Basic Population Genetics

Bruce Walsh
Second Bangalore School of
Population Genetics and Evolution
25 Jan- 5 Feb 2016

Outline

- Population genetics of random mating
- Population genetics of drift
- Population genetics of mutation
- Population genetics of selection
- Interaction of forces

Random-mating

Genotypes & alleles

- In a diploid individual, each locus has two alleles
- The genotype is this two-allele configuration
  - With two alleles A and a
  - AA and aa are homozygotes (alleles agree)
  - Aa is a heterozygote (alleles are different)
- With an arbitrary number of alleles A_1, .., A_n
  - A_iA_i are homozgyotes
  - A_iA_k (for i different from k) are heterozygotes

Allele and Genotype Frequencies

Given genotype frequencies, we can always compute allele Frequencies. For a locus with two alleles (A,a),

$$freq(A) = freq(AA) + (1/2)freq(Aa)$$

The converse is not true: given allele frequencies we cannot uniquely determine the genotype frequencies

If we are willing to assume random mating,

$freq(AA) = [freq(A)]^2$, Hardy-Weinberg
$freq(Aa) = 2*freq(A)*freq(a)$ proportions

For k alleles

- Suppose alleles are A_1 , .. , A_n
  - Easy to go from genotype to allele frequencies
  - $Freq(A_i) = Freq(A_iA_i) + (1/2) \Sigma freq(A_iA_k)$
- Again, assumptions required to go from allele to genotype frequencies.
- With n alleles, n(n+1)/2 genotypes
- Under random-mating
  - $Freq(A_iA_i) = Freq(A_i)*Freq(A_i)$
  - $Freq(A_iA_k) = 2*Freq(A_i)*Freq(A_k)$

Hardy-Weinberg

Importance of HW

- Under HW conditions, no
  - Drift (i.e., pop size is large)
  - Migration/mutation (i.e., no input ofvariation from other populations or new mutations)
  - Selection (no forces to systemically change allele frequencies)
- Under HW conditions, allele frequencies do not change
- Further, the distribution of genotypes in a population is stable under random mating

Derivation of the Hardy-Weinberg result

- Consider any population, where
  - Freq(AA) = X
  - Freq(Aa) = Y
  - Freq(aa) = Z
  - freq(A) = p = freq(AA) + (1/2) freq(Aa) = X + ½ Y
- What happens in the next generation from random mating?

Frequency of matings

female genotype	frequency	male genotype AA (X)	Aa (Y)	aa (Z)
AA	(X)	X^2	XY	XZ
Aa	(Y)	XY	Y^2	YZ
aa	(Z)	XZ	YZ	Z^2

Random Mating=independence

Genotype frequencies in next generation

Possible Matings	Frequency of Mating	Expected Frequency of Offspring AA	Aa	aa
AA x AA	X^2	1	0	0
AA x Aa	2XY	1/2	1/2	0
AA x aa	2XZ	0	1	0
Aa x Aa	Y^2	1/4	1/2	1/4
Aa x aa	2YZ	0	1/2	1/2
aa x aa	Z^2	0	0	1

Conditional Probabilities given genotypes of parents

Freq(AA) = 1* X^2 + ½*2XY + (1/4) Y^2 = (X + ½ Y) ^{2} = p^2.

Freq(aa) = 1* Z^2 + ½*2YZ + (1/4) Y^2 = (Z + ½ Y) ^{2} = q^2.

What about the next generation?

Possible Matings	Frequency of Mating	Expected Frequency of Offspring AA	Aa	aa
AA x AA	p^4	1	0	0
AA x Aa	$4p^3q$	1/2	1/2	0
AA x aa	$2p^2q^2$	0	1	0
Aa x Aa	$4p^2q^2$	1/4	1/2	1/4
Aa x aa	$4pq^3$	0	1/2	1/2
aa x aa	q^4	0	0	1

Freq(AA) = 1* p^4 + ½*4p^3q + (1/4) 4p^2q^2 = p^2 (p+q)^2 = p^2.

