

# Lecture 2

## Basic Population Genetics

Bruce Walsh. Aug 2004. Royal Veterinary and Agricultural University, Denmark

### Allele and Genotype Frequencies

The frequency  $p_i$  for allele  $A_i$  is just the frequency of  $A_iA_i$  homozygotes plus half the frequency of all heterozygotes involving  $A_i$ ,

$$p_i = \text{freq}(A_i) = \text{freq}(A_iA_i) + \frac{1}{2} \sum_{i \neq j} \text{freq}(A_iA_j) \quad (2.1)$$

The  $1/2$  appears since only half of the alleles in heterozygotes are  $A_i$ . Equation 2.1 allows us to compute *allele* frequencies from *genotypic* frequencies. Conversely, since for  $n$  alleles there are  $n(n+1)/2$  genotypes, the same set of allele frequencies can give rise to very different genotypic frequencies. To compute genotypic frequencies solely from allele frequencies, we need to make the (often reasonable) assumption of random mating. In this case,

$$\text{freq}(A_iA_j) = \begin{cases} p_i^2 & \text{for } i = j \\ 2p_i p_j & \text{for } i \neq j \end{cases} \quad (2.2)$$

Equation 2.2 is the first part of the **Hardy-Weinberg theorem**, which allows us (assuming random mating) to predict genotypic frequencies from allele frequencies. The second part of the Hardy-Weinberg theorem is that allele frequencies will remain unchanged from one generation to the next, *provided*: (1) infinite population size (i.e., no genetic drift), (2) no mutation, (3) no selection, and (4) no migration. Further, for an autosomal locus, a single generation of random mating gives genotypic frequencies in **Hardy-Weinberg proportions** (i.e., Equation 2) and the genotype frequencies forever remain in these proportions.

### Gamete Frequencies, Linkage, and Linkage Disequilibrium

Random mating is the same as gametes combining at random. For example, the probability of an  $AABB$  offspring is the chance that an  $AB$  gamete from the father and an  $AB$  gamete from the mother combine. Under random mating,

$$\text{freq}(AABB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \quad (2.3a)$$

For heterozygotes, there may be more than one combination of gametes that gives rise to the same genotype,

$$\text{freq}(AaBB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(aB|\text{mother}) + \text{freq}(aB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \quad (2.3b)$$

If we are only working with a single locus, then the gamete frequency is just the allele frequency, and under Hardy-Weinberg conditions, these do not change over the generations. However, when the gametes we consider involve two (or more) loci, recombination can cause gamete frequencies to change over time, even under Hardy-Weinberg conditions. At **linkage equilibrium**, the frequency

if a multi-locus gamete is just equal to the product of the allele frequencies. For example, for two and three loci,

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B) \quad \text{for 2 loci,} \quad \text{freq}(ABC) = \text{freq}(A) \cdot \text{freq}(B) \cdot \text{freq}(C) \quad \text{for 3 loci}$$

In linkage equilibrium, the alleles at different loci are independent — knowledge that a gamete contains one allele (say  $A$ ) provides no information on the allele from the second locus. More generally, loci can show **linkage disequilibrium** (LD), which is also called **gametic phase disequilibrium** as it can occur between unlinked loci. When LD is present,

$$\text{freq}(AB) \neq \text{freq}(A) \cdot \text{freq}(B)$$

Indeed, the disequilibrium  $D_{AB}$  for gamete  $AB$  is defined as

$$D_{AB} = \text{freq}(AB) - \text{freq}(A) \cdot \text{freq}(B) \quad (2.4a)$$

Rearranging Equation 2.4a shows that the gamete frequency is just

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B) + D_{AB} \quad (2.4b)$$

$D_{AB} > 0$  implies  $AB$  gametes are more frequent than expected by chance, while  $D_{AB} < 0$  implies they are less frequent.

We can also express the disequilibrium as a covariance. Code allele  $A$  as having value one, other alleles at this locus having value zero. Likewise, at the other locus, code allele  $B$  with value one and all others with value zero. The covariance between  $A$  and  $B$  thus becomes

$$\text{Cov}(AB) = E[AB] - E[A] \cdot E[B] = 1 \cdot \text{freq}(AB) - (1 \cdot \text{freq}(A)) \cdot (1 \cdot \text{freq}(B)) = D_{AB} \quad (2.5)$$

If the recombination frequency between the two loci is  $c$ , then the disequilibrium after  $t$  generations of recombination is simply

$$D(t) = D(0)(1 - c)^t \quad (2.6)$$

Hence, with loose linkage ( $c$  near  $1/2$ )  $D$  decays very quickly and gametes quickly approach their linkage equilibrium values. With tight linkage, disequilibrium can persist for many generations. As we will see in numerous instances throughout this course, it is the presence of linkage disequilibrium that allows us to map genes.

