

Contents

6 Natural Selection	1
6.1 Introduction	1
6.2 Theory for Asexuals	2
6.3 Theory for Diploids	5
6.3.1 Selection Schemes	8
6.4 Multiple Alleles	10
6.4.1 Haploid	10
6.4.2 Diploid	11
6.5 Fitness & Evolution	12
6.5.1 Haploid	12
6.5.2 Diploid	13
6.6 Statistical Estimation	14
6.6.1 Estimating Viability	15
6.6.2 Estimating Fitness	17

6 Natural Selection

6.1 Introduction

Natural Selection

Natural selection violates the Hardy-Weinberg assumptions by either

- changing **viability** based on genotype, or
- changing **fertility** based on genotype

The combination of viability and fertility of an individual defines its **fitness**.

It is also fundamental to the theory of evolution. It is the force that makes evolution “adaptive”, that makes evolution “progress” in time. To truly understand evolution, we need to know the

- environments and how they affect phenotypes to determine fitness,
- developmental processes that convert genotypes into phenotypes, and
- changes in genetic composition of the population (population genetics).

Questions About Natural Selection

- How do allele frequencies change in response to selection?
- How does selection acting on *diploid* individuals change allele frequencies over time?
- How much difference does a small change in fitness make?
- Under what conditions can an equilibrium be obtained?
- How does natural selection affect the overall fitness of the population? Is there improvement?

Detecting Selection

Positive selection (meaning an allele is favored by selection) at the molecular level has been a very hot area of research for much time.

- HIV-1 envelope gene (Nature **376**: 125)
- Major histocompatibility complex (Nature **335**: 167)
- Tumor suppressor gene BRCA1 (Nature Genetics **25**: 410)

Some current issues in detecting selection include:

- Detect selection at specific sites in a sequence when most sites are not subject to selection.
- Detect selection in non-coding regions of genomes, those regions that do not code for proteins.
- Detecting selection in certain lineages and not others. For example, finding those genome sites that have been selected in humans, but not chimpanzees for example.

Natural Selection

Definition: *viability* (v)

Viability is the probability of survival to adulthood and reproduction age.

Definition: *fertility*

Fertility is the propensity/ability of an individual at reproductive age to produce offspring. It can be summarized in multiple ways. The most detailed is

$$p_k = P(k \text{ offspring}) = P(k \text{ offspring} \mid \text{survival to adulthood})$$

but in large populations the mean

$$f = E(\text{number of offspring} \mid \text{survival to adulthood}) = \sum_k k p_k$$

is sufficient, since variability averages out in large, homogenous populations.

Definition: (*absolute*) *fitness* (W)

Fitness is a measure of an individual's ability to reproduce. Most simply,

$$W = v f$$

is the expected number of offspring a newborn will produce in its lifetime.

6.2 Theory for Asexuals

One-Generation Allele Frequency Update

Consider two genotypes A and a and let their viabilities be v_A and v_a , their fertilities be f_A and f_a , and define the absolute fitnesses as $W_A = v_A f_A$ and $W_a = v_a f_a$. Then, recursion equations for the population size of each type after one generation are

$$\begin{aligned} N_A(t+1) &= W_A N_A(t) \\ N_a(t+1) &= W_a N_a(t). \end{aligned}$$

The population frequency of genotype A is given by

$$\begin{aligned} p_A(t+1) &= \frac{N_A(t+1)}{N_A(t+1) + N_a(t+1)} = \frac{W_A N_A(t)}{W_A N_A(t) + W_a N_a(t)} \\ &= \frac{W_A p_A(t)}{W_A p_A(t) + W_a p_a(t)} = \frac{W_A}{\bar{W}(t)} p_A(t), \end{aligned}$$

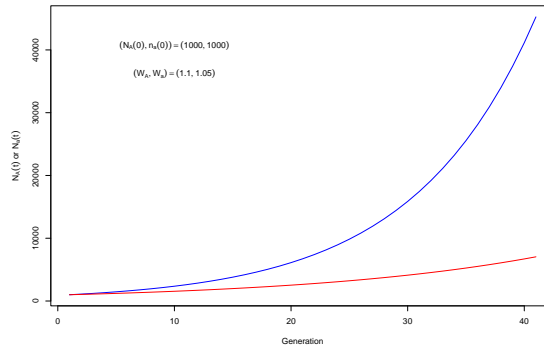
where $\bar{W}(t) = W_A p_A(t) + W_a p_a(t)$ is the average population fitness.

Effect on Population Size

The average population fitness can be written as a ratio of population sizes

$$\bar{W}(t) = W_A p_A(t) + W_a p_a(t) = W_A \frac{N_A(t)}{N(t)} + W_a \frac{N_a(t)}{N(t)} = \frac{N(t+1)}{N(t)}.$$

- $\bar{W}(t) > 1$, the population is growing, and
- $\bar{W}(t) < 1$, the population is shrinking.



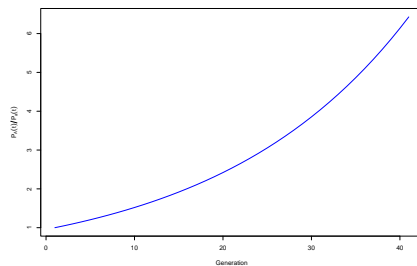
One-Generation Relative Allele Frequency Update

The genotype proportions are changing as well,

$$\frac{p_A(t+1)}{p_a(t+1)} = \frac{W_A p_A(t)}{W_a p_a(t)} = \frac{W_A}{W_a} \frac{p_A(t)}{p_a(t)} = w_A \frac{p_A(t)}{p_a(t)}$$

by a factor $w_A = \frac{W_A}{W_a}$.

Define the w_A as the **relative fitness** of a type A individual *with respect to* a type a individual. This allows us to define all relative fitnesses with respect to one type of individual. Note, in this case type a individuals are the standard and $w_a = 1$.



Environmental Effects

Relative fitnesses are useful in another way. Many common environmental effects are canceled in the ratio. As long as environmental effects are constant across genotypes (i.e. not genotype by environment interactions), then the same relative fitness will apply in multiple environments, even if the environments have an absolute fitness effect.

