Chapter 4

Neutral Mutations and Genetic Polymorphisms

The relationship between genetic data and the underlying genealogy was introduced in Chapter 1. Here we will combine the intuitions of Chapter 1 with the knowledge of the coalescent obtained in Chapter 3. Of course, we will also use the mathematical probability of Chapter 2 in generating predictions about levels and patterns of poymorphism in a sample of genetic data. In particular, we will now make extensive use of the Poisson distribution to represent numbers of mutations. We can do this with little error because mutation rates are very small, roughly 10^{-10} per base pair per replication event in eukaryote organisms (Drake *et al.*, 1998). When measured from sequence comparisons between species with divergence times known from the fossil record, estimates rates of substitution range from about 10^{-8} to about 10^{-10} per base pair per generation (Li, 1997). Mutation rates in microbes that use DNA as the genetic material vary over a broad range, from about 10^{-6} to 10^{-10} per base pair per replication event, while rates in RNA viruses may be as high as 10^{-4} (Drake *et al.*, 1998). Thus, mutation rates per generation are low, but numbers of mutations can become appreciable on the time scale of the coalescent which measures time in units of $N_e = N/\sigma^2$ generations. With these observations, and the additional fact that mutations in different generations occur independently, then the arguments of section 2.1.2 show that the number of mutations which occur over a branch or branches of a given length in a genealogy should be Poisson distributed with parameter equal to the expected number of mutations over that length of time.

As we saw in Chapter 3, for populations that are not too small, the times back to common ancestors among members of the sample are also well-modeled by a Poisson process. Thus, the world of simultaneous Poisson processes explored in section 2.2 provides a rich framework for thinking about mutation and coalescence together and, later, to include other processes such as recombination and migration. Because time is measured in units of N_e generations in the coalescent, mutation rates must be measured on a timescale proportional to this. For historical reasons, population geneticists use the mutation parameter $\theta = 2N_e u$, in which u is the mutation rate per generation, per locus or per site depending on the type of data under consideration. In the Wright-Fisher model, where $N_e = N$, the parameter θ is equal to twice the average number of mutations introduced into the population each generation, or twice the expected number of mutations along a single lineage over one unit of time on the coalescent time scale. Thus, mutation occur with rate $\theta/2$ one the coalescent time scale. The extra factor of two derives from the importance of the concept of heterozygosity, which was noted in Chapter 1. In particular, as we will show in Section 4.1.2, the expected number of pairwise differences in a sample is equal to θ defined in this way.

We can now add this mutation process with rate $\theta/2$ unit of time to the coalescent process.

First, given that the length of a genealogy or of some piece of a genealogy is equal to t, the number K of mutations, which is the sum of t independent Poisson($\theta/2$) random variates, is itself Poisson distributed with parameter $\theta t/2$:

$$P\{K = k|t\} = \frac{\left(\frac{\theta t}{2}\right)^k}{k!} e^{-\frac{\theta t}{2}} \qquad k = 0, 1, 2, \ldots , \tag{4.1}$$

and of course we have

$$E[K|t] = \text{Var}[K|t] = \frac{\theta t}{2}. \tag{4.2}$$

We will make extensive use of this result. It should be emphasized that the above applies to mutations that do not confer any selective advantage or disadvantage. Neutral mutations, because they do not alter patterns of reproductive success in the populations, do not affect the shape of genealogies. They are independent of the genealogical process. This is not true of mutations that affect fitness, which are considered in Chapters 5 and 6. However, even if the size and shape of the genealogy is determined by selection at some sites within a locus, equations 4.1 and 4.2 still hold for neutral mutations.

Neutral mutations create the genetic markers that reveal underlying genealogies. However, the fidelity with which they do this depends on how mutations occur, or on the kind of genetic data under consideration. Here the focus continues to be on the infinite-sites mutation model because it applies readily to DNA sequence data and because it offers the most direct view of the underlying genealogy. Most of the predictions that have been made about patterns of DNA sequence polymorphism, to which observed data are routinely compared, have been derived under the infinite-sites model. Other mutation models include the infinite-alleles model (Malécot, 1946; Kimura and Crow, 1964), various finite alleles models, such as those used for DNA substitutions over long periods of time reviewed in Li (1997), and the infinite allele or finite allele stepwise mutation models (Ohta and Kimura, 1973; Moran, 1975; Moran, 1976) that have recently been applied to data from repeat loci (Slatkin, 1995; Goldstein *et al.*, 1995). Section 4.2, presents results for the infinite-alleles mutation model. Importantly, equations 4.1 and 4.2 above hold for all these models. However, only under the infinite-sites model is there a one-to-one correspondence between mutations along the branches of the genealogy and polymorphic sites in a sample of DNA sequences. In this case it is straightforward to generate predictions about levels and patterns of polymorphism in a sample.

4.1 The Infinite Sites Model and Measures of DNA Sequence Polymorphism

Using the Poisson distribution of the number of mutations and the properties of coalescent genealogies obtained in Chapter 3, we can makes useful predictions about the shape of genetic variation. We will derive predictions about the three measures of genetic variation introduced in Chapter 1: the number S of segregating sites, the average number π of pairwise differences, and the numbers η_i and ξ_i of sites segregating in various frequencies among the members of the sample. The last two are referred to as the “folded” and the “unfolded” site frequencies, respectively. To make these predictions, it will be necessary to augment the descriptions of coalescent genealogies initiated in Chapter 3, typically using simple extensions of the ideas presented in that chapter. In addition, we continue until Chapter 6 to work under the assumption of no recombination at the locus under consideration. The consequence of this is that all the sites in the sequence share the same genealogy.

4.1.1 The Number Segregating Sites

The number S of segregating sites in a sample of size n is equal to the total number of mutations in the history of a sample. Thus, the aspect of the genealogy we are concerned with is T_{total}, the total length of the genealogy. Given T_{total}, the number of mutations on the genealogy is Poisson($\theta T_{\text{total}}/2$), and knowing the distribution 3.36 of T_{total}, we can use the formula 2.23 for the marginal distribution to obtain the distribution of S:

$$
\begin{aligned}
P\{S=k\} &= \int_0^\infty P\{S=k|t\} f_{T_{\text{total}}}(t)dt \\
&= \int_0^\infty \frac{\left(\frac{\theta t}{2}\right)^k}{k!} e^{-\frac{\theta t}{2}} \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} e^{-\frac{i-1}{2}t} dt \\
&= \left(\frac{\theta}{2}\right)^k \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} \int_0^\infty \frac{t^k e^{-\frac{\theta+i-1}{2}t}}{k!} dt \\
&= \left(\frac{\theta}{2}\right)^k \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} \left(\frac{2}{\theta+i-1}\right)^{k+1} \\
&= \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \left(\frac{i-1}{\theta+i-1}\right) \left(\frac{\theta}{\theta+i-1}\right)^k \qquad (4.3)
\end{aligned}
$$

(Tavaré, 1984). The distribution of S was first obtained by Watterson (1975), who found it in the form of a probability generating function. The step from the third to the fourth line above is achieved using the total probability of the gamma distribution 2.56.

Equation 4.3 is the most detailed prediction we can make regarding S. A graphical depiction of $P\{S=k\}$ is given in figure 4.1. Similar to the distribution of the size of the underlying genealogy, which is shown in figure 3.4, the distribution of S is L-shaped when n is small, then aquires a non-zero mode and assumes a characteristic shape as n increases. The distribution of the number of segregating sites, given in equation 4.3 and figure 4.1, has two related interpretations. First, it quantifies the stochastic variation associated with a single sample of size n from a population with a given value of θ. This interpretation is useful in the context of making inferences (*e.g.* maximum likelihood estimates of θ) from a sample of sequences. Second, $P\{S=k\}$ predicts what the distribution of the number of segregating sites should look like if identical-sized samples are taken from many independent (*i.e.* unlinked; see Chapter 6) loci which all have the same value of θ. This interpretation is what provided the theoretical comparison to the human single nucleotide polymorphism data in Table 1.1 (The International SNP Map Working Group, 2001).

