4.2. MUTATION AND SELECTION

4.2.i. Introduction

As we saw in **Section 1.3.iii**, mutation is an ever-present but usually very weak force, causing changes in DNA sequences to occur at a rate of the order of 10^{-9} to 10^{-8} per nucleotide site per generation in higher eukaryotes. This means that, even if selection has established the fittest possible sequence of a gene under the prevailing environmental circumstances, there will be some rate of occurrence of mutations to less fit variants. For example, the mean rate of mutation per generation for single-nucleotide substitutions in humans is estimated to be about 2×10^{-8} (see **Section 1.3.iii.b**). A typical human nuclear gene has about 500 codons, corresponding to 1500 nucleotides, and about 70% of nucleotide changes lead to a change in the amino acid sequence (Kryukov et al. 2007). This implies a rate of mutation to a changed form of the protein of $0.70 \times 1500 \times 2.0 \times 10^{-8} = 2.1 \times 10^{-5}$ per gene per generation. There is an additional, but substantially smaller, contribution from nonsense mutations, indel mutations, intron splice-site mutations and transposable elements, all of which severely disrupt protein structure (Kryukov et al. 2007).

As we saw in **Section 1.2.iv.a** and will discuss further in **Section 6.4.iv**, there is evidence that most of these amino acid changes are deleterious, but a majority have very small effects on fitness. Some, however, greatly reduce the fitness of their heterozygous carriers, causing Mendelian genetic diseases. These contribute a significant fraction of childhood diseases and mortality; the frequency of single-gene disorders among pediatric hospital admissions in North America is between 4% and 7% (Gelehrter et al. 1998). With at least 20,000 protein-coding genes in the human genome (http://www.ornl.gov/sci/techresources/Human_Genome), each new zygote must contain more than $2 \times 20{,}000 \times 2.1 \times 10^{-5} = 0.84$ new nonsynonymous mutations, of which at least 70% are probably sufficiently deleterious that selection is virtually certain to eliminate them from the population (see **Section 6.2.iii.a**), i.e., they are subject to *purifying selection*. This yields a net nonsynonymous deleterious mutation rate of more than 0.58 mutations per new zygote per generation.

There is also evidence for widespread purifying selection against mutations in noncoding sequences in the human genome (Keightley et al. 2005; ENCODE Project Consortium 2007). Since noncoding sequences are far more abundant in mammalian genomes than coding sequences, as many as 5% of the nucleotides in the genome may be subject to purifying selection (ENCODE Project Consortium 2007). The net *deleterious mutation rate* in humans may be as high as six mutations per individual per generation. This constant production of deleterious variants by mutation is of great importance for both medical and evolutionary genetics, and requires careful study by the approach initiated

by Haldane (1927b, 1937). This involves modeling the balance between the input of new, deleterious mutations and their elimination by selection. We will first consider this at the level of a single nucleotide site, then whole genes, and finally the whole genome. The extent to which deleterious mutations contribute to within-population variability in DNA sequences is discussed in **Section 6.4.iv**.

4.2.ii. Mutation–selection equilibrium

4.2.ii.a. *Autosomes with random mating*

How much variation can mutation maintain? We begin by studying the simple case of mutation at a single autosomal site in a randomly mating, diploid species. Assume a wild-type variant A_1 that mutates at rate u per generation to a deleterious alternative A_2. We also assume that A_1 is sufficiently common in the population that mutation in the reverse direction can be ignored (i.e., selection is so strong that the frequency of A_2, q, is very low). We saw in **Section 1.3.iii.c** that the change in the frequency of A_2 due to mutation is $\Delta q_{mu} = u(1 - q) - q$. If q and u are both small, their product can be neglected, and we have:

$$\Delta q_{mu} \approx u \tag{4.1}$$

Assume that the population is randomly mating, and that the fitnesses of A_1A_2 and A_2A_2 are $1 - hs$ and $1 - s$, respectively (using the notation introduced in **Section 2.1.ii.a**). Using the results of **Problem 3.3**, we find that the change due to selection is approximately:

$$\Delta q_s \approx -hs\,q \tag{4.2}$$

(compare this with **Equation B4.1.2** of **Section 4.1.i** for selection with migration.) At equilibrium, the changes due to mutation and selection are equal and opposite, and q^*, the equilibrium frequency of A_2, is thus:

$$q^* \approx \frac{u}{hs} \tag{4.3}$$

(compare this with **Equation B4.1.5.**) It is easily verified that this is a stable equilibrium (**Problem 4.2**). A similar result applies to haploid organisms, replacing hs with s.

