Population Genetics Basis of Quantitative Genetics

2010-10-24

Hossein Jorjani
Hossein.Jorjani@hgen.slu.se

Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences
Uppsala, Sweden

Population Genetics Basis of Quantitative Genetics
Compendium, Third version 2010-10-24

Instructions for reading this compendium

This compendium has been written in the same way as
Count Lazlo de Almásy wants Rudyard Kipling to be read to him.

You’d better read it in the same way as Kip reads it.

WARNING!

This compendium is under construction
and is not, by any means, complete.
However, in combination with lectures handouts,
it gives a fair coverage of the subject.

POPULATION GENETICS

Population genetics is a branch of genetics studying changes of allele and genotype frequencies in populations. Like any other branch of genetics, population genetics rests firmly on the foundations laid down by the Mendelian genetics.

In Mendelian genetics the inheritance of specific alleles in simple matings (crosses) or limited pedigrees is studied. The focus of interest in Mendelian genetics is the genotype of an individual. Therefore, we either follow the process of inheritance of certain specific alleles from parents with known genotypes to their offspring, or alternatively we trace back the origin of certain specific alleles from offspring with known genotypes to their parents.

In population genetics the frequency of specific alleles or genotypes in (usually) large groups of individuals is studied. The focus of interest in population genetics is the frequency of alleles or genotypes in populations. Therefore, we either follow the process of inheritance of certain specific alleles from one generation to the next, or alternatively we trace back the patterns of allele and genotype frequencies in one generation to the processes that have been at work in the previous generations. Even though individuals are building blocks of a group, the genotype of any specific individual is of less interest in comparison to the dynamics of change in allele and genotype frequency in the population.

What is important in population genetics is not what genotype any individual has, but how and why the frequency of alleles and genotypes in one generation or population differs from the frequency of alleles and genotypes in another generation or population. Population genetics is all about processes that are the causes of changes and the patterns that they create. It’s all about processes and patterns, patterns and processes.

DEFINITION

Population genetics is the study of allele and genotype frequencies across space (populations) and time (generations).

Population genetics is the science of patterns of allele and genotype frequencies and the processes that change these patterns.

LET'S STAND STILL AND TAKE A PICTURE

A very good start point for the study of population genetics would be to find a population in which frequency of alleles and genotypes are constant from generation to generation. If such a population existed, then we could examine it thoroughly.

Unfortunately, no real population is in steady-state equilibrium. There is always the possibility of something changing from generation to generation. What is even worse is that there might be a change of allele frequencies in two opposite directions, because two evolutionary processes with opposite effects may be at work. Then, what we may observe is the constancy of allele frequencies, but it doesn't mean that there hasn't been any change. We just cannot observe it. The effects of the two evolutionary processes have cancelled each other out. We should be careful and not fall into a false sense of "security".

Therefore, the start point for the study of population genetics has to be an imaginary population. In other words, we need a "model".

In population genetics two different, albeit related, concepts are used as the starting point for the study of evolutionary processes and patterns. The first one is the Hardy-Weinberg equilibrium (HWE) also called Hardy-Weinberg principle or Hardy-Weinberg law. The second one is the concept of idealized population. Here, I combine these two concepts with each other and, in anticipation of a better name, call it Hardy-Weinberg model HWM).

HARDY-WEINBERG MODEL: PART 1

Imagine a population with the following properties:

Population is large;
Organism is diploid;
Each locus has two alleles;
The locus is not sex-linked;
Allele frequencies in males and females are equal;
Genotypes can be distinguished clearly;
Reciprocal mating of gametes are equal (i.e. $A_1A_2 = A_2A_1$)
Segregation is normal;
Reproduction is sexual;
There are equal number of females and males;
Matings are at random;
There is no mutation;
There is no migration;
There is no selection;
Generations are non-overlapping.

DEFINITION

According to the Hardy-Weinberg model (HWM) if a population fulfills the above mentioned properties (assumptions), then allele frequencies and genotype frequencies are constant from generation to generation and given the information on any of them, the other one can be calculated. Such a population is usually called a population in Hardy-Weinberg equilibrium (HWE) or a "base population". Occasionally, a population in HWE may also be referred to as a population in "linkage equilibrium".

Allele and genotype frequencies in the parental generation

Let's put these words into symbols. Let the locus under consideration be called the A locus, with two alleles A_1 and A_2. Three distinct genotypes exist in this population A_1A_1, A_1A_2 and A_2A_2. There are a large number of females (N_f) and a large number of males (N_m) in the population. The population size can also be written as:

$$N = N_f + N_m \qquad [1]$$

Because the organism is diploid (i.e. each individual has two chromosomes), then the total number of alleles in the population is 2N. Now, let the frequency of the A_1 be equal to *p*. Or more formally:

$$f(A_1) = \frac{Number\ of\ A_1\ alleles}{2N} = p \qquad [2.1]$$

Equivalently, let the frequency of the A_2 be equal to *q*. Or more formally:

$$f(A_2) = \frac{Number\ of\ A_2\ alleles}{2N} = q \qquad [2.2]$$

Again because the organism is diploid and there are only two alleles in the population, the sum of frequencies of A_1 and A_2 alleles is equal to unity (1.0). In other words:

$$f(A_1) + f(A_2) = p + q = 1.0 \qquad [2]$$

From Equations 2.1 and 2.2 we can see that:

$$f(A_1A_1) = f(A_1) \times f(A_1) = p \times p = p^2 \qquad [3.1]$$

$$f(A_1A_2) = f(A_1) \times f(A_2) = p \times q = pq \qquad [3.2a]$$

$$f(A_2A_1) = f(A_2) \times f(A_1) = q \times p = pq \qquad [3.2b]$$

$$f(A_2A_2) = f(A_2) \times f(A_2) = q \times q = q^2 \qquad [3.3]$$

Because A_1A_2 and A_2A_1 are equivalent to each other we can write:

$$f(A_1A_2)+f(A_2A_1)=2[f(A_1)\times f(A_2)]=2[p\times q]=2pq \qquad [3.2]$$

Yet again, because the organism is diploid and there are only two alleles in the population, there are only three genotypes possible, and the sum of their frequencies adds up to unity (1.0). In other words:

$$f(A_1A_1)+f(A_1A_2)+f(A_2A_2)=p^2+2pq+q^2=1.0 \qquad [3]$$

Equations 2 and 3 are summary description of allele and genotype frequencies in a population. As such, they represent patterns and *per se* give us little indications about processes (it is in the comparison of observed patterns with the expected patterns, or observed patterns in different populations and/or generations, that we can infer processes).

Example 1: Cat coat color

SLU students participating in the "Animal Breeding and Genetics" course in the Academic year 2005-2006 registered cat coat color on three cats each. One of the cat coat colors is related to the so-called S locus which is responsible for the presence of white fur in cats. Animals of Genotype SS have no white fur. Animals of genotype Ss have less than 50% white fur. Finally, animals of genotype ss have more than 50% white fur. There were a total of 137 cats observed:

SS	27 cats	No white fur
Ss	72 cats	< 50% white fur
ss	38 cats	> 50% white fur

		Number of	
Genotype	Number of cats	S allele	s allele
SS	27	54	0
Ss	72	72	72
ss	38	0	76
Total	137	126	148

$$f(S)=\frac{54+72}{137\times 2}=0.460=46.0\%$$

$$f(s)=\frac{72+76}{137\times 2}=0.540=54.0\%$$

$$f(SS)=\frac{27}{137}=0.197=19.7\%$$

$$f(Ss)=\frac{72}{137}=0.526=52.6\%$$

$$f(Ss)=\frac{38}{137}=0.277=27.7\%$$

Comments on Example 1

You realize that the cat population under consideration in Example 1 is far from being large. We also don't know if the other assumptions of the HWM are fulfilled or not. We just used this example to show how allele and genotype frequencies are calculated.

At the time of reproduction, in the gonads and during meiosis, the diploid cells with two chromosome sets are divided into daughter cells with one chromosome set. The resulting daughter cells will further develop until they become gametes. Please notice that "reproduction" and "meiosis" are processes. The final result of these processes is the break-down of the genotype of the individual into distinct germinal cells that contain only one set of chromosomes, and therefore, only one allele for each locus. It is now easy to see that it is the alleles that are transferred to the offspring, and not the genotypes. When two gametes unite with each other to form a zygote, then a new genotype is formed in the offspring.

