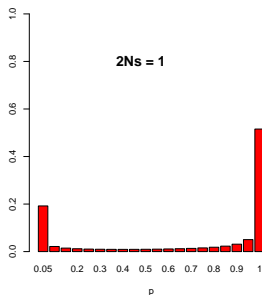
- When $2N\mu_1, 2N\mu_2 > 1$, mutation dominates drift and the stationary distribution is peaked about its mean (both alleles are common).
- When $2N\mu_1, 2N\mu_2 < 1$, drift dominates mutation and the stationary distribution is bimodal, with peaks at the boundaries (one allele is common and one rare).

With selection and mutation, the density of the stationary distribution is

$$\pi(p) = \frac{1}{C} p^{2\theta_{21}-1} (1-p)^{2\theta_{12}-1} e^{2\sigma p}.$$

Purifying selection has two consequences:

- It shifts the stationary distribution in the direction of the favored allele.
- It tends to reduce the amount of variation present at the selected locus.

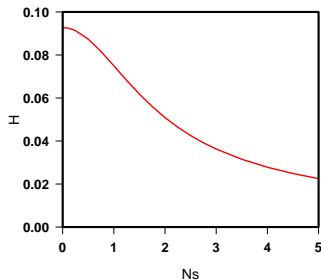


Genetic variation is often summarized by a statistic called the **heterozygosity** (H) or **nucleotide diversity** (π):

$$H = \mathbb{P} \{ \text{a random sample of two individuals contains two different alleles} \}$$

$$\equiv \int_0^1 2p(1-p)\pi(p)dp.$$

The figure below shows that directional selection reduces heterozygosity.



Purifying Selection and Polymorphism in Coding Regions

Prediction: If synonymous mutations are generally under weaker purifying selection than non-synonymous mutations, then we would expect synonymous diversity to be greater than non-synonymous diversity.

This is what is seen:

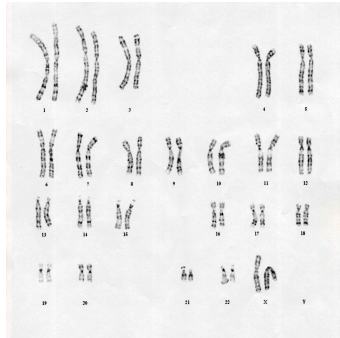
	syn (H)	non-syn (H)	ratio (syn/non-syn)
<i>D. melanogaster</i>	0.0054	0.00038	14.2
humans (US)	0.0005	0.0001	5

Selection in Diploid Populations

Thus far we have focused on directional or purifying selection in a haploid population. However, many organisms are **diploid** throughout much of their life cycle, i.e., most chromosomes are present in two copies per genome.

The Human Karyotype:

- 46 chromosomes
- 22 pairs of autosomes
- females have two X chromosomes
- males have one X and one Y
- haploid gametes (sperm and eggs) are produced by **meiosis**



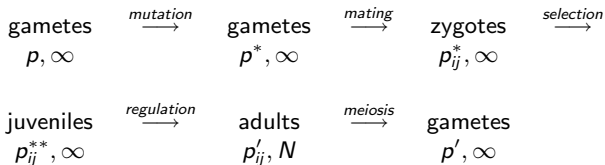
Genetic variation in diploids: If there are two alleles present at a locus, then there are three possible genotypes:

- **homozygotes:** A_1A_1 and A_2A_2
- **heterozygotes:** A_1A_2 ($= A_2A_1$)

To model selection in such a population, we need to know the fitness of each of the three diploid genotypes.

genotype	relative fitness
A_1A_1	$1 + s_{11}$
A_1A_2	$1 + s_{12}$
A_2A_2	$1 + s_{22}$

We will assume that the organism has the following life history:



Provided that mating is random, it suffices to track the changes in the **genetic** frequencies of A_1 from generation to generation. The transition probabilities for $p \rightarrow p'$ can be calculated by determining how p changes at each stage.