For a sample of size $n=2$, equation 4.3 reduces to a geometric distribution; see equation 2.41. Specifically, the number of events up to, and including the coalescent event which brings a sample of size $n=2$ to its MRCA is geometrically distributed with parameter $p=\theta/(\theta+1)$. In fact, a distribution of this sort applies during every coalescent interval in the history of a larger sample. We can see this by considering neutral mutation and coalescence as simultaneous, independent Possion processes. The results of section 2.2.1 become immediately useful. On the coalescent time scale, during the time when there are i lineages ancestral to the sample, the rate of mutation is $i\theta/2$ and the rate of coalescence is $i(i-1)/2$. Therefore, from equation 2.62, we have the

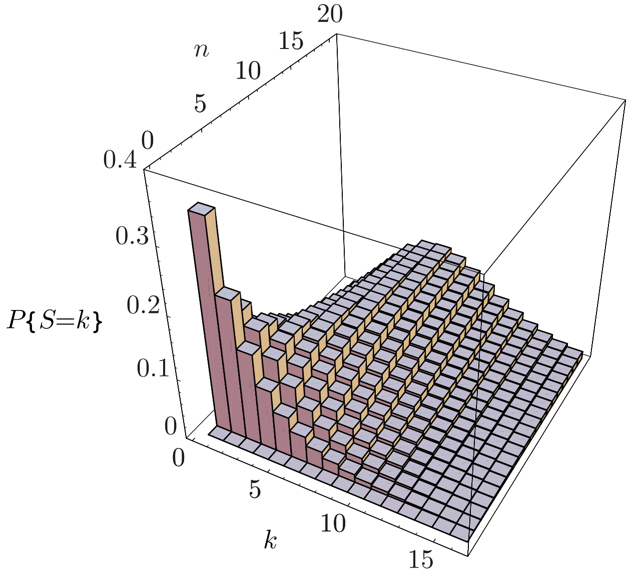


Figure 4.1: A series of histograms of the probability function of the number of segregating sites in a sample of n sequences. The mutation parameter is $\theta = 3$.

probability that a coalescent event is the first event to occur,

$$P\{\text{coalescence}|\text{event}\} \quad = \quad \frac{i(i-1)/2}{i\theta/2 \; + \; i(i-1)/2} \quad = \quad \frac{i-1}{\theta \; + \; i-1} \tag{4.4}$$

and the probability that a mutation event is the first event to occur,

$$P\{\text{mutation}|\text{event}\} \quad = \quad \frac{\theta}{\theta \; + \; i-1}. \tag{4.5}$$

From equation 2.64 it is clear that the distribution of the number of events up to, and including the first coalescent event among i lineages is geometrically distributed, so that we have

$$P\{S_i = k\} \quad = \quad \left(\frac{i-1}{\theta \; + \; i-1}\right)\left(\frac{\theta}{\theta \; + \; i-1}\right)^k \tag{4.6}$$

for the distribution of the number of segregating sites generated by mutations which occurred during the time there were i lineages ancestral to the sample. Since $S = \sum_{i=2}^{n} S_i$, we could obtain $P\{S = k\}$ as a convolution of the S_i, which is how Watterson (1975) approached the problem.

The consideration of coalescence and mutation as simultaneous, independent Poisson processes, as in section 2.2.1, will prove very useful in this chapter. As above, in this process every lineage mutates with rate equal to $\theta/2$ and each of the $i(i-1)/2$ possible pairs of lineages coalesces with rate equal to 1. However, we will often employ a different, but related method which is to condition on the lengths of branches, variously defined, and then to use the Poisson distribution 4.1. For example, we could obtain the moments $E[S]$ and $\text{Var}[S]$ from equation 4.3, but it is simpler in this case to condition on the total tree length T_{total} and to express $E[S]$ and $\text{Var}[S]$ in terms of $E[T_{\text{total}}]$, $\text{Var}[T_{\text{total}}]$, and the expected number $\theta/2$ of mutations per time unit. Although here T_{total} is a continuous rather than a discrete random variable, we can refer back to equations 2.31, 2.32, 2.33, and 2.34. We have

$$\begin{aligned} E[S] &= E[K]E[T_{\text{total}}] \\ &= \left(\frac{\theta}{2}\right)\left(2\sum_{i=1}^{n-1}\frac{1}{i}\right) \\ &= \theta\sum_{i=1}^{n-1}\frac{1}{i}, \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} \text{Var}[S] &= \text{Var}[K]E[T_{\text{total}}] \; + \; E[K]^2\text{Var}[T_{\text{total}}] \\ &= \left(\frac{\theta}{2}\right)\left(2\sum_{i=1}^{n-1}\frac{1}{i}\right) \; + \; \left(\frac{\theta}{2}\right)^2\left(4\sum_{i=1}^{n-1}\frac{1}{i^2}\right) \\ &= \theta\sum_{i=1}^{n-1}\frac{1}{i} \; + \; \theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}. \end{aligned} \tag{4.8}$$

These results are originally due to Watterson (1975) and are helpful in understanding patterns of genetic variation. First, the expected number of segregating sites is proportional to the expected total tree length, which again grows like $\log(n)$ when n is large. There is a diminishing return of increasing the sample size to discover more polymorphisms because the terms added to equation 4.7 become smaller and smaller as n increases. For example, sampling the third sequence will increase the number of polymorphisms discovered by 50% on average (*i.e.* will add a single new polymorphism for every two polymorphisms already discovered) while adding the 11th sequence will add only a single polymorphism to 28 already discovered, and adding a 101st sequence will add a single polymorphism to 518 already discovered.

Equations 4.7 and 4.8 imply that the shape of $P\{S = k\}$ might be Poisson in the limit of large sample size; see figure 4.1. That is, the mean number of segregating sites is equal to $\theta\sum_{i=1}^{n-1}1/i \approx \theta[\log(n) + \gamma]$ and the variance will be approximately the same since the second sum on the right in equation 4.8 converges to $\theta^2\pi^2/6$ as n goes to infinity while the first term continues to grow and is equal to $E[S]$. Indeed, S is approximately Poisson distributed for large samples, but it is not exactly so distributed (Watterson, 1975). This is similar to the fact that the distribution of T_{total} does not approach a Normal distribution in the limit of large sample size, but rather approaches the extreme value distribution given in equation 3.38.

4.1.2 Pairwise Differences

By conditioning on the genealogy, it is straightforward to make predictions about another of the commonly used measures of genetic variation: the average number of pairwise sequence differences among members of the sample, π, which was introduced in section 1.3. Expressions are available both for the expected value and the variance of π (Tajima, 1983). Their derivations illustrate the fact that the sampled lineages are exchangeable. Although it is possible to express

π in terms of the site frequencies η_i, which are the topic of the next section, we begin instead with equation 1.1 then take expectations to obtain

$$\begin{aligned} E[\pi] &= E\left[\frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij}\right] \\ &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[k_{ij}] \\ &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\theta}{2} E[2T_{ij}] \\ &= \frac{\theta}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[T_{ij}], \end{aligned} \tag{4.9}$$

in which T_{ij} is the coalescence time of sequence i and sequence j. In words, the expected value of π is equal to the average of the expected lengths of the lineages connecting each pair of sequences in the sample (up to their common ancestor) multiplied by the expected number of mutations per unit of time on the coalescent time scale. Figure 4.2 illustrates one such set of lineages, upon which a mutation, in the case depicted, would generate a difference between sequence 1 and sequence 8 in the sample.

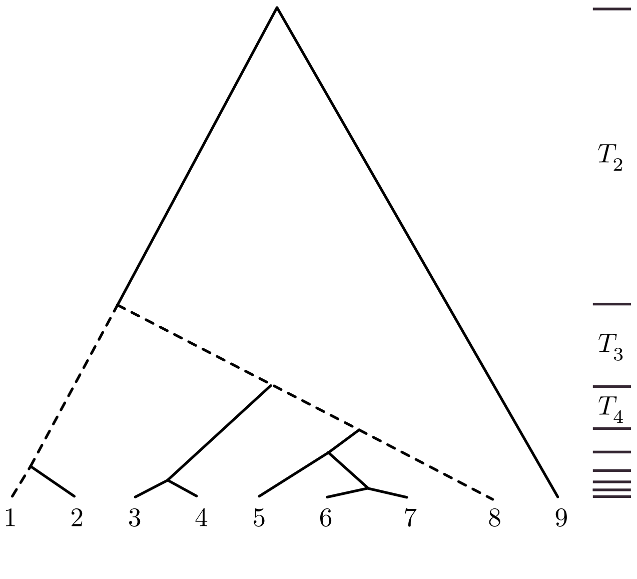


Figure 4.2: The (dashed) lineages connecting sequences 1 and 8 in a sample of size $n = 9$.

The members of the sample are exchangeable. This means that any labelling of them such as the one in figure 4.2 is arbitrary in the sense that it will not affect predictions about levels and patterns of polymorphism. In the present case, this means that $E[T_{ij}]$ must be the same for every pair of lineages. We can think of the expectation of $E[T_{ij}]$ being a marginal expectation with respect to all possible histories of the other members of the sample. Fundamentally, for example, when we compute $E[T_2]$ from equation 3.9 we are implicitly averaging over all possible histories of the $N-2$ other, unsampled sequences in the population. Thus, $E[T_{ij}]$ must not depend on the sample size, and from equation 3.10 must be equal to one for every pair.