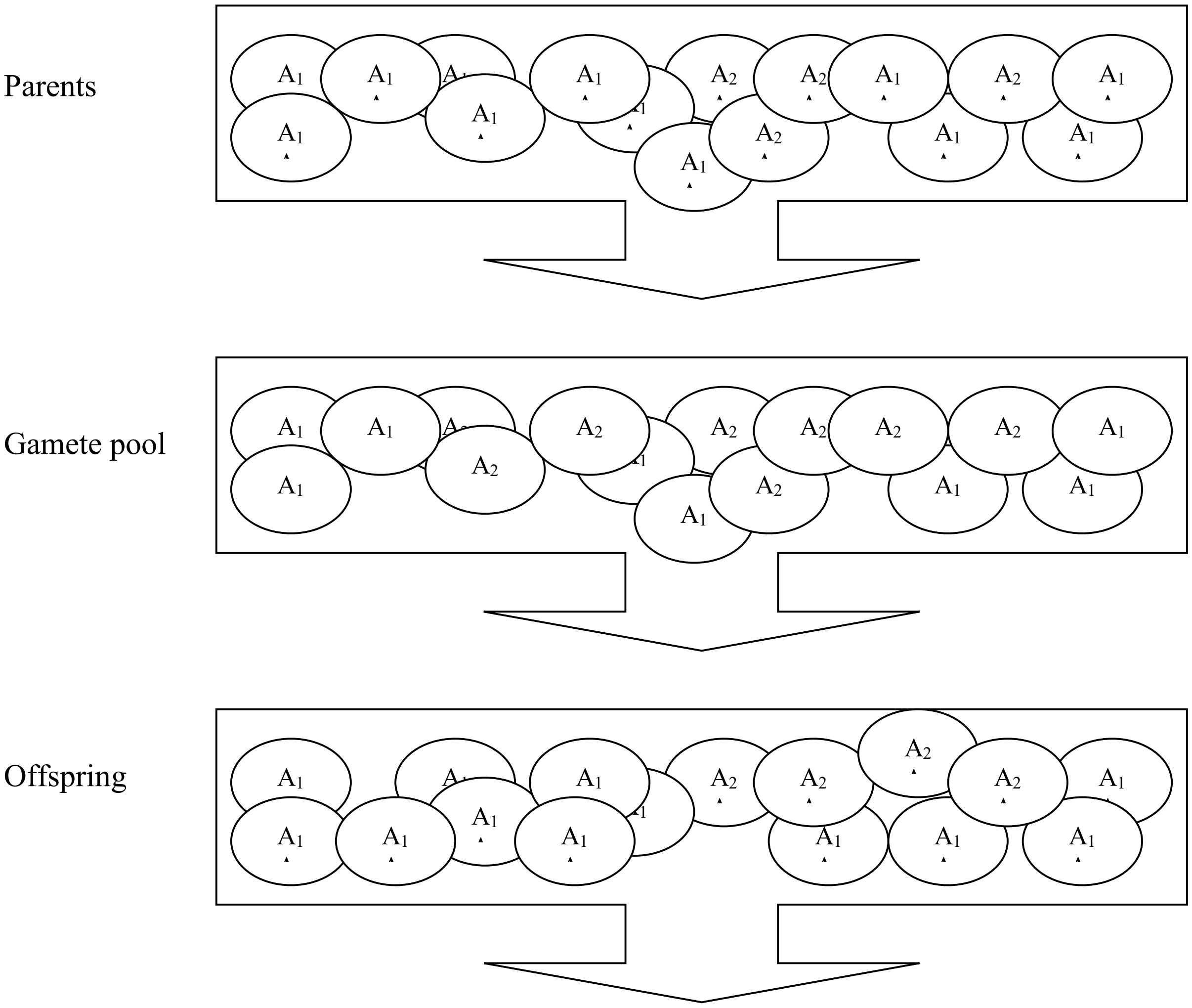

Figure 1 – Schematic representation of the reproductive process and break-down of genotypes in the parental generation and formation of new genotypes in the offspring generation.

Allele frequencies in the parental gametes

Given the fact that the organism under study is a diploid organism, it is obvious that the gametes carry only one set of chromosomes and therefore only one single allele for each locus. If a parent is of genotype A_1A_1 (A_2A_2), then all of its gametes will carry A_1 (A_2). On the other hand, if the parent is of genotype A_1A_2, then provided normal segregation and equal viability for the two sorts of gametes, half of the gametes will carry an A_1 allele and the other half an A_2 allele.

For the whole population, if all three genotypes (A_1A_1, A_1A_2 and A_2A_2) have equal fertility, then allele frequencies in the parental gametes should be equal to the allele frequencies in parental generation.

Allele frequencies in the uniting gametes

The next step in the process of reproduction is coming together of parents. If population is large and all male and female individuals of the parental generation have equal opportunity of mating with any member of the opposite sex, then the process of two gametes uniting with each other is like sampling one male and one female gamete from a vast gamete pool.

Allele frequencies in the zygotes

If the uniting male and female gametes sampled from the gamete pool are united with each other at random (random mating of parents) and they have equal fertilizing capacities, then allele frequencies in gametes are equal to the allele frequencies in the uniting parental gametes.

Genotype frequencies in the zygotes

In order to deduce the genotype frequencies in the zygotes, it is enough to multiply allele frequencies in the male and female individuals. This is shown in Table 1.

Table 1 – Allele frequencies in the uniting gametes and the resulting genotypes in the zygotes when the assumptions of the Hardy-Weinberg model are fulfilled.

		Female gametes and allele frequencies	
		A_1	A_2
		p	q
Male gametes and allele frequencies	A_1	A_1A_1	A_1A_2
	p	p^2	pq
	A_2	A_1A_2	A_2A_2
	q	pq	q^2

Genotype frequencies in the offspring generation

If all zygotes have equal viability, i.e. all genotypes have equal viability, then the genotype frequencies in the offspring generation will have the same proportions as in the parental generation.

Allele frequencies in the offspring generation

If all the newborn offspring have the same viability from birth to maturity, then allele frequencies in the offspring generation can be deduced by Equation 3. In other words allele and genotype frequencies will be the same in the parents and offspring.

Summary of the Hardy-Weinberg model

Starting from a parental population fulfilling assumptions of the HWM, and going through several steps, we could show that under certain circumstances an imaginary (model) population will have constant allele and genotype frequencies from generation to generation. A summary of these steps (adopted from Falconer & Mackay, 1996) is shown in Table 2.

Table 2 – Summary of the Hardy-Weinberg model

Step	Deduction from: to	Conditions
	Allele frequency in parents	
1a		Normal allele segregation Equal fertility of parents
	Allele frequency in all gametes	
1b		Equal fertilizing capacity of all gametes Large population
	Allele frequency in gametes forming zygotes	
2		Random mating Equal allele frequencies in ♂ and ♀ parents
	Genotype frequencies in zygotes	
3		Equal viability
4	Genotype frequencies in progeny	
	Allele frequency in progeny	

Consequences of the Hardy-Weinberg model

Summation of allele (Equation 2) and genotype (Equation 3) frequencies to unity has some interesting consequences. Some of these consequences are illustrated in Figure 2.

Allele and homozygote frequency of a rare allele: As can be seen in Figure 2, when the frequency of the less frequent allele is low, say 0.01 (or 1%) and lower, only a minority of the copies of the rare alleles can be seen in the homozygote individuals. For example, consider an allele with the frequency of 1%. According to Equation 3.3 frequency of the homozygotes carrying this allele would be:

$$f(A_2A_2) = f(A_2) \times f(A_2) = 0.01 \times 0.01 = 0.0001 = 10^{-4}$$

That is only 1 out of 10000 individual will be homozygote for this allele. Please notice that the numbers are not playing a trick on you. This is just a consequences of a fraction (say, 0.01) being raised to the power of 2.

Frequency of carriers of the rare allele: In the same way, Figure 2 shows that when an allele is rare, the majority of the copies of the allele can be found as heterozygotes. For a rare allele with the frequency of 0.01 the proportion of heterozygotes can be found from Equation 3.2; i.e.

$$f(A_1A_2)+f(A_2A_1)=2[f(A_2)\times f(A_1)]=2[q\times p]=2pq=2\times 0.99\times 0.01=0.0198$$

If we take the ratio of heterozygotes to homozygotes for this rare allele, we find that the heterozygotes are about 200 times more frequent that the homozygotes ($2pq/q^2=0.0198/0.0001=198\approx 200$).

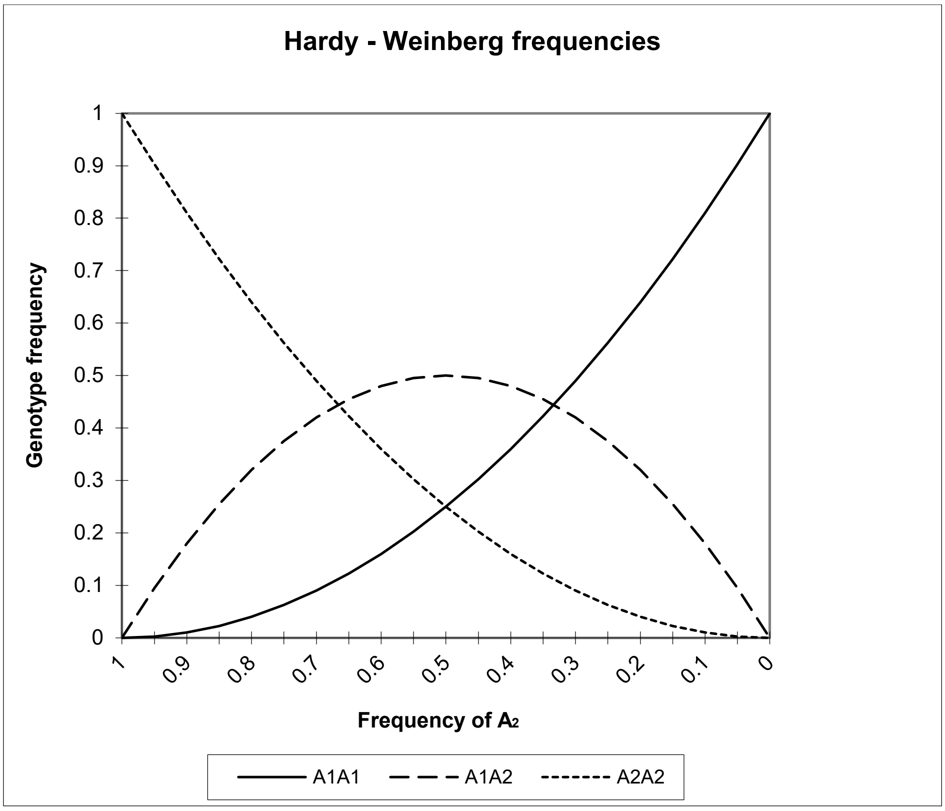

Figure 2 – Plot of genotype frequencies for a locus with two alleles in a population in the Hardy-Weinberg equilibrium.

Test for conformity to the Hardy-Weinberg equilibrium

In genetics (as in statistics) the visual inspection of the numerical results of a parameter/variable is not enough for making an inference. Numerical results must always be checked against theoretical expectations for that parameter/variable. For example, the numerical results obtained in the cat coat color example (with allele frequencies 46% and 54% and genotype frequencies 19.7%, 52.6% and 27.7%, see Example 1) cannot *per se* tell us if the observations are coming from a population in the HWE. To make such an inference we need a test, a proper statistical test.

Chi-Square test, χ^2

An appropriate statistical test to infer conformity to the HWE is the χ^2 test (the Greek letter chi). The equation to calculate the χ^2 test-statistics is as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad [4]$$

Where χ^2 is the test-statistics of interest, O is the observed number of individuals of certain genotype, E is the expected number of individuals of that genotype and Σ (the Greek letter sigma) stands for summation over all classes of observations. Calculated value of the χ^2 test-statistics should be compared with the critical values obtained from a true χ^2 distribution (for some values see Table 3). Degree of freedom (df) for the χ^2 value is the number of groups (n) minus 2 (df = n – 2).

Table 3 – Some critical values of χ^2 distribution

Degree of freedom (df)	χ^2-value	
	Probability (*P*) = 0.05	Probability (*P*) = 0.01
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345

If the calculated value of the χ^2 test-statistics is larger than the critical value, then it may be concluded that it is unlikely that the sample under consideration is coming from a population fulfilling HWM assumptions (unfortunately, this simple test cannot determine which of the assumptions have been violated).

On the other hand, if the calculated value of the χ^2 test-statistics is smaller than the critical value, then it may be concluded that we do not have enough reason to reject the likelihood of the sample under consideration coming from a population fulfilling HWM assumptions. (Please notice that we cannot prove that the sample is from a population fulfilling HWM assumptions.)