### The Effects of Population Structure

Many natural populations are *structured*, consisting of a mixture of several subpopulations. Even if each of the subpopulations are in Hardy-Weinberg proportions, samples from the entire population need not be. Suppose our sample population consists of  $n$  subpopulation, each in HW equilibrium. Let  $p_{ik}$  denote the frequency of allele  $A_i$  in population  $k$ , and let  $w_k$  be the frequency that a randomly-drawn individual is from subpopulation  $k$ . The expected frequency of an  $A_i A_i$  homozygote becomes

$$\text{freq}(A_i A_i) = \sum_{k=1}^n w_k \cdot p_{ik}^2 \quad (2.7a)$$

while the overall frequency of allele  $A_i$  in the population is

$$p_i = \sum_{k=1}^n w_k \cdot p_{ik} \quad (2.7b)$$

We can rearrange this as

$$\text{freq}(A_i A_i) = p_i^2 - \left( p_i^2 - \sum_{k=1}^n w_k \cdot p_{ik}^2 \right) = p_i^2 + \text{Var}(p_i) \quad (2.7c)$$

Hence, Hardy-Weinberg proportions hold only if  $\text{Var}(p_i) = 0$ , which means that all the subpopulations have the same allele frequency. Otherwise, the frequency of homozygotes is *larger* than we expect from Hardy-Weinberg (based on using the average allele frequency over all subpopulations), as

$$\text{freq}(A_i A_i) \geq p_i^2$$

While homozygotes are always over-represented, there is no clear-cut rule for heterozygotes. Following the same logic as above yields

$$\text{freq}(A_i A_j) = 2p_i p_j + \text{Cov}(p_i, p_j) \quad (2.8)$$

Here, the covariance can be either positive or negative.

Population structure can also introduce linkage disequilibrium (even among unlinked alleles). Consider an  $A_i B_j$  gamete and assume that linkage-equilibrium occurs in all subpopulations, then

$$\text{Freq}(A_i B_j) = \sum_{k=1}^n w_k \cdot p_{A_{ik}} \cdot p_{B_{jk}}$$

The expected disequilibrium is given by

$$\begin{aligned} D_{ij} &= \text{Freq}(A_i B_j) - \text{Freq}(A_i) \cdot \text{Freq}(B_j) \\ &= \sum_{k=1}^n w_k \cdot p_{A_{ik}} \cdot p_{B_{jk}} - \left( \sum_{k=1}^n w_k \cdot p_{A_{ik}} \right) \left( \sum_{k=1}^n w_k \cdot p_{B_{jk}} \right) \end{aligned} \quad (2.9)$$

Consider the simplest case of two populations, where the allele frequencies for  $A_i$  differ by  $\delta_i$  and by  $\delta_j$  for  $B_j$ . In this case, Equation 2.9 simplifies to

$$D_{ij} = \delta_i \cdot \delta_j \cdot [w_1(1 - w_1)] \quad (2.10)$$

Hence, in order to generate disequilibrium, the subpopulations must differ in allele frequencies at both loci. Further, the amount of disequilibrium is maximal when both subpopulations contribute equally ( $w_1 = 0.5$ ).

### Forces that Change Allele Frequencies: Genetic Drift

Under the Hardy-Weinberg assumptions, not only are genotype frequencies predictable from allele frequencies, but allele frequencies also remain unchanged from one generation to the next. Hardy-Weinberg is thus the answer to Fleming Jenkin's concern over blending inheritance: in the absence of other forces, the amount of standing genetic variation remains unchanged. However evolutionary forces do result in allele frequencies changing over, and we start of discussions of these forces by considering one of the most basic (and most subtle), **genetic drift**.

Genetic drift arises because populations are finite, and as a result of sampling  $2N$  gametes to form the  $N$  individuals for the next generation, changes (typically very small) in allele frequencies occur. Over long periods of time these small changes result (in the absence of any other forces) in all but one allele being lost from the population. To formally model genetic drift, suppose the current

allele frequency is  $p$  and the population size is  $N$ , then the allele frequency in the next generation is a random variable given by  $1/N$  times a Binomial random variable drawn from  $\text{Bin}(p, N)$ ,

$$\Pr(i \text{ copies} \rightarrow j \text{ copies}) = \frac{N!}{(N-j)!j!} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j} \quad (2.11)$$

