N	μ	s	m	r
∞	-	-	-	-

Genetic variation - From genotype frequencies to allele frequencies

The last lecture focused on mutation as the ultimate process introducing genetic variation into populations. We have covered basic terminology to describe variation in a population context, and touched upon how **allele frequencies** can be calculated from **genotype frequencies**. *Remember:* The genetic composition of a population is fully characterized by its genotype frequencies. If we know the genotype frequencies, we can calculate the frequencies of the underlying alleles. E.g. for a bi-allelic locus with alleles *A* and *a* we may find the following numbers of genotypes in a population:

	Genotype			
	AA	**Aa**	**aa**	**Total**
N_{ind}	100	200	100	400

We assume that the order of alleles in Aa does not play a role. The genotypic frequencies follow as

D = f_{AA} = Pr(AA) = 100/400 = .25
H = f_{Aa} = Pr(Aa) = 200/400 = .50;
R = f_{aa} = Pr(aa) = 100/400 = .25,

and sum up to 1. Note that the frequency *f* of a genotype equals the probability of drawing this genotype at random from a population (see below). By convention we denote the frequencies / probabilities of genotypes *AA*, *Aa*, *aa* as *D, H, R*, allele frequencies f_A and f_a as *p* and *q*.

The frequency of the *A* allele is then derived as follows:

p = D + H/2 = (100 + 0.5*200)/400 = 0.5,

where the factor of 0.5 reflects that only half the alleles in an *Aa* heterozygote are *A*.

Obviously, the frequency of allele *a* is

q = R+ H/2 = 1 – p = 1 - 0.5 = 0.5.

But can we also turn it around and predict genotype frequencies from allele frequencies? Recall the example of the 'spirit bears' where a recessive Tyr-to-Cys replacement at codon 298 is perfectly associated with the white morph (Ritland *et al*. 2001). Knowing that the frequency of the G allele is .33 on a particular island how many white spirit bears with genotype GG would we expect to encounter, how many black bears (AA, AG)?

To address this question we first need to consider that there are a number of processes that could potentially change genotype and/or allele frequencies in a population. These are:

mutation, **selection**, **segregation**, **recombination**, **genetic drift**, and **migration** between subpopulations. These basic evolutionary forces may depend on ecological conditions, and basic genetic factors, such as the breeding system. We will consider these forces one by one during the course of this lecture and start out with the easiest situation imaginable.

The Hardy-Weinberg-Equilibrium (HWE) – a central null model for population genetics

The first thing we need to know is: What happens if nothing happens? That is, we assume that there is no mutation, no selection, no genetic drift (i.e., population is infinitely large), no recombination (we study only single locus) and no migration (our population is isolated). We thus study the effects of plain inheritance (assuming of course no segregation distortion). We also imagine the easiest possible breeding system of a sexual population:

- infinitely large population
- random mating (panmictic)
- non-overlapping generations
- hermaphroditic (i.e. no differences among males and females, each contributes gametes at equal proportion)

These assumptions might sound artificial (and are) but this is the art of mathematical modeling: distill a problem down to its essence. For a start, we don't care about what happens in a population where only say equally coloured individuals tend to mate with each other and segregation differs by sex (as e.g. in the case of sex chromosomes). We just want to know: how do the principles of Mendelian inheritance alone affect allele and genotypic frequencies? In order to answer this question, we need to imagine an ideal population where all the "distractions" are ignored. So we take a look at the easiest non-trivial case: A single locus with 2 alleles, A and a, and genotype frequencies f_{AA} (=D), f_{Aa} (=H), and f_{aa} (=R). The allele frequencies, p for A, and q for a, can be expressed in terms of the genotype frequencies as above. We ask: What are the expected genotype frequencies for a given combination of allele frequencies in the following generation? If genotype frequencies were unaltered in the following generation, allele frequencies would be too and we could derive not only allele frequencies from genotype frequencies, but also infer the expected frequencies of genotypes from allele frequencies.

The result is given by the **Hardy-Weinberg Law** (also called principle or equilibrium) which we will derive below. Godfrey Hardy was introduced to the problem by the geneticist Reginald Punnett with whom he played cricket. By solving it he became the unwitting founder of a branch of applied mathematics. Wilhelm Weinberg developed the principle independently of the British mathematician and delivered an exposition of his ideas in a lecture in 1908, before the *Verein für vaterländische Naturkunde in Württemberg* about six months before Hardy's paper was published in English.