A useful way of understanding this result is to consider the situation when the deleterious mutant allele A_2 is present at a low frequency, so that the frequency of heterozygotes is approximately $2q$. Most selective elimination then involves heterozygotes, giving an average reduction in frequency of $hs/2$ per heterozygous individual per generation. The rate of elimination of A_2 by selection is thus approximately $2qhs/2 = qhs$. This must equal the rate per genera-

tion at which new mutations enter the population, which is approximately u when q is small.

The case of completely recessive mutations ($h = 0$) is slightly more complex (**Problem 4.3**). The equilibrium is now given by:

$$q^* \approx \sqrt{\frac{u}{s}} \tag{4.4}$$

For a given mutation rate and selection coefficient, the equilibrium frequency of a completely recessive variant is much higher than that for a corresponding variant with intermediate or complete dominance, reflecting the ineffectiveness of selection against rare recessive alleles (**Section 3.1.iii.c**).

Recall from **Section 2.2.iv** that most mutations with large effects are recessive with respect to their phenotypic effects. However, it is likely that mutations are rarely fully recessive with respect to their fitness effects. Mutations with recessive lethal effects have an estimated average heterozygous fitness effect of 2–3% in experiments on *Drosophila* (Crow 1993; García-Dorado and Caballero 2000), and mutations with small homozygous fitness effects (of the order of a few per cent) have much higher h values, in the range 0.1–0.4. While direct evidence is lacking for other organisms, there seems to be no reason to assume any radical difference from *Drosophila*. **Equation 4.3** is therefore probably more appropriate than **Equation 4.4** for most cases, even for apparently recessive mutations. The question of *why* mutations are generally or partially recessive with respect to their effects on phenotypes and fitness is one of the oldest in evolutionary genetics (Fisher 1928), and we discuss it in **Section 4.4.v** below.

4.2.ii.b. *Other cases*

Other biologically important situations can be modeled similarly to the autosomal case just explained. With X linkage and random mating, we need to include the possibility that the selection coefficients may differ between the sexes. We can denote the selection coefficients in males and females as s_f and s_m, and use **Equation 3.6** in **Section 3.1.v.b** (with the appropriate change of sign). The equilibrium allele frequency (weighting frequencies in eggs and sperm by 2/3 and 1/3, respectively) is:

$$q^* \approx \frac{3u}{\left(2hs_f + s_m\right)} \tag{4.5}$$

Under inbreeding with an inbreeding coefficient f, and assuming autosomal inheritance, **Equation 3.10** in **Section 3.1.v.c** gives:

$$q^* \approx \frac{u}{\left[h\left(1 - f\right) + f\right]s} \tag{4.6}$$

As would be expected intuitively, exposing mutations to selection in the hemizygous state (sex linkage) or in the homozygous state (inbreeding) greatly lowers their equilibrium frequencies, provided that the deleterious effects are not completely dominant.

4.2.ii.c. *Mutation and selection at multiple sites*

The equations just derived relate to the situation at a single nucleotide site, such as a mutation that alters an amino acid in a protein sequence. If we assume that selection at each site acts independently of selection at other sites, the equilibrium frequency of mutations at each site can be calculated from the results for each individual site considered in isolation. If we consider a gene whose coding sequence includes m nonsynonymous sites, the overall frequency of mutations in the gene is then simply the sum of the contributions from each site, given by **Equation 4.3** for the case of autosomal inheritance and random mating. If the strength of selection varies across sites, the equilibrium frequency q_g^* of deleterious nonsynonymous mutations present in the gene as a whole is:

$$q_g^* \approx \sum_i \frac{u_i}{h_i s_i} \tag{4.7a}$$

where the subscript i indicates a particular site and runs from 1 to m.

When the strength of selection varies across sites, but the selection coefficients and mutation rates at each site are independent of each other, this simplifies to:

$$q_g^* \approx \frac{u_g}{(hs)_H} \tag{4.7b}$$

where u_g is the mutation rate for the whole gene (i.e., the sum of the mutation rates at each site), and the subscript H denotes the *harmonic mean*, the reciprocal of the mean of the reciprocals.

This can be extended to the whole genome, by summing over all sites capable of mutating to deleterious variants. The mean number, $\overline{n}$, of deleterious mutations carried by a haploid genome is given by:

$$\overline{n} \approx \frac{U}{(hs)_H} \tag{4.7c}$$

where U is the total mutation rate per haploid genome to deleterious alleles for autosomal sites (a different expression must be used for X-linked sites, using the appropriate extension of **Equation 4.5**). If mutations are rare at individual sites, the number of mutations per diploid individual is $2\overline{n}$.