Genotype frequencies unchanged

Hardy-Weinberg

genotype	gen 0	gen 1	gen 2
P(AA)	X	p^2	p^2
P(Aa)	Y	$2pq$	$2pq$
P(aa)	Z	q^2	q^2

After one generation of random mating, genotype frequencies remain unchanged and are given by HW proportions

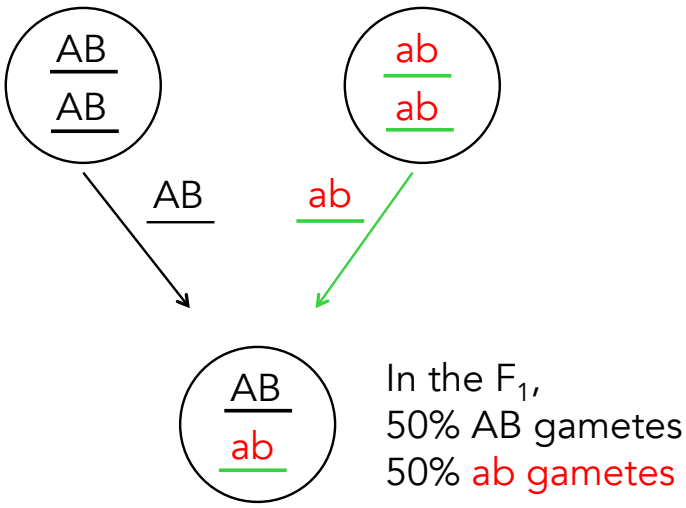
Assuming random mating, no migration, drift, or selection, then allele frequencies remain unchanged

More generally, for any number of alleles, freq(A_iA_i) = p_i^2, freq(A_iA_j) = $2p_ip_j$.

Multi-locus Hardy-Weinberg

Multi-locus HW

- When following multiple loci, we need to consider gametes, rather than alleles
  - For example, an AaBb parent gives four distinct gametes AB, Ab, aB, ab
  - While allele frequencies do not change under random mating, *gamete frequencies can*.
  - Concept of linkage disequilibrium



If A and B are unlinked, the F2 gamete frequencies are

AB 25% ab 25% Ab 25% aB 25%

Thus, even under HW conditions, *gamete frequencies change*

Linkage Disequilibrium

- Under linkage equilibrium, (LE), the frequency of gametes is the product of allele frequencies,
  - e.g. Freq(AB) = Freq(A)*Freq(B)
  - A and B are independent of each other
- When this does not occur, linkage disequilibrium (LD) is said to occur (also called gametic phase disequilibrium)
- The amount D_{AB} of disequilibrium for the AB gamete is given by
  - D_{AB} = Freq(AB) gamete - Freq(A)*Freq(B)
  - D > 0 implies AB gamete more frequent than expected
  - D < 0 implies AB less frequent than expected

Dynamics of D

- Under random mating in a large population, allele frequencies do not change. However, gamete frequencies do if there is any LD
- The amount of LD decays by (1-c) each generation
  - $D(t) = (1-c)^t D(0)$
- The expected frequency of a gamete (say AB) is
  - Freq(AB) = Freq(A)*Freq(B) + D
  - Freq(AB in gen t) = Freq(A)*Freq(B) + $(1-c)^t$ D(0)

The Decay of Linkage Disequilibrium

The frequency of the AB gamete is given by

$$\text{freq}(AB) = \text{freq}(A)\,\text{freq}(B) + D_{AB}$$

LE value

Departure from LE

If recombination frequency between the A and B loci is c, the disequilibrium in generation t is

$$D(t) = D(0)(1-c)^t$$

Initial LD value

Note that D(t) -> zero, although the approach can be slow when c is very small

Genotypic frequencies under HW

- Under multi-locus HW (random mating, no LD),
  - Freq(AABB) = Freq(AA)*Freq(BB)
  - i.e., can use single-locus HW on each locus, and then multiply the results
- When D is non-zero (LD is present), cannot use this approach
  - Rather, must follow gametes
  - Freq(AABB) = Freq(AB)*Freq(AB)

Population structure

Population Structure

Populations often show structure, with an apparently single random-mating population instead consisting of a collection of several random-mating subpopulations

Suppose there are n subpopulations, and let w_k be the probability that an random individual is from population k

Let p_{ik} denote the frequency of allele A_i in subpopulation k.