Suppose that the gametic frequency of A_1 in generation t is $p^N(t) = p$.

Mutation: Each A_i gamete mutates to A_j with probability μ_{ij} . This changes the frequency of A_1 from p to p^* :

$$p^* = p(1 - \mu_{12}) + (1 - p)\mu_{21}.$$

Random mating: Because mating is random and the number of gametes is assumed to be infinite, the frequencies of the diploid genotypes immediately following mating are in **Hardy-Weinberg** equilibrium:

genotype	frequency
A_1A_1	$p_{11} = (p^*)^2$
A_1A_2	$p_{12} = 2p^*(1 - p^*)$
A_2A_2	$p_{22} = (1 - p^*)^2$

Selection: Selection causes the frequency of each genotype to change in proportion to its relative fitness. If p_{ij}^* is the frequency of A_iA_j before selection, then the frequency p_{ij}^{**} after selection is

$$p_{ij}^{**} = p_{ij}^* \left(\frac{w_{ij}}{\bar{w}} \right),$$

where w_{ij} is the relative fitness of A_iA_j and \bar{w} is the mean fitness of the population:

$$\begin{aligned} \bar{w} &= (p^*)^2(1 + s_{11}) + 2p^*(1 - p^*)(1 + s_{12}) + (1 - p^*)^2(1 + s_{22}) \\ &= 1 + (p^*)^2s_{11} + 2p^*(1 - p^*)s_{12} + (1 - p^*)^2s_{22}. \end{aligned}$$

Consulting the table of relative fitnesses, we find:

genotype	frequency after selection
A_1A_1	$p_{11}^{**} = p_{11}^*(1 + s_{11})/\bar{w}$
A_1A_2	$p_{12}^{**} = p_{12}^*(1 + s_{12})/\bar{w}$
A_2A_2	$p_{22}^{**} = p_{22}^*(1 + s_{22})/\bar{w}$

Population regulation: Population regulation occurs as in the Wright-Fisher model: the N adults are randomly sampled from the juvenile cohort. However, because the species is diploid, we are sampling $2N$ genes in total.

Suppose that p'_{ij} denotes the frequency of $A_i A_j$ genotypes following population regulation. Then, the numbers of adults of each of the three genotypes has a Multinomial distribution:

$$N(p'_{11}, p'_{12}, p'_{22}) \sim \text{Multinomial}(N, p_{11}^{**}, p_{12}^{**}, p_{22}^{**})$$

Meiosis: The final stage is meiosis, during which each adult produces an effectively infinite number of haploid gametes. Whereas $A_1 A_1$ adults produce only A_1 gametes and $A_2 A_2$ adults produce only A_2 gametes, $A_1 A_2$ adults produce an equal mixture of A_1 and A_2 gametes. It follows that the gametic frequency of A_1 in generation $t + 1$ is equal to:

$$p^N(t + 1) = p' = p'_{11} + \frac{1}{2}p'_{12}.$$

To derive a diffusion approximation for this model, we must assume that selection and mutation are both of order $O(1/2N)$:

$$\begin{aligned}\mu_{ij} &\equiv \mu_{ij}^{(N)} = \frac{\theta_{ij}}{2N} \\ s_{ij} &\equiv s_{ij}^{(N)} = \frac{\sigma_{ij}}{2N}.\end{aligned}$$

With these scalings, a tedious but straightforward calculation shows that

$$\begin{aligned}\mathbb{E}_p[\delta] &= \frac{1}{2N} \left[\theta_{21}(1-p) - \theta_{12}p + \left\{ \frac{1}{2}(\sigma_{11} - \sigma_{22}) + (1-2p)(\sigma_{12} - \bar{\sigma}) \right\} p(1-p) \right] \\ &\quad + O(N^{-2}) \\ \mathbb{E}_p[\delta^2] &= \frac{1}{2N} p(1-p) + O(N^{-2}) \\ \mathbb{E}_p[\delta^e] &= O(N^{-2}) \text{ if } e \geq 3,\end{aligned}$$

where $\delta = p' - p$ and $\bar{\sigma} = \frac{1}{2}(\sigma_{11} + \sigma_{22})$.