Specifically, multiplicative environmental effects do not change relative fitness. (Example: a nearby water source increases viability probability for all genotypes; temperate climate allows the production of twice as many offspring)

$$w'_A = \frac{mW_A}{mW_a} = w_A.$$

Additive effects do not cancel and will change additive fitness. (Example: when the population density is high, each individual has 2 fewer offspring)

$$w'_A = \frac{v_A(f_A - 2)}{v_a(f_a - 2)} \neq w_A.$$

Selection Coefficient

It is sometimes more convenient to reparameterize as follows:

$$\begin{aligned} w_A &= 1 + s \\ w_a &= 1. \end{aligned}$$

Then, when there is no selection, $s = 0$. The quantity $s \in [-1, \infty]$ is called the **selection coefficient** favoring A . Or, we could parameterize as

$$\begin{aligned} w_A &= 1 \\ w_a &= 1 - s', \end{aligned}$$

where $s' \in [-\infty, 1]$ is the **selection coefficient** against a .

These two parameterizations are not equivalent even if $s = s'$.

$$\frac{p_A(t+1)}{p_a(t+1)} = w_A \frac{p_A(t)}{p_a(t)} = (1 + s) \frac{p_A(t)}{p_a(t)} \quad \Bigg| \quad \frac{p_A(t+1)}{p_a(t+1)} = \frac{1}{w_a} \frac{p_A(t)}{p_a(t)} = \frac{1}{1 - s'} \frac{p_A(t)}{p_a(t)}$$

One-Generation Change in Allele Frequency

We had that the genotype (allele) frequency at generation $t + 1$ is

$$p_A(t + 1) = \frac{p_A(t)W_A}{\bar{W}(t)} = \frac{p_A(t)w_A}{\bar{w}(t)},$$

where $\bar{w}(t) = \frac{\bar{W}(t)}{W_a}$ is the average relative fitness of the population.

How is the absolute genotype frequency changing in time?

$$\begin{aligned} \Delta p_A(t) = p_A(t + 1) - p_A(t) &= \frac{p_A(t)w_A}{\bar{w}(t)} - p_A(t) \\ &= p_A(t) \frac{w_A - \bar{w}(t)}{\bar{w}(t)} \end{aligned}$$

so the change in absolute genotype A frequency depends on the distance of A 's relative fitness from the overall mean fitness $\bar{w}(t)$ and the current absolute allele frequency $p_A(t)$.

Multi-Generation Relative Allele Frequency Update

Recall the per-generation change in relative allele frequencies:

$$\frac{p_A(t + 1)}{p_a(t + 1)} = w_A \frac{p_A(t)}{p_a(t)}.$$

We've solve these kind of recurrence relations before. By repeated substitution, we obtain

$$\frac{p_A(t)}{p_a(t)} = w_A^t \frac{p_A(0)}{p_a(0)}.$$

The above solution can be used to solve for t

$$t = \frac{\ln\left(\frac{p_A(t)}{p_a(t)}\right) - \ln\left(\frac{p_A(0)}{p_a(0)}\right)}{\ln w_A},$$

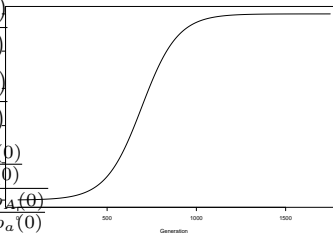
w_A	t	w_A	t
2	6.63	1.05	94.18
1.5	11.33	1.02	232.05
1.2	25.20	1.01	461.81
1.1	48.21	1.001	4597.42

telling how long it takes to change the allele frequency.

Table. Time t it takes to change allele frequencies from 1:1 to 99:1 in favor of A .

Multi-Generation Allele Frequency Update

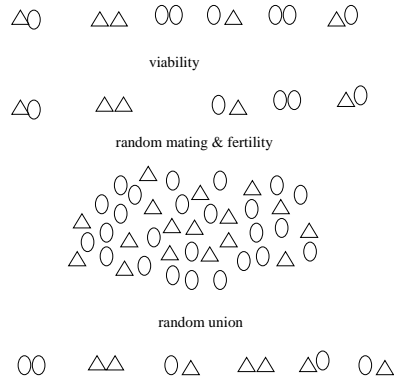
We can also solve the same equation for $p_A(t)$ to obtain the time-course of allele A change over many generations.

$$\begin{aligned} \frac{p_A(t)}{p_a(t)} &= w_A^t \frac{p_A(0)}{p_a(0)} \\ \frac{p_A(t)}{1 - p_A(t)} &= w_A^t \frac{p_A(0)}{p_a(0)} \\ p_A(t) \left(1 + w_A^t \frac{p_A(0)}{p_a(0)}\right) &= w_A^t \frac{p_A(0)}{p_a(0)} \\ p_A(t) &= \frac{w_A^t \frac{p_A(0)}{p_a(0)}}{1 + w_A^t \frac{p_A(0)}{p_a(0)}} \\ &= \frac{w_A^t p_A(0)}{p_a(0) + w_A^t p_A(0)}. \end{aligned}$$


6.3 Theory for Diploids

Diploid Selection

- Viability is a person-specific property.
- Fertility now appears to be a couple-specific property.
- View fertility as the person's ability to contribute to the reproductive gamete pool.
- A very fertile person will contribute more gametes than an unfertile person.
- Or, thought of in another, equivalent manner, more of an unfertile's gametes will *not* successfully unite.
- It no longer matters who is mating with whom.



Viability

The allele frequency at the start of generation t is the same as the allele frequency in the post-selection gamete pool of generation $t - 1$ that will randomly unite to form generation t .

