38-39. Genetic drift
[There is correction in the answer to problem 38.2.]
Reading pp. 768-9

Allele frequencies change after random mating in finite populations

The Hardy-Weinberg (HW) formula tells you that, in an infinitely large population, allele frequencies do not change after one generation of random mating. Real population are finite. As a result, allele frequencies change slightly because of random mating alone. This slight change is called **genetic drift**. In large populations, the change per generation is very small, but genetic drift is important in all populations because its effects accumulate over time and lead to the ultimate **loss** (p=0) or **fixation** (p=1) of alleles.

To understand genetic drift, you have to imagine you have a very large number of identical populations numbered 1 to *n*. Each population is initially the same: it has *N* diploid individuals and an allele A in frequency *p*. Call the other allele a. Its frequency is *q*=1–*p*. Then each population undergoes random mating. Call the frequency in population *i* after random mating p_i. The average frequency in the all the populations is

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i .$$

If A is **neutral**, meaning that it does not affect viability and reproduction, the average frequency will be *p*. In other words, **on average**, genetic drift does not change the frequency of neutral alleles. The allele frequency in each population will differ from the average, however, and the magnitude of the difference depends on *N*. The way to quantify the difference is with the **variance in frequency**, defined to be

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (p_i - p)^2 .$$

The theory of genetic drift tells us that the variance depends on the number of individuals in each population.

$$\sigma^2 = \frac{p(1-p)}{2N}.$$

Founder effect

Many populations are founded by only a few individuals. When this happens, allele frequencies in the newly founded population can differ substantially from those in the population from which the founders came. This type of genetic drift is called a founder effect. In evolutionary biology, a founder effect can create populations that differ from the source population and lead to rapid evolution of phenotypic differences. In human genetics, founder events can result in the relatively high frequency of alleles that are rare

elsewhere. The population in which HD was first mapped had the causative allele in almost 25% frequency because of a founder effect.

In a population started from a single fertilized female, the founder effect is very important. For a locus with two alleles, the chance she has one of the three genotypes is given by the HW frequencies (assuming random mating). The allele frequency in the newly founded population may differ substantially from the source population.

Mutation and genetic drift maintain genetic variability

If you continue with the hypothetical experiment, eventually each population will be **fixed for either A or a**. That will take roughly **4*N* generations**. In a particular population, you will not know whether A or a will be fixed. But in a large group of populations considered together, a fraction p will be fixed for A and a fraction q will be fixed for a. The **probability of fixation** of A is equal to its initial frequency if it is neutral.

Genetic drift alone will cause genetic variability to be lost. If there were no mutation, every locus would be fixed for one of the alleles it initially carries. Mutation will restore variability, however. A population will reach an equilibrium between variability introduced by mutation and variability lost to drift, a **mutation-drift balance**. The amount of genetic variability maintained by a mutation-drift balance can be calculated and can provide useful information about the population.

For DNA sequence data, there are two convenient ways to measure genetic variability. Suppose you have a sample of 3 sequences at 10 sites from a population:

I : ATGGCGATGG
II: ATTGCCATGG
III: ATGGCCATAT

Four of the ten sites are polymorphic (3, 6, 9,10), which means that there are four **segregating sites**, often denoted by S, which is one way to characterize variability in a set of sequences.

Another way to characterize the variability is by the **average pairwise difference** in sequence, usually denoted by π. In this data set π= (2+3+3)/3=2.3333.

Both S and π depend on the mutation rate and population size, but they do so in a slightly different way. The theory of genetic drift tells us that in a diploid population of **constant size**, N, and a **mutation rate per site** to neutral alleles μ, the expected value of π is $4N\mu L$, where L is the number of sites sequenced. If μ is known, you can estimate the population size from this formula, $N=\pi/(4\mu L)$. If $\mu=2\text{x}10^{-9}$, which is roughly the neutral mutation rate for autosomal loci in animals, our data set implies $N_\pi=2.333/(4\text{x}2\text{x}10^{-9}\text{x}10)=3\text{x}10^7$, where I have added the π to indicate it is the estimate of N made using π.

