Population and Quantitative Genetics – HS18

v1.2

Gleb Ebert

October 7, 2019

This document aims to summarize the lecture Population and Quantitative Genetics as it was taught in the autumn semester of 2018. It is heavily based on the slides and often contains passages verbatim. Unfortunately I cannot guarantee that it is complete or free of errors. You can contact me under `glebert@student.ethz.ch` if you have any suggestions for improvement. The newest version of this summary can always be found on my website: `http://www.glebsite.ch`

Contents

1 Molecular Markers, HWE, Genetic Variation

Population genetics apply Mendel's laws and other genetic principles to populations. It studies genetic variation within and between populations and species and the forces that result in evolutionary changes in populations and species through time. Population genetics are useful in studies of evolutionary processes, conservation, medicine, agriculture and other fields.

1.1 Useful Formulas

mean	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
variance	$V_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
standard deviation	$sd = \sqrt{V_x}$

1.2 Genetic Variation

"Nothing in biology makes sense except in the light of evolution"
— Theodosius Dobzhansky

Modern synthesis brought Mendelian genetics together with Darwin's theory of natural selection to help quantifying the genetic variation in natural populations.
The classical view of genome organization was that wild type alleles made up the whole genome with a few mutations in between. The balanced view however said that for each gene there are multiple alleles and that heterozygosity is possible.

1.3 Genetic variation

Experimental methods for detecting genetic variation (in chronological order):

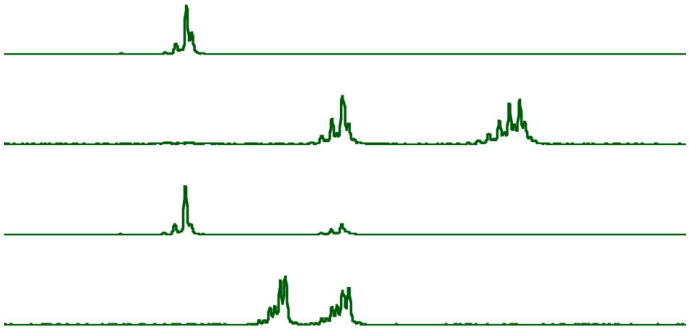
- Allozyme electrophoresis
- DNA (Sanger) sequencing
- Restriction fragment length polymorphism (RFLP)
- Simple sequence repeats (SSR or microsatellites)
- Amplified fragment length polymorphism (AFLP)
- Single nucleotide polymorphisms (SNPs)
- Next-generation sequencing (NGS)

1.3.1 Allozyme electrophoresis

Enzymes that differ in electrophoretic mobility as a result of allelic differences in a single gene are called **allozymes**. **Allozyme electrophoresis** separates these. Allozymes underestimate the levels of DNA polymorphism because they detect only a subset of existing amino acid replacement and they do not detect synonymous mutations. However, they may also overestimate polymorphisms because they represent mostly group 1 enzymes (common in tissues and body fluids) and enzyme polymorphisms may not be neutral and therefore not reflect polymorphism elsewhere in the genome. Allozyme electrophoresis was replaced by DNA electrophoresis.

1.3.2 Microsatellites

Microsatellites are highly polymorphic (in number of repeats) markers that are widely used in animals, plants and fungi to assess genetic variation. They are also known as **short tandem repeats (STRs)** or **simple sequence repeats (SSRs)**. Together with their longer cousins, the minisatellites, they are classified as **variable number tandem repeats (VNTRs)**. Like allozymes they are **codominant** markers. Homozygotes can be distinguished from heterozygotes.
$\ldots$GATCGA$\mathbf{(GC)}_7$TAGCCGAT$\ldots$



1.3.3 Amplified Fragment Length Polymorphisms (AFLPs)

In AFLP DNA is first digested by one or more digestion enzymes. Then adapters that are specific to the half-sites are ligated to the fragments. Some of these fragment are then selectively amplified with two primers each that are complimentary to the adaptor and the restriction site respectively. The amplification products are then visualized using gel electrophoresis. AFLPs are **dominant** markers. Homozygotes cannot be distinguished from heterozygotes.

1.3.4 Types of Polymorphisms

A polymorphism that does not alter the amino acid sequence is called **synonymous**. They are a consequence of the redundancy of the genetic code. **Non-synonymous** or **replacement** polymorphisms change the amino acid. If a polymorphism is **noncoding** or **silent**, it does not affect nucleotides in coding regions. Nucleotide polymorphisms can be devided into insertion/deletion polymorphisms (**indels**) and single-nucleotide polymorphisms (**SNPs**). A unique combination of linked genetic markers is often called a **haplotype**.

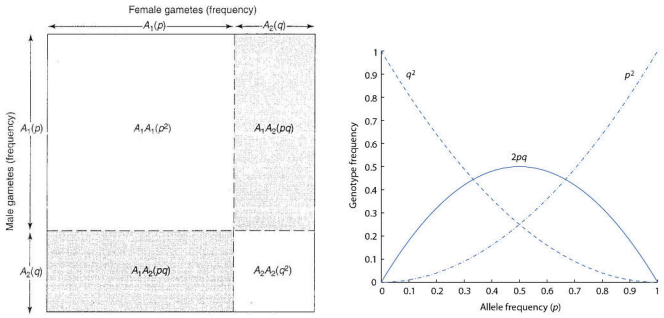
1.3.5 Genetic Variation in Natural Populations

A **population** is a group of interbreeding, same-species individuals that exist together in time and space. If it is **random-mating** (**panmictic**), the probability of mating between individuals of particular genotypes is equal to the product of their individual frequencies in the population.

1.4 The Hardy-Weinberg Principle

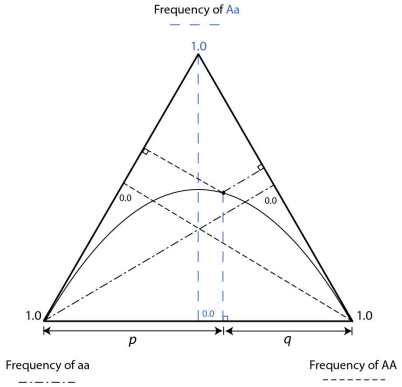
The Hardy-Weinberg principle describes allele frequencies in diploid populations. It applies after one generation of random mating. The principle assumes that allele frequency changing factors (natural selection, drift, ...) are absent. At low frequencies the majority of a certain allele occurs in heterozygous individuals.

$$(p+q)^2 = p^2 + 2pq + q^2 = 1$$



1.5 De Finetti Diagram

Below the De Finetti diagram for one locus with two alleles is shown. Each corner and the corresponding line leading out from it are a frequency coordinate of the respective genotype. The parabola describes HW expected genotype frequencies.



1.6 Estimating Allele Frequencies

Samples are taken from populations to estimate genotype and allele frequencies (maximum-likelihood approach). In a sample of N individuals with N_{11} individuals of genotype A_1A_1, N_{12} of A_1A_2 and N_{22} of A_2A_2 ($N_{11}+N_{12}+N_{22}$) the estimated genotype and allele frequencies are

$$\hat{P} = \frac{N_{11}}{N} \qquad \hat{H} = \frac{N_{12}}{N} \qquad \hat{Q} = \frac{N_{22}}{N}$$

$$\hat{p} = \frac{N_{11} + \frac{1}{2}N_{12}}{N} \qquad \hat{q} = \frac{N_{22} + \frac{1}{2}N_{12}}{N}$$

1.7 Dominance

Dominance refers to the effect on the phenotype of one allele relative to another recessive allele. To estimate the frequency of the recessive allele a, one has to assume that the proportion of genotype aa in the population equals $\frac{N_{22}}{N} = q^2$ and thus $\hat{q} = \sqrt{\frac{N_{22}}{N}}$. We therefore have to assume that the population is in HWE.