We can show that this is true, that $E[T_{ij}] = 1$, by conditioning on the relevant part of the genealogy of a sample of size n. Sequences i and j might have their most recent common ancestor at any of the $n-1$ coalescent events in the history of the sample. Writing CE(k) for the coalescent event which decreases the number of ancestral lineages from k to $k-1$ and MRCA(i,j) for the most recent common ancestor of sequences i and j, we have

$$E[T_{ij}] \quad = \quad \sum_{k=2}^{n} E[T_{ij}|\mathrm{MRCA}(i,j) \text{ is at } \mathrm{CE}(k)]P\{\mathrm{MRCA}(i,j) \text{ is at } \mathrm{CE}(k)\}. \tag{4.10}$$

The example in figure 4.2 is one in which the most recent common ancestor of the pair, sequences 1 and 8 in this case, occurs at the 3 → 2 coalescent event. The two terms on the right hand side of equation 4.10 are straightforward to compute. First, because the branching structure of the tree and the coalescence times are independent, the conditional expected time to common ancestry of the pairs is simply the sum of the expected lengths of the corresponding coalescent intervals:

$$E[T_{ij}|\mathrm{MRCA}(i,j) \text{ is at } \mathrm{CE}(k)] \quad = \quad \sum_{m=k}^{n} E[T_m] \quad = \quad \sum_{m=k}^{n} \frac{1}{\binom{m}{2}} \quad = \quad 2\left(\frac{1}{k-1} - \frac{1}{n}\right). \tag{4.11}$$

Next, the probability that sequence i and sequence j coalesce at the coalescent event which ends the time during which there were k lineages ancestral to the sample is equal to the probablity that a particular pair of lineages is not involved in any of the preceding coalescent events and then is involved in the $k \rightarrow k-1$ coalescent event:

$$P\{\mathrm{MRCA}(i,j) \text{ is at } \mathrm{CE}(k)\} \quad = \quad \frac{1}{\binom{k}{2}} \prod_{l=k+1}^{n} \left[1 - \frac{1}{\binom{l}{2}}\right] \quad = \quad \frac{2(n+1)}{k(k+1)(n-1)}. \tag{4.12}$$

Note that equation 4.12 does allow sequences i and j to coalesce with other lineages in the sample, as sequences 1 and 8 do in the genealogy in figure 4.2, they just cannot coalesce with each other. Putting 4.11 and 4.12 into equation 4.10, and simplifying, gives $E[T_{ij}] = 1$, and thus $E[\pi] = E[k_{ij}] = \theta$.

It is possible to derive Var[π] using similar considerations. This was done by Tajima (1983) who noted that the variance of π for a sample of size n can be computed by considering samples of just two, three, and four sequences. Again, k_{ij} is the number of differences between sequence

i and sequence j in the sample. We have

$$\begin{aligned}
\mathrm{Var}[\pi] &= \mathrm{Var}\left[\frac{1}{\binom{n}{2}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\right] \\
&= E\left[\left\{\frac{1}{\binom{n}{2}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\right\}^2\right] - E\left[\frac{1}{\binom{n}{2}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\right]^2 \\
&= \frac{1}{\binom{n}{2}^2} E\left[\left\{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\right\}^2\right] - E\left[k_{ij}\right]^2. \qquad (4.13)
\end{aligned}$$

We have just seen that $E[k_{ij}] = \theta$, so the second term on the right is simply θ^2. The expectation in the first term on the right in equation 4.13 can also be calculated:

$$E\left[\left\{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\right\}^2\right] = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n} E[k_{ij}k_{rs}]. \qquad (4.14)$$

Tajima (1983) recognized that there are only three kinds of terms in equation 4.14, depending on the number of distinct values among the subscripts, i, j, r, and s. These three cases for the expected product of pairwise differences, and the numbers of each kind, are shown in table 4.1.

Value	Number of terms	Condition
$E[k_{ij}^2]$	$\binom{n}{2}$	$i = r \neq j = s$
$E[k_{ij}k_{rj}](= E[k_{ij}k_{is}])$	$2\binom{n}{2}(n-1)$	$i = r \neq j \neq s$ or $i \neq r \neq j = s$
$E[k_{ij}k_{rs}]$	$\binom{n}{2}\binom{n-2}{2}$	$i \neq r \neq j \neq s$

Table 4.1: The three possible values of the expectation on the right in equation 4.14

As with the computation of $E[k_{ij}]$ in a sample of size n above, the expected values in table 4.1, are marginal expectations with respect to the histories of the other members of the sample. Because the samples are exchangeable, the three expected values in table 4.1 are the same for every subset of the n samples that satisifies the given condition. Therefore, $E[k_{ij}^2]$, $E[k_{ij}k_{rj}]$, and $E[k_{ij}k_{rs}]$ can be calculated by considering samples of just two, three, and four sequences, respectively. For example, $E[k_{ij}k_{rs}]$ is the expected product of the numbers of differences between two sequences labelled i and j and two other sequences labelled r and s, averaged over all possible genealogies of the sample, of size four, and all possible patterns of mutation on the genealogy. As with $E[k_{ij}]$, $E[S]$, and $\mathrm{Var}[S]$, $E[k_{ij}k_{rs}]$ can be expressed in terms of the moments of the branch lengths and numbers of mutations. The end result of these calculations is

$$\mathrm{Var}[\pi] = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2. \qquad (4.15)$$

Tajima (1983) used this result to argue that there is a large stochastic component to the average number of pairwise differences, even when the sample size is large. This is illustrated in figure 4.3 which compares the coefficient of variation of π to that of S. The coefficeint of variation is a standardized measure of dispersion, and is defined as the standard deviation, or the square root of the variance, divided by the expected value. Figure 4.3 shows that the coefficient of variation of S decreases as n increases. In fact, it approaches zero as n approaches infinity. In contrast, the coefficient of variation of π approaches a value greater than zero, specifically $\sqrt{1/(3\theta) + 2/9}$, as n approaches infinity. This has serious consequences for the estimation of θ from polymorphism data. In particular, the estimate based on π is inconsistent (Tajima, 1983; Donnelly and Tavaré, 1995), which means that the variance of the estimate does not approach zero as the sample size approaches infinity.

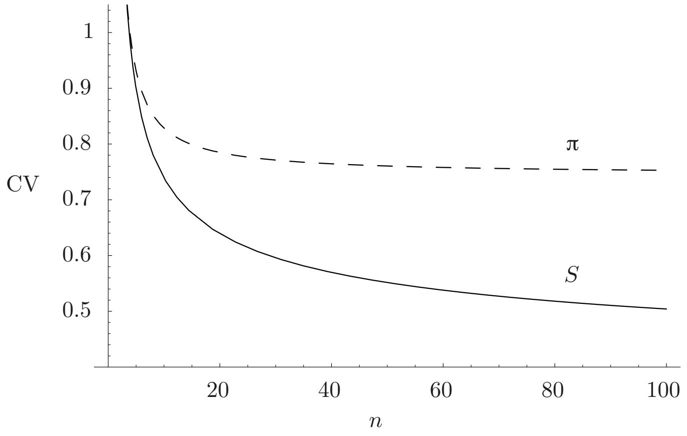


Figure 4.3: The coefficients of variation of π and S as a function of the sample of size n, with $\theta = 1$.

4.1.3 Site Frequencies

By considering the numbers of mutations on appropriate branches in the genealogy we can also make predictions about the site frequencies ξ_i and η_i. Again, ξ_i is the number of segregating sites where the mutant base is present on i sequences inη_i the sample and the ancestral base is found on the other $n - i$ sequences. Under the infinite-sites model, these are the result of mutations that occurred on branches in the genealogy which have i descendents in the sample. Unless sequence data are available from a closely-related species, it is impossible to distinguish the ancestral base from the mutant base, and η_i is the number of sites at which the less-frequent base is present on i sequences out of n. The analysis of the unfolded site frequencies ξ_i is more straightforward than the analysis of the folded site frequencies η_i. Equation 1.2 can be used to make predictions about η_i once the properties of the ξ_i are known. Much of current intuition in the field about how population-level processes shape genetic variation is based on the expected values of these quantities, and we will take up this topic in Section 4.3.

Let τ_i be the total length of branches that have i descendents in the sample. Then, by the Poisson($\theta\tau_i/2$) distribution of mutations given τ_i, and employing the same argument used above in equation 4.7, we have

$$E[\xi_i] = \frac{\theta}{2} E[\tau_i]. \tag{4.16}$$

Figure 4.4 shows an example of a mutation giving rise to a polymorphic site at which the mutant base is found in six copies in a sample of size nine. The branch on which the mutation happened is the only branch in the genealogy that could contribute to ξ_6 (or τ_6). In addition, there are nine branches that contribute to τ_1, three branches that contribute to τ_2, and one branch each that contribute to τ_3, τ_4, and τ_8. There are no branches in the genealogy in figure 4.4 that contribute to τ_5 or τ_7. Therefore, under infinite-sites mutation, the genealogy in figure 4.4 could generate data patterns ξ_1, ξ_2, ξ_3, ξ_6, and ξ_8, but could not generate patterns ξ_5 and ξ_7. Other genealogies will have different structures, and the expectations in equation 4.16 are taken over all possible genealogies, branch lengths, and numbers of mutations. This can be done in several different ways, and gives

$$E[\tau_i] = \frac{2}{i} \tag{4.17}$$

(Tajima, 1989; Fu and Li, 1993), so that $E[\xi_i] = \theta/i$. The variances and covariances of these patterns can also be obtained (Fu, 1995).