Let $p_A(t)$ be the frequency of allele A in the $(t - 1)$ th post-selection gametic pool or the t th generation at birth. Then, if there are N individuals in generation t , the expected numbers of each genotype are:

$$\overline{AA: p_A^2(t)N \quad Aa: 2p_A(t)p_a(t)N \quad aa: p_a^2(t)N}$$

Assign viabilities (probabilities of survival) to each genotype.

$$\overline{AA: v_{AA} \quad Aa: v_{Aa} \quad aa: v_{aa}}$$

The expected number to survive to reproductive age in generation t are:

$$\overline{AA: v_{AA}p_A^2(t)N \quad Aa: 2v_{Aa}p_A(t)p_a(t)N \quad aa: v_{aa}p_a^2(t)N}$$

Fertility

These individuals produce gametes in Mendelian proportions (all A for AA types, half A for Aa types, etc), but the absolute number is determined by selection. Assign fertilities to each genotype. Fertility here is the average number of gametes that make it into the gametic pool.

$$\overline{AA: f_{AA} \quad Aa: f_{Aa} \quad aa: f_{aa}}$$

So, each diploid genotype G that survives to adulthood will contribute on average f_G gametes to the t th generation gametic pool. So the numbers of each allele in the pool are:

Allele	Expected Counts
A:	$f_{AA}v_{AA}p_A^2(t)N + f_{Aa}v_{Aa}p_A(t)p_a(t)N$
a:	$f_{aa}v_{aa}p_a^2(t)N + f_{Aa}v_{Aa}p_A(t)p_a(t)N$

Gamete Pool

To obtain the proportions of each allele at the start of the $(t + 1)$ th generation (or equivalently in the post-selection gametic pool of the t th generation) we use

$$\text{proportion of } A = \frac{\text{expected number of } A}{\text{total expected number of alleles}}$$

so

$$\begin{aligned} p_A(t+1) &= \frac{f_{AA}v_{AA}p_A^2(t)N + f_{Aa}v_{Aa}p_A(t)p_a(t)N}{f_{AA}v_{AA}p_A^2(t)N + f_{Aa}v_{Aa}p_A(t)p_a(t)N + f_{aa}v_{aa}p_a^2(t)N + f_{Aa}v_{Aa}p_A(t)p_a(t)N} \\ &= \frac{f_{AA}v_{AA}p_A^2(t)N + f_{Aa}v_{Aa}p_A(t)p_a(t)N}{f_{AA}v_{AA}p_A^2(t)N + 2f_{Aa}v_{Aa}p_A(t)p_a(t)N + f_{aa}v_{aa}p_a^2(t)N} && \text{combine terms} \\ &= \frac{f_{AA}v_{AA}p_A^2(t) + f_{Aa}v_{Aa}p_A(t)p_a(t)}{f_{AA}v_{AA}p_A^2(t) + 2f_{Aa}v_{Aa}p_A(t)p_a(t) + f_{aa}v_{aa}p_a^2(t)} && \text{cancel } N \end{aligned}$$

Fitness

We define the **absolute fitness** of genotype G as $W_G = f_G v_G$.

We define the **relative fitness** of genotype G_1 with respect to genotype G_2 as $w_{G_1} = \frac{W_{G_1}}{W_{G_2}}$.

With these definitions and the identity $p_a(t) = 1 - p_A(t)$ in the two allele case, the equation can be rewritten as

$$p_A(t+1) = \frac{w_{AA}p_A^2(t) + w_{Aa}p_A(t)[1 - p_A(t)]}{w_{AA}p_A^2(t) + 2w_{Aa}p_A(t)[1 - p_A(t)] + w_{aa}[1 - p_A(t)]^2}$$

Note, the denominator is the average fitness of the population at the start of generation t . We define the notation

$$\bar{w}(t) = w_{AA}p_A^2(t) + 2w_{Aa}p_A(t)[1 - p_A(t)] + w_{aa}[1 - p_A(t)]^2.$$

One-Generation Allele Frequency Update

Or, another way to follow the absolute A allele frequency change in a single generation.

$$\begin{aligned} p_A(t+1) &= \frac{w_{AA}p_A^2(t) + w_{Aa}p_A(t)p_a(t)}{w_{AA}p_A^2(t) + 2w_{Aa}p_A(t)p_a(t) + w_{aa}p_a^2(t)} \\ &= p_A(t) \frac{p_A(t)w_{AA} + p_a(t)w_{Aa}}{\bar{w}(t)} = p_A(t) \frac{\bar{w}_A(t)}{\bar{w}(t)}. \end{aligned}$$

where $\bar{w}_A(t)$ is the mean fitness of A -carrying population. Specifically,

$$\begin{aligned} \bar{w}_A(t) &= p_A(t)w_{AA} + p_a(t)w_{Aa} \\ &= P(A \text{ from } AA | A) w_{AA} + P(A \text{ from } Aa | A) w_{Aa} \end{aligned}$$

and

$$P(A \text{ from } Aa | A) = \frac{\frac{1}{2}P(Aa)}{P(A)} = \frac{p_A(t)p_a(t)}{p_A^2(t) + p_A(t)p_a(t)} = p_a(t)$$

One-Generation Relative Allele Frequency Update

To determine how the relative proportions of each allele change in a single generation, we derive

$$\begin{aligned}
 \frac{p_A(t+1)}{p_a(t+1)} &= \frac{w_{AA}p_A^2(t) + w_{Aa}p_A(t)p_a(t)}{w_{Aa}p_A(t)p_a(t) + w_{aa}p_a^2(t)} \\
 &= \frac{p_A(t)}{p_a(t)} \times \frac{w_{AA}p_A(t) + w_{Aa}p_a(t)}{w_{Aa}p_A(t) + w_{aa}p_a(t)} \\
 &= \frac{p_A(t)}{p_a(t)} \times \frac{\bar{w}_A(t)}{\bar{w}_a(t)}.
 \end{aligned}$$