Example 2: χ^2 test for the cat coat color example

Based on the number of observations presented in Example 1, the allele frequencies of the two alleles at the **S** locus are 46% and 54% for the **S** and the **s** alleles, respectively.

Expected numbers of genotypes from these two allele frequencies can be calculated by multiplying the expected Hardy-Weinberg ratios by the total number of observations.

E(SS) = $(0.46)^2$ x 137 = 28.99
E(Ss) = 2 x 0.46 x 0.54 x 137 = 68.06
E(ss) = $(0.54)^2$ x 137 = 39.95

Please notice that the sum of the three expected values (28.99 + 68.06 + 39.95) is equal to 137, which is the total number of observed cats. The χ^2 value is:

$$\chi^2 = \frac{(27.00-28.99)^2}{28.99} + \frac{(72.00-68.06)^2}{68.06} + \frac{(38.00-39.95)^2}{39.95} = 0.46$$

Because the calculated value of the χ^2 with a df = 1 is smaller than the critical values presented in Table 3, we must conclude that we do not have enough reason to reject the likelihood of this cat sample coming from a population fulfilling HWM assumptions for the **S** locus.

Comment on Example 2
We really don't know if this sample is a random sample of the Swedish cat population. Therefore, we must be cautious about making generalized conclusions from this small sample. As a matter of fact, SLU students participating in the "Animal Breeding and Genetics" course in the Academic year 2004-2005 had the following observations on a total of 154 cats:

SS	16 cats	No white fur
Ss	100 cats	< 50% white fur
ss	38 cats	> 50% white fur

Exercise: Calculate the χ^2 value for the observations made in 2004-2005 and interpret the results. Also, discuss the discrepancy between the results of the two academic years.

Minor violations of the Hardy-Weinberg assumptions
Hardy-Weinberg model is a very robust model and resilient to violation of some of its assumptions.

Extension to more than two alleles per locus: Existence of more than two alleles per locus does not make any fundamental change. Hardy-Weinberg ratios are actually expansion of the following simple algebraic equation:

$$(p+q)^2 = p^2 + 2pq + q^2 \qquad [5]$$

For extension to more than two alleles two options are available. Imagine that in a locus there are three alleles with frequencies p, q and r. An easy option to handle this situation is to consider the two less frequent alleles as one allele and work out the ratios according to Equation 5. However, the more appropriate option is to extend Equation 5 to three alleles:

$$(p+q+r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr \qquad [6]$$

You can see that extension to four, five or more alleles per locus is as easy as extension to three alleles (in equations similar to Equation 5). One important point to remember is that when more alleles are involved, especially if some of them are very rare, randomness of sampling and size of the sample become more crucial.

Different allele frequencies in males and females: Imagine an experiment in which males from one population and females from another population are put together for mating. Then, the assumption of equal allele frequencies in the two sexes may not be fulfilled.

Table 4 – Allele frequencies in the uniting gametes and the resulting genotypes in the zygotes when the male and female allele frequencies are different (i.e. assumptions of the hardy-Weinberg model are not fulfilled).

		Female gametes and allele frequencies	
		A_1 p_f=0.60	A_2 q_f=0.40
Male gametes and allele frequencies	A_1 p_m=0.30	A_1A_1 $p_mp_f = 0.18$	A_1A_2 $p_mq_f = 0.12$
	A_2 q_m=0.70	A_1A_2 $q_mp_f = 0.42$	A_2A_2 $q_mq_f = 0.28$

From the two allele frequencies for the A_1 (A_2) allele we can conclude that the average A_1 (A_2) allele in the mixed population is:

$$\bar{p} = \tfrac{1}{2} p_m + \tfrac{1}{2} p_f \qquad [7.1]$$

$$\bar{q} = \tfrac{1}{2} q_m + \tfrac{1}{2} q_f \qquad [7.2]$$

In the example given in Table 4, the average allele frequencies are:

$\bar{p} = \tfrac{1}{2} p_m + \tfrac{1}{2} p_f = \tfrac{1}{2}\, 0.30 + \tfrac{1}{2}\, 0.60 = 0.45$, and
$\bar{q} = \tfrac{1}{2} q_m + \tfrac{1}{2} q_f = \tfrac{1}{2}\, 0.70 + \tfrac{1}{2}\, 0.40 = 0.55$

If the mixed population was fulfilling the HWM requirements, then from the average allele frequency for allele A_1 we may expect the genotype frequency for the A_1A_1 genotype in the zygotes (and consequently in the offspring) to be 0.2025, while the actual frequency of the A_1A_1 genotype is 0.18.

We can calculate the actual allele frequencies from the genotype frequencies (and confirm that the actual allele frequencies are equal to the average of allele frequencies among the two parental sexes) as follows:

$p = 0.18 + \tfrac{1}{2}\, 0.12 + \tfrac{1}{2}\, 0.42 = 0.45$, and
$q = 0.28 + \tfrac{1}{2}\, 0.12 + \tfrac{1}{2}\, 0.42 = 0.55$

A close look at the genotype frequencies shows that if the assumption of random mating is fulfilled, there must be equal number of male and female gametes (offspring) among all individuals carrying the A_1A_1 genotype. The same is true for the other genotypes. Therefore, even though the two parental sexes had different allele frequencies, the males and females in the offspring generation have the same allele frequencies. The conclusion is that a (mixed) population violating the assumption of equal allele frequencies in males and females will achieve equal allele frequencies after one generation of random mating and will remain in the HWE thereafter.

Extension to sex-linked loci: One of the assumptions of the HWM was that the locus under consideration is not sex-linked. The reason is that the homogametic sex (e.g. the female in most mammalian species) carries two alleles and the heterogametic sex only one. Therefore, counting of alleles in the population will show that two-third of all alleles are found in the homogametic sex and one-third in the heterogametic sex, or more formally (assuming the female to be the heterogametic sex):

$$\bar{p} = \tfrac{1}{3} p_m + \tfrac{2}{3} p_f \qquad [8.1]$$

$$\bar{q} = \tfrac{1}{3} q_m + \tfrac{2}{3} q_f \qquad [8.2]$$

If the allele frequencies are the same in males and females the population can already be in the HWE or reaches the HWE already after one round of random mating. However, if the allele frequencies are different in the two sexes, then a very interesting pattern emerges. The reason is that in the offspring generation all males receive their only allele from their dam, and females receive their two alleles from both their sires and dams. Therefore, allele frequency among the male and female progeny (p'_m and p'_f, respectively) is

$$p'_m = p_f \qquad [9.1]$$

$$p'_f = ½(p_m + p_f) \qquad [9.2]$$

To see the interesting pattern we should look at the difference between allele frequencies in females and males, i.e.

$$p'_f - p'_m = ½(p_m + p_f) - p_f = -½(p_f - p_m) \qquad [9.3]$$

In other words, the difference in allele frequencies in the offspring generation is half of the difference in the parental generation, but in the opposite direction! Let's illustrate Equation 9.3 in a graph. Figure 3 illustrates an extreme case in which allele frequency for an allele (say, A_1) in the females is $p_f = 1.0$ and in the males is $p_m = 0.0$. In each generation the allele frequency in the males is equal to that of the females in the previous generation (Equation 9.1), and the allele frequency in the females is the average of allele frequencies in the males and females of the previous generation (Equation 9.3). Allele frequencies in the males and females fluctuate around the average of two sexes and become reduced by half in each generation, until the two sexes achieve equal allele frequency. When equal allele frequency in the two sexes is achieved, the population as a whole fulfills the HWM requirements and continues to remain in the HWE thereafter.

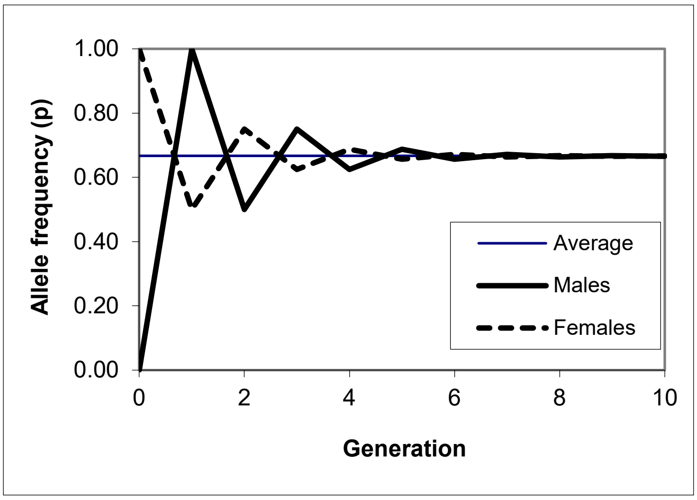

Figure 3 – Gradual achievement of equal allele frequency in a population with extreme allele frequency difference in males and females.

Extension to more than one locus: We have seen that the HWE is achieved after one generation of random mating for any autosomal locus. However, for two (or more) autosomal loci considered together the HWE cannot be achieved just after one generation. The reason is that there are, almost always, allele combinations, i.e. genotypes, that cannot be produced after just one generation of random mating. As an example think about two autosomal loci: locus A (with two alleles A_1 and A_2) and locus B (with two alleles B_1 and B_2). Further, assume that we have two populations, each one homozygous for one of the alleles in each locus. Therefore, one population is entirely made of individuals with $A_1A_1B_1B_1$ genotype and the other population is entirely made of individuals with $A_2A_2B_2B_2$ genotype. It can easily be seen that in the first round of mating between these two populations there are only two types of gametes available: A_1B_1 and A_2B_2. The other two possible types of gametes (A_1B_2 and A_2B_1) cannot be produced until the next generation. Let's illustrate this in a Table.

Table 5 – Joint consideration of two loci and possible gametes/genotypes

Parental generation genotypes	$A_1A_1B_1B_1$		$A_2A_2B_2B_2$
Parental generation gametes	A_1B_1		A_2B_2
Offspring generation genotypes		$A_1A_1B_1B_1$, $A_2A_2B_2B_2$ $A_1A_2B_1B_2$	
Offspring generation gametes		A_1B_1,A_1B_2, A_2B_1, A_2B_2	

The genotype produced from A_1B_1/A_2B_2 gametes is sometimes called coupling heterozygotes (or non-recombinant genotype). Conversely, the genotype produced from A_1B_2/A_2B_1 gametes is sometimes called repulsion heterozygotes (or recombinant genotype).