First, it is important to recognize that due to the assumptions given above genotypes represent **random variables** (just as head or tail when tossing a coin) with associated probabilities given by the genotype frequencies (for a fair coin Pr(head)=0.5, Pr(tail)=0.5). If two variables are random, we can multiply their probabilities to get the **joint probability** of more than one outcome (e.g. the probability of tossing two heads in a row Pr(H,H)=Pr(H)Pr(H)=0.25). Analogous to the coin example, the probability of mating between individuals of a given

genotype equals the product of the genotype frequencies (=genotype probabilities). In diploid individuals, we further need to consider the **First Mendelian Law of Segregation** to assess which offspring genotypes are formed in which proportion from the parental combination (e.g. *Aa* x *Aa*= ¼ *AA*, ½ *Aa*, ¼ *aa*). The expected mating frequencies and resulting offspring genotypes are then given by:

Mother	**Father**	**Frequency**	**Offspring**		
			AA	**Aa**	**aa**
AA	AA	D^2	1	0	0
AA	Aa	DH	½	½	0
AA	aa	DR	0	1	0
Aa	AA	HD	½	½	0
Aa	Aa	H^2	¼	½	¼
Aa	aa	HR	0	½	½
aa	AA	RD	0	1	0
aa	Aa	RH	0	½	½
aa	aa	R^2	0	0	1

We obtain the genotype frequencies of the offspring as:

D'= Pr(AA)'= Pr(AAxAA)+ ½ Pr(AA xAa) + ½ Pr (Aa x AA) + ¼ Pr(AaxAa) =
D^2 + DH + ¼ H^2 =
$(D+H/2)^2$

Recalling that the allele frequency *p* (for allele *A*) could be derived as *D + H/2* (see above) we see that the probability of genotype *AA* in the offspring is directly related to the allele frequency in the parental population: **D'=Pr(AA)'=p^2**

Similarly we can obtain the probability for an *aa* homozygote as **R' = Pr(aa)'=q^2**. For the heterozygote we calculate

H' = ½ DH + DR + ½ HD + ½ H^2 + ½ HR + RD + ½ RH =
½ H^2 + HD + 2DR + HR = 2(D + H/2)(R+ H/2)

Given
p = D+H/2
q = R+H/2

this results in **H' =Pr(Aa)= 2pq.**

The same result can also be obtained simpler by recognizing that if genotypes mate at random, gametes therefore also unite at random to form zygotes. Under the simplifying assumptions given above, the probability that a gamete carries the *A* allele and is transmitted to the next generation is given by its frequency *p*, because according to Mendel's First Law all alleles have the same probability of transmission. Due to the assumption of random mating we can simply multiply the probabilities of paternal and maternal transmission to obtain the joint probability that an individual in the population is of type *AA*. This probability is accordingly $p \times p = p^2$. Following the same logic, we find the probability for a homozygous

individual of type *aa* as *q x q*=q^2. Heterozygote individuals can receive an *a* allele from the father and an *A* allele from the mother, each with probability *q* and *p*. A heterozygous offspring can likewise receive the *A* allele from the father and the *a* allele from the mother. Hence, the probability for a heterozygous offspring of type *Aa* or *aA* (the order does not matter) is given by *qp*+*pq*=*2pq*.

Summarizing,

Genotype	**AA**	**Aa**	**aa**
Frequency	p^2	2pq	q^2

These are the famous Hardy-Weinberg proportions relating allele frequencies to expected genotypes (visualized in **Figure 1**).

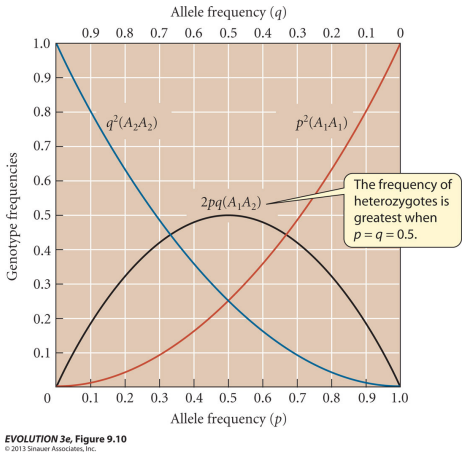

Figure 1: Hardy-Weinberg genotype frequencies as a function of allele frequencies at a locus with two alleles. Heterozygotes are the most common genotype in the population if the allele frequencies are between 1/3 and 2/3

There are several important points about this result.