One-Generation Change in Allele Frequency

$$\begin{aligned}
 p_A(t+1) - p_A(t) &= \frac{w_{AA}p_A^2(t) + w_{Aa}p_A(t)[1 - p_A(t)]}{w_{AA}p_A^2(t) + 2w_{Aa}p_A(t)[1 - p_A(t)] + w_{aa}[1 - p_A(t)]^2} \\
 &\quad - p_A(t) \\
 &\quad \vdots \\
 &= p_A(t)p_a(t) \frac{\bar{w}_A(t) - \bar{w}_a(t)}{\bar{w}(t)} \quad * \\
 &= p_A(t) \frac{\bar{w}_A(t) - p_A(t)\bar{w}_A(t) - p_a(t)\bar{w}_a(t)}{\bar{w}(t)} \\
 &= p_A(t) \frac{\bar{w}_A(t) - \bar{w}(t)}{\bar{w}(t)}
 \end{aligned}$$

where we have recognized that

$$\bar{w}(t) = p_A(t)\bar{w}_A(t) + p_a(t)\bar{w}_a(t).$$

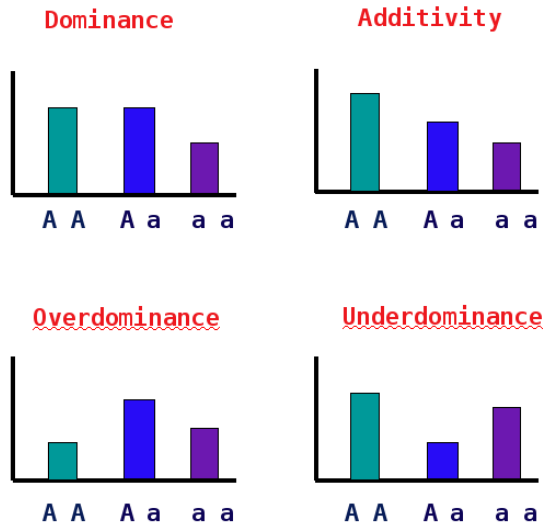
6.3.1 Selection Schemes

Kinds of Fitness

Here, we consider various types of fitness schemes for the 3 genotypes AA , Aa , and aa . Here $p = p_A(t-1)$.

Type	Fitnesses	\bar{w}_A	\bar{w}
Geometric	$(1+s)^2 : 1+s : 1$	$(1+s)(1+sp)$	$(1+sp)^2$
Additive	$1+2s : 1+s : 1$	$1+s+sp$	$1+2sp$
Recessive	$1+s : 1 : 1$	$1+sp$	$1+sp^2$
Dominant	$1+s : 1+s : 1$	$1+s$	$1+2sp(1-p) + sp^2$
Overdominance	$1-s : 1 : 1-t$	$1-sp$	$q-sp^2-t(1-p)^2$
Underdominance	$1+s : 1 : 1+t$	$1+sp$	$q+sp^2+t(1-p)^2$
Type	$\frac{p_A(t)}{p_a(t)}$	$\Delta p_A(t)$	
Geometric	$\frac{p(1+s)}{1-p}$	$\frac{sp(1-p)}{1+sp}$	
Additive	$\frac{p(1+s+sp)}{(1-p)(1+2sp)}$	$\frac{sp(1-p)}{1+2sp}$	
Recessive	$\frac{p(1+sp)}{1-p}$	$\frac{sp^2(1-p)}{1+sp^2}$	
Dominant	$\frac{(1+s)p}{1+2sp(1-p)+sp^2}$	$\frac{sp(1-p)^2}{1+2sp(1-p)+sp^2}$	
Overdominance	$\frac{(1-sp_A)p_A}{p_a - tp_a^2}$	$\frac{p_A p_a [t - (s+t)p_A]}{1-sp_A^2 - tp_a^2}$	
Underdominance	$\frac{(1+sp_A)p_A}{p_a + tp_a^2}$	$\frac{p_A p_a [(s+t)p_A - t]}{1+sp_A^2 + tp_a^2}$	

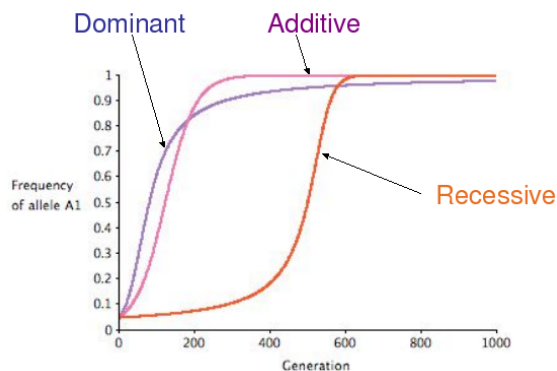
Visualizing Dominance Types



Implications

- **Multiplicative (Geometric) fitness.** Population behaves as if reproducing asexually.
- **Additive fitness.** Behaves almost like a locus with multiplicative fitness, but the approximation is only good for $s < 0.2$.
- **Recessive allele.** Selection *against* a recessive allele becomes increasingly less effective as the frequency of the allele declines. If recessive allele a starts at frequency $p_a(0) = 0.5$, initially a will quickly disappear. It takes 999,998 generations to reduce from 0.5 to 0.000001 when the allele is recessive and *lethal* and only 1375 generations if multiplicative fitness applies with relative weak selection $s = -0.1$.
- **Dominant allele.** If A is lethal, it takes one generation to eliminate all A . Otherwise, it is the recessive case reversed.

Temporal Allele Change by Dominance Types



Over or Under Dominance

$$\Delta p_A = \frac{p_A p_a [t - (s + t)p_A]}{1 - s p_A^2 - t p_a^2}$$

When is $\Delta p_A = 0$?

- $p_A = 0$ or $p_a = 0$.
- $t - (s + t)p_A = 0$ or $p_e = \frac{t}{t+s}$ is the stationary frequency.

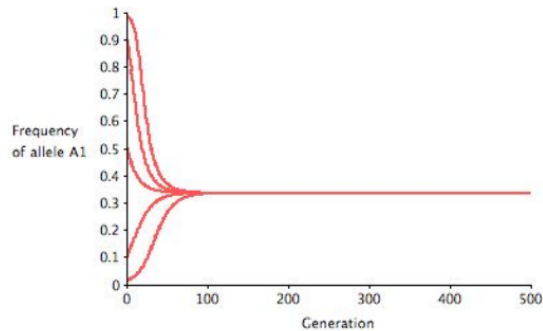
For an overdominant allele, the allele frequency p_A will approach equilibrium p_e from anywhere.

For an underdominant allele, the allele frequency p_A will approach 0 if $p_A(0) < p_e$ or 1 if $p_A(0) > p_e$. In other words, eventually one allele will be lost from the population and which one is lost depends on the starting state of the population at time 0.