One consequence of the absence of certain genotypes in the parental gametes is that achieving equilibrium would depend on the time (in generations) needed for formation of all possible genotypes in gametes. This phenomenon, lack of certain gamete types and the delay in achieving the Hardy-Weinberg equilibrium is called gametic phase disequilibrium, or more commonly known as linkage disequilibrium. Please notice that the term "linkage disequilibrium" has very little to do with the actual physical linkage of adjacent loci. However, with the help of actual physical linkage it is easier to see how the linkage disequilibrium works. The following illustration depicts the same two populations discussed above.

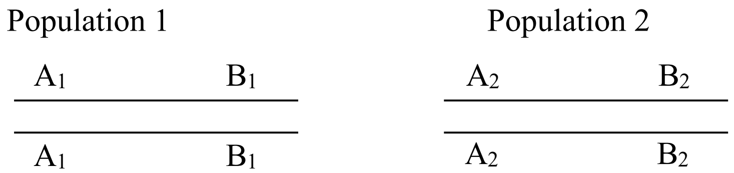

Figure 4 – Schematic representation of two bi-allelic loci on one chromosome in two populations

If the two loci A and B are residing on the same chromosome, but very far from each other, then the probability of a crossover, and consequently, recombination, between them can be very high. However, if these two loci are very close to each other, then it may take many generations for a crossover to occur between them. Measuring the amount of disequilibrium requires calculation of expected and observed genotype frequencies.

Table 6 – Gametic (phase) linkage disequilibrium

Alleles	A_1	A_2	B_1	B_2
Allele frequencies	p_A	q_A	p_B	q_B
Gametic types	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Expected frequencies	p_Ap_B	p_Aq_B	q_Ap_B	q_Aq_B
Observed frequencies	r	s	t	u
Disequilibrium	+D	-D	+D	-D

If the population is in equilibrium, then the difference between the expected and observed frequencies should be zero. For example, if $p_A = 0.3$, $q_A = 0.7$, $p_B = 0.4$ and $q_A = 0.6$, then the expected frequencies are as follows:

Alleles	A_1	A_2	B_1	B_2
Allele frequencies	0.3	0.7	0.4	0.6

Gametic types	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Expected frequencies	0.12	0.18	0.28	0.42
Observed frequencies	0.12	0.18	0.28	0.42
Disequilibrium	0.00	0.00	0.00	0.00

However, if the population is not in equilibrium, then the observed frequencies (r, s, t and u) would be different from the expected. For example:

Alleles	A_1	A_2	B_1	B_2
Allele frequencies	0.3	0.7	0.4	0.6
Gametic types	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Expected frequencies	0.12	0.18	0.28	0.42
Observed frequencies	0.18	0.12	0.22	0.48
Disequilibrium	+0.06	-0.06	-0.06	+0.06

The amount of disequilibrium is defined as the difference between the coupling heterozygotes produced by the original gametic types (A_1B_1 and A_2B_2) and the repulsion heterozygotes produced by the newly formed gametic types (A_1B_2 and A_2B_1). The frequencies of the coupling and repulsion heterozygotes are equal 2ru and 2st, respectively. The disequilibrium is defined as half the difference between these two, i.e.

$$D = ru - st \qquad [10]$$

The amount of disequilibrium measured by D is very much dependent on the allele frequencies in the population. Therefore, comparison of disequilibrium across populations with D is not possible. To make across population comparisons we can standardize the D and obtain a new measure of disequilibrium (r^2), as follows:

$$r^2 = \frac{D^2}{p_A \times q_A \times p_B \times q_B} \qquad [11]$$

For the example above, D = (0.18)*(0.48) – (0.12)*(0.22) = 0.06 and r^2 = (0.06)*(0.06) / ((0.30)*(0.70)*(0.40)*(0.60)) = 0.071.

There are evolutionary processes (e.g. selection, assortative mating, and sampling process in small populations) that lead to linkage disequilibrium and the loci involved are inherited in such a way “as if they were physically linked together”. As a matter of fact, two loci in linkage disequilibrium can be located on two different chromosomes and yet be consistently inherited together “as if they were physically linked together”, because some evolutionary process is holding them together!

Generally, random mating, i.e. the absence of evolutionary processes, reduces linkage disequilibrium gradually. The reduction in linkage disequilibrium depends on the initial disequilibrium (D_0) and recombination frequency (c), as

$$D_t = D_0\,(1 - c)^t \qquad [12]$$

where D_t is the amount of disequilibrium at generation t, and c is the recombination frequency. For unlinked loci c = ½ (because of independent assortment of loci). Therefore, in absence of physical linkage, the disequilibrium reduces by half after each generation of random mating until it disappears.

Example 3: Decline of linkage disequilibrium

In the following graph, the decline of linkage disequilibrium for five different values of recombination frequency has been shown. From left to right the value of c = 0.5 (no linkage), 0.4, 0.3, 0.2 and 0.1, respectively.

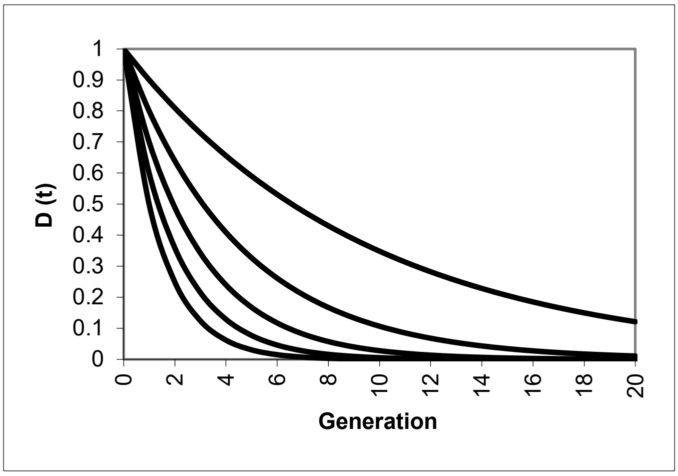

Figure 5 – Gradual decline of linkage disequilibrium.

Major violations of the Hardy-Weinberg assumptions

A close look at the assumptions of the Hardy-Weinberg model reveals that these assumptions are two sorts, processes and patterns.

Table 7 – Decomposing the Hardy-Weinberg assumptions into population genetics processes and patterns.

Processes	Patterns
Large population;	Large population;
Random mating;	Diploid organism;
No mutation;	Two alleles per locus;
No migration;	Non-sex-linked loci;
No selection.	Equal allele frequencies in ♂ and ♀;
	Unequivocal distinction of genotypes;
	Equality of reciprocal matings;
	Normal segregation;
	Sexual reproduction;
	Equal number of ♂ and ♀;
	Non-overlapping generations.

Let's avoid being carried away. The things listed here as patterns may act and be called as processes in other branches of genetics. However, for our purposes a pattern is defined as the allele and genotype frequency at a single stage of life of an organism in a single population, in a single generation, and a process is defined as anything causing a change in allele and genotype frequencies.

In any case, violation of certain assumptions of the HWM (that is the existence of evolutionary processes) leads to changes in allele and genotype frequencies from one stage of life to another. Therefore, populations subjected to these evolutionary processes will have different patterns of allele and genotype frequencies in space and time. And this is a major difference between patterns and processes: violation of pattern-assumptions cannot lead to changes in allele and genotype frequencies, while violation of process-assumptions does certainly lead to changes in allele and genotype frequencies.

From a population genetics point of view (according to most textbooks) there are four evolutionary processes, also called evolutionary forces: genetic drift, mutation, migration and selection. In my opinion putting sampling process (a consequence of small population size) and random mating in one category (genetic drift) does not give a fair picture of these two processes. Therefore, I choose to consider sampling process and random mating as two different processes (even though they have related and sometimes identical effects).

In the following sections major violations of the Hardy-Weinberg assumptions, one at the time, are described. In describing each of the major violations, we assume that all the other assumptions hold. Describing combinations of two or more violations is beyond the scope of this compendium and will not be covered.

MUTATION

Up to now we have assumed that DNA replication from parental germinal cells to the gametes proceeds without any “mistake”. However, we know that each individual’s DNA goes through a large number of replications and especially for production of gametes there are potentially a very large number of replications involved. Therefore, it seems inevitable that the DNA replication is accompanied with some mistakes.

DEFINITION

For our purposes, we can define mutation as any mistake in DNA replication.

Assume a population homozygous for an allele (say, A_1). Further, assume that a mutation can change the A_1 allele to A_2. It is obvious that if the mutation to A_2 is a unique event, then it would not have any measurable effect in the population. However, if mutation to A_2 is a recurrent event with a probability of μ (the Greek letter mu), then no matter how small the value of μ is, all individuals will eventually become homozygous for A_2. The change in allele frequency (Δ_p, that is the difference between A_1 allele frequency in the present, p_t, and the previous generation, p_{t-1}) is as follows:

$$\Delta p = p_t - p_{t-1} = (p_{t-1} - \mu p_{t-1}) - p_{t-1} = -\mu p_{t-1} \qquad [13]$$

The actual mutation rate per locus per generation is usually very small, about 10^{-5} to 10^{-6}, meaning that out of 100,000 to 1,000,000 individual, on average one individual carries a mutation for the locus under consideration. Please notice that *i*) each species (especially humans and farm species) has a large number of loci and *ii*) for quantitative traits the actual mutation rates per trait may be 10^3 to 10^4 times larger than the mutation rates per locus.

If the mutation event from A_1 to A_2 continues at the same rate during a long time, the accumulated effect of mutation on allele frequency can be calculated from the following approximate equation:

$$p_t = p_0 \, e^{-t\mu} \qquad [14]$$

In Figure 6, using Equation 14 the accumulated effect of a recurrent mutation has been shown in a population for different mutation rates (from 10^{-2} to 10^{-6}). With $\mu=10^{-2}$ it takes only 100 generation for the new allele (A_2) to become fixed in the population, while with $\mu=10^{-5}$ passing of 1,000,000 generations is necessary for A_2 to become fixed.