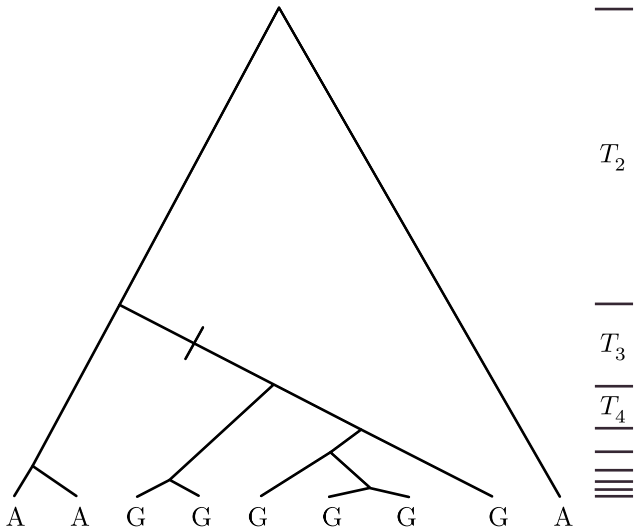


Figure 4.4: Example of a mutation generating a polymorphic site in frequency 2/3 in a sample of size $n = 9$.

We can use an approach that parallels the derivation of expected average pairwise differences above to obtain $E[\xi_1]$, the expected number of *singletons* in the sample. Note that singleton polymorphisms must have resulted from mutations that occurred on the external branches of the genealogy. Every genealogy has n external branches, and the joint distribution of the lengths of these is constrained by the structure of the tree. However, the expected number of singletons does not depend on these complicated correlations. Let $\tau_1^{(i)}$ be the length of the branch leading to sequence i in the sample. Then, τ_1 is equal to the sum of these, or $\sum_{i=1}^{n} \tau_1^{(i)}$, and we have

$$E[\tau_1] \quad = \quad E\left[\sum_{i=1}^{n} \tau_1^{(i)}\right] \quad = \quad nE[\tau_1^{(i)}]. \tag{4.18}$$

Further, $E[\tau_1^{(i)}]$ is the same for every sequence $i = 1, 2, \dots, n$ because the lineages are exchangeable. By conditioning on the coalescent event at which lineage i joins the genealogy and writing FCA(i) for the first common ancestor event that involves lineage i, we have

$$E[\tau_1^{(i)}] \quad = \quad \sum_{k=2}^{n} E[\tau_1^{(i)}|\text{FCA}(i) \text{ is at CE}(k)]P\{\text{FCA}(i) \text{ is at CE}(k)\}. \tag{4.19}$$

and both of the terms on the right can be computed. First, similarly to equation 3.46,

$$P\{\text{FCA}(i) \text{ is at CE}(k)\} \quad = \quad \frac{k-1}{\binom{k}{2}} \prod_{j=k+1}^{n} \left(1 - \frac{j-1}{\binom{j}{2}}\right) \quad = \quad \frac{2(k-1)}{n(n-1)}. \tag{4.20}$$

In words, the probability that one particular lineage joins the genealogy at the $k \rightarrow k-1$ coalescent event is equal to the probability that it does not coalesce with any of the other lineages, from the present back to the time when there were k lineages, and then the next coalescent event is between that lineage and one of the other $k-1$ lineages. Next, the expected length of the branch, conditional on the lineage joining the genealogy at this point, is identical to equation 4.11 above. Putting this and equation 4.20 into equation 4.19 gives

$$\begin{aligned} E[\tau_1^{(i)}] \quad &= \quad \sum_{k=2}^{n} \frac{2(k-1)}{n(n-1)} \cdot 2\left(\frac{1}{k-1} - \frac{1}{n}\right) \\ &= \quad \frac{4}{n(n-1)} \sum_{k=2}^{n} \left(1 - \frac{k-1}{n}\right) \\ &= \quad \frac{4}{n(n-1)} \left(n - 1 - \frac{\binom{n}{2}}{n}\right) \\ &= \quad \frac{2}{n} \end{aligned} \tag{4.21}$$

Finally, using equation 4.18, we have $E[\tau_1] = 2$ which is in agreement with equation 4.17 and shows that the expected number of polymorphic sites at which the mutant base found on just a single sequence in the sample is $E[\xi_1] = \theta$.

Fu (1995) and Griffiths and Tavaré (1998) used similar considerations to obtain the expected values of the full spectrum of site frequencies. The expected values of the unfolded site frequencies are

$$E[\xi_i] \quad = \quad \frac{\theta}{i} \qquad\qquad 1 \leq i \leq n-1, \tag{4.22}$$

and do not depend on the sample size n while the expected values of the folded site frequencies

$$E[\eta_i] \quad = \quad \theta \frac{\frac{1}{i} + \frac{1}{n-i}}{1 + \delta_{i,n-i}} \qquad\qquad 1 \leq i \leq [n/2]\,. \tag{4.23}$$

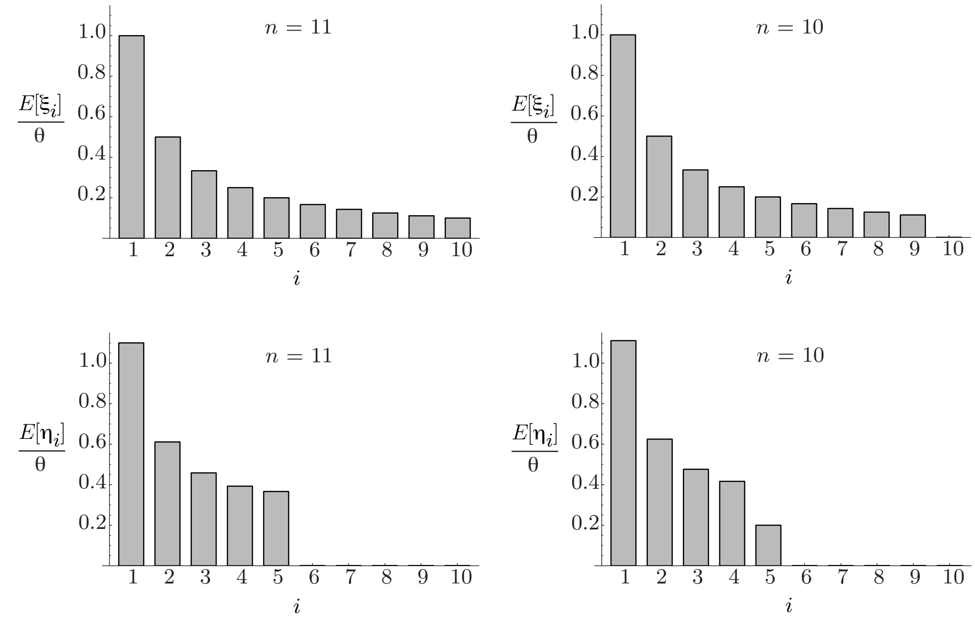


Figure 4.5: The relative expected numbers of polymorphic sites ξ_i and η_i in an odd-sized sample ($n = 11$) and in an even-sized sample ($n = 10$).

do depend on n. Again, $[n/2]$ means the largest integer less than or equal to $n/2$, and $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ if $i \neq j$ (see equation 1.2). Griffiths and Tavaré (1998) considered the expected proportion of sites segregating at different frequencies in the sample, and found the following general formula:

$$\frac{E[\xi_i]}{E[S]} = \frac{(n-i-1)!(i-1)!\sum_{k=2}^{n-i+1} k(k-1)\binom{n-k}{i-1} E[T_k]}{(n-1)!\sum_{k=2}^{n} kE[T_k]} \quad , \quad 1 \leq i \leq n-1. \quad (4.24)$$

Equation 4.24 links the expected site frequencies to the expected lengths of coalescent intervals via the probabilities that branches which exist during the time when there are k lineages ancestral to the sample have i descendents in the sample. It is general in the sense that it holds for any model in which the branching structure of genealogies is the same as in the standard coalescent model, *i.e.*, random-joining or random-bigurcating, while the expected values $E[T_k]$ need not be the same as those in Kingman's coalescent.