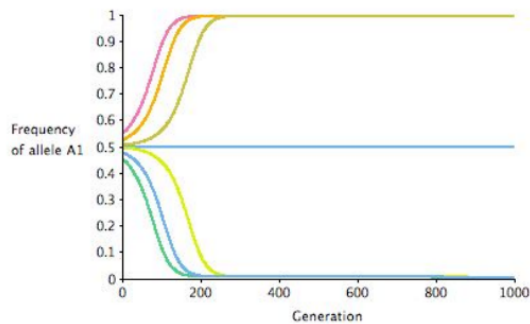
Visualizing Overdominance

$$p_e = \frac{t}{t+s}$$

For $s = 0.2$ and $t = 0.1$, we have



Visualizing Underdominance



6.4 Multiple Alleles

6.4.1 Haploid

Haploid - Multiple Alleles

For the asexual haploid case, we have k different alleles A_i , for $i = 1, \dots, k$ and a corresponding relative fitness w_i for each type. For one generation, the recursion equation for allele frequency is

$$p_i(t+1) = \frac{p_i(t)w_i}{\bar{w}(t)}$$

where

$$\bar{w}(t) = \sum_i p_i(t)w_i.$$

The change in allele frequency in a single generation is

$$\Delta p_i(t) = p_i(t) - p_i(t-1) = p_i(t) \frac{w_i - \bar{w}(t)}{\bar{w}(t)}.$$

6.4.2 Diploid

Diploid - Multiple Alleles

For the diploid case, we have k different alleles A_i , for $i = 1, \dots, k$, and a relative fitness for each genotype w_{ij} , for $i < j = 1, \dots, k$.

The allele frequency change in one generation is by direct analogy:

$$p_i(t+1) = \frac{\sum_{j=1}^k p_i(t)p_j(t)w_{ij}}{\sum_{i=1}^k \sum_{j=1}^k p_i(t)p_j(t)w_{ij}} = \frac{p_i(t) \sum_{j=1}^k p_j(t)w_{ij}}{\bar{w}(t)} = \frac{p_i(t)\bar{w}_i(t)}{\bar{w}(t)}.$$

Rearrangement provides the standard equations for change in allele frequency in one generation

$$\Delta p_i(t) = p_i(t) - p_i(t-1) = p_i(t) \frac{\bar{w}_i(t) - \bar{w}(t)}{\bar{w}(t)}.$$

Equilibrium - Multiple Alleles

The equation

$$\Delta p_i(t) = p_i(t) \frac{\bar{w}_i(t) - \bar{w}(t)}{\bar{w}(t)}.$$

allows us to identify equilibria from which allele frequency will not stray. $\Delta p_i(t) = 0$ whenever $p_i = 0$ or whenever $\bar{w}_i(t) - \bar{w}(t) = 0$. If alleles $\{3, 6, 17\}$ are to persist at equilibrium in a population then the following three equations must be satisfied

$$\begin{aligned} \bar{w}_3(t) &= \bar{w}(t) \\ \bar{w}_6(t) &= \bar{w}(t) \\ \bar{w}_{17}(t) &= \bar{w}(t), \end{aligned}$$

with all other frequencies $p_1 = p_2 = p_4 = \dots = 0$.

The existence of a solution to the above equations does not mean that the population will approach and stay at this equilibrium.

Stable Equilibrium - Multiple Alleles

It turns out there is a simple condition for stability.

If for all alleles i that are absent from the equilibrium (i.e. $p_i = 0$), then the equilibrium is stable if the following condition is satisfied at the equilibrium

$$\bar{w}_i < \bar{w},$$

So, take the underdominance example where the fitnesses are

$$\begin{array}{ccc} 11 & 12 & 22 \\ \hline 1+t & 1 & 1+s \end{array}$$

To show $p_1 = 1$ is stable, we must show $\bar{w}_2 < \bar{w}$.

$$\begin{aligned} \bar{w}_2 = p_2\bar{w}_{22} + p_1\bar{w}_{12} &< \bar{w} = p_2^2\bar{w}_{22} + 2p_2p_1\bar{w}_{12} + p_1^2\bar{w}_{11} \\ \bar{w}_{12} = 1 &< \bar{w}_{11} = 1+t \end{aligned}$$

6.5 Fitness & Evolution

Fitness and Evolution

We are interested in whether the mean fitness of the population is increasing in time.

The mean absolute fitness \bar{W} cannot always increase for then the population would explode to ∞ .

In fact, if $\bar{W}(0) > 1$, then $\bar{W}(t) \rightarrow 1$ at which point the population will no longer expand. Or, if $\bar{W}(0) < 1$, then the population will go extinct.

Indeed, if we are going to look for improvement in fitness, we need to look at relative mean fitness \bar{w} , relative to some fixed standard.

6.5.1 Haploid

Mean Relative Fitness - Asexual

For asexual haploid populations, mean relative fitness is

$$\bar{w}(t) = \sum_{i=1}^k p_i(t)w_i.$$

How does this change over time?

$$\begin{aligned} \Delta\bar{w}(t) = \bar{w}(t) - \bar{w}(t-1) &= \sum_{i=1}^k p_i(t)w_i - \sum_{i=1}^k p_i(t-1)w_i \\ &= \sum_{i=1}^k w_i \frac{p_i(t-1)w_i}{\bar{w}(t-1)} - \bar{w}(t-1) \\ &= \frac{1}{\bar{w}(t-1)} \left[\sum_{i=1}^k p_i(t-1)w_i^2 - \bar{w}^2(t-1) \right] \\ &= \frac{\text{Var}(w)}{\bar{w}(t-1)} \end{aligned}$$

Interpretation - Temporal Change in \bar{w}

- The more variation in fitness between alleles (genotypes), the faster the change in the mean relative fitness of the population.

- The mean relative fitness is always increasing (or staying constant). It never declines.
- The rate of progress in fitness is proportional to the square of the relative fitness (or selection coefficient). In contrast, the rate of change in allele frequency is directly proportional to relative fitness (or selection coefficient).