1) Notice first that the genotype frequencies sum to one: $p^2 + 2pq + q^2 = 1$, as they should.

2) Note next that the allele frequencies do not change as a result of the genetic transmission mechanism. This can be see as follows:

$p' = D' + H'/2 = p^2 + 2pq/2 = p(p + q) = p$

Likewise, q' = q. This is a very important result, since it shows that allelic variation is not reduced by inheritance. This is independent of the phenotype. In particular, the frequency of an allele does not decrease just because it is recessive (which has been puzzling scientists at the time).

3) Even if genotype frequencies cannot be predicted, in general, from allele frequencies, the genotype frequencies after a single generation of random mating are simple functions of the allele frequencies. This has important practical consequences: Assuming **Hard-Weinberg**

equilibrium (HWE) allows us to calculate with allele frequencies instead of genotype frequencies – making things much easier and allowing direct assessment of micro-evolutionary changes in allele frequencies.

4) Once Hardy-Weinberg equilibrium is reached, also the genotype frequencies do not change any further.

5) For loci with more than two alleles, the same reasoning holds: Assume n alleles $a_i,\ldots, a_n$ with p_k the frequency of *k*th allele a_k. Then the frequency of the homozygote genotype a_ka_k is p_k^2 and the frequency of heterozygote genotype a_ka_i is $2p_kp_i$. We can then express the following quantities in terms of allele frequencies:

Total homozygosity (total frequency of homozygotes): $G = \sum_{k=1}^{n} p_k^2$

Total heterozygosity (total frequency of heterozygotes): $H = 1- \sum_{k=1}^{n} p_k^2$

Application: Assume a disease is caused by recessive autosomal allele *a*, and it is know that Pr(*aa*) = 0.0001 in the population (proportion of affected individuals). What is then the frequency of non-affected carriers of the disease gene (i.e. *Aa* heterozygotes)? The first step is to calculate the allele frequency *p* of the 'disease' allele *a*. Assuming Hardy- Weinberg proportions, $Pr(aa) = p^2 = 0.0001$, so $p = 0.01$. We then obtain the frequency of heterozygotes as $Pr(Aa) = 2pq = 2(0.01)(0.99) = 0.02$. So although the disease looks very rare (only 1 in 10,000 have it), in fact 2% of the population are carriers. Note that we have obtained this result under the assumptions of no mutation and no selection.

Having established a **null model** of which genotype frequencies to expect under random mating, large population size (no genetic drift) and no selection, we now can test for deviations in natural populations. These deviations are what is really of interest to us, as they may inform us about the underlying microevolutionary processes. Let's go back to the 'Spirit bears'. On the island Princess Royal Ritland *et al.* (2001) counted a total of 52 bears with 26, 17 and 9 bears having genotype AA, AG (both black) and GG (white), respectively. Is that what we would expect under Hardy-Weinberg equilibrium?

Let us first calculate the allele frequency *p* for allele *A*: $p = (26 + 17/2)/52 = 0.663$. The frequency for the allele *a* is then $q = 1 - p = 0.337$. We next find the expected genotype counts under the HWE. $E(AA)=52 \times (.663^2) = 22.86$; $E(AG) = 52 \times 2\ (0.663)(0.337) = 23.34$; $E(GG) = 52 \times (0.337^2) = 5.91$. If we compare the observed to the expected frequencies, we see that the observed number of heterozygotes is lower than expected in favour of an access of homozygotes both for black and white forms.

	AA	**AG**	**GG**
observed	26	17	9
expected HWE	22.86	23.34	5.91

To test whether this deviation could have just arisen by chance (sampling error) we can calculate the Chi-Square(X^{2-}) test statistic which provides a goodness-of-fit test for categorical data (count data).