The overall frequency of allele A_i is $$p_i = \sum_{k=1}^{n} w_k * p_{ik}$$

The frequency of A_iA_i in the population is just

$$\text{freq}(A_iA_i) = \sum_{k=1}^{n} p_{ik}^2 \cdot w_k$$

Expressed in terms of the population frequency of A_i,

$$\begin{aligned}\text{freq}(A_iA_i) &= p_i^2 + \left(\sum_{k=1}^{n} p_{ik}^2 \cdot w_k - p_i^2\right) \\ &= p_i^2 + \text{Var}(p_i)\end{aligned}$$

Thus, unless the allele has the same frequency in each population ($\text{Var}(p_i) = 0$), the frequency of homozygotes exceeds that predicted from HW

Population structure also generates disequilibrium

Again suppose there are k subpopulations, each in linkage equilibrium

The population frequency of A_iB_j gametes is

$$\text{Freq}(A_iB_j) = \sum_{k=1}^{n} w_k * pA_{ik} * pB_{jk}$$

The population-wide disequilibrium becomes

$$\begin{aligned}D_{ij} &= \text{Freq}(A_iB_j) - \text{Freq}(A_i) * \text{Freq}(B_j) \\ &= \sum_{k]=1}^{n} w_k * p_{A_{ik}} * p_{B_{jk}} - \left(\sum_{k=1}^{n} w_k * p_{Aik}\right)\left(\sum_{k=1}^{n} w_k * p_{Bik}\right)\end{aligned}$$

Consider the simplest case of k = 2 populations

Let p_i be the frequency of A_i in population 1,
$p_i + \delta_i$ in population 2.

Likewise, let q_j be the frequency of B_j in population 1,
$q_j + \delta_j$ in population 2.

The expected disequilibrium becomes

$$D_{ij} = \delta_i * \delta_j * [\, w_1 (1 - w_1) \,]$$

Here, w_1 is the frequency of population 1

F_{ST}, a measure of population structure

- One measure of population structure is given by Wright's F_{ST} statistic (also called the fixation index)
- Essentially, this is the fraction of genetic variation due to between-population differences in allele frequencies
- Changes in allele frequencies can be caused by evolutionary forces such as genetic drift, selection, and local adaptation
- Consider a biallelic locus (A, a). If p denotes overall population frequency of allele A,
  - then the overall population variance is p(1-p)
  - Var(p_i) = variance in p over subpopulations
  - F_{ST} = Var(p_i)/[p(1-p)]

Example of F_{ST} estimation

Population	Freq(A)
1	0.1
2	0.6
3	0.2
4	0.7

Assume all subpopulations contribute equally to the overall metapopulation

Overall freq(A) = p = (0.1 + 0.6 + 0.2 + 0.7)/4 = 0.4

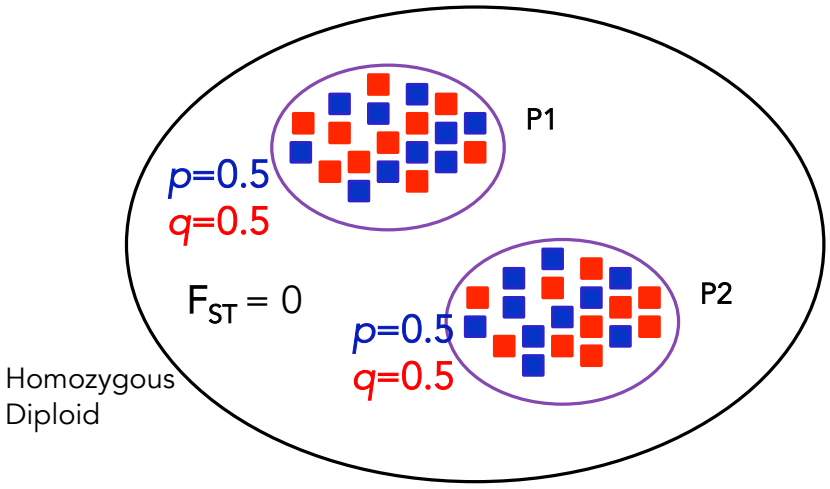
$Var(p_i) = E(p_i^2) - [E(p_i)]^2 = E(p_i^2) - p^2$