In order to estimate the value of $2\overline{n}$ for a species, we would need to know the abundance of mutations within the genome, and the proportion of those

that are sufficiently strongly selected that their frequencies can be well predicted from the infinite-population formulae used here. We discuss this problem in detail in **Section 6.4.iv**, where we show how to combine the theory of selection and genetic drift with genomic data on the frequencies of amino acid variants in populations to obtain estimates of the selection coefficients against deleterious amino acid variants. The results imply values of about 700 deleterious amino acid mutations per individual in human populations, and around 10-fold more in *Drosophila*, mostly with very small selection coefficients (< 0.001). Deleterious mutations in noncoding sequences probably contribute even more than this. Since most of these mutations are rare, a different set of mutations is present in each individual in the population. There is thus an enormous number of slightly deleterious mutations segregating in a natural population of an outbreeding organism.

If the frequencies of variants at one site are independent of those at other sites (**Section 1.2.v.b**), there will be a Poisson distribution of the number of mutations per haploid genome (**Appendix A2.v.d**). This is because we can treat the number of mutations carried in a given haploid genome as a random draw from a binomial distribution with mean $\overline{n}$; if the total number of sites m is large compared with $\overline{n}$, the Poisson approximation to the binomial distribution will apply. The corresponding frequency distribution for diploid individuals with random mating is Poisson with mean $2\overline{n}$. We will use this result frequently later in the book.

An important implication of these results is that individuals completely free of deleterious mutations are very unlikely to be found in natural populations. With $2\overline{n} = 700$, the value of the zero term of the Poisson distribution is $\exp(-700) = 10^{-304}$, so that a population size of 10^{304} would be required for just one mutation-free individual to be produced. Even with $2\overline{n}$ as low as 100, a population size of 10^{43} would be needed.

4.2.ii.d. *Estimating mutation rates from equilibrium frequencies*

If estimates of selection coefficients are available, and the population is at equilibrium, **Equations 4.3**, **4.5**, or **4.7b** can be used to estimate mutation rates (**Problem 4.4**). Such estimates are often called *indirect estimates*. In practice, this method is usually limited to mutations with conspicuous phenotypic effects, which are likely to be only a subset of all mutations that affect a protein sequence, so it substantially underestimates overall mutation rates. The first such estimate was for hemophilia (Haldane 1935), and this approach still forms the basis for many of the published estimates of human mutation rates for genes causing diseases. It is especially useful for recessive X-linked disorders, since these are probably eliminated largely through their effects on male carriers. The results indicate that the net rate of mutation per gene per generation for such mutations is typically around 10^{-5} (Haldane 1949c; Vogel and Motul-

sky 1997, Chapter 9). This is in rough agreement with what is expected from the estimate of the human mutation rate from DNA sequence data (**Section 4.2.i** above).

4.3. GENETIC LOAD

4.3.i. Introduction

What is the overall effect of deleterious mutations on the mean fitness of the population? This question was first asked by Haldane (1937) and later explored in a famous paper by Muller (1950), who coined the term "load" for the reduction in fitness caused by the presence of deleterious mutations in a population. Genetic load was defined by Crow (1958) as the proportional reduction in mean fitness of the population below that for the genotype with the highest possible fitness, the *optimal genotype*:

$$L = \frac{\left(w_M - \bar{w}\right)}{w_M} \tag{4.8}$$

where w_M is the fitness of the optimal genotype, and $\bar{w}$ is the population mean fitness (note that the optimal genotype, such as an entirely mutation-free individual, may be so rare that it is never observed in the population).

The load effectively measures the fraction of the population that fails to survive or reproduce because of selective differences among its constituent genotypes; this is sometimes referred to as the amount of *selective death*. There must also always be environmental sources of death or reproductive failure, even for the optimal genotype. We will now derive expressions for the genetic load for several different ways in which variation in fitness can arise.