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

$$\chi^2 = \frac{(22.86-26)^2}{22.86} + \frac{(23.34-17)^2}{23.34} + \frac{(5.91-9)^2}{5.91} = 3.77$$

The degrees of freedom (df) are difficult to calculate for this particular situation. For k alleles there are k(k+1)/2 possible genotypes, i.e. categories in a Chi-Square test. But there are k constraints, because the allele frequencies in the expected categories have to match the observed allele frequencies. The degrees of freedom are then calculated as df= k(k+1)/2 – k=k(k-1)/2. For two alleles, as in this case, we thus have one degree of freedom. A X^2-value of 3.76 for 1 df corresponds to a type I error probability of 0.052. While not significant at the traditional significance threshold of p <0.05, the probability that the observed genotypes are consistent with the assumption of HWE is rather low. When combining several populations, Ritland et al. (2001) found that the heterozygote deficiency is overall statistically significant.

Deviations from Hardy-Weinberg-Equilibrium

This is where biology starts. Which factors may be responsible for heterozygote deficiency in these black bear populations? There are many factors that can cause deviations from HWE.

1. Assortative Mating

Individuals may be more likely to mate with individuals from the same, or similar, genotype. This is called **assortative mating**. The opposite situation is called **negative assortative mating** or **dis-assorative mating**.

Consider the following scenario of full assortative mating: If *AA* individuals prefer to mate with other *AA* individuals, and *aa* individuals prefer to mate with other *aa* individuals, *AA* and *aa* individuals will rarely mate. As a consequence there will be fewer heterozygous individuals in the next generation than predicted by HWE. Assume a population initial in HWE with alleles *A* and *a* at frequency p=0.5 and q=0.5 and corresponding genotype frequencies D=0.25, H=0.5 and R=0.25. Assuming full assortative mating the genotype frequency of *AA* individuals in the next generation will be: 0.25 (*AA* will only mate with *AA* with probability 1) x 0.25(0.5) (one quarter of the heterozygous choosing randomly will result in AA genotype) = 0.375 = *D'*. Using similar arguments we can see that the frequency *R'* of *aa* offspring is 0.375. The frequency of heterozygotes is reduced by half to 1-*D'*-*R'*=0.25. The allele frequencies, however, are still p=0.5 and q=0.5! After many rounds of assortative mating the population will eventually be fully depleted of heterozygotes.

2. Inbreeding

Inbreeding occurs as a result of mating between relatives having one or more ancestors in common. The effect of such matings is very much the same as for assortative mating. If such matings are more common then expected under random mating, homozygote genotypes will rise in frequency and heterozygotes will become depleted. The most extreme form of inbreeding is **selfing** as is common in many plant species.

An important difference is that inbreeding affects the whole genome, while assortative mating only affects those loci that determine the trait relevant for mating preference. Given enough recombination, assortative mating does not affect loci elsewhere in the genome. Historically, deviations from HWE have been primarily attributed to inbreeding. This is why we still measure deviations from HWE as the **inbreeding coefficient** F. If consanguineous mating is a consistent feature of a population, F will increase over generations at a rate that depends on how closely related individuals are, on average.

3. Population structure

When deriving the HWE we assumed one large panmictic population. Inadvertently pooling individuals from two populations *pop1* and *pop2* can lead to strong deviations from HWE. Assume that *pop1* only has individuals with genotype *AA*, *pop2* only individuals of genotype *aa*. If we pooled them at equal proportions allele frequencies would appear to be p=0.5 and q=0.5. However, we would not observe any heterozygous individual, leave alone 50% as predicted by the HWE. **Population structure** is thus a very important parameter to take into account before considering other potential forces distorting Hardy-Weinberg proportions.

4. Selection

Selection refers to differential fitness (survival, mating success, etc.) among individuals due to their phenotypes. Considering that phenotypes are eventually determined by the underlying genotype, it is not difficult to see that selection can cause deviations from HWE. Note, however, that deviations from HWE can only be detected after selection has been acting

5. Genetic drift and mutation

Small population sizes (genetic drift) and mutation can also influence genotype frequencies and cause slight deviations from HWE. However, the effect of these factors is generally small and only causes random deviations from HWE that do not accumulate over time.

Literature: (Futuyma 2005; Barton *et al.* 2007; Nielsen & Slatkin 2013)

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH (2007) *Evolution*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Futuyma DJ (2005) *Evolution*. Sinauer Associates. Chapter 9.

Nielsen R, Slatkin M (2013) *An Introduction to Population Genetics: Theory and Applications*. Macmillan Education, Sunderland, Mass. Chapter 1.

Ritland K, Newton C, Marshall HD (2001) Inheritance and population structure of the white-phased "Kermode" black bear. *Current Biology*, **11**, 1468–1472.