Population Genetics Course

Finite Populations and Molecular Evolution and Variation

(sections 4 and 5 are most relevant to the course; the others are for information only)

1. Change in inbreeding coefficient in a Wright-Fisher population of size *N*

Let the inbreeding coefficient in generation t be F_t. Consider a randomly chosen individual (if mating is at random, this is equivalent to choosing a random pair of alleles at a locus) of this generation. The chance that one of the two alleles at a locus is a copy of a given allele in the preceding generation is $1/2N$, since there are $2N$ distinct alleles at the locus, so the chance that both alleles are derived from this allele is $(1/2N)^2$.

There are $2N$ alleles present in the population in generation t-1, so the chance that the two chosen alleles are copies of the same allele from generation t-1, and hence are *identical by descent*, is $2N \times (1/2N)^2 = 1/2N$.

There is thus a probability $1-1/2N$ that the pair are copies of two different alleles in generation t-1, in which case their probability of identity by descent is F_{t-1}.

We therefore obtain the recurrence relation:

$$F_t = 1/(2N) + (1-1/2N)F_{t-1} \qquad (1)$$

Subtracting both sides from 1, and defining P as $1-F$, we get

$$P_t = 1-F_t = (1-1/2N)\, P_{t-1}$$

so that

$$P_t = (1-1/2N)^t P_0 \approx P_0 \exp -(t/2N) \qquad (2)$$

2. The Wahlund Effect

Suppose that we have a set of populations, each of which is in Hardy-Weinberg equilibrium, but which differ in their allele frequencies at a biallelic locus. The mean and variance of the set of allele frequencies can be written as $E\{p\}$ and σ_p^2, respectively (E denotes the operation of taking a mean, often called an *expectation*).

Under drift in a set of isolated populations that all started with the same initial allele frequency p_0, $E\{p\} = p_0$.

Consider the mean frequency of the A_1A_1 homozygotes; this is the mean of p_i^2, where p_i is the frequency of *A* in the *i*th population. But the *variance* of allele frequency is equal to the mean of $(p_i - E\{p\})^2$, which can be rewritten as

$$\sigma_p^2 = E\{p^2\} - (E\{p\})^2 \qquad (3)$$

where $E\{p^2\}$ is the mean of the p_i^2. (you can check this for yourself!).

Hence, the mean frequency of A_1A_1 is equal to $\sigma_p^2 + (E\{p\})^2$. A similar calculation shows that the mean frequency of A_2A_2 *homozygotes* is equal to $\sigma_p^2+(E\{q\})^2$ (note that $\sigma_q^2= \sigma_p^2$, since $q = 1-p$).

Since genotype frequencies sum to 1, the mean frequency of *heterozygotes*, A_1A_2 is equal to $2\ E\{p\}E\{q\}) - 2\sigma_p^2$.

The mean frequency of heterozygotes over all populations is thus *reduced below* that expected under Hardy-Weinberg equilibrium in a single population with alllele frequency $E\{p\}$, while the homozygote frequencies are *increased*.

This is the *Wahlund Effect.*

3. Relation between *F* and σ_p^2

In the above set of populations, there is a probability $E\{p\}$ that a randomly chosen gene is A_1 in state. By definition, the probability that the other gene at the same locus of the individual from which this gene was drawn is *identical by descent* is *F*, in which case the individual is A_1A_1.

There is a probability 1-*F* that the other gene is *non-identical by descent*, in which case there is a probability $E\{p\}$ that it is A_1. Hence, the probability that a random individual from the set of populations has genotype A_1A_1 (which is equivalent to the above mean frequency of A_1A_1) is

$$F\,E\{p\} + (1-F)\,(E\{p\})^2 = (E\{p\})^2 + F\,E\{p\}E\{q\} \qquad (4)$$

Comparison of this with the earlier formula shows that

$$\sigma_p^2 = F\,E\{p\}E\{q\}$$

i.e. $$F = \sigma_p^2 / (E\{p\}E\{q\}) \qquad (5)$$

This relation is very widely used in describing genetic differences between populations, as we will see in a later lecture.

Importantly, it implies that, under random genetic drift in a Wright-Fisher population, the variance in gene frequency between populations *increases in parallel* with the approach to homozygosity of each population.

4. Fixation probabilities and the rate of molecular evolution

The results derived in **section 1** imply that a population subject to a purely neutral process of pure genetic drift will approach *complete homozygosity* for any alleles that may be segregating initially, with probability one. Since drift cannot alter mean allele frequencies, this means that a population that was initially segregating for an allele present at frequency p_0 will end up either fixed for this allele (with probability p_0) or having lost it (probability 1-p_0), so that the mean allele frequency remains at p_0.