6.5.2 Diploid

Mean Relative Fitness - Diploid

For the diploid case, the mean relative fitness is defined as

$$\bar{w}(t) = \sum_i \sum_j p_i(t)p_j(t)w_{ij} = p^2w_{11} + 2p(1-p)w_{12} + (1-p)^2w_{22},$$

where we have reduced to the two allele case and dropped dependence on t . To see how this mean relative fitness changes with respect to allele 1, take the derivative

$$\begin{aligned} \frac{\partial \bar{w}(t)}{\partial p} &= 2pw_{11} + 2(1-p)w_{12} - 2pw_{12} - 2(1-p)w_{22} \\ &= 2[pw_{11} + (1-2p)w_{12} - (1-p)w_{22}]. \end{aligned}$$

Next, we will relate this to the change in allele frequency in a single generation.

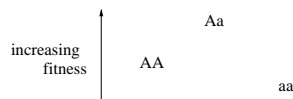
Mean Relative Fitness - Diploid

The change in allele frequency in a single generation is

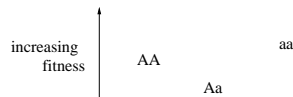
$$\begin{aligned} \Delta p(t) &= p(t) - p(t-1) \\ &= \frac{p^2(t-1)w_{11} + p(t-1)[1-p(t-1)]w_{12}}{\bar{w}(t-1)} - p(t-1) \\ &= \frac{p^2(t-1)w_{11} + p(t-1)[1-p(t-1)]w_{12} - p(t-1)\bar{w}(t-1)}{\bar{w}(t-1)} \\ &= \frac{p^2w_{11} + p[1-p]w_{12} - p\{p^2w_{11} + 2p[1-p]w_{12} + [1-p]^2w_{22}\}}{\bar{w}} \\ &= p(1-p) \frac{pw_{11} + (1-2p)w_{12} - (1-p)w_{22}}{\bar{w}} \\ &= \frac{p(1-p)}{2\bar{w}} \frac{\partial \bar{w}}{\partial p} \end{aligned}$$

Allele frequency is stable wherever the mean relative fitness curve ($\bar{w}(p)$) is flat.

Selection on Diploids Cannot be Perfect



- Maximum relative mean fitness in the overdominant case occurs when all individuals are heterozygous, but as soon as this population mates, homozygotes will appear and decrease the population mean fitness.



- Maximum relative mean fitness in the underdominant case is maximized when the less fit allele is eliminated from the population, but if the population starts with more of the less fit allele than the more fit allele, the more fit allele will be eliminated first.

6.6 Statistical Estimation

Selection and HWE

How do we detect evidence of selection?

Selection is a violation of the HW assumptions, therefore it should lead to HW disequilibrium. Theoretically, tests for HWE could be used to find evidence of selection (confounded with multiple other possible causes).

Let's examine this more precisely for the two allele, single locus case.

- At the time of gamete union, the alleles in the post-selection gametic pool unite randomly so that instant HWE applies.
- However, we are not likely to observe individuals at gamete union (when the zygote forms). We sample them in adulthood or later in life after viability probabilities have had a chance to change the genotype frequencies.

Viability and HWE

After viability selection, the genotype probabilities are obtained as usual

$$\begin{aligned} P(11 \mid \text{survival to sampling}) &= \frac{P(11, \text{survival to sampling})}{P(\text{survival to sampling})} \\ &= \frac{v_{11}p_1^2}{v_{11}p_1^2 + 2v_{12}p_1p_2 + v_{22}p_2^2} \\ &= \frac{v_{11}p_1^2}{\bar{v}}. \end{aligned}$$

Given these genotype probabilities, the allele frequencies at sampling time are

$$\begin{aligned} p'_1 &= P(11 \mid \text{survival to sampling}) + \frac{1}{2}P(12 \mid \text{survival to sampling}) \\ &= \frac{v_{11}p_1^2 + v_{12}p_1p_2}{\bar{v}} \end{aligned}$$

So, HWE is true if

$$\begin{aligned} (p'_1)^2 &= \left(\frac{v_{11}p_1^2 + v_{12}p_1p_2}{\bar{v}} \right)^2 = \frac{v_{11}p_1^2}{\bar{v}} \\ 2p'_1p'_2 &= 2 \frac{v_{11}p_1^2 + v_{12}p_1p_2}{\bar{v}} \frac{v_{22}p_2^2 + v_{12}p_1p_2}{\bar{v}} = \frac{v_{12}2p_1p_2}{\bar{v}} \\ (p'_2)^2 &= \left(\frac{v_{22}p_2^2 + v_{12}p_1p_2}{\bar{v}} \right)^2 = \frac{v_{22}p_2^2}{\bar{v}} \end{aligned}$$

Taking the first equation, the relationship that must hold is

$$\begin{aligned} v_{11}^2p_1^4 + 2v_{12}v_{11}p_1^3p_2 + v_{12}^2p_1^2p_2^2 &= \bar{v}v_{11}p_1^2 \\ &= v_{11}^2p_1^4 + 2v_{11}v_{12}p_1^3p_2 + v_{11}v_{22}p_2^2p_1^2 \end{aligned}$$

and this is only true if

$$v_{12}^2 = v_{11}v_{22}$$

It turns out all three equations lead to the same condition $v_{12}^2 = v_{11}v_{22}$.

The biological implication is that certain kinds of fitness differences will not be detectable as a deviation from HWE.

- No fertility difference is visible from tests of HWE.
- Viability probabilities satisfying the above condition, for example if 1 in 2 genotype 11 individuals survive to sampling, 1 in 8 genotype 22 individuals survive to sampling, and 1 in 4 genotype 12 individuals survive to adulthood, then the condition is satisfied, but selection *is* present.

6.6.1 Estimating Viability

Method I: Estimating Relative Viability

Suppose that the population is at a stable equilibrium (overdominance for example). Then, allele frequencies are not changing in time $p'_1 = p_1$. (There is an implicit assumption that there are no fertility effects.)