The theory of genetic drift also tells us that S depends on population size: $S = 4N\mu L\sum_{i=1}^{n-1}1/i$, where n is the number of sequences compared. The formulas for π and S are almost the same. They both depend on $4N\mu L$, but the formula for S depends on the sample size n while the formula for π does not. You can estimate N using S as well, $N = S/(4\mu L\sum_{i=1}^{n-1}1/i)$. For the sample data set, S=4 and N is estimated to be N_S=4/(4x2x 10^{-9}x10x1.5)=3.3x10^7, where the S indicates the estimate made using S.

In the sample data set, $N_\pi \approx N_S$. When that is the case, there is support for the hypothesis that the model on which these two estimates are based actually does describe the population from which the samples were taken. When $N_\pi < N_S$, you have evidence that the model does not fit. Consider the following (contrived) data set.

I-IX : ATGGCGATGG
X: TGCGTAACTT

In this case, S=8 and N_S=8/(4x2x10^{-9}x10x2.83)=3.353x10^7. π=72/45=1.6 and N_π=1.6/(4x2x10^{-9}x10)=2x10^7. Therefore, $N_\pi < N_S$. In this data set all the segregating sites are singletons. In a population of constant size, you would not expect so many singletons. But in a population that has grown in size rapidly, there would be more singletons and you would expect to find $N_\pi < N_S$.

This provides a way to test whether the data fit the assumption of constant population size or fit the assumption of past growth. Tajima's D is a statistic that is proportional to $N_\pi - N_S$, so a negative value indicates past growth. Figure 1c from Garrigan and Hammer (2006) shows the distribution of D values for human populations.

The effects of genetic drift accumulate and lead to fixation or loss of alleles

If you imagine that this group of populations continues to undergo random mating for a large number of generations, then eventually A will be lost or fixed. If A is neutral**,** the probability that it will be fixed is equal to the initial frequency, p. Roughly speaking, fixation or loss will occur within $4N$ generations.

Garrigan D, Hammer MF (2006) Reconstructing human origins in the genomic era. Nature Reviews Genetics 7:669-680. http://www.nature.com/nrg/journal/v7/n9/abs/nrg1941.html

Additional problems

36.1 Suppose the frequency of allele A in a large population is p and the genotypes are at the HW frequencies. Suppose one self-fertile individual is moved from this population to found a new population. What is the probability that the new population is fixed for A or is fixed for the other allele.

Ans. The new population will be fixed for A if the founder happens to be AA, which will occur with probability p^2. The founder will be aa with probability q^2.

36.2 Suppose the new population is founded by two individuals chosen at random.

a. What is the probability that the new population will have at least one copy of A if p=0.001?

Ans. There will be at least one copy of A if both founders are not aa, which will occur with probability $1–q^4$. With p=0.001, that probability is $1–0.999^4 \approx 1–0.996=0.004$.

b. If A is present in the new population, what is its minimum frequency initially?

Ans. 0.25.

I: TGCGTAACTT
II: TACGTAACTT
III: TGCATAACAT
IV: TCCGTAACTT
V: TGCGTAACCT

36.3 In the above data set compute N_S and N_π.

Ans. Work out in section

The rate of substitution of neutral alleles is the mutation rate

Suppose there is a mutation at a site that changes the nucleotide present. A new single nucleotide polymorphism (SNP) is created at that site. The new nucleotide is a new allele, and its frequency immediately after it arises by mutation is

$$p = \frac{1}{2N}.$$

If it is neutral, the probability that it will ultimately be fixed because of genetic drift is $1/(2N)$.

The probability that a new nucleotide appears at a site is μ, the mutation rate per site. Because there are $2N$ chromosomes, the probability that there is a mutation on any chromosome at a particular site is $2N\mu$. If a mutation occurs, the probability that the new nucleotide is fixed because of genetic drift is $1/(2N)$. Therefore, the overall probability that a new nucleotide at a site is created by mutation and is fixed by genetic drift is

$$K = 2N\mu \times \frac{1}{2N} = \mu$$

independently of N.