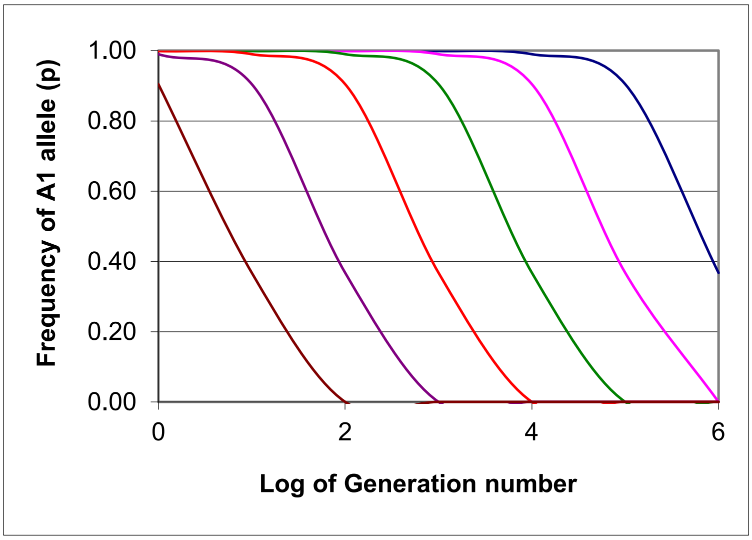

Figure 6 – Cumulative effect of mutation on the frequency of the A_1 allele. The X-axis shows the number of generations on a logarithmic scale (e.g. 2 stands for the allele frequency after 10^2=100 generations). Mutation rates are from 10^{-1} to 10^{-6}, for the six graphs from left to right, respectively.

A special and interesting case is when there is a mutation that changes A_2 back to A_1. Let the mutation rate from A_2 to A_1 be denoted by ν (the Greek letter nu). It can be seen that after some (or many many) generations an equilibrium will arise under which there is a seemingly constancy of allele frequencies (but don't be fooled by it!). At equilibrium change of allele frequency for A_1 (that is p) at the mutation rate μ will be equal to the change of allele frequency for A_2 (that is q) at the mutation rate ν. Or formally

$$p\mu = q\nu \qquad [15]$$

Equation 15 can be re-arranged to calculate the frequency of any of the alleles involved. For example, the frequency of A_2 at the equilibrium would be

$$q = \frac{\mu}{\mu + \nu} \qquad [16]$$

Example 4 – Bi-directional mutation balance

Consider a forward mutation rate (μ) of 10^{-6} and a backward mutation rate (ν) of 10^{-7}. At equilibrium the frequency of the A_2 allele (q) would be equal to $(10^{-6}) / (10^{-6} + 10^{-7}) \approx 0.91$.

Comment on Example 4
The initial frequency of allele A_1 is not important at all. Even if the population at the start of this process in homozygote for A_1, the end result still will be the same, i.e. q = 0.91 for the above mentioned mutation rates.

Exercise: Try some different values for μ and ν to see what would be the equilibrium value for the A_2 allele.

MIGRATION

Up to now we have assumed that the populations under consideration are closed populations in the sense that no individual ever leaves the population nor any individual from other populations enters to it. However, it is conceivable that many individual emigrate from a population or immigrate to it. Effects of emigration are very similar to the effects of selection and therefore, they will be not discussed here.

DEFINITION

Migration, or specifically immigration, is the process of individuals from a population with different genetic constitution (that is different allele and genotype frequency) entering the population under consideration.

Assume a group of individuals from a donor population with allele frequencies p_m and q_m (for A_1 and A_2, respectively) immigrate to a recipient (host) population with initial allele frequencies p_0 and q_0 (for A_1 and A_2 alleles, respectively). In the absence of other evolutionary processes, it is easy to conclude that after immigration, the host population is composed of a proportion of individual from the donor population, m, and the rest (1-m) from the original host population. Therefore, the new allele frequency for A_1 in the host population is:

$$p_1 = mp_m + (1-m)\,p_0 = m\,(p_m - p_0) + p_0 \qquad [17]$$

Consequently, the difference between the new and the old A_1 allele frequency in the host population is:

$$\Delta_p = p_1 - p_0 = m(p_m - p_0) \qquad [18]$$

When the migration process continues over generations, the cumulative effect of migration (M) can be obtained from

$$M = \frac{\Delta p_{total}}{p_m - p_0} \qquad [19]$$

where M stands for the total migration.

Example 5 – effect of repeated generations of migration

Consider a recipient population with $q_0 = 1.0$ (the recipient population is homozygote for the A_2 allele), and a donor population with $q_m = 0.0$ (the donor population is homozygote for the A_1 allele). If at each generation, the number of individuals from the donor population immigrating to the recipient population is 1% (m = 0.01), then after about 600 generations almost all individuals in the donor population will be homozygote for the A_1 allele (See Figure 7).

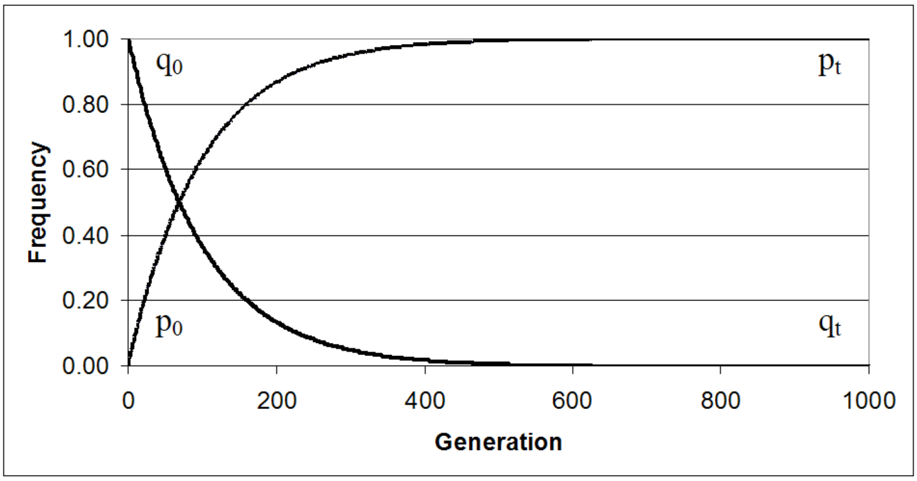

Figure 7 – Cumulative effect of 1% migration from an A_1A_1 population to an A_2A_2 population (p_t and q_t are the frequency of the A_1 and A_2 alleles, respectively, in the recipient population at generation t)

Comment on Example 5

First, please notice that the actual value of p_t at Generation 1000 is not exactly 1.0, but rounded to two decimal places. Second, a migration rate of 1% per generation is indeed a very high value in comparison with a mutation rate of 10^{-6}. To appreciate the strength of 1% migration per generation, compare the x-axis of Figures 6 and 7.

Exercise: Try some different values for m, p_0 and p_m to see what would be the accumulated effect of migration after a large number of generations of continuous migration.

GENERAL COMMENT ON NOTATIONS

There is nothing sacrosanct about the symbols used here is this compendium or any textbook. The same is true for equations. Any equation is just algebraic shorthand for describing the effects of a process, and symbols or notations are just equivalent to words. As such, any concept can be expressed in many different ways. Of course, some may be considered more beautiful than the others, but that is just a matter of taste. As an example, consider Equation 17. I can choose to define the frequency of the A_1 allele in the donor population as P instead of p_m, define the frequency of the A_1 allele in the host population as p_t instead of p_0, and finally define the new allele frequency as p_{t+1}. As the result Equation 17 changes to:

$$p_{t+1} = (1-m)\,p_t + mP = p_t + m\,(P - p_t) \qquad [17]$$

I can also choose to study the changes in the allele frequency for A_2 instead of A_1, and then Equation 17 changes to:

$$q_1 = mq_m + (1-m)\,q_0 = m\,(q_m - q_0) + q_0 \qquad [17]$$

As you see, the choice of letters, symbols, notations, alleles, and so on are just eccentric trivialities and entail no importance of their own. They are just tools.

SELECTION

Up to now we have assumed that all alleles and genotypes of the parental generation contribute equally likely to the gamete pool of the offspring generation, and allele and genotype frequencies in the offspring generation are only dependent on the allele and genotype frequencies in the parental generation. However, it is very important to realize that some alleles or allele combinations (genotypes) may actually be associated with some advantages or disadvantages to the individual. In other words, there might be some differences between individuals in their ability to produce fertile and viable gametes or survive at different stages of life that depends on the alleles that they carry.

DEFINITIION

Selection is the process of differential contribution of individuals of the parental generation to the gamete pool of the offspring generation. In population genetics, and for all practical purposes, "selection" is synonymous to "natural selection" acting on individuals with different abilities with regard to fertility and viability.

The contribution of each individual can be measured by a mathematical concept called "fitness", which is the multiplication of individual's fertility and viability values.

Consider a locus with two alleles A_1 and A_2 (Table 8). Let the individuals carrying the three genotypes A_1A_1, A_1A_2 and A_2A_2 have fitness values $1-s_1$, 1 and $1-s_2$, where s_1 and s_2 are the selection coefficients acting against A_1A_1 and A_2A_2, respectively. Please notice that the choice of value of 1 for the heterozygotes is arbitrary. What is important is that, we would like to study relative fitness values and we have to choose one of the genotypes as our reference point. We could have chosen the fitness of A_1A_1 or A_2A_2 as the reference point. Please also notice that, the values assigned to s_1 and s_2 can be positive or negative. (Choice of A_1A_2 as the reference point creates some complexity and probably some confusion. However, it gives a desirable generality.)

If both s_1 and s_2 are positive (that is the selection acts against homozygotes), then the heterozygotes are at advantage. If s_1 is positive and s_2 negative, then A_2A_2 is at advantage, A_1A_1 at disadvantage and individuals carrying A_1A_2 have intermediate fitness values.

Selection acting on the three genotypes leads to changes in the relative contribution of them and consequently leads to changes in allele and genotype frequencies (see Table 8).

Table 8 – Effect of selection on allele and genotype frequencies

	Genotypes			
	A_1A_1	A_1A_2	A_2A_2	Total
Initial frequency	p^2	2pq	q^2	1
Selection coefficient	s_1	0	s_2	
Fitness	$(1-s_1)$	1	$(1-s_2)$	
Contribution	$(1-s_1)\ p^2$	2pq	$(1-s_2)\ q^2$	$1-s_1p^2-s_2q^2$

Frequency of the A_2 allele after considering the effect of selection is:

$$q_1 = \frac{q - s_2q^2}{1 - s_1p^2 - s_2q^2} \qquad [20]$$

The change in frequency of A_2 can be calculated from:

$$\Delta_q = q_1 - q_0 = \frac{pq(s_1p - s_2q)}{1 - s_1p^2 - s_2q^2} \qquad [21]$$

Equation 18 is a general equation and can be simplified for different situation. For example, assume that A_1 is completely dominant over A_2, i.e. fitness of A_1A_1 is equal to the fitness of A_1A_2. In such a case the value of s_1 is equal to zero. Substituting $s_1=0$ in Equation 20 gives:

$$q_1 = \frac{q - s_2q^2}{1 - s_2q^2} \qquad [22]$$

Exercise: Try to derive equations for selection against the dominant allele.