The net result is that the mean change in allele frequency is zero (if the current frequency is  $p$ , the expected frequency in the next generation is also  $p$ ). However, the variance in the change in allele frequency is  $p(1-p)/2N$ . Summarizing,

$$E(\Delta p|p) = 0, \quad \sigma^2(\Delta p) = \frac{p(1-p)}{2N}$$

This sampling generates a random walk, a walk that stops when the allele being followed reaches frequencies zero (allele is lost) or one (allele is **fixed**). If the starting frequency of an allele is  $p$ , its ultimate probability of fixation is also  $p$ . Hence, if allele  $A_1$  has frequency 0.1 and allele  $A_3$  has frequency 0.05, then the probability neither is fixed by drift is  $1 - 0.1 - 0.05 = 0.85$ .

Thus, under drift an allele is eventually either lost or fixed, with the time scale for this process scaling with  $N$ . In particular, the expected time to fixation is  $4N$  generations, with a standard error also on the order of  $4N$  generations.

### Coalescence Theory

There is a very rich statistical theory associated with genetic drift. In particular, over the last 15 years or so, the problem has been framed using the very powerful approach of coalescence theory, which follows the distribution of time back to a **common ancestor** for alleles being drawn from a sample. The idea is that under drift, one can eventually trace all existing alleles in a population back to a single DNA molecule from which they all descend. If the mutation rate is high relative to the population size (see below), the alleles may show considerable sequence variation. However, the strength of the coalescent approach is that we first deal with the genealogy (i.e., the full age distribution) of the alleles in a sample, and then superimpose our particular mutation model on this sample.

For two randomly-drawn sequences from an ideal population of size  $N$ , the time back to their most recent common ancestor follows a geometric distribution with success parameter  $q = 1/(2N)$ , so that

$$\Pr(\text{Coalescence in generation } t) = \left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right) \simeq \frac{1}{2N} \exp\left(-\frac{t}{2N}\right) \quad (2.12)$$

The mean coalescence time is  $E[t] = 2N$  generations with variance  $\sigma^2(T) = 4N^2$ . Hence, the probability that two randomly-chosen alleles have a common ancestor within the last  $\tau$  generations is

$$1 - \Pr(\text{no common ancestor in last } \tau \text{ generations}) = 1 - (1 - q)^\tau \quad (2.13)$$

### Forces that Change Allele Frequencies: Mutation

Mutation is another evolutionary force that can change allele frequencies. Historically, models of mutation were rather simplistic, with an allele simply mutating back and forth between two states, i.e., allele  $M$  mutates to  $m$  and vice-versa. Under this simple model, if  $\mu$  is the mutation rate from  $M$  to  $m$  and  $\nu$  the back mutation rate from  $m$  to  $M$ , then the change in allele frequency over one generation is obtained as follows. If  $p$  is the current frequency of  $M$ , the probability that it does not

mutate to  $m$  is  $(1 - \mu)$ , while the chance that the  $1 - p$  of the alleles that are  $m$  mutate to  $M$  is  $\nu$ . Putting these together given the new frequency  $p'$  as,

$$p' = (1 - \mu)p + \nu(1 - p) \quad (2.14a)$$

Thus, allele frequencies change, but on the order of the mutation rate, which are on the order  $10^{-4}$  to  $10^{-9}$  per generation (i.e., very slowly). The allele frequencies change until an **equilibrium** value is reached where  $p' = p$ . Substituting into Equation 2.14 gives the equilibrium value  $\tilde{p}$  as satisfying

$$\tilde{p} = (1 - \mu)\tilde{p} + \nu(1 - \tilde{p}), \quad \text{or} \quad \tilde{p} = \frac{\mu}{\mu + \nu} \quad (2.14b)$$

In 1964, Crow and Kimura produced a much more realistic model of gene mutation, motivated by the structure of a DNA sequence. Their **infinite-alleles model** assumes that since the DNA sequence for a typical gene consists of up to several thousand nucleotides, that any particular mutation is unlikely to be recovered by a back mutation. Rather, each new mutation likely gives a different DNA sequence, and hence a new allele (if we are scoring alleles from DNA sequencing). This generates a very large (essentially infinite) collection of alleles. Crow and Kimura were interested in the balance between genetic drift removing variation and mutation introducing new variation. Their analysis showed that the expected heterozygosity  $H$  at the mutation-drift equilibrium is

$$H = \frac{4N\mu}{1 + 4N\mu} \quad (2.15)$$

We can use coalescence theory to see where their result comes from. A heterozygote occurs when the two alleles in a random individual differ in sequence. The expected time back to the common ancestor for two randomly-chosen chromosomes is given by Geometric( $1/2N$ ), which has an expected value of  $2N$  generations. Hence, if mutations follow a Poisson distribution with a (per copy, per generation) mutation rate  $\mu$ , an approximation for the expected number of mutations is  $2 \cdot 2N \cdot \mu$ . The “extra” 2 follows since the expected number of mutation from one allele back to the MRCA is  $2N\mu$ , and likewise the expected number of mutations from allele two back to the MRCA is also  $2N\mu$ . Hence, if  $4N\mu > 1$ , we expect most individuals to be heterozygotes (high levels of polymorphism), while if  $4N\mu < 1$ , most will be homozygotes (low polymorphism).