$Var(p_i) = [(0.1^2 + 0.6^2 + 0.2^2 + 0.7^2)/4] - 0.4^2 = 0.065$

Total population variance = $p(1-p) = 0.4(1-0.4) = 0.24$

Hence, $F_{ST} = Var(p_i) / [p(1-p)] = 0.065/0.24 = 0.27$

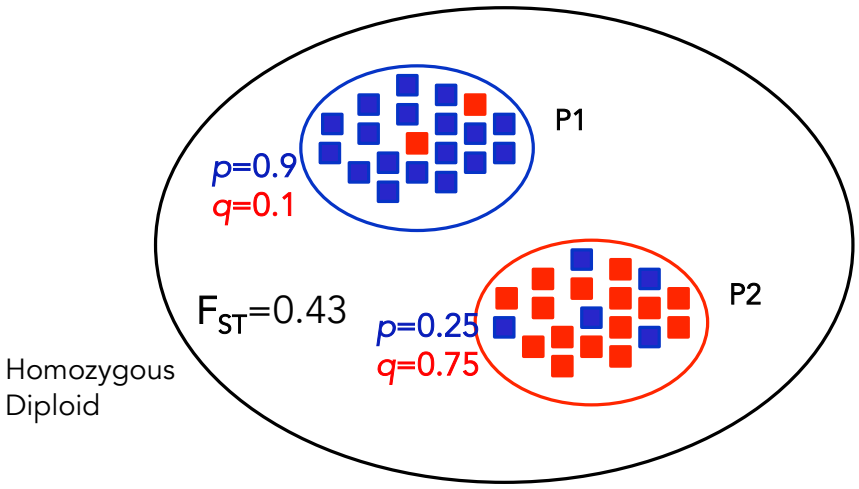
Graphical example of F_{ST}



No population differentiation

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

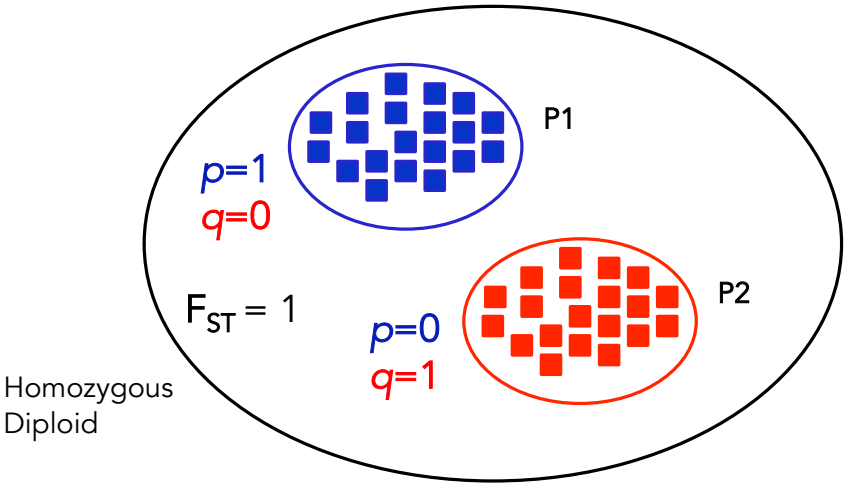
Graphical example of F_{ST}



Strong population differentiation

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

Graphical example of F_{ST}



Complete population differentiation

Modified from Escalante et al. 2004. Trends Parasitol. 20:388-395

Drift

Genetic Drift

Random sampling of 2N gametes to form the N individuals making up the next generation results in changes in allele frequencies.

This process, originally explored by Wright and Fisher, is called Genetic Drift.

Suppose there are currently i copies of allele A, so that freq(A) = p = i/(2N)

What is the probability that, following a generation of random sampling, the freq of A is j/(2N)?

This probability follows binominal sampling,

$$\Pr(i \text{ copies} \rightarrow j \text{ copies}) = \frac{N!}{(N-j)!j!}\left(\frac{i}{N}\right)^{j}\left(\frac{N-i}{N}\right)^{N-j}$$

p = i/N 1- p

Hence, if the current allele frequency is p, the expected allele frequency in the next generation is also p, but with sampling variance p(1-p)/(2N)

Thus, with N is large, the changes in allele frequency over any generate are expected to be rather small

However, the cumulative effects of generations of such sampling are very considerable.

Eventually, any random allele will either be lost from the population or fixed (frequency one).

If the allele has initial frequency p, then

Pr(Fixation) = p

Pr(loss) = 1- p

The expected time to fixation is on order of 4N generations.