Once fixation occurs, there has been a **substitution** of one nucleotide by another at that site. If you could compare sequences before and after the substitution, you would detect a different nucleotide at that site. By comparing sequences of different species, you can count the number of substitutions that have occurred. Furthermore, if you know when in the past those species had a common ancestor, you can estimate the **substitution rate** and hence the underlying mutation rate for individual nucleotides. The formula is

$$K = \frac{D}{2TL}$$

where D is the number of differences found when comparing L sites, and T is the time separating two species. The 2 is needed because substitutions can occur in both species that are descended from the common ancestor, [Note: this formula is only approximate because it does not take account of the possibility that more than one substitution can occur per site.]

Be careful to **not confuse a mutation with a substitution**. A mutation appears in the offspring of a single individual. Most mutations will be lost and but a few will be fixed. When a mutation has become fixed, there has been a substitution.

One way to estimate the mutation rate is to compare third positions of codons that are 4-fold degenerate, meaning that any of the 4 nucleotides present result in the same amino acid. The third positions of codons for Proline, Valine, Alanine and several other amino

acids are **4-fold degenerate**. Because a mutation of this type does not change the amino acid coded for, it is reasonable to assume that it is neutral. In the gene coding for β-globin, there are 78 4-fold degenerate codons (L=78). When sequences from mice and humans are compared, 30 of these codons are found to differ at the third position (D=30). The fossil record indicates that the most recent common ancestor of humans and rodents was present about 80 million years ago ($T=8\text{x}10^7$). Using the formula, you find $K=2.42\text{x}10^{-9}$ per year. That is an estimate of the mutation per site per year. That estimate is very close the to currently accepted average, which is based on the analysis of a much large number of genes and species pairs.

Lower than expected substitution rates indicate constraints on evolution

The

Substitution rates of silent and replacement mutations differ

If the mutation of a nucleotide does not result in a change in the amino acid coded for, it is a **silent mutation**. A substitution of a silent mutation is called a **silent substitution**. The mutation rate μ is estimated from the rate of silent substitutions.

If a mutation in a coding sequence results in a change in amino acid, it is a **replacement mutation**. Because of the genetic code, all mutations at the second-codon position are replacement mutations. When coding sequences of the same gene in different species are compared, the rate of replacement substitution is almost always lower than the silent rate, sometimes much lower. If you found 13 differences at 102 second-codon positions of β-globin in humans and mice, you would estimate the rate of replacement substitution to be $0.8\text{x}10^{-9}$, which is approximately the correct value for that gene. For insulin, the replacement rate is $0.13\text{x}10^{-9}$. For two different histone genes, it is less than 10^{-13}. A lower rate of replacement substitutions indicates that some replacement mutations are not neutral. Instead some of them are so deleterious that they cannot be fixed. The reduction in substitution rate from the neutral rate indicates how sensitive the function of a protein is to changes in the amino acid sequence. Furthermore, detecting protein subunits for which rates are very low can help identify the most important parts of proteins.

Substitution rates in non-coding DNA

flanking sequence

There appears to be strong purifying selection in some parts of the genome not coding for proteins. Numerous (481) segments of 200 bases or more in noncoding regions of the human genome have been found to be perfectly conserved between humans, mice and rats. There were no substitutions detected. These regions have been called **ultraconserved** These observations suggest that that ultraconserved regions have had an important function in mammals that is currently unknown.

There have been functional studies of unltraconserved elements.

Reference: Ultraconserved elements in the human genome 2004 Bejerano et al *Science* 304: 1321-1325 (http://www.sciencemag.org/cgi/content/abstract/304/5675/1321)

Additional problem.

If $\mu=2x10^{-9}$ per site per year, how many substitutions would you expect to see if you compared 200 neutral sites in humans and mice ($T=8x10^7$ years)?

Answer: For neutral sites, $K=\mu$. Therefore, $D=2TL\mu=2x8x10^7x200x2x10^{-9}=32$.