Some rare ($< 10\%$) recessive alleles cause human diseases like albinism, cystic fibrosis or sickle-cell anemia. The frequency of carriers of this allele is $\hat{H} = 2\hat{p}\hat{q}$.

1.8 Testing Hardy-Weinberg Proportions

The chi-squared test quantifies the quality of the fit between observed and expected genotype frequencies (with $k = \#$ genotype classes).

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

The resulting value and the degrees of freedom $df = k - 1 -$ (# parameters estimated from the data) give us the probability for the difference between observed and expected or a more extreme case.
The sample size should be over 50 and the expected number in all classes should be greater than 5. Otherwise an **exact test** should be used.

1.9 Measures of molecular genetic variation

- **HW-expected heterozygosity / gene diversity** $H_E = 1 - \sum_{i=1}^{n} \hat{p}_i^2$
- **Observed heterozygosity** (no HW) $H_O = \sum_{i<j}^{n} \hat{P}_{ij}$
- **Effective number of alleles** (good measure of **true allelic diversity**): $\hat{n}_e = \frac{1}{1-\hat{H}}$
- **average number of pairwise differences per site** $\pi = \sum \frac{p_{ij}}{\#\text{ comparisons}}$

2 Genetic Drift

Genetic drift is the **random alteration of allele frequencies** that results from the sampling of gametes from generation to generation in finite populations. It has the same expected effect on all loci in the genome.

2.1 Wright-Fisher Model

The model is a simplified view of reproduction, sampling 2N gametes from an infinite gamete pool. Major assumptions include:

- equal sex ratio
- equal fitness
- non-overlapping generations
- constant population size

2.2 Allele fixation

The chances of fixation are equal to the initial allele frequency. Over replicate generations, the mean allele frequency does not change, but the distribution of the allele frequencies changes.
In very large populations, random changes in allele frequency will be minor, but in small populations, genetic drift may cause large fluctuations in allele frequencies across generations and can result in chance fixation or loss of alleles and increased autozygosity (IBD).

3 Populations

The **census size** is the total amount of individuals of in a population. The **breeding population size** is the number of sexually mature individuals. The **effective population size** is thought to be the appropriate measure for evolutionary studies it is often quite different (typically lower) than the breeding population size due to complicating factors such as variation in sex ratio, offspring number per individual (family size) and numbers of breeding individuals across generations.
N_e refers to an ideal population of size N in which all parents have an equal expectation of being the parents of any progeny individual (Poisson-distributed family size). It can also be understood as the size of an idealized Wright-Fisher population that would produce the same amount of inbreeding, allele frequency variance, or heterozygosity loss as the (empirical) population under study.

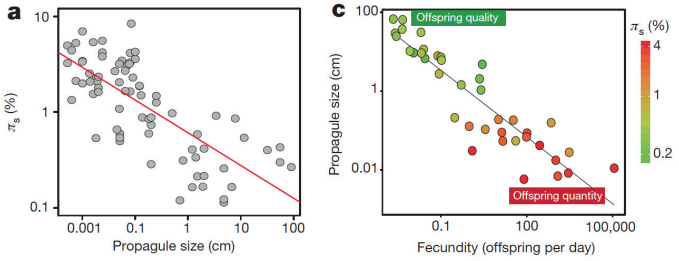
$$N_e = \frac{4N_f N_m}{N_f + N_m}$$

3.1 Bottleneck & founder effect

A **bottleneck** is a period during which only few individuals survive to continue the existance of the population. The **founder effect** describes a population that has grown from a few founder individuals. Populations descended from a small founder group may have low genetic variation or by chance have a high or low frequency of particular alleles (consequences of low N).
N_e is determined by the harmonic mean across generation:
$\frac{1}{N_e} = \frac{1}{3}\left(\frac{1}{100} + \frac{1}{10} + \frac{1}{100}\right) = 0.04 \qquad \Rightarrow N_e = 25$

3.2 Molecular data

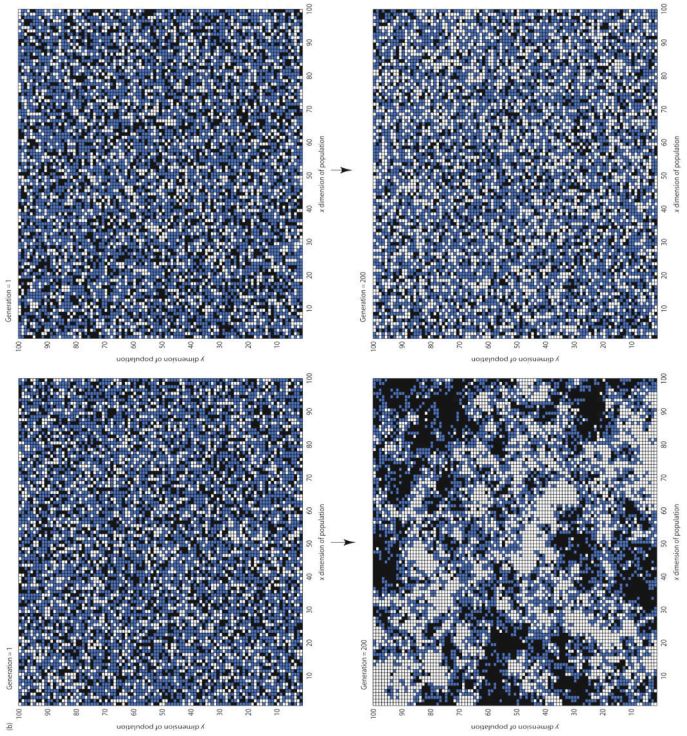
In practice, information about long-term N_e is available from estimates of the population mutation parameter $\theta = 4N_e\mu$. One of the two commonly used estimators of the parameter θ is π, the **average pairwise nucleotide diversity**. High nucleotide diversity implies high N_e and low nucleotide diversity implies low N_e. Examples: $N_{e,\text{Human}} \sim 5 * 10^4$, $N_{e,D.\ melanogaster} > 10^6$



The geographic spread of a species is inversely correlated with genetic diversity.

3.3 The Orthodox Paradigm

When populations are subdivided because of geographical, ecological, or behavioral factors, genetic connectivity among subpopulations is often reduced and depends on the amount of genetically effective gene flow. Gene flow indicates movement of individuals or gametes between groups that results in genetic exchange.



The upper graph shows alleles after mating with a random individual of the whole population (99x99 mating neighbourhood). The mating neighbourhood of the lower graph is only 3x3.

3.4 F-statistics

F-statistics are a measure of the **deficit of heterozygotes** relative to expected Hardy-Weinberg proportions in the specified base population. The F parameters are thus inbreeding coefficients for different specified base populations. F_{ST} is also known as **fixation index** and ranges from 0 (all subpopulations have equal allele frequencies) to 1 (all subpopulations are fixed for one or the other allele). F is also widely used as a measure of **allelic differentiation between subpopulations**, regardless of the number of alleles. Two populations can have the same F_{ST} while not having any allels in common. There is a lack of distinction between **fixation** and **differentiation**.

3.5 Wahlund's Principle

Population substructuring is not always obvious, and as a consequence, a sample may sometimes consist of individuals from different subpopulations. If subpopulations are lumped together and there are differences in allele frequencies among these subsamples, there will be a **deficiency of heterozygotes** and an **excess of homozygotes**, even if HW proportions exist within each subsample. The difference between expectation and reality is the effect size.