The second class of mutational models that is currently popular are the **stepwise mutational models** for the change in microsatellites. Recall that microsatellites (or STRs) are scored by the number of repeats of a basic sequence (i.e. an ACACAC is three repeats of the AC unit). When a mutation occurs, the repeat number changes, typically by plus or minus one. Hence, two sequences with (say) 10 repeats could be the same sequence which has not mutated or could be sequence which have converged by mutation (i.e. a nine could mutate to a 10 and an 11 could mutate to a 10). Hence, **identity in state** (the sequences being identical) under the stepwise model does not imply **identity by descent**. In the infinite alleles model, identity in state does imply identity by descent, as no two alleles have the same state unless they have a common ancestor and have suffered no mutation. The basic symmetric single-step mutation model has the following structure: if the current number of repeats is  $i$ , then with probability  $\mu$  the allele remains in state  $i$  in the next generation. Otherwise with probability  $\mu/2$  it mutates to state  $i + 1$  or with probability  $\mu/2$  to state  $i - 1$ . The analysis of even this apparently simple model is rather involved, eventually requiring the use of Type II Bessel Functions.

### Forces that Change Allele Frequencies: Selection

The final evolutionary force we will consider is natural selection, wherein not all genotypes leave the same expected number of offspring. Such differences in **fitness** result in some alleles being lost, others fixed. Let  $W_{ij}$  denote the fitness of genotype  $G_{ij}$ , which is the expected number of offspring that  $G_{ij}$  leave. To see the effects of selection, consider the simple case of one locus with two alleles,  $A$  and  $a$ . Assume the genotype frequencies are in Hardy-Weinberg before selection (as occurs with random mating). Following selection, some of the genotypes leave more offspring than others,

Genotypes	$AA$	$Aa$	$aa$
Frequency before selection	$p^2$	$2p(1-p)$	$(1-p)^2$
Fitness	$W_{AA}$	$W_{Aa}$	$W_{aa}$
Frequency after selection	$p^2 W_{AA} / \bar{W}$	$2p(1-p) W_{Aa} / \bar{W}$	$(1-p)^2 W_{aa} / \bar{W}$

where

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

$\bar{W} = E[W_{ij}]$  is the **mean population fitness**, the average fitness of a randomly-chosen individual. Hence, if  $W_{ij} > \bar{W}$ , then the genotype  $G_{ij}$ , on average, leaves more offspring than a randomly-chosen individual ( $\bar{W}$ ). Hence, the weighting  $W_{ij} / \bar{W}$  is the contribution following selection. To obtain the allele frequency  $p'$  following selection, since  $\text{Freq}(A) = \text{freq}(AA) + (1/2)\text{freq}(Aa)$ ,

$$p' = \frac{p^2 W_{AA} + p(1-p) W_{Aa}}{\bar{W}} = p \frac{p W_{AA} + (1-p) W_{Aa}}{\bar{W}} \quad (2.16)$$

The rankings of the fitnesses for the genotypes determine the ultimate fate of an allele. If  $W_{XX} \geq W_{Xx} > W_{xx}$ , then allele  $X$  fixed and allele  $x$  is lost. If  $W_{Xx} > W_X, W_{xx}$  then we have **overdominance** and selection maintains both alleles  $X$  and  $x$ .

A more general expression when there are  $n$  alleles at a locus is

$$p'_i = p_i \frac{W_i}{\bar{W}}, \quad W_i = \sum_{j=1}^n p_j W_{ij}, \quad \bar{W} = \sum_{i=1}^n p_i W_i \quad (2.17)$$

Here,  $W_i$  is the **marginal fitness** of allele  $i$ , the mean fitness of a random individual carrying a copy of allele  $i$ . If  $W_i > \bar{W}$  (A random individual carrying  $i$  has a higher fitness than a random individual), and the frequency of allele  $i$  increases. If  $W_i < \bar{W}$ , allele  $i$  decreases. If  $W_i = \bar{W}$ , the frequency of  $i$  does not change. At an equilibrium point, the marginal fitnesses for all segregating alleles are equal, i.e.  $W_i = \bar{W}$  for all  $i$ .