Example in R

Effective Population Size

Actual vs. effective population size

- The actual size of a population N is far less important than the effective number of breeding individuals, or the effective population size, N_e, which sets the strength of drift.
- N_e replaces N in the previous results

Effective Population size, N_e

When the population is not ideal (changes over time, unequal sex ratio, uneven contribution from individuals), we can still compute an effective population size N_e which gives the size of an ideal population that behaves the same as our population

We will consider N_e under
population bottlenecks
unequal sex ratio
unequal contribution for all individuals

N_e under varying population size

If the actual population size varies over time, the effective population size is highly skewed towards the smallest value

If the populations sizes have been N(1), N(2), …, N(k), the effective population size is given by the harmonic mean

$$N_e = \frac{k}{\sum_{i=1}^{k} \frac{1}{N(i)}}$$

Suppose the population sizes are 10000, 10000, 10000, 100.

N_e becomes 399

If 10000 is replaced by 10^9, N_e becomes 400

N_e under unequal sex ratios

When there are different number of males (N_m) and females (N_f), the effective population size is skewed towards the rarer sex

$$N_e = \frac{4N_m * N_f}{N_m + N_f}$$

For example, suppose we used 2 pollen plants to fertilize a 1000 seed plants. What is N_e in this case?

N_e = (4*2*1000)/(2 + 1000) = 8

N_e under unequal individual contributions

Not all individuals contribute equally to the next generation. What effect does this have on N_e?

Let σ^2_0 be the variance in offspring number for individuals in the population, then

$$N_e \simeq \frac{2N}{\sigma_O^2/2 + 1}$$

If contributions follow a Poisson distribution with a mean of 2 offspring per parent (male + female replace each other), then $\sigma^2_0 = 2$, and $N_e = N$

Key results for drift

- Expected change in allele frequency = 0
- Variance from drift = $p(1-p)/[2N_e]$
- Prob(fixation) = p
- Expected time to fixation for a new mutation ~ $4N_e$ generations
- Time scale for drift ~ $4N_e$
- Number of new mutations per generation = 2Nu
- Expected # of new mutations destined to become fixed $2Nu(1/[2N]) = u$
  - Time between successful new mutations ~1/u

Mutation

Mutation

The second force that can change allele frequencies is mutation.

Mutation is a key fundamental force in that it introduces new variation into the population.

The simplest model is that a gene mutates back and forth between two states, A and a.

Let Pr(A -> a) = μ and Pr(a -> A) = ν

The new frequency, p′, of allele A following a single generation of mutation is

$$p' = p(1-\mu) + (1-p)\nu$$

Randomly-picked allele is A and this allele DOES not mutate

Random allele is a, but this allele mutates to A

Allele frequencies change on the order of the mutation rate (typically VERY slow), on order of 10^{-3} to 10^{-9}

At equilibrium, $p' = p = \mu/(\mu+\nu)$

This 2-state model is very unrealistic given what we know about DNA and gene structure

Crow and Kimura's (1964) Infinite Alleles Model

Each new mutation leads to a allele

Crow & Kimura looked at the balance between mutation introducing new variation and genetic drift (finite population size) removing it.

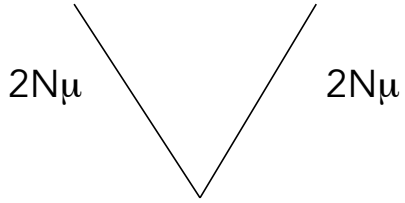
At equilibrium, the expected heterozygosity H of the population is

$$H = 4N_e u/(1+ 4N_e u)$$

We can use coalescent theory to see where this result comes from.

Consider two randomly-chosen alleles. They are different if they have experienced a mutation since their most recent common ancestor (i.e., their coalescence)

For a population of size N, this time has expected value 2N. This gives the expected number of mutations as $4N\mu$.