	Initial subpopulations	Fused population
Allele freq. q	0.4 and 0.0	$\frac{0.4+0.0}{2} = 0.2$
Var. in q	$\frac{(0.4-0.2)^2+(0.0-0.2)^2}{2} = 0.04$	0
Freq. of aa	$\overline{q^2} = \frac{0.16+0.0}{2} = 0.08$	$0.2^2 = 0.04$
Freq. of Aa	$\overline{p^2} = \frac{0.36+1.0}{2} = 0.68$	$0.8^2 = 0.64$

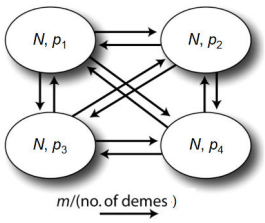
$H_I = \frac{1}{n}\sum_{i=1}^{n} \hat{H}_i$	The average observed heterozygosity within each subpopulation.
$H_S = \frac{1}{n}\sum_{i=1}^{n} 2p_i q_i$	The average expected heterozygosity of subpopulations assuming random mating withing each subpopulation.
$H_T = 2\hat{p}\hat{q}$	The expected heterozygosity of the total population assuming random mating within subpopulations and no divergence of allele frequencies among subpopulations.

Wright's **fixation indices** measure the consequences of population subdivision. For two levels of population organization they look as follows (IS = individual subpopulation, ST = subpopulation total, IT = individual total):

$F_{IS} = 1 - \frac{H_I}{H_S}$	The average difference between observed and HW-expected heterozygosity within subpopulations due to nonrandom mating.
$F_{ST} = 1 - \frac{H_S}{H_T}$	The difference between the average expected heterozygosity of subpopulations and the expected heterozygosity of the total population. Reduction in heterozygosity due to divergence in allele frequency among subpopulations.
$F_{IT} = 1 - \frac{H_I}{H_T}$	The average difference between observed heterozygosity within subpopulations and the expected heterozygosity of the total population, due possible to nonrandom mating and allele frequency divergence among subpopulations.

3.6 General island model

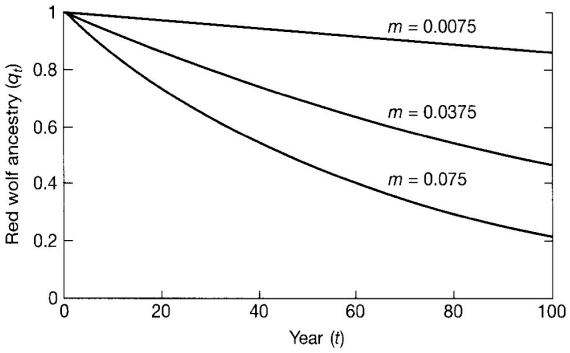
The general island model describes interactions between multiple demes. A **deme** is a local group of individuals from the same taxon that interbreed with each other and share a gene pool. In the graph below, N is the number of individuals per deme and m is the proportion of immigrants.



Wright showed, that **at equilibrium between drift and gene flow** (for small m) $F_{ST} \approx \frac{1}{4Nm+1}$

3.7 Continent island model

This model deals with unidirectional gene flow. Real-world examples include species on islands with nearby large land masses, or aquatic species in ponds with a nearby lake as the source of gene flow. An example is the hybridization between red wolves with coyotes. Depending on the mating rate of red wolves with coyotes, red wolf ancestry will disappear sooner or later.



3.8 Stepping-stone model

For the (linear) stepping-stone model we assume that subpopulations are arranged in a one-dimensional spatial pattern and gene flow is restricted to adjacent subpopulations ($m/2$). A more generally applicable version is the two-dimensional stepping-stone model, with migrants being exchanged between the 4 adjacent demes ($m/4$). Stepping-stone structure leads to **isolation by distance**.

3.9 Jost's D

D_{Jost} is a measure of relative differentiation.

$$H_{ST} = \frac{H_T - H_S}{1 - H_S} \quad \Longrightarrow \quad D_{Jost} = \frac{H_T - H_S}{1 - H_S}\left(\frac{n}{n-1}\right)$$

Applying D_{Jost} to the finite-island model:
$D_{Jost} \approx \frac{\mu n}{m}$ for moderate n.

n	N	m	μ	G_{ST}	D
5	100	0.01	0.001	0.127	0.282
5	1′000			0.014	0.282
5	10′000			0.001	0.282
10	10′000			0.002	0.469
20	10′000			0.002	0.651
40	10′000			0.002	0.793
80	10′000			0.002	0.886
160	10′000			0.002	0.940

$$D_{Jost} = 1 - \frac{J_{between}}{J_{within}}$$

When does this really matter?

- SNPs in **pairwise** comparisons: possibly little
- Markers with high mutation rates: a lot
- Theory development hampered by continued reliance on F_{ST}: role of degree of population subdivision (n demes) and μ?

4 Mutations

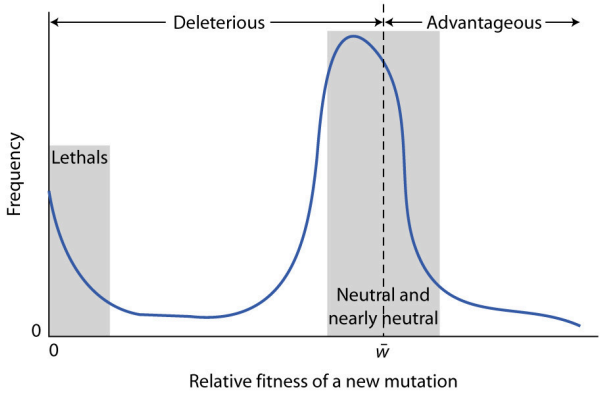
Mutations are the **original source of genetic variation**. They may involve changes in a single nucleotide, part of a gene, part of a chromosome, a whole chromosome, or entire sets of chromosome. Mutations can be induced by specific **mutagens** – UV light, chemicals, radiation. Such specific mutagens typically cause certain types of mutations. For **spontaneous mutations** the immediate cause for the mutation is unkown.

4.1 Transposable elements

Transposable elements are pieces of DNA that are capable of moving and replicating themselves withing the genome of an organism, often causing spontaneous visible mutants.

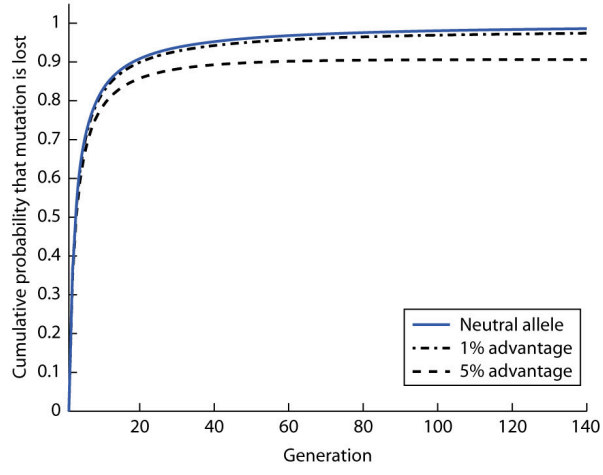
4.2 Fitness effets of mutations

Relative fitness of a given mutant depends on the environment and the alleles at other loci, i.e. the **genetic background**. The distribution of fitness effects is approximately bimodal – most mutations are either very **deleterious** (i.e. they cause lethality or near lethality) or **neutral / nearly neutral**. **Advantageous** mutations are presumably very rare but important.



4.3 Fate of a single mutation

When a new mutation occurs it is the only copy in the entire population and a single individuals is heterozygous for the mutation: A_1A_2. If coalescence is assumed the chance that a new allele is fixed is $\frac{1}{2N_e}$. Conversely, the probability that it is lost is $1 - \frac{1}{2N_e}$. In the real world the expected time to fixation for a rare allele is $T_1(p) \sim 4N_e$ generations.



4.4 Infinite-alleles model

The IAM proposes that mutation will increase the number of alleles and genetic drift will reduce it. It assumes that each mutation creates a new, unique allele. The expected equilibrium heterozygosity for the infinite-alleles neutral model is

$$H_e = \frac{4N_e\mu}{4N_e\mu + 1} = \frac{\theta}{\theta + 1}$$

θ is the **population mutation parameter** and is defined as $\theta = 4N_e\mu$. The equilibrium assumes that the distribution of alleles remains constant but allele frequencies and identities change constantly.