It follows that the processes $(p^N(\lfloor 2Nt \rfloor) : t \geq 0)$ converge to a diffusion process with infinitesimal variance and drift coefficients

$$v(p) = p(1-p)$$

$$m(p) = \theta_{21}(1-p) - \theta_{12}p + \left\{ \frac{1}{2}(\sigma_{11} - \sigma_{22}) + (1-2p)(\sigma_{12} - \bar{\sigma}) \right\} p(1-p)$$

Remarks:

- We have rescaled time by a factor of $2N$ rather than N because there are $2N$ genes in a diploid population with N individuals.
- The infinitesimal variance of the diffusion approximation is then the same as that for a haploid Wright-Fisher model with N individuals:

$$v(p) = p(1-p)$$

- The infinitesimal drift of the diffusion approximation can be written as

$$m(p) = \theta_{21}(1-p) - \theta_{12}p + \sigma(p)p(1-p)$$

where $\sigma(p) \equiv \frac{1}{2}(\sigma_{11} - \sigma_{22}) + (1-2p)(\sigma_{12} - \bar{\sigma})$.

- If $\sigma_{12} = \bar{\sigma}$, then $\sigma(p) = \frac{1}{2}(\sigma_{11} - \sigma_{22})$ is constant, as in the haploid model. (This is called **genic** selection.)
- Otherwise, selection is **frequency-dependent**.

The **marginal fitness** of an allele is equal to the average of the fitnesses of the genotypes containing that allele, weighted by the frequencies of those genotypes:

$$\begin{aligned} w_{A_1} &= p(1 + s_{11}) + (1-p)(1 + s_{12}) = 1 + ps_{11} + (1-p)s_{12} \\ w_{A_2} &= (1-p)(1 + s_{22}) + p(1 + s_{12}) = 1 + ps_{12} + (1-p)s_{22} \\ w_{A_1} - w_{A_2} &= \frac{1}{2}(s_{11} - s_{22}) + (1-2p)(s_{12} - \bar{s}) \end{aligned}$$

Inbreeding Depression and Recessive Deleterious Alleles

Suppose that A_1 is deleterious compared to A_2 and that the selection coefficients have the form

$$\sigma_{11} = -\sigma$$

$$\sigma_{12} = -h\sigma$$

$$\sigma_{22} = 0,$$

where $\sigma > 0$ and $h \in [0, 1]$.

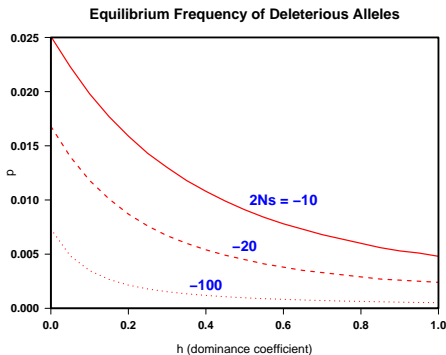
The constant h is called the **dominance coefficient** because it quantifies the contribution of the A_1 allele to the fitness of the heterozygote. A_1 is said to be

- **dominant** if $h \in (1/2, 1]$
- **recessive** if $h \in [0, 1/2)$
- **additive** if $h = 1/2$.

In this case, the stationary distribution of the diffusion process has density

$$\pi(p) = \frac{1}{C} p^{2\theta_{21}-1} (1-p)^{2\theta_{12}-1} e^{-\sigma p + (1-2h)\sigma p(1-p)}.$$

Because the exponent is a decreasing function of h , recessive deleterious alleles tend to be more common than dominant deleterious alleles.