Expected number = $2N\mu + 2N\mu = 4N\mu$

Number is Poisson-distributed, mean = $4N\mu = \theta$

Pr(k mutations) = $\theta^k \exp(-\theta)/k!$

Pr(0 mutations) = $\exp(-\theta)$

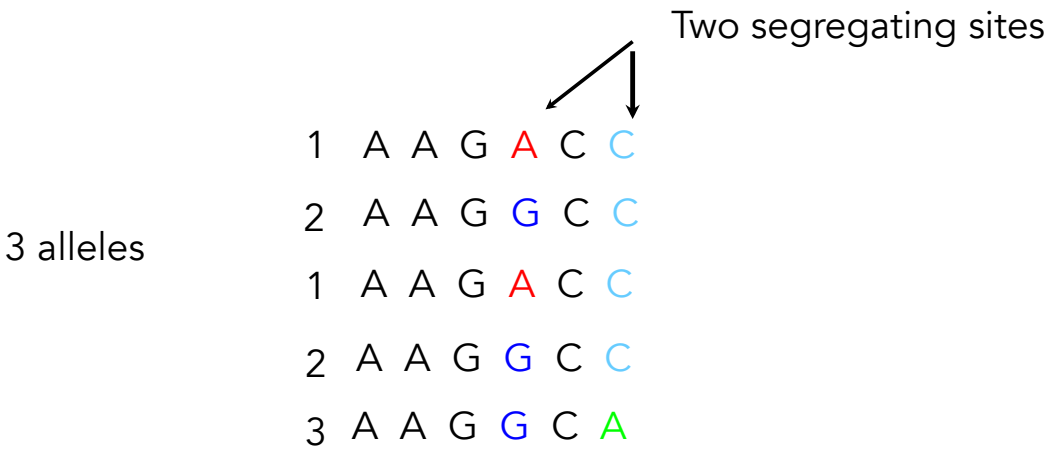
If $4N\mu = \theta >> 1$, most randomly-drawn alleles will be different. Lots of heterozygosity, H near one.

If $4N\mu << 1$, most randomly-drawn alleles will be identical. Almost no heterozygosity, H near zero.

Infinite sites model

- Infinite alleles model typically used for the analysis of haplotypes
- Infinite sites model assumes each new mutation occurs at a new site (i.e., nucleotide)
- Used in the analysis of (unphased) DNA sequence data

Infinite alleles vs. infinite sites



Measures of variation under infinite-sites

- Number S of segregating sites in a sample
  - $E(S) = \theta * a(n)$, $a(n) = \Sigma^{n-1} 1/i$
- Nucleotide diversity π = average number of pairwise differences
  - $E(\pi) = \theta$

Stepwise mutational models

A final popular class of mutational models that is motivated by DNA structure are stepwise mutation models

Consider a microsatellite locus (an STR).

$ACCACCACCACCACC = (ACC)_5$

$ACCACCACCACCACCACCACC = (ACC)_7$

Such loci are scored by their repeat number (5 and 7 above). The mutation process on such repeats is that they typically change repeat size by +/- one

Such loci are called STR (simple tandem repeats) or SSR (simple sequence repeats) in the literature

Hence, if two alleles have the same repeat size (say 5), this could result from:

The alleles sharing a common ancestor recently enough so that no mutations have occurred.

Or the alleles could have experienced several mutations since their most recent common ancestor. For example, one allele could have been a 6 that mutated to a 5, and the other a 4 that mutated to a 5.

Under stepwise mutation, identity in state (alleles have the same sequence) does not imply identity by descent (no mutation since their most recent common ancestor, MRCA)

The symmetric single-step mutation model is the most widely used.

Given allele is in state (repeat number) i,

$$\Pr(\text{stays at } i \text{ after one generation}) = 1-\mu$$

$$\Pr(i \rightarrow i+1) = \mu/2$$

$$\Pr(i \rightarrow i-1) = \mu/2$$

The analysis of even this simple model can be rather involved (requiring the use of Type II Bessel functions)

Selection

Fitness

- Some genotypes (on average) leave offspring than others
- Fitness, denoted by W, is a measure of the number of offspring.
  - $W_{AA} = 10$, $W_{Aa} = 5$
  - For every offspring that an Aa leaves, on average, AA leaves twice as many
- What matters is the ratios of fitnesses, not their actual values (the relative fitness of one genotype with respect to another).
  - Same allele-frequency dynamics as
  - $W_{AA} = 1$, $W_{Aa} = 0.5$ or $W_{AA} = 2$, $W_{Aa} = 1$

One locus with two alleles

Genotype	AA	Aa	aa
Frequency (before selection)	p^2	$2p(1-p)$	$(1-p)^2$
Fitness	W_{AA}	W_{Aa}	W_{aa}
Frequency (after selection)	$\frac{p^2 W_{AA}}{\overline{W}}$	$\frac{2p(1-p) W_{Aa}}{\overline{W}}$	$\frac{(1-p)^2 W_{aa}}{\overline{W}}$

Where $\overline{W} = p^2 W_{AA} + 2p(1-p) W_{Aa} + (1-p)^2 W_{aa}$

is the mean population fitness, the fitness of an random individual, e.g. $\overline{W} = E[W]$