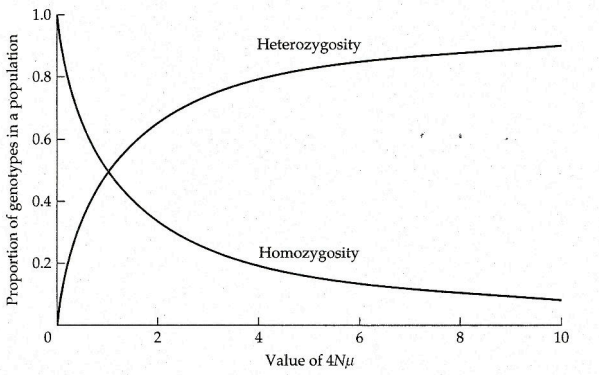


FIGURE 4.7 Plot of average homozygosity and average heterozygosity for the infinite-alleles model. Intermediate values of heterozygosity are maintained over only a small range of $\theta = 4N_e\mu$.

4.5 Stepwise mutation model

The SMM assumes that mutation occurs only to adjacent states (e.g. the number of repeats does not increase by more than one with each mutation). In contrast to the IAM mutation may produce alleles that are already present in the population. Thus both generation of variation and the equilibrium level of heterozygosity should be lower.

4.6 Infinite-site model

The infinite-sites model is usually used when working with DNA sequences. Every nucleotide mutation is assumed to occur at a previously unmutated site. Thus every segregating (polymorphic) site can only be two of four nucleotides.

4.7 Finite-site model

The finite-sites model allows for mutations to occur at already mutated sites. Multiple mutations can obscure patterns of relatedness (leading to homoplasy).

5 Linkage disequilibrium and recombination

Gametic phase disequilibrium or linkage disequilibrium (LD) is the nonrandom association of alleles at different loci into gametes (haplotypes). Gamete frequencies and disequilibrium can be influenced by selection, inbreeding, genetic drift, gene flow and mutation. The level of recombination between loci and N_e (**population recombination parameter** $\rho = 4N_ec$ with **recombination rate** c) strongly affect the extent of linkage disequilibrium . Linkage disequilibrium can occur between closely linked as well as unlinked loci (even across chromosomes).

5.1 Measuring linkage disequilibrium

Assume a large random-mating population with discrete generations segregating for two allels each at loci A (alleles A_1 and A_2) and B (alleles B_1 and B_2). Gamete frequencies are given by x_{ij} values. The frequency of alleles A_1 and B_1 are given by p_1 and q_1 respectively. Additionally $p_1+p_2=1$, $q_1+q_2=1$ and $\sum x_{ij}=1$.

Gamete	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

If the association between alleles within gametes is random the frequency of each gamete is equal to the product of the frequencies of the alleles it contains (left). Non-random association of alleles lead to a daviation D that is added to the expected frequencies (right). x_{11} and x_{22} are so-called coupling gametes while x_{12} and x_{21} are repulsion gametes.

	Random association	Non-random association
x_{11}	$= p_1q_1$	$= p_1q_1 + D$
x_{12}	$= p_1q_2$	$= p_1q_2 - D$
x_{21}	$= p_2q_1$	$= p_2q_1 - D$
x_{22}	$= p_2q_2$	$= p_2q_2 + D$

D is called the **linkage disequilibrium parameter** and is a measure of the deviation from random association between alleles at different loci.

$$D = x_{11} - p_1q_1 \qquad \text{(observed - expected)}$$
$$D = x_{11}x_{22} - x_{12}x_{21}$$

D is thus the product of the frequencies of the coupling gametes minus the product of the frequicies of the repulsion gametes. D has a **maximum value of 0.25** when there are only coupling gametes ($x_{11}=x_{22}=0.5$) and a **minimum value of −0.25** when there are only repulsion gametes ($x_{12}=x_{21}=0.5$). The decay of LD is proportional to the population recombination parameter $\rho = 4N_ec$.
Changes in gamete frequencies can take place only through recombination (with rate c; $c_{\max=0.5}$) in **double heterozygotes**.

Genotypes	Gametes
A_1B_1/A_1B_1	A_1B_1
A_1B_1/A_1B_2	$A_1B_1,\ A_1B_2$ (50% each)
A_1B_1/A_2B_2	$A_1B_1,\ A_2B_2,\ A_1B_2,\ A_2B_1$ (25% each)

$$D_t = (1-c)^t D_0 \qquad D' = \frac{D}{D_{\max}} \qquad r^2 = \frac{D^2}{p_1p_2q_1q_2}$$

t being the number of generations. Other measures of linkage disequilibrium include D', which allows comparisons of LD levels irrespective of how close we are to equilibrium, and r^2, which is the squared allele frequency correlation within gametes (range = [0, 1]) and takes allele frequency differences into account.

5.2 Population admixture

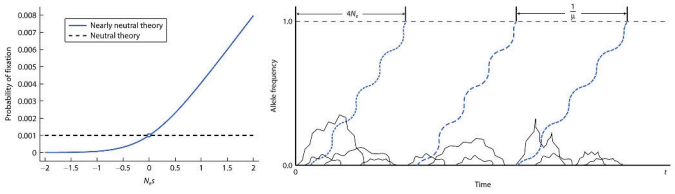
Population admixture can generate strong gametic disequilibrium when source allele frequencies are divergent. This is also called the **two-locus Wahlung effect**. The two populations are assumed to be at their respective gametic equilibrium and the mixture population consists of an equal number of gametes from the two source populations.

Gam. / D	Gam. freq.	Pop. 1	Pop. 2	Mix
A_1B_1	g_{11}	0.01	0.81	0.41
A_2B_2	g_{22}	0.81	0.01	0.41
A_1B_2	g_{12}	0.09	0.09	0.09
A_2B_1	g_{21}	0.09	0.09	0.09
D		0.0	0.0	0.16
D'		0.0	0.0	0.16/0.25 = 0.64

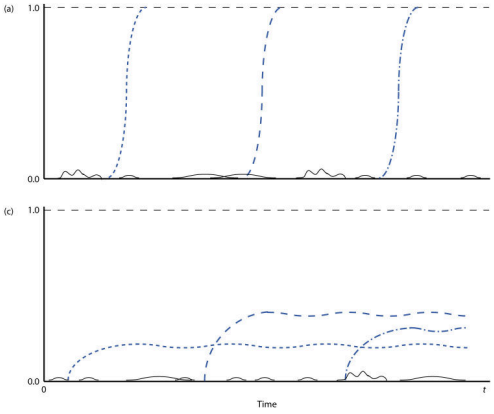
6 Neutral theory and coalescent

6.1 Neutral theory of molecular evolution

The neutral theory of molecular evolution states that genetic variation is primarily influenced by mutation generating variation and genetic drift eliminating it. Different molecular variants have almost identical relative fitnesses, i.e. they are neutral with respect to each other. The actual definition of selective neutrality depends on whether changes in allele frequency are primarily determined by genetic drift – when $s < \frac{1}{2N}$ (with s being the **selection coefficient**). The neutral theory has provided the null hypothesis for examining the amount and pattern of molecular genetic variation. It was later generalized to form the **nearly neutral theory**, which states that $|2N_es| \approx 1$ (context-dependent "weak selection").