### Interaction of Selection and Drift

Finally, consider the interactions between drift and selection. A classic result, due to Kimura (1959), is that if the genotypes  $AA : Aa : aa$  have **additive fitnesses**,  $1 + 2s : 1 + s : 1$ , then the probability  $U(p)$  that allele  $A$  is fixed given it starts at frequency  $p$  is

$$U(p) = \frac{1 - \exp(-4Nsp)}{1 - \exp(-4Ns)} \quad (2.18)$$

Note that if  $s > 0$ , we expect (in an infinite population) that  $A$  is fixed by selection, while if  $s < 0$ ,  $A$  is lost. However, when the population size is finite, if selection is sufficiently weak relative to drift, the allele can behave as if it essentially neutral. In particular, if  $4N|s| \ll 1$ ,  $U(p) = p$  and thus the allele behaves as if it essentially neutral. Conversely, if  $4N|s| \gg 1$ , selection dominates, with  $A$  having a very high probability of fixation when  $4Ns \gg 1$  and essentially a zero probability of fixation when  $4Ns \ll -1$ .

Finally, a most interesting case is when a highly-favored allele ( $4Ns \gg 1$ ) enters the population as a single copy  $p = 1/(2N)$ . Here  $U = 2s$ . Note that this is independent of the actual population size, so that even in a very large population, a favored allele introduced as a single copy still has a small probability of fixation. If  $s = 0.1$ , a 10% advantage (which is huge in evolutionary terms), the fixation probability for a single copy is only 20%.

## Lecture 2 Problems

1. Suppose loci  $A$  and  $B$  are linked, with  $c = 0.25$ . Further, suppose  $\text{freq}(AB) = 0.1$ ,  $\text{freq}(A) = 0.5$  and  $\text{freq}(B) = 0.5$ . Assume a random mating population.
  - a. Under Hardy-Weinberg, what is the frequency of an  $AA$  homozygote? A  $BB$  homozygote?
  - b. Assuming gametes combine at random, what is the expected frequency of an  $AABB$  individual assuming the above gamete frequencies.
  - c. What is the initial disequilibrium for the  $AB$  gamete,  $D_{AB}$ ?
  - d. After four generations of recombination, what is the disequilibrium,  $D_{AB}(4)$ ? What is  $\text{freq}(AB)$ ? What is  $\text{freq}(AABB)$ ?
  
2. Consider a locus with four alleles with the following allele frequencies and marginal fitnesses (for these frequencies)
 

Allele	1	2	3	4
Frequency	0.1	0.2	0.3	0.4
$W_i$	1.1	0.9	2.0	0.8

  - a. Compute  $\bar{W}$ .
  - b. What is the frequency of allele 3 after selection?
  - c. What is the frequency of allele 4 after selection?
  
3. For populations of size 50 and 500, compute the probabilities that two randomly-chosen alleles have a most recent common ancestor of less than 50, 500, and 2000 generations.

## Solutions to Lecture 2 Problems

1.

a.  $0.5^2 = 0.25$  for both homozygotes. Hence, one might expect  $\text{freq}(AABB) = 0.25^2 = 0.0625$

b.  $\text{freq}(AABB) = \text{freq}(AB)^2 = 0.1^2 = 0.01$

c.  $D_{AB}(0) = \text{freq}(AB) - \text{freq}(A) \cdot \text{freq}(B) = 0.1 - 0.5 \cdot 0.5 = -0.15$

d.  $D_{AB}(4) = (1 - c)^4 D_{AB}(0) = -0.15(1 - .25)^4 = -0.047,$

$\text{freq}(AB)(4) = \text{freq}(A)\text{freq}(B) + D_{AB}(4) = 0.20.$

$\text{freq}(AABB) = \text{freq}(AB)(4) \cdot \text{freq}(AB)(4) = 0.04$

2.

a.  $\bar{W} = 0.1 \cdot 1.1 + 0.2 \cdot 0.9 + 0.3 \cdot 2.0 + 0.4 \cdot 0.8 = 1.21$

b.  $p'_3 = 0.3 \cdot (2.0/1.21) = 0.496$

b.  $p'_4 = 0.4 \cdot (0.8/1.21) = 0.264$

2.  $\Pr(\text{MCRA} < \tau) = 1 - (1 - 1/[2N])^\tau$

N	Pr(< 50)	Pr(< 500)	Pr(< 200)
50	0.395	0.993	1.000
500	0.049	0.394	0.865