Long term effect of selection: Prediction of long-term effect of selection on a single locus with two alleles is very simple. One just needs to use Equation 18 repeatedly. However, the emerging patterns are very diverse and depend on the values of s_1, s_2, p_0 and q_0. Figure 7 illustrates the effect of 25 generations of selection with the values s_1= -0.25, s_2 =0.25, p_0 =0.05 and q_0 =0.95. Please notice that, the difference in the fitness values of the two homozygous genotypes is equal to 50% of the fitness value for the heterozygous genotype.

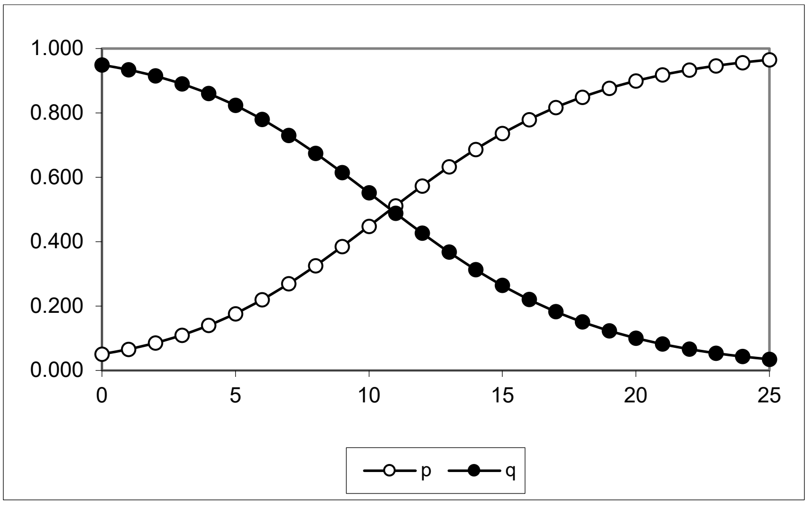

Figure 8 – Strong selection against A_2 allele (for more details see the text).

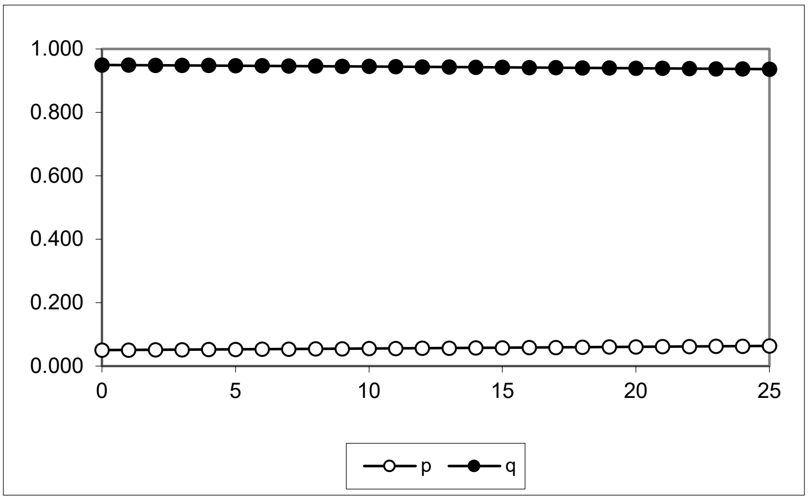

Figure 9 – Weak selection against A_2 allele (for more details see the text).

Changing the fitness values to s_1= -0.01, s_2 =0.01 (Figure 9) leads to a pattern that is completely different. With the fitness values, as is depicted in Figure 8 (s_1= -0.25, s_2 =0.25), it seems that 25 generation is enough to change the frequency of A_1 allele to a value very close to 1.0. On the other hand, with the difference in fitness values as small as 2% (Figure 9), it seems that 25 generations is a very short time to observe any large change.

One special case of predicting the time (generations) needed to achieve a certain amount of change is when the selection is against a rare recessive allele. Number of generations required can be calculated as:

$$t = \frac{1}{q_t} - \frac{1}{q_0} \qquad [23]$$

HARDY-WEINBERG MODEL: PART 2

In the first part of describing the Hardy-Weinberg model the emphasis was on how a large random mating population with no mutation, migration and selection reaches to (or stays at) an equilibrium state in which allele and genotype frequencies are constant from parental generation to offspring generation. In other words, the emphasis was on a single cycle of reproduction, as if there are always only two generations to consider. The model was very robust and minor violations of its minor pattern-assumptions posed no serious threats to it. We also considered some major violations of the model assumptions (one at a time) and could conclude that violating any major process-assumptions will destroy the equilibrium.

One thing that we did not do was to examine the assumptions to see if they were realistic. Here, in Part 2 of describing the HWM, without going into any detail, it is important to realize that none of the assumptions of the HWM is far from reality as the assumption of large population size (and its consequences such as the absence of sampling effect and random-mating). The reason is that a population can be kept close (to avoid migration), or in a very desirable and uniform environment (to minimize natural selection), and so on. However, constructing and maintaining a large population and preventing the sampling process is next to impossible.

The emphasis in this part is to study the effects of changing the population size from large to small and examine the consequences of repeated cycles of reproduction when the population size is small.

Consider the population designated as the “base population” in Figure 10. The base population not only fulfills all assumptions of the Hardy-Weinberg model, but has one extra (explicit) feature, namely all individuals in the base population are unrelated. Further, although it is not strictly necessary, assume that all alleles in the base population are unique. The uniqueness of alleles means that every individual has two alleles that are not only different from each other, but are different from all other alleles in the population. Some of these alleles may have similar effects on the trait under consideration, i.e. they are identical by function. However, it is their origin which is unrelated to the origin of all the other alleles.

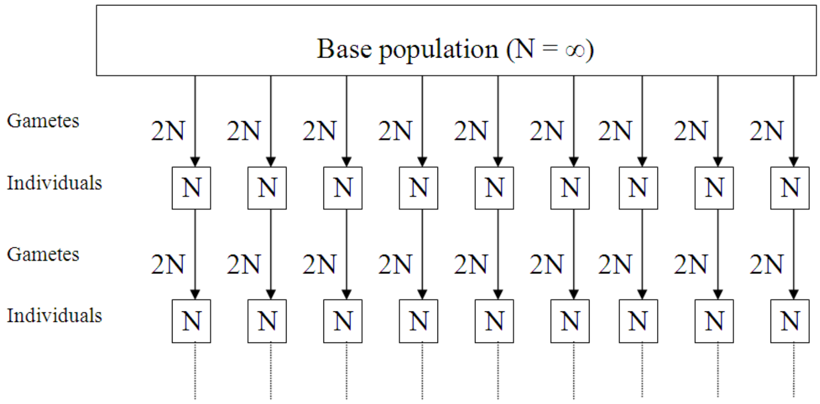

Figure 10 – Division of a population fulfilling the assumptions of the Hardy-Weinberg model into a number of small sub-populations, and thereby violating the assumption of large population size.

Now, imagine that we sample, just randomly, a small number of individuals, say N, to form a small sub-population. The process of sampling can be repeated many times until we have a large number of sub-populations. Let all sub-populations have the same size and follow the assumptions of the HWM, except for the assumption of large population size.

Each of the sub-populations, or lines (or idealized populations), formed in the manner described above, has only N members and carries only 2N alleles. At each round of reproduction, out of the large number of gametes that the N individuals generate, only 2N gametes are sampled to form the individuals of the next generation. And the small and constant population size continues generation after generation. The small population size entails two processes at two levels: the level of locus (sampling process) and the level of individuals (mating process).

These two processes, combined with each other, create several patterns: random drift, divergence among sub-populations, increased homozygosity within each sub-population, etc.

SAMPLING

At the locus level, there is the sampling process. Each of the alleles may not get sampled, so that it is lost from the sub-population. Because all alleles are unique (and there is no mutation and migration to re-generate them or re-introduce them), if an alleles is lost from the sub-population, it is lost for ever. On the other hand, each of the alleles may get sampled, so that new copies of it are generated and passed on to the next generation.

DEFINITION

Sampling is the process of random passage of alleles from parental generation to the offspring generation.

Randomness of this process lies in the fact that different alleles are not under any discernible selection pressure, and are passed down to the next generation, or not, just by chance.

Because there are no selective differences between different alleles, allele frequencies in each of the sub-populations are subject to random fluctuations. Consider a base population in which the frequencies of the two alleles, f(A_1)=p_0 and f(A_2)=q_0, are equal to each other, that is 0.50. Therefore, each of sub-populations starts with an expected allele frequency of 0.50. In other words, the average of allele frequencies in all sub-populations are $\overline{p}=p_0$ and $\overline{q}=q_0$. At the time of reproduction, allele frequencies among the gametes in one of the sub-populations may change to 0.45 and 0.55 for the A_1 and A_2 alleles, respectively, just by chance. In a different sub-population the allele frequencies may change to 0.60 and 0.40 for the A_1 and A_2 alleles, respectively. Because the sampling process is a random process, there is no way to predict if any allele frequency will go up, or down. What can be predicted is how large is the variance of change in allele frequencies among all sub-populations. The reason is that the process of picking up two alleles at random is a binomial process, very similar to picking up two colored balls from a bag containing red and blue balls (or tossing a coin, if you like).

Using the properties of the binomial distribution we can say that the sampling consists of picking up an allele at a time. Probability of picking up A_1 is, on average, p_0 and probability of picking up A_2 is, on average, q_0 and we are going to do the sampling 2N times. Therefore, the variance of change in an allele frequency (say Δ_q) is:

$$\sigma^2_{\Delta_q} = \frac{p_0 q_0}{2N} \qquad [24]$$

Presence of N in the denominator of Equation 24 shows that the variance of change in allele frequency very much depends on the size of sub-populations. The smaller the population, the larger the variance becomes. This means that if the sub-populations are small, the amount of change in an allele frequency in any single sub-population can be quite large. When the sampling process continues for many generations, the allele frequencies in any sub-population fluctuate up and down. Depending on the size of sub-populations, sooner or later, frequency of one the alleles gets closer and closer to 1.0, and eventually that allele gets fixed in that sub-population. The emerging pattern from the sampling process is that the sub-populations diverge from each other (see Figure 11).