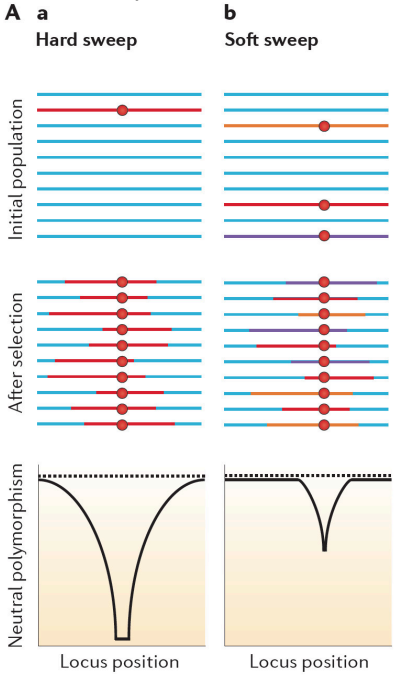
Selectively neutral mutations take an average $4N_e$ individuals to become fixed and the time between such fixations is on average $\frac{1}{\mu}$ (μ is the mutation rate).



The dwell time of new mutations under directional selection (top) and balancing selection (bottom).

6.1.1 Selective Sweeps

Selective sweeps happen when a beneficial allele carries other neutral alleles close to it along through hitchhiking. It results in less diversity in the beneficial alleles vicinity.



6.2 Coalescent theory

Coalescent events mark the timepoint of the **most recent common ancestro (MRCA)** of two instances in a population.

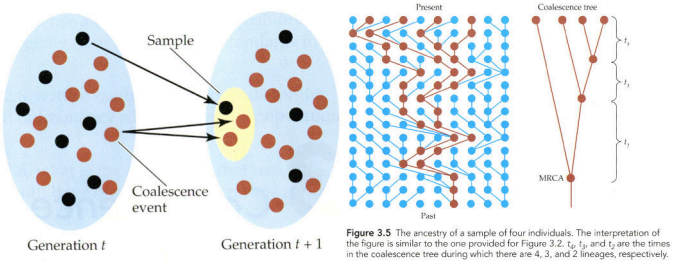


Figure 3.5 The ancestry of a sample of four individuals. The interpretation of the figure is similar to the one provided for Figure 3.2. t_4, t_3, and t_2 are the times in the coalescence tree during which there are 4, 3, and 2 lineages, respectively.

$$E(T_k) = \frac{2N}{\frac{k(k-1)}{2}}$$

T_k is the expected time in which there are k lineages

6.2.1 Site frequency spectrum

A **singleton** is a mutation that occurs only once in a population. A **doubleton** occurs twice. The graph below is called a **site frequency spectrum (SFS)**

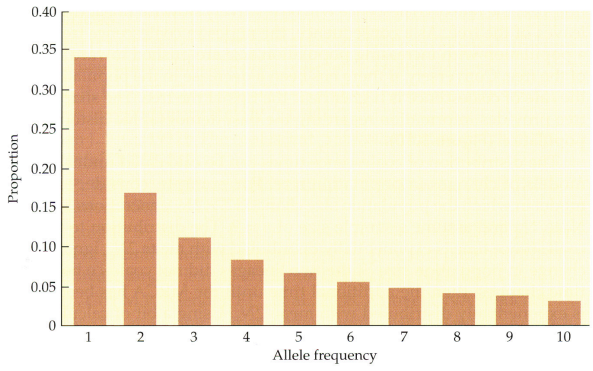
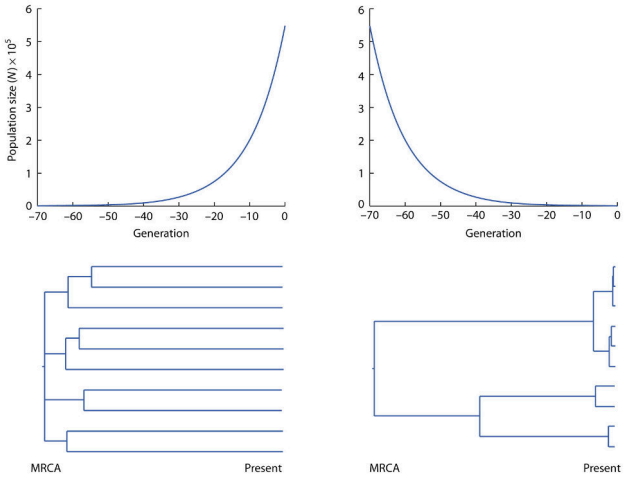


Figure 3.9 The expected site frequency spectrum (SFS) for a sample of $n = 10$ haploid individuals under the standard neutral coalescence model with infinite sites mutation.

6.2.2 Effects of exponential population growth and shrinkage



6.2.3 Watterson's estimator of the population mutation parameter

The population mutation parameter is $\theta = 4N_e\mu$. Watterson's estimator is defined as follows:

$$\theta_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

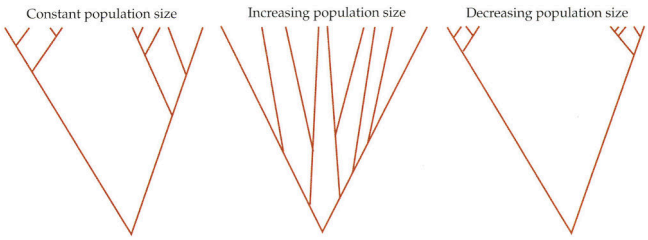
θ_W (θ_S) only depends on the number of segregating sites S, taking into account the number of sampled sequences n. Dividing by the total sequence length in bp yields the per-site θ_W.

6.2.4 Tajima's D

Tajima's D statistic is a test of selection or non-constant N_e

$$d = \pi - \theta_W \qquad D = \frac{d}{Var(d)^{\frac{1}{2}}}$$

- θ_W and π should estimate the same parameter
- $E(D) \approx 0$ under neutrality and constant N_e
- excess of low-frequency polymorphism: $\theta_W > \pi, \quad D < 0$
- excess of intermediate-frequency SNPs: $\theta_W < \pi, \quad D > 0$



$\theta_\pi \approx \theta_W$	$\theta_\pi < \theta_W$	$\theta_\pi > \theta_W$
$D \approx 0$	$D < 0$	$D > 0$

Differences in the shape of genealogies are the basis of Tajima's D test. Changes in N_e over time change the probablility of coalescence over time as well. SFS' depend on the change of N_e over time. E.g. population growth favours singletons compared to a constant N_e. The SFS for non-synonymous SNPs is more biased towards singletons than the one for synonymous SNPs, as the former are selected against. This phenomenon is called **purifying selection**.

7 Quantitative traits

"Quantitative characters are those differences between individuals that are of degree rather than of kind, that are quantitative rather than qualitative."
— Falconer and MacKay

Nowadays the majority of improvements in yield of agricultural products are based on breeding for quantitative traits. The three leading causes of mortality in industrialized nations, heart disease, cancer and diabetes, are all quantitative traits. The response to infectious diseases is a quantitative trait as well. Examples from evolutionary biology and ecology include beak size in Darwin's finches as well as the changing migration habits in response to climate change of the European blackcap.
The majority of traits under selection are quantitative and the alleles at all loci contributing to the phenotypic variation act predominantly additively.

7.1 Categories of quantitative traits

- **Continuous traits** show an uninterrupted gradient from one phenotype to the next (e.g. height).
- **Categorical traits** have their phenotype determined by counting (e.g. number of offspring).
- **Treshold traits** cause only two or a few phenotypic classes (e.g. diabetes).

7.2 Genetic basis of quantitative traits

In **Mendelian traits** described by the dominance model one allele A is contributing the entire phenotypic difference (dominance effect). The additive model says that each allele is contributing a part of the phenotypic variance (additive effect). This is also called the **multiple factor hypothesis**.
If n is the number of genes involved in a quantitative trait, then $(2n+1)$ will determine the total number of phenotype classes.