The expected site-frequency spectrum has the characteristic shape shown in figure 4.5. Singletons are expected to be the most abundant kind of polymorphism, followed by doublets, which are expected to be half as numerous as singletons, then by triplets, and so on. The folded site-frequency spectrum looks different when n is odd, and the highest sample frequency class corresponds to two unfolded patterns, than when n is even, and highest sample frequency class corresponds to just one unfolded pattern. Again, these expected values are taken over all possible genealogies and all possible arrangements of the mutations on the sequences, so they tell us little about what to expect in a sample from a single locus, expecially one with limited recombination. As more and more independent loci are sampled, the site frequencies in the sample will approach these expectations if the assumptions of the standard coalescent model are true. Clearly, the site-frequency counts ξ_i or η_i themselves carry no information about linkage patterns or about recombination (see Chapter 6). For example, a sample in which a single sequence posesses two mutant bases and a sample in which two different sequences each possess one mutant base both give $\xi_1 = 2$. We will return to these notions in Section 4.3 when we consider the potential for the site-frequency spectrum to capture deviations from the standard coalescent model.

4.2 The Infinite Alleles Model and the Ewens Sampling Formula

One of the most important results of theoretical population genetics is the Ewens sampling formula (Ewens, 1972), which gives the probabilities of allelic configurations of a sample under the same conditions that yield the coalescent but with the additional assumption of infinite-alleles mutation. As a measure of its novelty and impact, one recent probability text devotes an entire chapter to "Ewens Distributions" (Johnson *et al.*, 1997). Ewens discovered the sampling formula by computing patterns of identity by descent in a sample. Recall, from Chapter 1, that the infinite-alleles model assumes that every mutation introduces a new allele into the population. This idea was first put forward by Malécot (1946) and was considered later by Kimura and Crow (1964). In the decade or so following the first use of gel electrophoresis to measure the genetic diversity of populations (Lewontin and Hubby, 1966; Harris, 1966), there was a flurry of work on the forward-time diffusion of allele frequencies under the infinite-alleles model; see Ewens (2004). At the same time, there was a great deal of work on an alternative mutation model for electrophoretic alleles: the charge-state, or stepwise mutation, model (Ohta and Kimura, 1973; Moran, 1975; Moran, 1976). These two lines of work played a vital role in revealing the genealogical structures underlying the Ewens sampling formula and other results, and laid the foundations of the coalescent (Kingman, 2000).

Under the infinite-alleles model of mutation, Ewens (1972) derived a formula for the probability that a sample of n gene copies contains k alleles and that there are $a_1, a_2, \ldots, a_n$ alleles represented $1, 2, \ldots, n$ times in the sample:

$$P\{k, a_1, a_2, \ldots, a_n\} = \frac{n!\theta^k}{\theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!} \tag{4.25}$$

in which $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$. Karlin and McGregor (1972) gave a rigorous mathematical proof of equation 4.25. Equation 4.25 is called the Ewens sampling formula. Note that the sum of allele counts is equal to the total number of alleles,

$$\sum_{j=1}^{n} a_j = k, \tag{4.26}$$

and that equation 4.25 applies only for configurations that satisfy

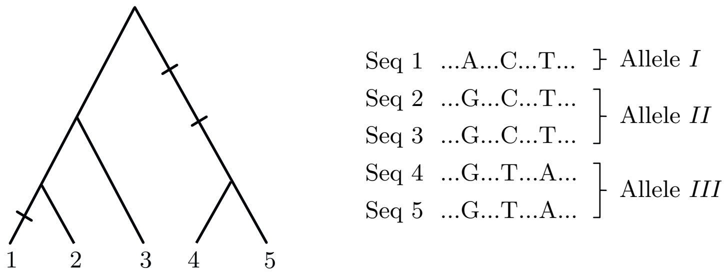
$$\sum_{j=1}^{n} j a_j = n \tag{4.27}$$

otherwise $P\{k, a_1, a_2, \ldots, a_n\}$ is equal to zero. For an example of this notation, if a sample of size $n = 10$ contained four alleles labelled I, II, III, and IV, and these were in the configuration (I, II, II, I, IV, III, I, I, I, I) for the ten sampled items, then

$$(a_1, a_2, \ldots, a_{10}) = (2, 1, 0, 0, 0, 1, 0, 0, 0, 0) \tag{4.28}$$

and this of course satifies equations 4.26 and 4.27. Equation 4.25 gives the probability of all such configurations, *i.e.* regardless of the order in which the alleles are observed.

There are many ways to interpret the assumption of infinite-alleles mutation, but perhaps the most sensible is in its relationship with the infinite-sites model without intragenic recombination. The infinite-sites model assumes that every mutation occurs at a previously unmutated site, and this is a good starting point for DNA sequences, which typically comprise a very large number of nucleotide sites each with a very low rate of mutation. An allele is a unique string of nucleotides at such a locus. These are often referred to as haplotypes, and it is clear that every mutation under the infinite-sites model creates a new haplotype, or allele. Simply counting numbers of haplotypes ignores much of the information in the data, but it might sometimes be desirable to do so. It is useful here, as a consideration of haplotypes sheds light on the Ewens sampling formula. Figure 4.6 shows a genealogy of sample of five sequences, upon which three mutations have occurred. The three mutations produced three polymorphic sites in the sequence data on the right in the figure. Because two of the mutations occurred on the same branch in the genealogy, three alleles were produced. If all three mutations occurred on the same branch of the tree, the sample would contain just two alleles, and if all three happened on different branches, the sample would contain four alleles.



Seq 1 ...A...C...T... } Allele *I*
Seq 2 ...G...C...T...
Seq 3 ...G...C...T... } Allele *II*
Seq 4 ...G...T...A...
Seq 5 ...G...T...A... } Allele *III*

Figure 4.6: Infinite-sites mutations and infinite-alleles data.

Thus, infinite-sites mutation produces infinite-alleles, haplotype data within the coalescent framework when each lineage is followed back only to the most recent mutation event. Using this notion it is straightforward to obtain the distribution of the number k of alleles in a sample. This marginal distribution $P\{k\}$ can be obtained from the full Ewens sampling formula by summing over all $(a_1, a_2, \ldots, a_n)$ that satisfy $\sum_{j=1}^{n} a_j = k$, but the following is more intuitive. Recall equations 4.4 and 4.5, which give the probabilities that the first event looking back among i lineages is a coalescent event or that it is a mutation event, respectively. Because a mutation guarantees that a lineage and all of its descedents will be of a unique allelic type, there is no need to follow lineages beyond the first mutation event looking back. Thus, both mutation and coalescence have the same effect on the sample: they decrease the number of lineages by one. Then, the following algorithm produces a random draw from $P\{k\}$:

1. Start with $i = n$ lineages and $k = 0$.
2. $k \rightarrow k + 1$ with probability $\theta/(\theta + i - 1)$.
3. Subtract one lineage: $i \rightarrow i - 1$.
4. Stop if $i = 0$, otherwise return to step 2.

The above is identical to tossing a series of n coins with increasing probabilities of success, in this case mutation, given by $\theta/(\theta + i - 1)$ for $i = n, n - 1, \ldots, 2, 1$. Note that, in contrast to the usual situation in coalescent theory, it will sometimes be necessary to follow the lineage ancestral to the MRCA of the sample back to an inevitable mutation event in order to guarantee that a sample with no polymorphic nucleotide sites contains a single allele, which will be in count n with probability equal to one.

Analogously to the way in which, in Section 2.1.2, the binomial distribution results from the expansion of $(p+1-p)^n$, the distribution of the number of alleles in the sample is obtained from the expansion of

$$1 = \left(\frac{\theta}{\theta+n-1} + \frac{n-1}{\theta+n-1}\right)\left(\frac{\theta}{\theta+n-2} + \frac{n-2}{\theta+n-2}\right)\cdots\left(\frac{\theta}{\theta+1} + \frac{1}{\theta+1}\right)\frac{\theta}{\theta}.$$

In particular, for there to be k alleles in the sample, there must be k sucesses, or mutations, in these n coin tosses. Therefore, we have

$$P\{k\} = \frac{s_n^{(k)}\theta^k}{\theta_{(n)}} \tag{4.29}$$

where $s_n^{(k)}$ is the coefficient of θ^k in the expansion of $\theta_{(n)}$. The numbers $s_n^{(k)}$ are the unsigned Stirling numbers of the first kind, and these satisfy

$$x_{(n)} = \sum_{k=1}^{n} s_n^{(k)} x^k. \tag{4.30}$$

Equation 4.30 shows that $\sum_{k=0}^{n} P\{k\} = 1$ as required for $P\{k\}$ to be a probability function. Unsigned stirling numbers of the first kind are generated recursively using $s_n^{(1)} = (n-1)!$ and

$$s_n^{(k)} = s_{n-1}^{(k-1)} + (n-1)s_{n-1}^{(k)}, \tag{4.31}$$

for $k = 2, 3, \ldots, n-1$, and with $s_n^{(n)} = 1$. Again, Abramowitz and Stegun (1964) are a good reference for Stirling numbers. Note that Stirling numbers of both kinds come in signed and unsigned varieties, leading Johnson *et al.* (1997) to list four kinds of Stirling numbers, and that the notation for Stirling numbers are highly variable.