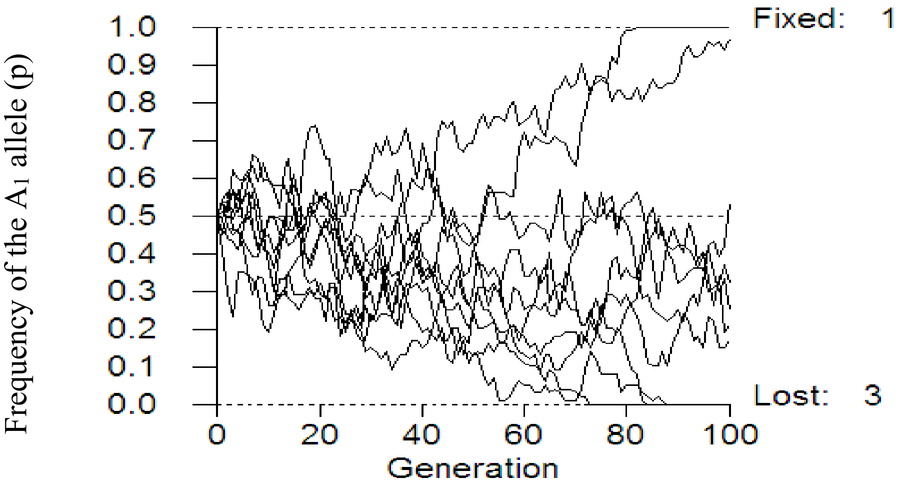

Figure 11 – Divergence of sub-population under sampling process is a small population.

MATING PROCESS & INBREEDING

At the individual level, there is the mating process. Each individual has the opportunity of mating with any other individual (this is called panmixia), and because there is no selection, all matings lead to production of equally fertile and healthy offspring. However, because population size is constant, not all individuals can mate or produce offspring and consequently some of the individuals cannot contribute to the production of offspring. Eventually, and inevitably, the small population size leads to the non-randomness of matings.

It is extremely difficult to define random mating. In the most abstract way, random mating can be defined as a correlation of zero between mates (correlation with respect to phenotype, genotype or relationship). One thing that is even more difficult to do is classification of different sorts of non-random mating.

For our purposes, let (for the moment) define inbreeding as a form of non-random mating and a consequence of small sub-population size. In other words, in a small sub-population, matings depart from random mating, in the sense that the mating individuals tend to become more related to each other. The matings happens progressively among relatives.

DEFINITION

Inbreeding is defined as the process of two identical copies of a single ancestral allele being passed down to a descendent.

It is the "identity by descent" that is of central importance in the inbreeding process.

Now that we have defined inbreeding, we need a way of measuring the inbreeding (let's call it the inbreeding coefficient, F). Conceptually, the inbreeding coefficient can be measure in the following manner. At the time of reproduction in an idealized population (sub-population) a limited number of gametes are going to be sampled (to be exact 2N gametes). Each of the N individuals in this generation has two unique alleles, each of which has been copied a large number of times (each gamete contains one copy of one of these two alleles). Therefore, in the sub-population we have the following alleles: A_1 and A_2 in the first individual, A_3 and A_4 in the second individual, A_{2N-1} and A_{2N} in the N^{th} individual. The question of interest is: What is the probability of sampling a second gamete containing an allele (say A_1) when the first sampled gamete already contains the same allele (that is A_1)? The answer is:

$$\frac{1}{2N} \qquad [25]$$

The reason is that, there is virtually an infinite number of gametes. The fact that we have already sampled one gamete containing A_1 does not change the frequency of A_1 gametes. The frequency of A_1 gamete was $1/2N$ and it remains at $1/2N$. After one round of reproduction (that is, in the next generation, let's call it Generation 1), there are still 2N alleles left in the sub-population. However, these 2N alleles are not unique anymore. Some of them are unique, and some are not. The proportion of non-unique alleles is $1/2N$ and the proportion of unique alleles is, obviously, $1-(1/2N)$. At the time of reproduction for the individuals of Generation 1, to produce the individuals of Generation 2, the sampling of gametes needs to be repeated. Again, the question to be asked is: What is the probability of sampling a second gamete containing an allele when the first sampled gamete already contains the same allele? The answer is that we have to calculate two probabilities. The first probability is for the new copies that may be produced in this round of reproduction, the new inbreeding, which is $1/2N$. The second probability is for the old copies that are being passed down from the previous generation. Therefore, in Generation 2 the inbreeding coefficient (F_2) is:

$$F_2 = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_1 \qquad [26]$$

or more generally:

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1} \qquad [27]$$

It can be concluded that the change in inbreeding coefficient (ΔF, also called rate of inbreeding) in an idealized population is:

$$\Delta F = \frac{1}{2N} \qquad [28]$$

Please notice that the above description was a "conceptual" way of measuring the inbreeding coefficient. In the above description, we assumed that two gametes from an individual can unite with each other, as if the organism was self-fertilizing hermaphrodite. That is not a realistic situation.

However, the general results (Equations 26 and 27) are still valid for the absolute majority of species, including all farm animals. The only difference is that for sampling of two gametes containing two identical copies of one ancestral allele we have to wait for two generation. In a realistic situation, all alleles in Generation 0 are assumed to be unique, and multiple copies of them are created in Generation 1. It is in Generation 2 that the possibility of two uniting gametes, having identical copies of an ancestral allele, exists for the first time.

Measuring inbreeding coefficient (F)

In a way, the inbreeding coefficient of an individual can be measured only, and only, when unequivocal knowledge of identity by descent for all loci is available, and this is almost always impossible.

However, there are many different methods to calculate the average inbreeding coefficients for all loci of an individual or the average inbreeding coefficient for all individuals of a population with respect to a locus. These methods can be, roughly, divided in two groups: Exact methods (which are based on the information about the pedigree), and approximate methods (which are based on the information about the population structure).

Measuring F based on pedigree information by path method

For calculating the inbreeding coefficient of an individual based on the pedigree information we need to go back in the pedigree of that individual, through one of its parents, until we reach a common ancestor, and then to the other parent. Using the following equation the inbreeding coefficient is calculated:

$$F_X = \sum \left(\frac{1}{2} \right)^n \qquad [29]$$

where n is the number of individual starting from one parent, through the common ancestor, to the other parent, and Σ is for summation over all common ancestors. Consider the following simple pedigree:

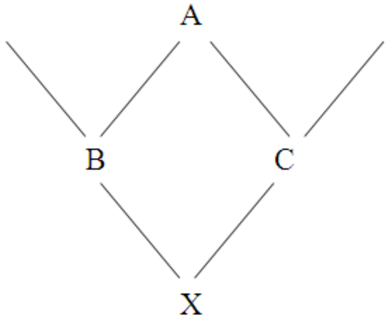

Figure 12 – A simple pedigree for an individual with one common ancestor

In the pedigree in Figure 12 the animal of interest is X, whose parents B and C have only one common ancestor, A. There are three individuals to count B, C and A. Therefore, $F_X = (1/2)^3 = 1/8 = 0.125$.

In the pedigree in Figure 13, the parents of the animal of interest, E, have two common ancestors, A and B, both of which should be taken into account.

There are two loops and three individuals in each loop. One loop consists of C, A and D, while the other loop consists of C, B and D. Therefore, F_E = $(1/2)^3 + (1/2)^3 = 1/4 = 0.25$.

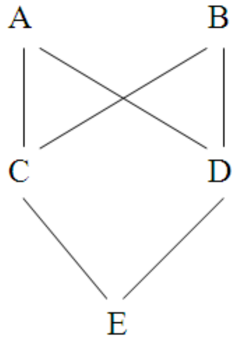

Figure 13 – A simple pedigree for an individual with two common ancestor

Equation 29 can be used for those individuals whose common ancestor is not inbred. If the common ancestor is itself an inbred individual, the inbreeding coefficient of the ancestor should be taken into account also. Consider the following pedigree:

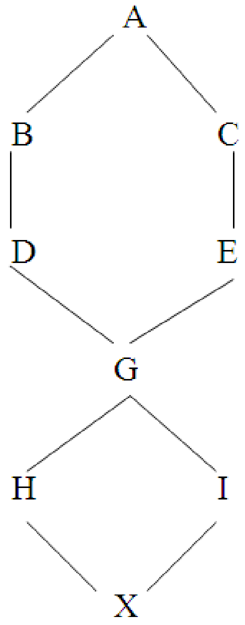

Figure 14 – A pedigree for an individual whose common ancestor is an inbred individual

Equation to be used for such pedigrees is:

$$F_X = \sum \left(\frac{1}{2}\right)^n (1 + F_G) \qquad [30]$$

For individual X the inbreeding coefficient should be calculated as follows:

$F_x = (1/2)^3\ [1 + (1/2)^5] = (1/2)^3 + (1/2)^{3+5} = (1/8) + (1/256) = (33/256)$

Relationship

Calculation of inbreeding coefficient in an individual reveals also the "coefficient of relationship" between the parents of the individual. The coefficient of relationship, also called "additive genetic relationship", "theoretical correlation" or simply "relationship", measures the average proportion of [common] alleles that are identical by descent in those two individuals. The relationship between any two individuals, designated as "a" is twice the inbreeding coefficient of their offspring. For individuals H and I in Figure 14, the relationship is calculated as:

$$a_{HI} = 2F_X \qquad [31a]$$

Equation 30 can also be used to calculate the relationship between two individuals. However, instead of starting from the offspring, we start from one of the individuals, go through the common ancestor and to the other individual and count the number of paths.

Equation 31a can be re-arranged to so that the inbreeding coefficient of any individual can be calculated from the additive genetic relationship of its parents, i.e.

$$F_X = ½ a_{HI} \qquad [31b]$$

Measuring F based on pedigree information by tabular method

Measuring F using the path method is very easy as long as the pedigree is simple. However, if the pedigree becomes deep (more than 2-3 generations), or if the pedigree becomes large (more than 20-30 individuals), or if the pedigree becomes complicated (more than 2-3 common ancestors), then measuring F by the path method will no longer be easy. The problem with the path method, one could say, is that it is not easy to program it, to computerize it.

An alternative method of measuring F is by using the so-called tabular method. Description of tabular method is, a bit, cumbersome. So, please give it a chance. As an example, consider the simple pedigree depicted in Figure 13. Let's make a list of all individuals and their parents. Further, order the individual from old to young (as much as possible, so that no offspring appear in the list before their parents).

Table 9 – Tabular representation of the pedigree in Figure 13.

Individual	Sire	Dam

A	-	-
B	-	-
C	A	B
D	A	B
E	C	D

Before starting the calculations draw a table for all individuals. Let the first column be filled with the names (or ID's or numbers) of individuals. Similarly, let the first row be filled with the names of individuals and their parents.