4.3.ii. Mutational load

4.3.ii.a. *Autosomal inheritance with random mating*

The load under mutation at a single nucleotide site in a large, randomly mating population can be found as follows (Haldane 1937). Mutations with heterozygous effects on fitness are mostly eliminated from the population as heterozygotes, as we saw in **Section 4.2.ii.a** above. Since the frequency of heterozygotes is approximately $2q$, the reduction in fitness to the population, measured relative to the fitness of wild-type homozygotes, is $L \approx 2qhs$. If the population is at equilibrium, we can use the value of q from **Equation 4.3** and obtain:

$$L \approx 2u \tag{4.9}$$

Similarly, for the case of completely recessive mutations (eliminated exclusively as homozygotes), we have (**Problem 4.5.i**):

$$L \approx u \tag{4.10}$$

The difference between the two cases reflects the fact that a single "selective death" eliminates two mutations in the case of recessivity (when selection acts only on mutant homozygotes), but only one when heterozygotes are the main source of selective elimination, so that selection is twice as efficient with recessivity. The remarkable result that the load is independent of the selection coefficient can be understood as follows. Although weakly selected mutations cause fewer selective deaths among their carriers than strongly selected mutations, they rise to higher equilibrium frequencies. Provided that these frequencies are still sufficiently low that the relevant approximations are valid, the two effects exactly cancel out (Haldane 1937).

4.3.ii.b. *Other cases*

Equation 4.10 applies in the case of haploids. With X-linked inheritance, and equal selection on the two sexes ($s_f = s_m = s$), the load (averaging over males and females) is $3u/2$, exactly intermediate between the two autosomal cases (**Problem 4.5.ii**). If selection acts only on one sex, the load for this sex for nonrecessive mutations is $3u$, and the overall load is half of this value (**Problem 4.5.ii**).

In an inbreeding population with autosomal inheritance, **Equation 4.6** of **Section 4.3.ii.b** gives:

$$L \approx \frac{u\left[2(1-f)+f\right]}{\left[h(1-f)+f\right]} \tag{4.11}$$

This shows that, for nonrecessive mutations, the equilibrium load decreases as the inbreeding coefficient increases, reflecting the more efficient elimination of mutations in the homozygous state. The lower the dominance coefficient h, the faster the rate of decline in the load; if h is small but nonzero, even a relatively small amount of inbreeding greatly reduces the load. If the inbreeding coefficient is close to 1, the load approaches the mutation rate, independently of the dominance coefficient (see **Section 4.3.ii.c** below).

4.3.ii.c. *Multiple sites*

The theory just outlined shows that, in all cases, the mutational load at a site is of the same order as the mutation rate. Since mutation rates per nucleotide site are extremely small in organisms other than RNA viruses and some mitochondrial genomes (**Section 1.3.iii.b**), it might at first sight seem that the mutational load is negligible. However, this ignores the fact that a genome usually includes a very large number of sites capable of producing deleterious variants (**Section**

4.2.i). To assess the genetic load, we thus need to determine its genome-wide value. This can easily be done by applying the assumption already used above, that mutations at different sites affect fitness independently of each other. In addition, we assume that they are distributed independently of each other in the population.

The first assumption is equivalent to the assumption of *multiplicative fitnesses*, as explained in **Box 4.4**. **Box 4.4** also shows how to obtain the mean fitness and the total load on the population. For nonrecessive autosomal muta-

Box 4.4 GENETIC LOAD DUE TO VARIATION AT MULTIPLE SITES

If mutations at different sites reduce the probability of survival or reproduction independently by an amount s_i for the i^{th} site, the net fitness of an individual, relative to the fitness of a mutation-free individual, is the product of $(1 - s_i)$. This follows from the rule that the probability of an event caused by the co-occurrence of a set of independent events is given by the product of the probabilities of each event (**Appendix A2.ii**). More intuitively, the overall chance of successful survival or reproduction is like a race with many hurdles. The net probability of getting to the end of the course is the product of the chances of *not* falling at each successive hurdle encountered in the race.

The same principle can be applied to the mean fitness of the population, replacing s_i by the load L_i for the i^{th} site:

$$\bar{w} = (1 - L_1)(1 - L_2)\cdots(1 - L_m) \qquad \text{(B4.4.1a)}$$

so that:

$$\ln \bar{w} = \sum_i \ln(1 - L_i) \qquad \text{(B4.4.1b)}$$

Since the individual loads are small, $\ln(1 - L_i)$ can be approximated by $-L_i$ (**Appendix A1.ii.c**), so that $\ln \bar{w}$ is approximately $-\Sigma_i L_i$. We therefore have:

$$\bar{w} \approx \exp -\sum_i L_i \qquad \text{(B4.4.2)}$$

The total load, L, is simply $1 - \bar{w}$.

tions in a randomly mating population, we can substitute from **Equation 4.9** into **Equation B4.4.2** to obtain the load:

$$L \approx 1 - \exp(-2U) \tag{4.12a}$$

where $2U$ is the mean number of new deleterious mutations in a zygote (the genomic diploid deleterious mutation rate).