Table 4.2 shows all the possible realizations of the algorithm given above, for the case of $n = 4$, and illustrates how the coefficients $s_n^{(k)}$ fall out of this analysis. In a similar manner, by keeping track of the numbers of descendents of each ancestral lineage back to the first mutation event along each lineage, it is possible to construct a proof of the full Ewens sampling formula, equation 4.25, but we do not pursue this here. From the analogy to coin tossing, or to Bernoulli trials, the expected number of alleles in the sample is given by the sum of the probabilites of mutation, or

$$\frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1}. \tag{4.32}$$

This equation resembles equation 4.7 for the expected number of segregating sites in the sample. In particular, if θ is very small, then equation 4.32 becomes equal to one plus the expected number of segeregating sites. This makes intuitive sense because when the mutation rate is very small there will typically be either zero mutations or one mutation in the history of the sample, and if there is one segregating site then there are two alleles. It is also possible to show, although less obviously, that the probabilities of one segregating site from equation 4.3 and of two alleles from equation 4.29 become identical in the limit of small θ.

Pattern	Probability	# Alleles, k	$P\{k\}$
1111	$\frac{\theta}{\theta+3}\frac{\theta}{\theta+2}\frac{\theta}{\theta+1}\frac{\theta}{\theta}$	4	$\frac{\theta^4}{(\theta+3)(\theta+2)(\theta+1)\theta}$
1101	$\frac{\theta}{\theta+3}\frac{\theta}{\theta+2}\frac{1}{\theta+1}\frac{\theta}{\theta}$	3	$\frac{6\theta^3}{(\theta+3)(\theta+2)(\theta+1)\theta}$
1011	$\frac{\theta}{\theta+3}\frac{2}{\theta+2}\frac{\theta}{\theta+1}\frac{\theta}{\theta}$		
0111	$\frac{3}{\theta+3}\frac{\theta}{\theta+2}\frac{\theta}{\theta+1}\frac{\theta}{\theta}$		
1001	$\frac{\theta}{\theta+3}\frac{2}{\theta+2}\frac{1}{\theta+1}\frac{\theta}{\theta}$	2	$\frac{11\theta^2}{(\theta+3)(\theta+2)(\theta+1)\theta}$
0101	$\frac{3}{\theta+3}\frac{\theta}{\theta+2}\frac{1}{\theta+1}\frac{\theta}{\theta}$		
0011	$\frac{3}{\theta+3}\frac{2}{\theta+2}\frac{\theta}{\theta+1}\frac{\theta}{\theta}$		
0001	$\frac{3}{\theta+3}\frac{2}{\theta+2}\frac{1}{\theta+1}\frac{\theta}{\theta}$	1	$\frac{6\theta}{(\theta+3)(\theta+2)(\theta+1)\theta}$

Table 4.2: Breakdown of the Ewens(4,θ) distribution. The patterns are the results, in order, of the coin tosses, with 1 = mutation and 0 = coalescence.

One very interesting property of the Ewens distribution is that

$$P\{a_1, a_2, \ldots, a_n | k\} \quad = \quad \frac{P\{k, a_1, a_2, \ldots, a_n\}}{P\{k\}} \quad = \quad \frac{n!}{S_n^k} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}. \tag{4.33}$$

Given that there are k alleles in the sample, the distribution of allele counts does not depend on θ. Thus, k is a sufficient statistic for θ. This means that there is no added information about θ in the allele counts. The maximum likelihood estimator of θ is given by equating the observed number of alleles in the sample with its expected value 4.32 and solving. The book chapter mentioned above — chapter 41 in Johnson, Kotz, and Balakrishnan (1997) — provides a good review of these and other properties of the Ewens sampling formula. Equation 4.33 is one of a very few such results in population genetics. Another is that the number of segregating sites is a sufficient statistic for θ under the assumption of independence among sites (Ewens, 1974). There is a great deal to be done in terms of advancing our understanding of the information content of measures of sequence polymorphism concerning the various factors that shape genetic variation, as the next section illustrates.

4.3 Deviations from the Standard Model: Testing "Neutrality"

It was emphasized in Chapter 3 that the standard neutral model includes a number of assumptions. From this model flow numerous predictions about the shapes of genealogies and about patterns of DNA sequence polymorphism. These predictions are the backdrop to our modern understanding and interpretation of genetic variation. Of course, they are valid only for populations that meet the underlying assumptions, chiefly that there is no selection, no population subdivision, and no changes in effective population size over time. Additional assumptions include that the sample size is much smaller than the effective size of the population, and, for many of the predictions above, that mutations occur according to the infinite-sites model without intra-locus recombination. Most of the rest of this book is devoted to extensions of the coalescent approach to accommodate deviations from these assumptions and to include such well-known biological phenomena as natural selection and population sibdivision. However, it is possible even at this point to grasp the major effects that these processes and events have on sequence data by understanding the ways in which they shape genealogies relative to the standard model. The connection between genealogies and genetic data is clear when each polymorphism is due to a single mutation event, *i.e.* when the infinite-sites mutation model applies. In this case, the numbers of different kinds of polymorphic sites reflect the lengths of corresponding branches in the genealogy of the sample, mediated by the random, Poisson process of mutation. Readers are referred back to figures 4.2, 4.4, and 4.6.

Of the many different measures of genetic variation that are possible, this chapter has focussed on the total number of polymorphisms (segregating sites, or SNPS) and on the decompositon of segregating sites into the site-frequency spectrum. Much of current intuition about the structure of genetic varition and most of the tests proposed to detect deviations from the standard model are based either directly or indirectly on the site-frequency spectrum. Two other kinds of measures were considered above: pairwise differences, which are in fact a simple function of the site frequencies, and haplotype numbers and counts, to which the Ewens sampling formula applies. This section introduces introduces the commonly-used “neutrality” tests (Tajima, 1989; Fu and Li, 1993), which are based on site frequencies. As noted above, site-frequency counts ignore the way in which the polymorphism are distributed among the sequences in the sample, so-called *linkage disequilibrium*, which can be a potentially rich source of information (Hudson *et al.*, 1994; Fu, 1996; Kelly, 1997; Andolfatto *et al.*, 1999; Machado *et al.*, 2001; Sabeti *et al.*, 2002; Beaumont *et al.*, 2003; Przeworski, 2003). The standard neutrality tests also ignore any differences in patterns of polymorphism among different genetic loci when these are included in a sample. By considering the effects of population history and demography on gene genealogies, this section presents some intuitions about variability in the number of segregating sites among loci and, to a lesser extent, about linkage disequilibrium; see also Wakeley (2004).

4.3.1 Test Statistics Based on Site Frequencies

Tajima (1989) noticed that the average number of pairwise differences π and the number of segregating sites S could be used to test the standard neutral model. The intuition behind this is that since $E[\pi] = \theta$ and $E[S] = \theta a_1$, where $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$, then the expected value of the difference $\pi - S/a_1$ is equal to zero under the standard neutral model. Significant deviations from zero should cause the model to be rejected. Tajima (1989) proposed the test statictic

$$D \quad = \quad \frac{\pi - S/a_1}{\sqrt{\widehat{\mathrm{Var}}[\pi - S/a_1]}}. \tag{4.34}$$

The denominator of Tajima’s D is estimated from the data using the formula

$$\widehat{\mathrm{Var}}(\pi - S/a_1) \quad = \quad e_1 S \;+\; e_2 S(S-1),$$

in which

$$e_1 \quad = \quad \frac{1}{a_1}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_1}\right) \quad , \quad e_2 \quad = \quad \frac{1}{a_1^2 + a_2}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{na_1} + \frac{a_2}{a_1^2}\right),$$

with a_1 as above and $a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$. The denominator of Tajima’s D is an attempt to normalize for the effect of sample size on the critical values. The coefficients e_1 and e_2 follow from the computation of

$$\mathrm{Var}(\pi - S/a_1) \quad = \quad \mathrm{Var}(\pi) \;-\; 2\mathrm{Cov}(\pi, S)/a_1 \;+\; \mathrm{Var}(S)/a_1^2 \tag{4.35}$$

(see equation 2.28) in the manner of section 4.1 above (Tajima, 1989).

Tajima (1989) suggested that the distribution of D might be approximated by a beta distribution, and provided tables of critical values for the rejection of the standard neutral model. The upper (lower) critical value is the value above (below) which the observed value of the statistic cannot be explained by the null model. As with any statistical test, it is necessary to specify a significance level α, which represents the acceptability of rejecting the null model just by chance when it is true. Very roughly speaking, values of Tajima's D and the other statistics given below are significant at the 5% level ($\alpha = 0.05$) if they are either greater than two or less than negative two. Tajima's D is not exactly beta-distributed and critical values are often determined using computer simulations (see Chapter 8). In a key paper on this subject, Simonsen *et al.* (1995), in addition to proposing several new statistics and exploring the sensitivity of the various tests to deviations from the null model, describe how critical values should be determined in light of the fact that the parameter θ must be estimated from the data.