Table 9.1 – The A-matrix of the pedigree in Figure 13.

j	-,-	-,-	A,B	A,B	C,D
i	A	B	C	D	E
A					
B					
C					
D					
E					

The table shown above is actually the famous A-matrix or the "additive genetic relationship matrix" (sometimes also called "Wright's numerator relationship matrix").

The diagonal values of the A-matrix, a_{ii}, are equal to the relationship of an individual with itself, which is,

$$a_{ii} = 1.0 + \frac{1}{2} a_{sd} \qquad [32]$$

where a_{sd} is the coefficient of relationship between sire and dam of the individual.

The off-diagonal values of the A-matrix, a_{ij} (where i ≠ j), are equal to the average relationship of individual i with the parents of individual j, which is,

$$a_{ij} = \frac{1}{2}(a_{is_j} + a_{id_j}) \qquad [33]$$

where s_j and d_j are sire and dam of individual j.

In practice the cells of the A-matrix can be filled as in the following steps:

Step 1
Start with the individuals of the "base" population, i.e. individuals whose parents are unknown (in this example, individuals A and B). These individuals are assumed to be non-inbred and unrelated to each other. Let the diagonal and upper-diagonal values for these individuals to be 1.0 and 0.0, respectively. It can be seen that these values are results of Equations 32 and 33, respectively, for animals with unknown ancestry.

Table 9.2 – The A-matrix of the pedigree in Figure 13.

j	-,-	-,-	A,B	A,B	C,D
i	A	B	C	D	E
A	1.0	0.0			
B		1.0			
C					
D					
E					

Step 2

Starting from the first individual of the "base" population, use Equation 33 to calculate the relationship of the individual in row i with the individual in column j. The relationship between individual A and individuals C, D and E are shown below:

$$\begin{aligned} a_{AC} &= ½(a_{AA} + a_{AB}) = ½(1.0 + 0.0) = 0.5 \\ a_{AD} &= ½(a_{AA} + a_{AB}) = ½(1.0 + 0.0) = 0.5 \\ a_{AE} &= ½(a_{AC} + a_{AD}) = ½(0.5 + 0.5) = 0.5 \end{aligned}$$

You realize that when we go from left to right, all elements that we need for any cell have been already calculated. In this example, a_{AA} and a_{AB} that are needed for individuals C and D have already been calculated in Step 1. And a_{AC} and a_{AD} that are needed for individual E have just been calculated when we were dealing with individual C and D. Calculation of relationship between individual B and individual C, D and E follow the same method as for individual A. Therefore, the first two rows of the A-matrix can now be filled.

Table 9.3a – The A-matrix of the pedigree in Figure 13.

j	-,-	-,-	A,B	A,B	C,D
i	A	B	C	D	E
A	1.0	0.0	0.5	0.5	0.5
B		1.0	0.5	0.5	0.5
C					
D					
E					

Step 3

Starting with the diagonal value for the first individual of "non-base" population, use Equation 33 and work through the row from left to right using Equation 33. Values for individual C are shown below:

$$\begin{aligned} a_{CC} &= 1.0 + ½a_{AB} = 1.0 + ½(0.0) = 1.0 \\ a_{CD} &= ½(a_{CA} + a_{CB}) = ½(0.5 + 0.5) = 0.5 \\ a_{CE} &= ½(a_{CC} + a_{CD}) = ½(1.0 + 0.5) = 0.75 \end{aligned}$$

Table 9.3b – The A-matrix of the pedigree in Figure 13.

j	-,-	-,-	A,B	A,B	C,D

i	A	B	C	D	E
A	1.0	0.0	0.5	0.5	0.5
B		1.0	0.5	0.5	0.5
C			1.0	0.5	0.75
D					
E					

Values for individual D are shown below:

$$a_{DD} = 1.0 + \tfrac{1}{2} a_{AB} = 1.0 + \tfrac{1}{2}(0.0) = 1.0$$
$$a_{DE} = \tfrac{1}{2}(a_{DC} + a_{DD}) = \tfrac{1}{2}(0.5 + 1.0) = 0.75$$

Table 9.3c – The A-matrix of the pedigree in Figure 13.

	j	-,-	-,-	A,B	A,B	C,D
i		A	B	C	D	E
A		1.0	0.0	0.5	0.5	0.5
B			1.0	0.5	0.5	0.5
C				1.0	0.5	0.75
D					1.0	0.75
E						

Now it's time to calculate the diagonal value for the last individual, E, as follows:

$$a_{EE} = 1.0 + \tfrac{1}{2} a_{CD} = 1.0 + \tfrac{1}{2}(0.5) = 1.25$$

Table 9.3d – The A-matrix of the pedigree in Figure 13.

	j	-,-	-,-	A,B	A,B	C,D
i		A	B	C	D	E
A		1.0	0.0	0.5	0.5	0.5
B			1.0	0.5	0.5	0.5
C				1.0	0.5	0.75
D					1.0	0.75
E						1.25

Because the A-matrix is symmetric, we can just copy the upper-diagonal values to the lower-diagonal values and the A-matrix is complete.

Table 9.3e – The A-matrix of the pedigree in Figure 13.

	j	-,-	-,-	A,B	A,B	C,D
i		A	B	C	D	E
A		1.0	0.0	0.5	0.5	0.5
B		0.0	1.0	0.5	0.5	0.5
C		0.5	0.5	1.0	0.5	0.75
D		0.5	0.5	0.5	1.0	0.75
E		0.5	0.5	0.75	0.75	1.25

Combining Equations 31b and 32, you can see that the diagonal value for any individual is:

$$a_{ii} = 1.0 + ½ a_{sd} = 1.0 + F_I \qquad [34]$$

Therefore, the inbreeding coefficient for individual E, $F_E = 0.25$.

As mentioned earlier, the tabular method seems, a bit, cumbersome. However, because all calculations are performed in small steps and any values needed for younger individual have been calculated earlier, it is very easy to write a computer program can calculate additive genetic relationships and inbreeding coefficients for the most complicated pedigrees, even if the pedigree is very large.

Measuring F based on effective population size

In situations where the information on the pedigree is not available, or the pedigree is too extensive (i.e. pedigree contains a very large number of individuals) we may calculate an approximation of the inbreeding coefficient. This is done through calculation of a parameter called "effective population size". The concept of effective population size is a confusing concept for many people. In order to facilitate its understanding, we can use some analogy from ordinary (!) life. Consider two large international corporations (Company A and B), each with 1000 shareholders. In Company A, all shareholders own the same number of shares, and consequently can influence the decisions equally. In company B, one of the shareholders owns 51% of the shares. Do you think that each of the shareholders in Company B can influence the decisions equally? In Company A, there are 1000 decision makers. In Company B, there is effectively, only one decision maker. In population genetics, the gamete pool (collection of all alleles) is like a company. The question is: "Are all parents contributing equally to the gamete pool?"

DEFINITION

For a real population not fulfilling the assumptions of the idealized population, the effective population size, N_e, *is defined as the size of an idealized population that would lead to the same rate of inbreeding (ΔF) as in the real population.*

In other words, if the rate of inbreeding, or the sampling variance, or the probability of identity by descent could be calculated for a real population, then we could back calculate the size of an idealized population with equal rate of inbreeding, or the sampling variance, or the probability of identity by descent. The size of that idealized population is the "effective population size" for the real population.

If the effective population size is known for a real population, then we can use the following equation for calculation of rate of inbreeding (and many other inbreeding related parameters):

$$\Delta F = \frac{1}{2N_e} \text{ or } N_e = \frac{1}{2\Delta F} \qquad [35]$$

Effective population size depends on the population structure, .i.e. number of individuals of each gender, mating ratios, family size and so on. The problem is that there is no really general equation that can provide the effective population size for all sorts of population structures. The closest you can get to a general equation is the following [horrible looking] equation:

$$\begin{aligned}\frac{1}{N_e} = & \frac{1}{16M}\left[2+\sigma^2_{mm}+\frac{2M}{F}\sigma_{mm,mf}+(\frac{M}{F})^2\sigma^2_{mf}\right] \\ & +\frac{1}{16F}\left[2+(\frac{M}{F})^2\sigma^2_{fm}+\frac{2F}{M}\sigma_{fm,ff}+\sigma^2_{ff}\right]\end{aligned} \qquad [36]$$

where M and F are the number of males and females, respectively, σ^2 and σ are the variance and covariance of family size, respectively, and m and f indicate the path from parent to offspring. For example, $\sigma_{mm,mf}$ stands for the covariance between “number of sires to sons” and “number of sires to daughters”. Equation 36 can be simplified for some situations (unfortunately, not all situations).

When the number of males (N_m) and females (N_f) are different, Equation 36 can be approximated to:

$$N_e \approx \frac{4N_mN_f}{N_m+N_f} \qquad [37]$$

If there is a variation in family size, but the variation is equal for males and females, Equation 36 is approximated to:

$$N_e \approx \frac{4N}{V_k+2} \qquad [38]$$

And if the variation in family size is different for the two sexes, then we can use the following equation:

$$N_e \approx \frac{8N}{V_{km}+V_{kf}+4} \qquad [39]$$

However, when the numbers of individuals in successive generations are different, Equation 36 cannot be used. A close look at Equation 31 shows that for calculation of effective population size we need to calculate the harmonic mean, and this is exactly what we need to do to calculate the effective population size when the numbers of individuals in generations 1 to t (N_1, N_2, …, N_t) are different:

$$\frac{1}{N_e} \approx \frac{1}{t}\left[\frac{1}{N_1}+\frac{1}{N_2}+\ldots+\frac{1}{N_t}\right] \qquad [40]$$

Using Equations 37 to 40, together with the general Equation 36, one can calculate the rate of inbreeding in a population and consequently an average inbreeding coefficient for all individuals.

Interpretation of the inbreeding coefficient

As mentioned before, the inbreeding coefficients calculated from the pedigree or the population structure, are average values and can be interpreted in two ways (or at two levels). At the level of loci, the inbreeding coefficient can be interpreted as the average proportion of all loci in an individual that contain two alleles that are identical by descent. At the population level, the inbreeding coefficient can be interpreted as the average proportion of all individuals that carry two identical copies of an ancestral allele in a certain locus.

ACKNOWLEDGEMENTS

I have used, freely, ideas from many sources. I am, of course, indebted to these sources, and their authors. Two sources have had major effects on my understanding of the subject. The first is Falconer and Mackay (1996) Introduction to quantitative genetics. The second is the collection of class notes by Larry Schaeffer. In conveying the ideas described in my sources, I hope I have not exceeded the boundaries of ethically scientific conduct.