7.3 Basic statistics for quantitative genetics

The **central limit theorem** states that if the sum of the variables has a finite variance, then it will be approximately normally distributed. Almost any set of measurements will follow a normal distribution if enough measurements are taken.
Variance is calculated as $\sigma^2 = \frac{\sum(x-\overline{x})^2}{n}$. **Standard deviation** is defined as $\sigma = \sqrt{\frac{\sum(x-\overline{x})^2}{n}}$.
68% of the distribution is within 1 standard deviation of the mean, 95% are within 2 and 99.7% are within 3 standard deviations. Mean and standard deviation are enough to describe a normal distribution. The lower the variance, the narrower the bell-shaped normal curve.

8 Phenotypic variation

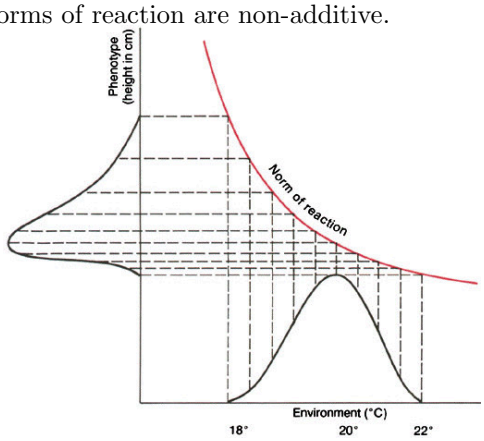
The phenotype itself is the sum of a genetic and an environmental component. The environment contributes to phenotypic variation in quantitative traits.

$$P = G + E \quad \Longrightarrow \quad V_P = V_G + V_E$$

The environment does not however have a constant contribution to the phenotype at all times.

$$V_P = V_G + V_{G\times E}$$

The norm of reaction is a pattern of phenotypes under a variety of environmental conditions (environmental distribution is transformed into the phenotypic distribution). Many actual norms of reaction are non-additive.



$$V_P = V_G + V_E + V_{G\times E}$$

8.1 Genotypic variation

Genotypic variation can be divided into components

$$V_G = V_A + V_D + V_I$$
$$\Rightarrow\ V_P = V_A + V_D + V_I + V_E + V_{G\times E}$$

8.1.1 Additive genetic variance

V_A is the porportion of the total genotypic variance V_G caused by the **sum of phenotypic effects of alleles** when they are assembled into genotypes. When gene action is additive, V_A of a population depends on allele frequencies. It's higher when alleles are at intermediate allele frequencies than when they are near fixation of loss.

8.1.2 Dominance genetic variance

V_D is the proportion of V_G caused by the **deviation of genotypic values** from their values under additive gene action caused by the combination of alleles assembled into a single-locus genotype.
Consider a single locus with two alleles, A_1 and A_2. Call genotypic values as follows: $A_1A_1 = -a$, $A_2A_2 = +a$ and $A_1A_2 = d$. The midpoint between $+a$ and $-a$ is 0.

- No dominance ($d = 0$): A_1A_2 is at the midpoint.
- A_1 is dominant to A_2: $d > 0$
- A_2 is dominant to A_1: $d < 0$
- Dominance is complete: $A_1A_2 = A_1A_1$ or A_2A_2
- Over-dominance: $A_1A_2 < A_1A_1$ or $> A_2A_2$

8.1.3 Population mean

The sum M is both the mean genotypic and phenotypic value for the population.

Genotype	Value	Freq. × Val.
A_1A_1	$+a$	p^2a
A_1A_2	d	$2pqd$
A_2A_2	$-a$	$-q^2a$
	$M =$	$a(p-q) + 2dpq$

8.1.4 Epistatis or interaction genetic variance

V_I is the proportion of V_G due to the deviation of genotypic values from their values under additive gene action caused by interactions between and among loci.

9 Heritability

Heritability is the proportion of phenotypic variance in a population that is due to genetic differences. It is not the same for a given trait in different environments. Heritability does not say anything about what genes influence a phenotype. It can only explain the amount of genetic variation that causes phenotypic variation.

9.1 Broad Sense Heritability

Values for **BSH** range from 0 to 1:

$$H^2 = V_G/V_P$$

To calculate BSH, one can take two approaches:

1) Fix the genotype to estimate V_E (selfing organisms, clones, monozygotic twins) $\Rightarrow V_P = V_E$.
Measure V_P in many genetically different individuals in the same environment, then obtain V_G by subtraction: $V_G = V_P - V_E$

2) Inbreed parents to homozygosity and calculate V_P in P1, P2, F1 and F2 generations. As inbred lines are genetically uniform, $V_P = V_E$.

$$V_{P1} = V_{P2} = V_{F1} = V_E$$
$$V_{F2} = V_P = V_G + V_E$$
$$V_G = V_{F2} = V_E$$

9.2 Narrow Sense Heritability

The ratio V_A/V_P expresses the extent to which phenotypes are determined by the alleles transmitted from the parents and is called **the heritability in the narrow sense** (NSH), or simply the heritability.

$$h^2 = V_A/V_P \quad \Rightarrow \quad h^2 = \frac{V_A}{V_A + V_D + V_I + V_E}$$

Regression analysis is used to quantify the relationship between variables that are correlated (e.g. the relationship between height of fathers and sons). The regression line can be represented with the equation:

$$y = a + bx$$

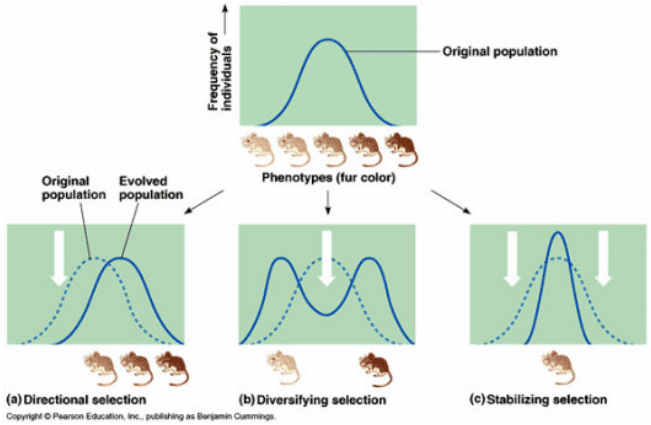
where the x and y values represent the two variables, b is the slope of the line, also called the **regression coefficient**, and a is the y-intercept. b can be calculated using the following equation:

$$b_{xy} = \frac{Cov(x, y)}{V(x)}$$

- If $b = 1$, V_A is the only component that influences variation
- If $b = 0$, V_A does not influence variation
- If b is between 0 and 1, V_A and other components influence variation

10 Response to selection

10.1 Phenotypic response to selection



10.1.1 Selection differential

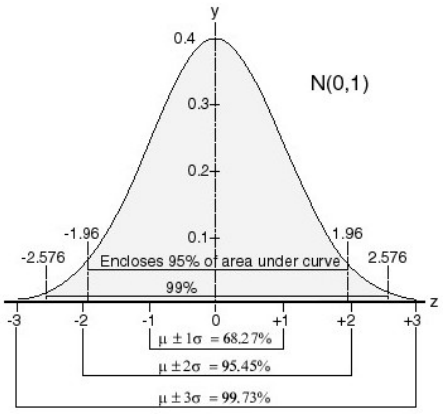
The selection differential S describes the strength of selection. It is the difference between the mean of the selected parents μ^* and the phenotypic mean of the initial population μ. The selection differential can be interpreted as the **within-generation change** in phenotypic mean due to selection.

$$S = \mu^* - \mu$$

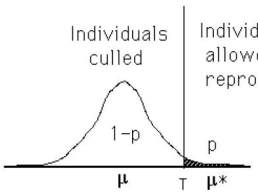
10.1.2 Selection intensity

The selection differential S is not particularly informative when trying to compare the strength of selection on different traits and/or in different populations. A much more useful measure is the selection intensity i, which is the selection differential expressed in fractions of phenotypic standard deviations.

$$i = S/\sigma_Z$$



10.1.3 Truncating selection



Under truncating selection, the upper- or lowermost fraction p of a population is selected to reproduce. Following from the properties of the normal distribution a good approximation of the intensity i is:

$$i \simeq 0.8 + 0.41 * ln\left(\frac{1}{p} - 1\right)$$

10.1.4 Response to selection

The response to selection R describes the difference between the mean phenotypic value of the original population μ and the mean of the next generation μ_0 that originated from the selected parents by random mating. The **between-generation change** in the mean due to the reproduction of the selected parents is:

$$\text{(observed) } R = \mu_0 - \mu$$

10.2 Genetic response to selection

Selecting a fraction of the phenotypes means selecting a fraction of the genotypes / alleles of the populations if the character is genetically determined. If the selected parents mate randomly and their offspring shows a change in allele frequencies, evolution took place.