Two other commonly-employed test statistics that behave in a manner similar to Tajima's D are the statistics of Fu and Li (1993),

$$D^* = \frac{S/a_1 - \frac{n-1}{n}\eta_1}{\sqrt{\widehat{\mathrm{Var}}[S/a_1 - \frac{n-1}{n}\eta_1]}}, \tag{4.36}$$

$$F^* = \frac{\pi - \frac{n-1}{n}\eta_1}{\sqrt{\widehat{\mathrm{Var}}[\pi - \frac{n-1}{n}\eta_1]}}, \tag{4.37}$$

where η_1 is the number of singletons in the folded site-frequency spectrum. These statistics are based on the same intuition as Tajima's D, namely that a comparison between different measures of polymorphisms that have the same expected value under the standard neutral model can be the basis for a test. Fu and Li's D^* and F^* make the two other possible pairwise comparisons once the number of singletons is included as a third measure.

Because the three measures S, π, and η_1 are simple functions of the unfolded site-frequency counts ξ_i, deviations of the three statistics D, D^*, and F^* can be understood in terms of the overrepresentation or underrepresentation of polymorphisms in different frequencies in the sample or, equivalently, of different types of branches in the genealogy (see equation 4.16). We have the relationships

$$S = \sum_{i=1}^{n-1} \xi_i, \tag{4.38}$$

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i, \tag{4.39}$$

$$\eta_1 = \frac{\xi_1 + \xi_{n-1}}{1 + \delta_{1,n-1}}, \tag{4.40}$$

in which ξ_i is again the number of polymorphic sites that have i copies of the mutant base and $n = i$ copies of the ancestral base among the sample of size n, and $\delta_{i,j} = 1$ if $i = j$ and zero otherwise.

Tajima's (1989) statistic D and the several statistics proposed subsequently by Fu and Li (1993) and by Simonsen *et al.* (1995) were among the first practical benefits garnered from the coalescent. They provided direct tests of the standard neutral model using the information in molecular sequence data. While here we will focus on the statistics D, D^*, and F^* designed for DNA sequence data, it is important to recognize the pre-coalescent precursor to these tests, namely the Ewens-Watterson test (Ewens, 1972; Watterson, 1977; Slatkin, 1982), which is based on deviations from the predictions of the Ewens sampling formula concerning the homoygosity of the population. Although D, D^*, and F^* are very widely used, and despite their groundbreaking start, it is clear that these and related statistics are of limited utility with respect to question of detecting selection. In particular, there are only two ways in which these statistics can deviate from the neutral prediction of zero — they can be too big either in the positive direction or in the negative direction — yet the standard neutral model includes a long list of assumptions. Only one of these assumptions is about natural selection, so it is wrong to think of these tests as tests of neutrality alone. Simonsen *et al.* (1995) studied the sensitivity of these tests to a variety of deviations from the standard neutral model.

The response of D, D^*, and F^* to deviations from the standard neutral model can be understood from the way each is related to the site frequencies ξ_i, that is via equations 4.38, 4.39, and 4.40. The sign of each test statistic is determined only by the sign of the numerator because the denominator is always taken to be positive. Tajima (1997) used 4.38, 4.39, and 4.40 to write the numerators of D, D^*, and F^* in terms of the site frequencies. We have, respectively,

$$\pi - \frac{S}{a_1} = \sum_{i=1}^{n-1} \left(\frac{2i(n-i)}{n(n-1)} - \frac{1}{\sum_{j=1}^{n-1} \frac{1}{j}} \right) \xi_i \tag{4.41}$$

$$\frac{S}{a_1} - \frac{n-1}{n}\eta_1 = \left(\frac{1}{\sum_{j=1}^{n-1} \frac{1}{j}} - \frac{n-1}{n} \right) \frac{\xi_1 + \xi_{n-1}}{1+\delta_{1,n-1}} + \sum_{i=2}^{n-2} \frac{\xi_i}{\sum_{j=1}^{n-1} \frac{1}{j}} \tag{4.42}$$

$$\pi - \frac{n-1}{n}\eta_1 = \left(\frac{2i(n-i)}{n(n-1)} - \frac{n-1}{n} \right) \frac{\xi_1 + \xi_{n-1}}{1+\delta_{1,n-1}} + \sum_{i=2}^{n-2} \frac{2i(n-i)}{n(n-1)} \xi_i. \tag{4.43}$$

The point of these complicated-looking equations is that the numerators of D, D^*, and F^*, are linear combinations of the site-frequency counts, ξ_i for $i = 1, \dots, n-1$, with coefficients that depend on n and i. Thus, for a given sample size n, each ξ_i makes either a positive or a negative contribution to each test statistic. The magnitudes of these contributions are easily computed for any n and i using the equations above. If we replace ξ_i with its the standard neutral expectation θ/i, then equations 4.41, 4.42, and 4.43 become equal to zero. On the other hand, if the site-frequency spectrum is different than the standard neutral prediction, then all three statistics will deviate from zero.

Figure 4.7 plots the coefficients of ξ_i in the numerator of Tajima's D and of Fu and Li's D^* for two different sample sizes: $n = 10$ and $n = 30$. The corresponding graphs for Fu and Li's F^* are similar to those for D^* except that the coefficients for $\xi_2, \dots, \xi_{n-2}$ depend on i. The graphs in figure 4.7 are symmetric about $n/2$ because these test statistics were designed for data in which the ancestral and mutant bases at polymorphic sites could not be distinguished. Although the detailed behavior of each statistic is different, their basic response to deviations from the site-frequency predictions of the standard neutral model is the same: they become negative values when there is an excess of either low-frequency or high-frequency polymorphisms and deficiency of middle-frequency polymorphisms. However, what constitues a low or a high frequency polymorphism is different for the different statistics. For D^* and F^* only the most extreme frequency counts ξ_1 and ξ_{n-1} make a negative contribution. Further, the two panels on the right in figure 4.7 show that all the middle frequencies make the same contribution to D^*.

For Tajima's D, there is more than just one low and one high frequency class and, interest-

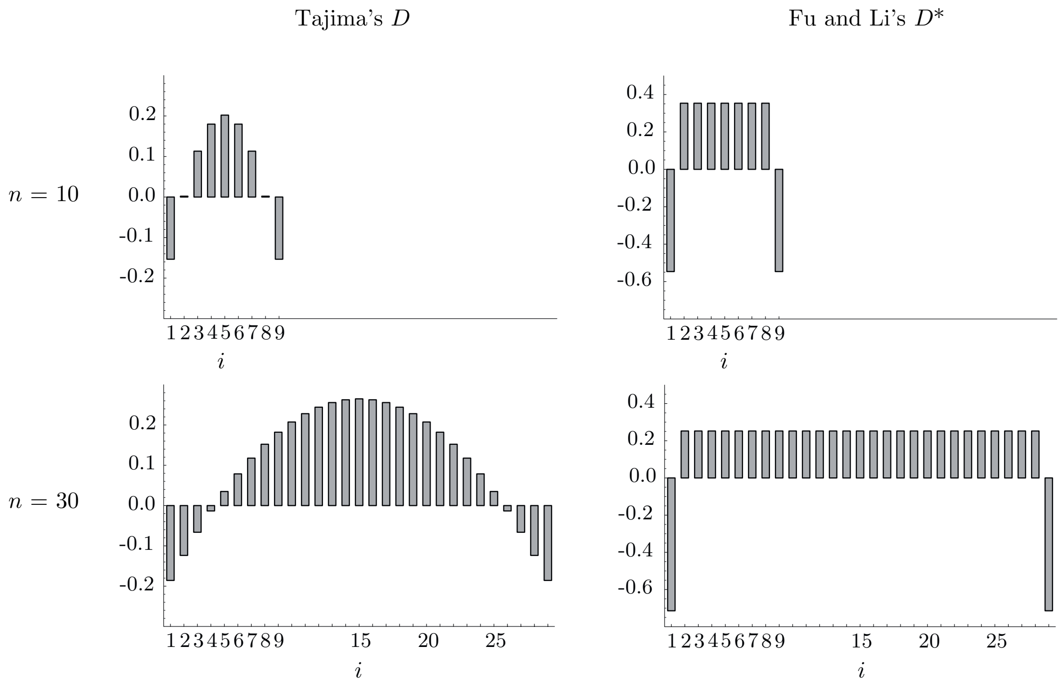


Figure 4.7: Graph of the coefficients of ξ_i in the numerator of Tajima's D and Fu and Li's D^* for two different sample sizes: $n = 10$ and $n = 30$.

ingly, site frequencies which make a positive contribution for smaller samples turn out to make a negative contribution for larger samples. From equation 4.41 we can see that the sign of ξ_i's contribution to D depends on whether $2i(n-i)/(n(n-1))$ is greater than or less than $1/a_1$. The term $2i(n-i)/(n(n-1))$ is largest when i is close to $n/2$, that is for the middle-frequency polymorphisms, while the term $1/a_1$ is a constant and does not depend on i. This creates the potential for the contribution of ξ_i be positive for some sample sizes and negative for others. For example, in the top left panel of figure 4.7 ξ_3 makes a positive contribution to D for a sample of size ten, while in the bottom left panel ξ_3 makes a negative contribution to D for a sample of size thirty. This makes intuitive sense — it seems safe to call $3/30 = 0.1$ a low frequency, while $3/10 = 0.3$ does not seem low at all — but it means that the behavior of Tajima's D in response to deviations from the standard neutral model are less straightforward to predict than those of D^* and F^*. This somewhat complicated response to data may help to explain the finding of Simonsen *et al.* (1995), that D has greater power than D^* and F^* to detect deviations from the standard neutral model.