Many species have mechanisms that reduce the likelihood of inbreeding. These have probably evolved to avoid **inbreeding depression** caused by recessive deleterious alleles.

- Many deleterious alleles are **loss-of-function** mutations that are recessive because a single functional copy of the gene produced enough of the protein.
- Recessive deleterious alleles can rise to significant frequencies because they are shielded from selection in heterozygotes.
- If the frequency of such an allele is $p \ll 1$, then the frequency of deleterious homozygotes in an outbred population is $p^2 \ll 1$.
- In contrast, the frequency of such homozygotes in a cross between two sibs will be (approximately) $2p \times \frac{1}{2} = p \gg p^2$.
- Thus, inbred individuals are much more likely to suffer from heritable diseases caused by recessive deleterious alleles.

Overdominance and Balancing Selection

If the fitness of the heterozygote is greater than the fitness of either homozygote, i.e., if $\sigma_{12} > \sigma_{11}, \sigma_{22}$, then the heterozygote is said to be **overdominant**. In this case, there is an intermediate frequency

$$\bar{p} = \frac{\sigma_{12} - \sigma_{22}}{2\sigma_{12} - \sigma_{11} - \sigma_{22}} \in (0, 1),$$

such that

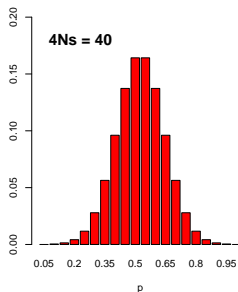
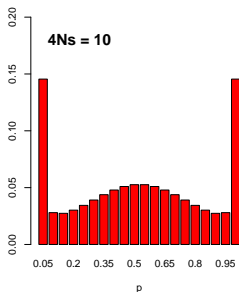
- $\sigma(\bar{p}) = 0$ (both alleles are equally fit)
- $\sigma(p) > 0$ if $p < \bar{p}$ (A_1 is more fit)
- $\sigma(p) < 0$ if $p > \bar{p}$ (A_2 is more fit)

Thus, A_1 tends to rise in frequency when rare and tends to decrease when common. This kind of selection is called **balancing selection** because it maintains genetic variation in the population.

The tendency of balancing selection to maintain variation can be seen in the density of the stationary distribution for this diffusion:

$$\pi(p) = \frac{1}{C} p^{2\mu_1-1} q^{2\mu_2-1} e^{(2\sigma_{12}-\sigma_{11})p(2\bar{p}-p)}.$$

Symmetric Balancing Selection: In the following histograms, $\sigma_{11} = \sigma_{22} = 0$, $2\sigma_{12} = 4Ns$, and $4N\mu = 0.1$.



Example: The classic example of overdominance is the **sickle cell** mutation that is prevalent in some human populations with a high incidence of malaria infections. This is an amino-acid changing mutation which causes hemoglobin molecules to clump together.

There are two alleles - A which is the non-sickle-cell ('wild type') allele and S which causes sickling of red blood cells. The diploid genotypes and their phenotypes are:

- AA : These individuals have normal hemoglobin, but are susceptible to malaria infections (which can be fatal in children and pregnant women).
- AS : These individuals have a mild form of anemia but are very resistant to malaria infection.
- SS : These individuals suffer from a very severe anemia.

In regions with a high incidence of malaria, the benefits of the resistance to malaria conferred by the *AS* genotype outweigh the costs of the mild anemia, and *AS* heterozygotes have higher fitness than either homozygote.

The viabilities of the three genotypes in malarial regions have been estimated to be (Cavalli-Sforza and Bodmer, 1971):

<i>SS</i>	<i>AS</i>	<i>AA</i>
0.2	1.1	1

Using these fitnesses, the model predicts an equilibrium frequency for *S* of $\bar{p} \approx 0.1$, and the observed frequency is about 0.09 averaged across West Africa.

In contrast, in regions with little or no malaria, the sickle cell mutation is deleterious and is usually very rare.