10.2.1 Breeders' equation

The response to selection R depends on the strength of within-generation selection S and on the fraction of the offspring's phenotypic value that can be predicted from the parental value, i.e., the **heritability** of the character.

$$\text{(expected) } R = h^2 S \text{ or } R = b_{OP} S$$

This relationship is often called the **breeders' equation** and shows that the heritability of a character is the link between the within-generation change S and the between-generation response R.

10.2.2 Fisher's fundamental theorem of natural selection

"The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time"
— Ronald Fisher

Populations will respond to selection as long as there is additive genetic variance on which to act. $h^2 = V_A/V_P$ so if $V_A = 0$ then $h^2 = 0$ and therefore $R = 0$ because $R = h^2 S$.

10.3 Asymmetrical responses to selection

Drift can cause the cumulative response in one direction to be greater than the other. This often cannot explain a repeated bias in response in one direction across replicate lines and must then be rejected as a null hypothesis. If there is stronger natural selection for the trait in one direction than the other, then natural selection will aid artificial selection in one direction and hinder it in the other.

10.4 Long-term responses to selection

The outcome of selection over a long period is unpredictable for many reasons. First the outcome depends on the properties of the individual genes contributing to the response and this cannot be determined by observation at the outset. Second, mutation produces new variation whose nature cannot be predicted. Without the creation of new variation by mutation, the response to selection cannot continue indefinitely. Eventually all segregating genes in a population will come to fixation by the selection or accompanying inbreeding. The response is expected to slowly diminish and eventually cease. At this point, the population is at its **selection limit**.

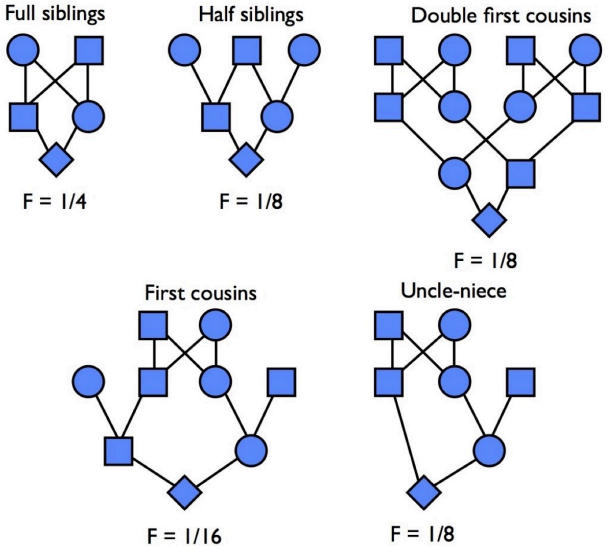
11 Inbreeding and heterosis

11.1 Inbreeding

Inbreeding is the mating of closely related individuals (e.g. sibling, cousins, self-fertilized organisms). It tends to increase the number of individuals in a population that are homozygous for a certain locus und therefore makes recessive traits appear more often. This overall decrease in fitness is called **inbreeding depression**. In other words, inbreeding is **non-random (assortative) mating** that results in deviations from HW-expectations. **Self-fertilization** is the most extreme form of inbreeding. Complete self-fertilization results in only three mating types (three genotypes on a diploid locus).
A sort of opposite of inbreeding is **outcrossing**, where two breeds are crossed. Neither inbreeding nor outcrossing result in allele frequency changes but both change genotype frequencies.
Even though the loss of heterozygosity leads to a loss of genetic diversity and the risk of heritable diseases increases, inbreeding is used to amplify desired traits in plants and animals.

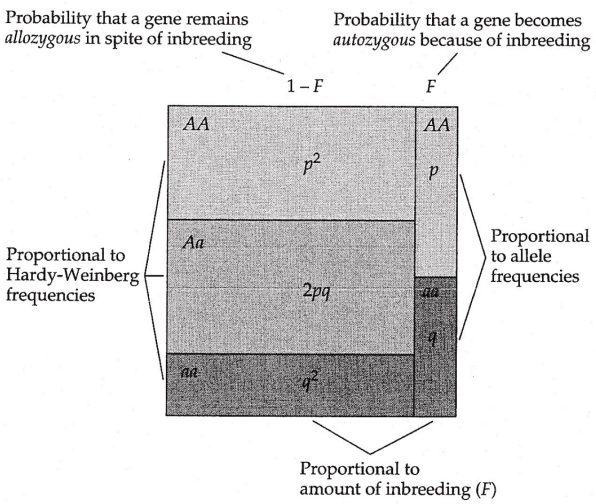
11.2 Inbreeding coefficient

The inbreeding coefficient F is the probability for two alleles of a homozygote to be **identical by descent (IBD)**, or in other words that two alleles originate from he same ancestral allele. In a random-mating population $F = 0$ while in a completely inbred population $F = 1$.



11.3 Inbreeding vs. random mating

Take a population with alleles A (frequency = 0.5) and a (frequency = 0.5). If random mating occurs and all other influences on allele frequencies are ignored, the population will stay at $AA = 0.25$, $Aa = 0.5$ and $aa = 0.25$. However after only one generation of inbreeding (only individuals with the same genotype mate) allele frequencies change to the following: $AA = 0.375$, $Aa = 0.25$ and $aa = 0.375$. With each further inbred generation the Aa genotype will become less and less common until it disappears from the population.



Genotype	With inbreeding coefficient F	With $F = 0$ (random mating)	With $F = 1$ (complete inbreeding)
AA	$p^2(1-F)+pF$	p^2	p
Aa	$2pq(1-F)$	$2pq$	0
aa	$q^2(1-F)+qF$	q^2	q

pF and qF are **autozygous** (from the same parent) alleles, while the rest are **allozygous** (from different parents).

Possible effects of inbreeding include

- Reduced fertility both in litter size and sperm viability
- Higher infant and child mortality
- Increased occurence of genetic disorders
- Smaller adult size
- Fluctuating facial asymmetry
- Loss of immune system function
- Increased cardiovascular risks

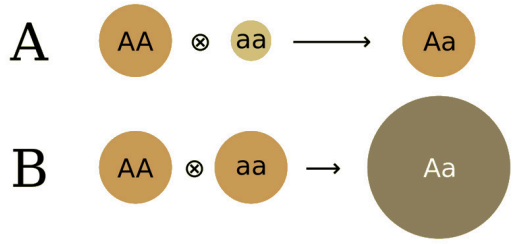
11.4 Changes of mean trait values

Inbreeding changes the mean trait value of quantitative traits.

$$\mu_F = \mu_0 - 2dpqF = a(p-q) + 2dpq - 2dpqF$$

The mean value will only change if $d \neq 0$, i.e. there is dominance. When looking at a single locus, the mean value will increase or decrease, depending on whether $d > 0$ or $d < 0$, to be closer the value of recessive alleles. The magnitude of the change depends on the allele frequency, being the greatest when $p = q = 0.5$.