Then the update equations from birth to sampling

$$\begin{aligned} p'_1 &= \frac{v_{11}p_1^2 + v_{12}p_1p_2}{\bar{v}} \\ p'_2 &= \frac{v_{22}p_2^2 + v_{12}p_1p_2}{\bar{v}} \end{aligned}$$

can be rewritten and to reveal two relationships among the parameters

$$\begin{aligned} p_1 &= \frac{v_{11}p_1^2 + v_{12}p_1p_2}{\bar{v}} & p_2 &= \frac{v_{22}p_2^2 + v_{12}p_1p_2}{\bar{v}} \\ \bar{v}p_1 &= v_{11}p_1^2 + v_{12}p_1p_2 & \bar{v} &= v_{22}p_2 + v_{12}p_1. \\ \bar{v} &= v_{11}p_1 + v_{12}p_2 \end{aligned}$$

Method I: Estimating Relative Viability

There is the other usual relationship $p_1 + p_2 = 1$, so there are two free parameters p_1 and v_{11} out of the total $p_1, p_2, v_{11}, v_{12}, v_{22}$.

The likelihood is

$$L(n_{11}, n_{12}, n_{22}) = \binom{n}{n_{11}n_{12}n_{22}} \left(\frac{v_{11}p_1^2}{\bar{v}} \right)^{n_{11}} \left(\frac{v_{12}2p_1p_2}{\bar{v}} \right)^{n_{12}} \left(\frac{v_{22}p_2^2}{\bar{v}} \right)^{n_{22}}$$

The data consists of genotype counts n_{11}, n_{12}, n_{22} as always, so Bailey's method applies with $n = n_{11} + n_{12} + n_{22}$.

$$\begin{aligned} n_{11} &= n \frac{v_{11}p_1^2}{\bar{v}} \\ n_{12} &= n \frac{v_{12}2p_1p_2}{\bar{v}} \\ n_{22} &= n \frac{v_{22}p_2^2}{\bar{v}} \end{aligned}$$

Method I: Testing Relative Viability

Previously, we had tested HWE by looking for significantly non-zero D_1 . Back then we also used the sampling distribution under the alternative to perform power and sample size calculations. If we could relate our parameterization in terms of v_{11} to the one in terms of D_1 , then we could borrow all our work then and apply

it to this case.

$$\begin{aligned} D_1 &= P_{11} - p_1^2 \\ &= \frac{v_{11}p_1^2}{\bar{v}} - \left(\frac{v_{11}p_1^2 + v_{12}p_1p_2}{\bar{v}} \right)^2 \\ &= \frac{p_1^2 p_2^2 (v_{11}v_{22} - v_{12}^2)}{\bar{v}}. \end{aligned}$$

If HW disequilibrium is caused by viability selection, then the test for $H_0 : D_1 = 0$ is the same as testing $H_0 : v_{11}v_{22} = v_{12}^2$.

Method II: Estimating Relative Viability

Consider two classes of individuals that you think are under different selection pressures. If you can design an experiment in which you expect equal proportions of both types, then you can examine the proportions in your data for agreement with your expectation.

An example: Suppose you suspect that 11 and 12 individuals have different fitness. If you can design a controlled cross between 11 and 12 individuals, then you expect 50% of the offspring to be 11 and 50% to be 12. Any deviation from the expected 1:1 ratio could be evidence of selection.

$$\begin{aligned} P(12 \mid \text{survival}) &= \frac{v_{12}P_{12}}{v_{11}P_{11} + v_{12}P_{12} + v_{22}P_{22}} \\ &= \frac{v_{12}0.5}{v_{11}0.5 + v_{12}0.5} \\ &= \frac{v_{12}}{v_{11} + v_{12}}. \end{aligned}$$

Let's suppose we're measuring relative viability with respect to 11 individuals, so define $v = v_{12}$ and of course $v_{11} = 1$. Then

$$P(12 \mid \text{survival}) = \frac{v}{1 + v}.$$

You observe n_{11} type 11 individuals and n_{12} type 12 individuals. A Method of Moments estimator for v is obtained using equation $n_{12} = (n_{11} + n_{12}) \frac{v}{v + 1}$ or $\hat{v} = \frac{n_{12}}{n_{11}}$.

It turns out that \hat{v} can be quite biased, and a better estimator is

$$\hat{v} = \frac{n_{12}}{n_{11} + 1}.$$

Method II: Bias & Variance of \hat{v}

The new estimator \hat{v} is still a little biased

$$E(\hat{v}) = v \left[1 - \left(\frac{v}{v + 1} \right)^n \right]$$

which can be seen by recognizing that the sampling distribution is

$$\text{Binomial}(n_{11} + n_{12}, P[12 \mid \text{survival}]).$$

The estimator

$$\hat{v} = \frac{n_{12}}{n_{11} + 1}$$

is a ratio of functions of the same order in the counts, so Fisher's approximation applies. Bootstrap and jackknife are also valid. The Fisher's approximate formula is

$$\text{Var}(\hat{v}) \approx \frac{v(v+1)^2}{n}$$

Method II: Example

Using

$$\hat{v} = \frac{n_{12}}{n_{11} + 1}$$

$$\text{Var}(\hat{v}) \approx \frac{\hat{v}(\hat{v} + 1)^2}{n}$$

how can one test for significantly different viabilities of type 11 and 12 individuals?

To make a very specific example, suppose I hypothesize the following viabilities where $u < t$.

$$\begin{array}{ccc} 11 & 12 & 22 \\ \hline t & u & t \end{array}$$

This is an example of what kind of selection?

What are my two types of individuals?

How can I design a cross such that my two types of individuals are expected to appear in equal frequencies?

$$\begin{array}{cc} 11 & | & 12 \\ \hline 18 & & 22 \end{array}$$

$$\hat{v} \in (0.44, 1.88)$$

6.6.2 Estimating Fitness

Adding Fertility

11	12	22
Pre-Selection		
p_1^2	$2p_1p_2$	p_2^2
↓		
Viability Selection		
$\frac{v_{11}p_1^2}{\bar{v}}$	$\frac{2v_{12}p_1p_2}{\bar{v}}$	$\frac{v_{22}p_2^2}{\bar{v}}$
↓		
Fertility Selection		
$\frac{f_{11}v_{11}p_1^2}{W}$	$\frac{2f_{12}v_{12}p_1p_2}{W}$	$\frac{f_{22}v_{22}p_2^2}{W}$
↓		
Gametic Pool		
$p'_1 = \frac{p_1^2 v_{11} f_{11} + p_1 p_2 v_{12} f_{12}}{W}, p'_2 = 1 - p'_1$		

where p_1 and p_2 are the allele frequencies in the gametic pool.

where $\bar{v} = v_{11}p_1^2 + v_{12}2p_1p_2 + v_{22}p_2^2$ is mean viability.

where \bar{W} is the usual mean absolute fitness.