- Suppose fitnesses are 1 : 1.2 : 1.4 for the genotypes qq : Qq : QQ, and p = freq(Q)= 0.2

	qq	qQ	QQ
Freq	$0.8^2 = 0.64$	$2*0.8*0.2 = 0.32$	$0.2^2 = 0.04$
Fitness	1	1.2	1.4
Freq*fit	0.64	0.384	0.056

Mean fitness = 0.64 + 0.384 + 0.056 = 1.08

Freq(Qq after selection) = 0.384/1.08 = 0.356

Freq(QQ after selection) = 0.04/1.08 = 0.037

New freq (Q) = (1/2)* 0.356 + 0.037 = 0.215

The new frequency p′ of A is just
freq(AA after selection) + (1/2) freq(Aa after selection)

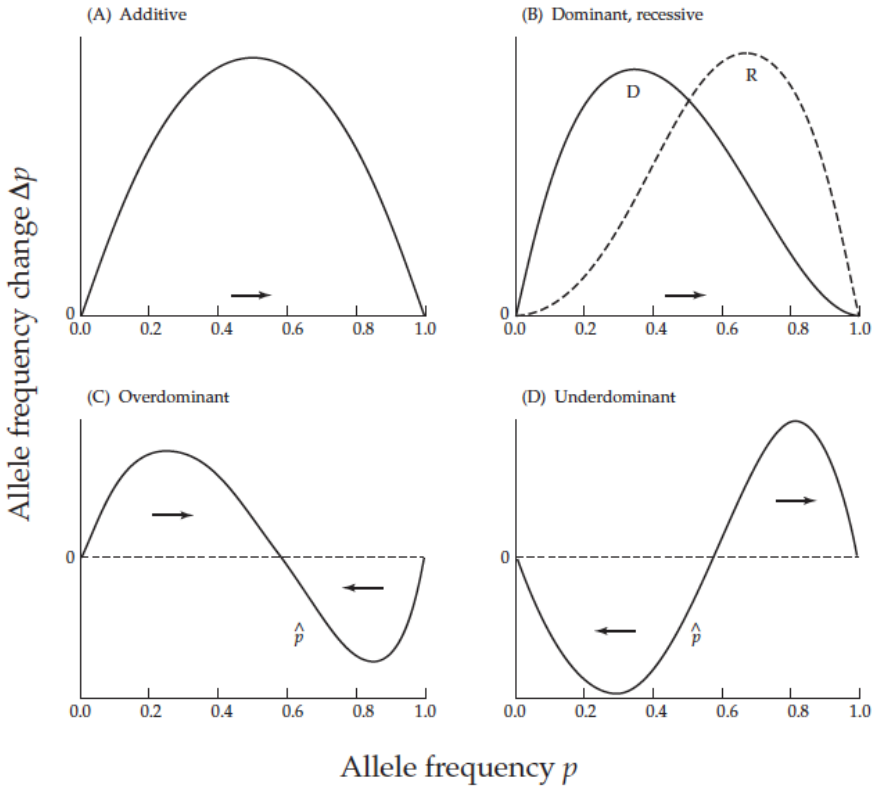
$$p' = \frac{p^2 W_{AA} + p(1-p)W_{Aa}}{\overline{W}} = p\,\frac{pW_{AA} + (1-p)W_{Aa}}{\overline{W}}$$

The fitness rankings determine the ultimate fate of an allele

If $W_{AA} \geq W_{Aa} > W_{aa}$, allele A is fixed, a lost

If $W_{Aa} > W_{AA}, W_{aa}$, selection maintains both A & a

Overdominant selection



Genotypes	aa	Aa	AA
fitness	1	1+(1+h)s	1+2s

$$\Delta p = \frac{sp(1-p)[1+h(1-2p)]}{\overline{W}}$$

For additive selection (h=0)

$$p_t = \frac{p_0}{p_0 + (1-p_0)e^{-st}}$$

$$t_{p_t,p_0} \simeq \frac{1}{s} \ln\left(\frac{p_t\,(1-p_0)}{p_0\,(1-p_t)}\right)$$

Time scale for selection ~ 1/s

General expression for selection with n allelles

Let p_i = freq(A_i), W_{ij} = fitness A_iA_j

$$p_i' = p_i \frac{W_i}{\overline{W}}, \qquad W_i = \sum_{j=1}^{n} p_j W_{ij}, \qquad \overline{W} = \sum_{i=1}^{n} p_i W_i$$

W_i = marginal fitness of allele A_i

$\overline{W}$ = mean population fitness = $E[W_i] = E[W_{ij}]$

If $W_i > \overline{W}$, allele A_i increases in frequency

If a selective equilibrium exists, then $W_i = \overline{W}$ for all segregating alleles.