This approach can be extended to inbreeding populations by using **Equation 4.11** (Lande and Schemske 1985; Charlesworth and Charlesworth 1998):

$$L \approx 1 - \exp\left\{-\frac{U\left[2(1-f)+f\right]}{\left[h(1-f)+f\right]}\right\} \tag{4.12b}$$

Figure 4.7 displays the dependence of the load on the dominance coefficient and inbreeding coefficient for a genome-wide deleterious mutation rate of 0.1 per haploid genome.

These theoretical results raise the question of the magnitude of the mutational load in nature. As we saw in **Section 4.2.i**, $2U$ for humans is at least 0.58 per generation. **Equation 4.12a** then implies that the mean fitness of the population is less than 56% of that of a mutation-free individual (i.e., the load is greater than 44%). The mutational load places some limits on the size of the functional portion of the genome that an organism can sustain. The mean fit-

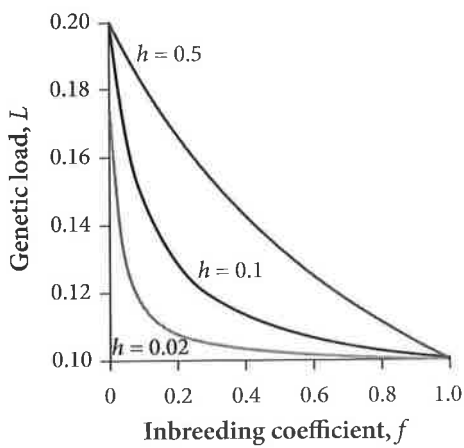

FIGURE 4.7 The predicted equilibrium genetic load, L, due to mutation in populations with different inbreeding coefficients, f, using **Equation 4.12b**. The mutation rate, U, is equal to 0.1, and results for three different values of the dominance coefficient, h, are shown. The loads for a randomly mating population and a completely inbred population are close to 0.2 and 0.1, respectively, independent of h (unless h is very close to 0). The rate of decline in L as f increases from 0 is greatest for mutations with the smallest h values.

ness from **Equation 4.12a** is an upper limit, since it considers only mutational load (and ignores both environmental sources of mortality or sterility, and also contributions from other kinds of genetic load; see below).

A population with two sexes must produce a mean of two offspring per individual for the population size to be stable. The absolute mean fitness of the population (the average number of zygotes contributed to the next generation by a new zygote) must therefore equal or exceed 2 if the population is to remain in existence. A load of 44% then implies that the absolute mean fitness for mutation-free individuals must be 3.6. If mutations in noncoding sequences are taken into account (**Section 4.2.i**), the load is much higher.

With the human reproductive capacity seen in hunter–gatherer societies, which represents the state under which our species evolved, this level of successful offspring production is only just sustainable. A value of 8 for the lifetime expected number of offspring for women at the start of their reproductive life is at the high end of the range (Howell 1976); this must be discounted by early-life mortality. Of course, strictly multiplicative fitness effects are unlikely, especially as selection may often be competitive, so that the proportion of successful individuals depends on the availability of the resources for which they are competing, rather than on their genes. Non-multiplicativity of the fitness effects of different mutations may considerably mitigate the problem of genetic load, as we discuss in **Section 10.2.iv.d**. Nevertheless, the long-term survival of populations of organisms with a functional genome much larger than the human genome seems unlikely, unless their mutation rate per nucleotide is much lower than the estimates above, or their reproductive capacity is much higher than for humans.

4.3.iii. Segregational load

Variability maintained by balancing selection also creates a genetic load, the *segregational load*. Consider the case of heterozygote advantage (**Section 2.1.ii.c**). The formula for the equilibrium variant frequencies (**Equation 2.3**) leads to the following expression for the load (**Problem 4.6**):

$$L = \frac{st}{(s+t)} \tag{4.13}$$

In contrast to the mutational load, the segregational load is of the same order of magnitude as the selection coefficients against the homozygotes (Morton et al. 1956). Unless selection is very weak, the load increases very rapidly with the number of polymorphisms maintained by balancing selection, if multiplicative fitnesses are assumed (**Problem 4.6.ii–iii**). It follows that a genome can have only a few sites subject to the intensity of balancing selection estimated for the malaria resistance polymorphisms discussed in **Section 2.1.ii.d**.