11.5 Dominance vs. overdominance



The size of the circles depicts the expression levels.

Scenario **A** shows the **dominance hypothesis**. Allele A is dominant while a is both recessive as well as deleterious. The superiority of hybrids, also called **heterosis**, is attributed to the suppression of the undesirable a allele. If this hypothesis is the main cause for the fitness advantage, fewer genes should be under-expressed in the heterozygous offspring compared to their parents. Expression levels for any heterozygous gene should also be comparable to the ones from the dominant homozygous ancestral gene. Inbreeding reduces genetic variability and increases the chance of an individual being homozygous for allele a. The genetic variance for fitness is caused by rare deleterious alleles that are (partly) recessive. They persist in populations because of recurrent mutation. Most copies of these alleles in the base population are in heterozygotes. Inbreeding then increases the frequency of homozygotes causing inbreeding depression.

$$Aa = AA > aa$$

Scenario **B** shows the **overdominance hypothesis**. Here heterosis manifests in the heterozygote Aa having inreased fitness over both its homozygous parents. This can happen if two inbred strains are crossed. If this hypothesis is the main cause for the fitness advantage, the heterozygous offspring should show over-expression in certain genes over their homozygous parents. Since some inbred lines have means for fitness traits equal to the base population, the overdominance hypothesis cannot be generally true.

$$Aa > AA > aa$$

In both scenarion the descendants of the original parents tend to have higher heterozygosity due to selection. The main difference lies in the impossiblity of obtaining homozygotes as vigorous as heterozygotes if single-gene overdominance is important to inbreeding depression.

11.6 Heterosis

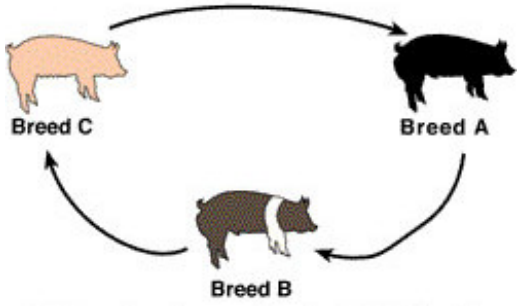
To recap, heterosis is the increase in fitness of hybrids between two inbred lineages. It is the reverse of inbreeding depression. Heterosis works against fixation and leads to more genetic variability. Populations that exhibit heterosis are better suited to adaptation. This is often exploited by breeders to enhance productivity of plants or animals. Some plants even have alleles that are lethal when homozygous so only heterozygous individuals survive.

$$H_{F_1} = \sum_{i=1}^{n} (\delta p_i)^2 d_i$$

Heterosis in F_1 depends on dominance. If $d = 0$ then no inbreeding depression and thus no heterosis is possible. Again, as with inbreeding depression, **directional dominance** is required for heterosis. If some loci are dominant in one direction and some in the opposite one, their effects will tend to cancel each other out and no heterosis may be observed. The absence of heterosis is not sufficient to conclude that no dominance exists. H is proportional to the square of the **difference in gene frequency** between populations. It is greatest when alleles are fixed in one population and absent in the other (so that $|\delta_i| = 1$) or in other words, heterosis is bigger the higher the genetic distance. $H = 0$ if $\delta = 0$. H is specific to each particular cross and must be determined empirically, since we do not know the relevant loci nor their allele frequencies.

11.6.1 Maximising and maintaining heterosis

To maximally exploit heterosis, $F1$ hybrids should be used as the heterotic advantage decreases in $F2$ hybrids. **Terminal crosses**, which do not reproduce further, are often used in plant breeding. As this method is not practical for animals, two other strategies can be used to balance the cost of breeding $F1$ hybrids and the decrease in performance of $F2$ hybrids: The first is the use of **synthetics**, where n parental lines with superior combining ability are chosen and a random-mating population is formed by making all $\frac{n(n-1)}{2}$ pairwise intercrosses between the lines. The second is the method of **rotational crossbreeding**. Here two, three or four (in theory there is no limit) different breeds can be used. Let's take a three-breed rotation as an example: A female with a father from breed A is bred with a male from breed B. Their female progeny is bread with males from breed C. The females of this generation are then bread with males from breed A. In other words, each generation of females is bread with another breed than their female parent generation was.



12 Formulas useful in quantitative genetics

$N = D^2/8VarG$

Standard deviation = $[Var(x)]^{1/2}$

$M = a(p - q) + 2dpq$

$h^2 = b_{AP}$

$S = \mu^* - \mu$

$R = h^2S$

$R = i\rho_{PA}\sigma_A$

$R = 2N_e i V_M/\sigma_P$

$i = S/\sigma_z$

$R = \mu_O - \mu$

$\mu_F = \mu_0 - 2Fpqd$

$M = \Sigma a(p - q) + 2\Sigma dpq$

$\alpha_1 = q[a + d(q - p)]$
$\alpha_2 = -p[a + d(q - p)]$

$V_P = V_A + V_D + V_{AA} + V_{AD} + V_{DD} + V_E + V_{IGE}$

$Cov = rV_A + uV_D$

$h^2 = V_A/V_P$
$H^2 = V_G/V_P$

$$r(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}}$$

$$P(x_1 \le x \le x_2) = \int_{x_1}^{x_2} p(x)\,dx$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$$a = \overline{y} - b\,\overline{x}$$

$$b = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right)$$

$$\mu_F = \mu_0 - 2F\sum_{i=1}^{k} p_i\,q_i\,d_i$$

$$R = h^2 S\frac{\sigma_z}{\sigma_z} = h^2\sigma_z i$$

$$\mu_F = \mu_0 - 2F\sum_{i=1}^{k} p_i\,q_i\,d_i = \mu_0 - B\,F$$

$$i \simeq 0.8 + 0.41\,\ln\left(\frac{1}{p} - 1\right)$$

$$\widehat{h}_r^2 = \frac{R}{S}$$

$$\mu_{F_1} = \mu_{P_1} + (\delta p)a$$

$$H_{F_1} = \mu_{F_1} - \frac{\mu_{P_1} + \mu_{P_2}}{2} = (\delta p)^2 d$$

$$H_{F_1} = \sum_{i=1}^{n}(\delta p_i)^2\,d_i$$

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}\right)$$

$$s \simeq \frac{a}{\sigma_z}\,i$$

$$t_{PHS} = \frac{\mathrm{Cov(PHS)}}{\mathrm{Var}(z)} = \frac{\mathrm{Var}(s)}{\mathrm{Var}(z)}$$

$$t_{FS} = \frac{\mathrm{Cov(FS)}}{\mathrm{Var}(z)} = \frac{\mathrm{Var}(s) + \mathrm{Var}(d)}{\mathrm{Var}(z)}$$

$$H_{F_2} = F_2 - \overline{P} = \left(F_1 - \frac{H}{n}\right) - \overline{P} = (F_1 - \overline{P}) - \frac{H}{n} = H\left(1 - \frac{1}{n}\right)$$

$$b_{O\overline{P}} = \frac{\frac{1}{2}V_A}{\frac{1}{2}V_P} = \frac{V_A}{V_P}$$

Table 10.2

Relatives	*Covariance**	*Regression (b) or correlation (t)*
Offspring and one parent	$\frac{1}{2}V_A$	$b = \frac{1}{2}h^2$
Offspring and mid-parent	$\frac{1}{2}V_A$	$b = h^2$
Half sibs	$\frac{1}{4}V_A$	$t = \frac{1}{4}h^2$
Full sibs	$\frac{1}{2}V_A + \frac{1}{4}V_D + V_{Ec}$	$t \ge \frac{1}{2}h^2$

*The contributions of epistatic interactions are ignored, and so are the possible environmental contributions to relatives other than full sibs.