Wright's formula

$$\Delta p_i = \frac{p_i(1-p_i)}{2\overline{W}} \frac{\partial \overline{W}}{\partial p_i}$$

Correct for 2 alleles

$$\Delta p_i = \frac{p_i}{2\overline{W}} \left(\frac{\partial \overline{W}}{\partial p_i} - \sum_{j=1}^{n} p_j \frac{\partial \overline{W}}{\partial p_j} \right)$$

Correct for n alleles

Can express Wright's formula in matrix form

$$\Delta p_i = \sum_{j=1}^{n} G_{ij} \cdot \frac{\partial \ln \overline{W}}{\partial p_j},$$

$$G_{ij} = \begin{cases} p_i(1-p_i)/2 & i = j \\ -p_i p_j/2 & i \neq j \end{cases}$$

$$\Delta \mathbf{p} = \frac{1}{\overline{W}} \mathbf{G} \boldsymbol{\nabla} \overline{W} = \mathbf{G} \boldsymbol{\nabla} \ln(\overline{W}) \qquad \boldsymbol{\nabla} \overline{W} = \begin{pmatrix} \partial \overline{W}/\partial p_1 \\ \vdots \\ \partial \overline{W}/\partial p_n \end{pmatrix}$$

Selection and mutation

Selection vs. mutation

- Most new mutations are deleterious, so an interesting question is that is their expected frequency under mutation (introducing them) and selection removing them. We assume a very large population
- Since we expect their frequency to be very low, back mutation is ignored

Fitness are 1 : 1-hs :1- s

$$\tilde{p} = \frac{\mu}{hs}, \quad \text{provided} \quad h \gg \sqrt{\mu/s}$$

Gives p = 2u/s for additive (h = ½)

Gives p = u/s for dominant (h = 1)

For a recessive (h = 0) $\tilde{p} = \sqrt{\frac{\mu}{s}}$

What is the reduction in fitness at equilibrium (the mutational load)?

For a recessive

$$\overline{W} = 1 - s \cdot \text{freq}(aa) = 1 - s\left(\sqrt{\frac{\mu}{s}}\right)^2 = 1 - \mu$$

More generally,

$$\begin{aligned} \overline{W} &= 1 - 2hs\,\tilde{p}(1-\tilde{p}) - s\tilde{p}^2 \\ &\simeq 1 - 2hs\,\tilde{p} = 1 - 2hs\left(\frac{\mu}{hs}\right) = 1 - 2\mu \end{aligned}$$

Note that the fitness reduction is independent of the strength of selection, and only a function of the mutation rate. This is known as Haldane's principal

Selection and drift

When is drift stronger than selection?

- Recall (for additive selection), that the time scale for fixation scales as 1/s, while for drift it scales as $4N_e$.
- Hence, when the time scale for selection is quicker ($1/s < 4N_e$ or $4N_e s > 1$), selection dominates
- When drift is faster ($1/s > 4N_e$ or $4N_e s < 1$), drift dominates

More formal treatment

- The time-scale analysis can be made more formal by considering u(p), the probability that an allele is fixed given it starts are frequency p.
- Kimura (using diffusion approximations) showed for an additive allele (1:1+s :1+2s) that

$$u(p) = \frac{1 - \exp(-4N_e sp)}{1 - \exp(-4N_e s)}$$

Starting with a single copy,

$$u\left(\frac{1}{2N}\right) \simeq \frac{1}{2N} \quad 4N_e|s| \ll 1$$

Behaves as a neutral allele, as u(p) = p)

Fixation probability small, even when favored!

$$u\left(\frac{1}{2N}\right) \simeq \quad 2s\left(\frac{N_e}{N}\right) \quad 4N_e s \gg 1$$

N_e under unequal individual contributions (cont)

$$N_e \simeq \frac{2N}{\sigma_O^2/2 + 1}$$

If all individuals contribute EXACTLY the same number of offspring, $\sigma^2_0 = 0$, and $N_e = 2N$, so that the effective pop size is twice the actual size