Comparing Across Generations

One of the few things we can actually observe is the ratio of homozygotes to heterozygotes in each generation. In the first generation,

$$R_{11} = \frac{v_{11}p_1^2}{v_{12}2p_1p_2}$$

$$R_{22} = \frac{v_{22}p_2^2}{v_{12}2p_1p_2}$$

In the next generation, we have

$$R'_{11} = \frac{v_{11}p_1'^2}{v_{12}2p_1'p_2'} = \frac{v_{11} (p_1^2 v_{11} f_{11} + p_1 p_2 v_{12} f_{12})^2}{2v_{12} (p_1^2 v_{11} f_{11} + p_1 p_2 v_{12} f_{12}) (p_2^2 v_{22} f_{22} + p_1 p_2 v_{12} f_{12})}$$

$$= \frac{v_{11} (p_1^2 v_{11} f_{11} + p_1 p_2 v_{12} f_{12})}{2v_{12} (p_2^2 v_{22} f_{22} + p_1 p_2 v_{12} f_{12})} = \frac{v_{11}}{2v_{12}} \frac{2R_{11} \frac{f_{11}}{f_{12}} + 1}{2R_{22} \frac{f_{22}}{f_{12}} + 1}.$$

Comparing Across Generations - Viability

From the previous slide, we have

$$R'_{11} = \frac{v_{11}}{2v_{12}} \frac{2R_{11} \frac{f_{11}}{f_{12}} + 1}{2R_{22} \frac{f_{22}}{f_{12}} + 1}.$$

When there are no fertility effects $f_{12} = f_{11} = f_{22} = 1$, and the relative viability is estimated from the observable ratios as

$$\frac{v_{11}}{v_{12}} = 2R'_{11} \frac{2R_{22} + 1}{2R_{11} + 1}$$

$$\frac{v_{22}}{v_{12}} = 2R'_{22} \frac{2R_{11} + 1}{2R_{22} + 1}$$

When there are fertility effects, the equation cannot be rearranged to provide a relative fitness estimate $\frac{v_{11}f_{11}}{v_{12}f_{12}} = w_{11}$.

Comparing Across Generations - Fitness

In particular, if there are fertility differences, one can show there is a linear relationship between $\frac{v_{11}f_{11}}{v_{12}f_{12}} = w_{11}$ and w_{22} , where relative fitnesses are measured with respect to the heterozygote 12.

Instead of considering genotype ratios like R_{11} , consider allele ratio $x = \frac{p_1}{p_2}$. As before, follow the genotype proportions in time, now lumping viability and fertility selection into one.

$$\begin{array}{ccc} 11 & 12 & 22 \\ \frac{x^2 p_2^2}{x^2 p_2^2} & \frac{2x p_2^2}{2x p_2^2} & \frac{p_2^2}{p_2^2} \\ \downarrow & & \\ \frac{W_{11} x^2 p_2^2}{W} & \frac{W_{12} 2x p_2^2}{W} & \frac{W_{22} p_2^2}{W} \end{array}$$

In the next generation,

$$x' = \frac{p_1'}{p_2'} = \frac{W_{11} x^2 + W_{12} x}{W_{22} + W_{12} x}$$

Comparing Across Generations - Fitness

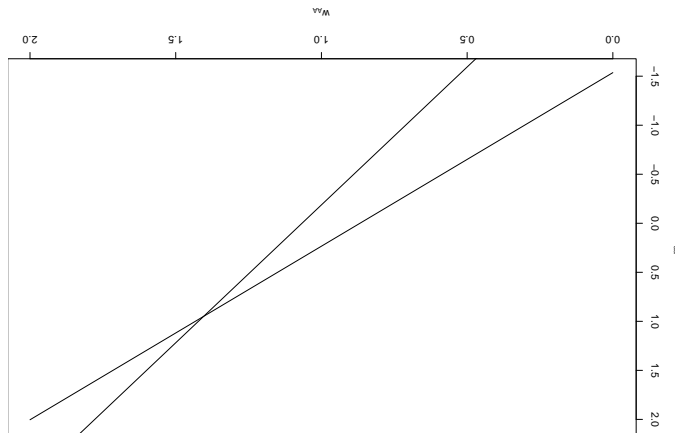
Rearrangement of the above equation yields,

$$\frac{W_{22}}{W_{12}} = w_{22} = \frac{x^2}{x'} w_{11} + \frac{x}{x'} - x.$$

Collect x and x' once and you will obtain a linear relationship between w_{11} and w_{22} . Collect a second pair x, x' and you will obtain another linear relationship between w_{11} and w_{22} . Assuming that w_{11} and w_{22} are not changing between the two experiments but x and x' do change because of different allele frequencies in the two sample populations, the intersection of the two lines provides estimates of w_{11} and w_{22} .

What does it mean if the lines do not intersect or do not intersect in the positive quadrant of the w_{11} by w_{22} plane?

Example - Plot



Maximum Likelihood Estimation of Selection

We have previously derived (and used already today) recurrence relations for allele frequencies over time given relative fitnesses w and starting allele frequencies p

$$p_u(t+1) = \frac{p_u^2(t)w_{uu} + \sum_{v \neq u} p_u(t)p_v(t)w_{uv}}{\bar{w}(t)}.$$

Suppose you observe allele counts over multiple generations $\{n_1(1), n_2(1), \dots, n_1(2), n_2(2), \dots\}$.

Then, the likelihood for the fitness model is

$$L(\{w_{uv}\}) \propto \prod_t \prod_u (p_u(t))^{n_u(t)}.$$

Numerical methods are required to maximize this likelihood over w_{uv} and p_u .