Under neutrality, therefore, an allele with initial frequency p_0 has a *probability of fixation* of p_0.

Another way of looking at this, for the case of a new mutation present as a single copy, is as follows. There are $2N$ potentially distinct alleles at an autosomal locus in a population of N breeding adults. Only one out of these $2N$ alleles will be the ancestor of all alleles at the locus at some time in the future, when full homozygosity has been attained.

The probability that a given allele is the lucky one is thus $1/2N$, under neutrality. The probability of fixation of a new neutral mutation, which is present initially as a single copy, is thus $1/2N$.

If there is a mutation rate u to new neutral variants at a particular locus, the expected number of new mutations that arise in the population per generation is $2Nu$. But the probability that any one of these is ultimately fixed in the population is $1/2N$, so that the net rate per generation at which new mutations arise that ultimately become fixed is

$$K = 2Nu \times (1/2N) = u \qquad (6)$$

In a steady-state situation, the mean number of mutations that become fixed each generation must also equal K, so that K is often referred to as the *rate of substitution*– over a period of T generations, KT mutations are expected to become fixed. This quantity can be related to

data on DNA sequence divergence between species with known divergence dates, on the hypothesis of neutral molecular evolution.

5. The theory of the coalescent process; sequence diversity

Suppose we consider a pair of alleles at a locus sampled from a Wright-Fisher population of size N at a given point in time. From what has been shown in **section 1**, it is evident that there is a probability of $1/2N$ that the two alleles are replicates of a single ancestral allele in the previous generation. If this happens, the two alleles are said to *coalesce.* In general, the probability that two alleles coalesce t generations back in time is:

$$Pr(t) = (1-1/2N)^{t-1}(1/2N) \qquad (7)$$

since this is the probability that no coalescent events happen over t-1 generations, followed by a coalescent in generation t.

This is the well-known *geometric distribution.* From the properties of this distribution, the mean time to coalescence of a pair of alleles is $2N$ generations, and the variance of the coalescent time is approximately $(2N)^2$ if N is large.

Under the *infinite sites* model of variability, at most one mutation can occur at a nucleotide site on the genealogical tree connecting two sampled alleles, whose expected length is twice the mean coalescent time i.e. $4N$. If the mutation rate per site is u, then the frequency with which two alleles differ at a given site is equal to the product of the mutation rate and the time separating them i.e.

$$\pi \approx \theta \qquad (8)$$

where $\theta = 4Nu$ is the *scaled mutation rate*.

This relates the *pairwise nucleotide site diversity* measure of **Lecture 2** to the properties of a finite population with mutation.

This reasoning can be extended to samples of arbitrary size k. If the population size is fairly large, it is reasonable to assume that at most one coalescent event can occur among a set of alleles in a given generation, so that the genealogy connecting the alleles is a bifurcating tree with a steadily decreasing number of nodes (see diagram on p. 208 of the textbook). With i alleles present at a given point in time, the total number of ways in which coalescent events can occur is $j_i = i(i-1)/2$, so that the probability of an event is $j_i\, 2N$ instead of $2N$.

Coalescent events thus occur *fast at first*, and then *more and more slowly* as the number of distinct alleles decreases.

The *mean time* back to the *common ancestor* of a sample of k alleles can be shown to be $4N(1-1/k)$, so that, for a large sample of alleles, about half the total average time to the common ancestor is taken up with the coalescence of the last remaining pair of alleles.

The mean sum of lengths of all the branches in a gene tree can found as follows. The mean time to the next coalescence when there are i alleles present immediately after a coalescent event is, by the results derived above, $2N/j_i = 4N/i(i-1)$.

There are i branches connecting these alleles to the next coalescent, so that the net contribution of this section of the tree to the *mean total length* of the tree is $i \times 4N/i(i-1) = 4N/(i-1)$.

The mean total length of the tree is obtained by summing over all values of i, from the initial value of n down to the final value of $i = 2$. It is thus equal to $4Na_k$, where a_k is the sum from $i = 1$ to $k-1$ of $(1/i)$.

Under the *infinite sites* model, where each new mutation is at a unique position in a given DNA sequence, the expected number of sites in a sequence of length m that are segregating in the sample, $E\{S\}$, is equal to the expected number of mutations (mu) that occur in the sequence per generation, times the total length of the tree i.e.

$$E\{S\} = m\, a_k \theta \qquad (9)$$

This provides a useful method for estimating θ from data on DNA sequence variation, by equating the observed number of segregating sites to the expectation given by equation (9), as we discussed in lecture 1. The method assumes *equilibrium under neutrality*, with a constant population size.