4.3.2 Demographic History and Patterns of Polymorphism

From the results of the previous section, it is clear that the effects of deviations from the standard neutral model on Tajima's D and on Fu and Li's D^* and F^* can be predicted from an understanding of how alternative demographic processes and events affect the site frequencies ξ_i. With reference to genealogies, we can consider how alterations in the site-frequency spectrum result from either differences in the structure of genealogical trees or the distributions of coalescence times, or some combination of the two. This section considers the potential for non-selective deviations from the standard model to alter the site-frequency spectrum and cause D, D^*, and F^* to deviation from the "neutral" prediction of zero. A scenario invloving natural selection is considered in Section 4.4 in the context of some sequence data from *Drosophila simulans*. In these two sections, rather than deriving equations, we will take a heuristic approach and build upon the intuition gained in this chapter and the last concerning the shapes and sizes of genealogies. Using a simple graphical framework, shown in figure 4.8, we can move beyond site frequencies to generate some powerful qualitative statements about other patterns of DNA sequence polymorphism. In particular, this section considers the effect of alternative demographic scenarios on the variance of the number of segregating sites, and on the tendency for different loci in a genome to show the same patterns of polymorphism among the members of the sample. The latter is a measure of covariation in patterns of polymorphism among loci, or linkage disequilibrium (see Chapter 6).

Our intuition about the connection between genealogies and $E[\xi_i]$ is summarized as follows. Branches in the tree which have i descendents in the sample can contribute polymorphisms to ξ_i, so any process that increases the representation of such branches in the genealogy should also increase $E[\xi_i]$ relative to the predictions of the standard coalescent model. This will in turn cause D, D^*, and F^* to become positive or negative, depending on the value of i. Changes in either the branching structure of the genealogy or the relative lengths of the coalescence times can change the representation of branches which have i descendents in the sample. For example, if one member of the sample is barred from coalescing for some period of time into the past, we can predict there will be an excess of singletons: the mutations that occur on the branch leading to the isolated sample. Even when the branching structure is the same as under the standard model, differences in the coalescence times can affect the site frequencies, but somewhat more subtly, through the ability of branches at a given level in the genealogy to contribute to the site frequency counts. For example, mutations on any of the branches spanning the $n \rightarrow n-1$ coalescent interval must produce singleton polymorphisms (ξ_1), so stretching out this interval relative to the standard model will cause D, D^*, and F^* to become negative. At the opposite extreme, stretching out the final, $2 \rightarrow 1$, coalescent interval will will cause D, D^*, and F^* to be positive since the two branches that exist during this interval contribute equally to ξ_1 through ξ_{n-1} (recall equation 3.39). In general, branches that exist during the $k \rightarrow k-1$ coalescent interval have the potential to contribute to ξ_1, ξ_2, up to ξ_{n-k+1}. These effects of coalescence times are captured in equation 4.24.

Similarly, we have developed an intuition about the effects of genealogies on $\mathrm{Var}[S]$. Fundamentally, this follows from equation 4.8 which expresses $\mathrm{Var}[S]$ in terms of θ, $E[T_{\text{total}}]$ and $\mathrm{Var}[T_{\text{total}}]$. Here we will focus on the influence of $\mathrm{Var}[T_{\text{total}}]$, which means we imagine that θ and $E[T_{\text{total}}]$ are fixed. Under this assumption, there is a direct, linear relationship between $\mathrm{Var}[S]$ and $\mathrm{Var}[T_{\text{total}}]$. Changes in $\mathrm{Var}[T_{\text{total}}]$ are due to changes in the distribution of coalescence times, as opposed to changes in the branching structure of the tree (although we will see in a moment that the two can become conflated when we deviate from the standard model). For example, consider an ancestral process in which the expected lengths of the coalescence times were the same as in the standard coalescent, but in which the variances of coalescence times were different. If the coalescence times became less variable than the exponential distribution of equation 3.9, then $\mathrm{Var}[T_{\text{total}}]$ would be smaller and so would $\mathrm{Var}[S]$. If somehow they became more variable, then $\mathrm{Var}[T_{\text{total}}]$ and $\mathrm{Var}[S]$ would become relatively larger. It is important to say again here — see the text below equation 4.3 — that $\mathrm{Var}[S]$ is the variance we would expect to observe among loci if we took samples from many independent loci, if all the loci had the same value of θ. Although the aim of this section is to illustrate the effects of genealogical processes only, note that differences in θ among loci would have a strong effect on $\mathrm{Var}[S]$.

We have not so far explicitly developed an intuition about covariation in patterns of polymorphism among loci, or linkage disequilibrium, and although this is mostly deferred to Chapter 6, some of the pieces are already in place. For example, in Chapter 1 we learned how the patterns at different sites within a single, non-recombining locus are always compatible: they will never fail the four-gamete test if the infinite-sites mutation model holds. If there are enough polymorphic sites in the sample, there will be sets of sites which show the same pattern of polymorphism because they are due to mutations which occurred on the same branch in the genealogy. Note that a pattern of polymorphism in this context means a particular bi-partition of the sample, and not simply a frequency class which may contain many different bi-partitions. Now, what is the chance that two or more loci, whose genealogies are independent, will show the same pattern of polymorphism? The only way this can occur under infinite-sites mutation is if both genealogies have a branch which partitions the sample in exactly the same way. Except for the n external branches of the genealogy, which induce singleton bi-partitions and which are all present on every genealogy, the random-joining structure of genealogies introduced in Section 3.3.2 makes the chance of shared bi-partitions at different loci very low under the standard coalescent for any but the smallest samples.

Armed with these intuitions, we can predict the effects of deviations from the standard neutral model on the shape of DNA sequence polymorphism. Figure 4.8 shows the hypothetical genealogies of two samples from two different genetic loci. We will assume that the two loci are not physically linked, *i.e.* not on the same chromosome. Under the standard coalescent, and the scenarios in figure 4.8, the genealogies of such loci are independent. They are random draws from the distribution over all possible genealogies, a distribution which will depend on the demographic history of the population. We will see in Chapter 6, that the genealogies of physically linked loci will also be independent if the loci are sufficiently far apart on the chromosome. The thick lines in the figure represent population boundaries and the thin lines show the genealogies of the samples at each locus. Scenario (a) is population growth, in which lineages encounter an ancestral population much smaller than the current population as we follow them back into the past. Scenario (b) is population decline, in which the current population is much smaller than the ancestral population. For the sake of simplicity, we assume that these changes in population size occurred instantaneously. Scenarios (c) and (d) are two different kinds of population subdivision. Scenario (c) is equilibrium migration, in which limited *gene flow* has occurred across a permeable boundary between the populations since a very long time in the past. Scenario (c) is isolation without gene flow, in which the two populations are derived from a single ancestral population at some time in the past.

Due to the dependence of the time scale of the coalescent process on the population size, namely that each pair of lineages coalesces with rate equal to one when time is measured in units of N/σ^2 generations, (a) and (b) differ from the standard coalescent in the relative lengths of the coalescent intervals. However, because the lineages are still exchangeable, the branching structure is the same as in the standard coalescent. In (a), the lineages coalescence on a longer time scale in the recent past, so there will be very few coalescent events between the present and the time the growth occurred. Once back in ancestral population, the effective size drops and the coalescent time scale shortens. One effect of this is to increase the relative lengths of the most recent coalescent intervals, which increases number of low-frequency polymorphisms, so D, D^*, and F^* become negative. In addition, genealogies at different loci tend to be similar in size. In addition, the sizes of the two genealogies will tend to be more similar than under the standard model with constant population size, making Var$[S]$ relatively smaller. Finally, the shortening of the more ancient coalescent intervals means a smaller proportion of polymorphisms will have high mutant counts, so the chance that non-singleton bi-partitions is even lower than under the standard model; see Slatkin (1994). Scenario (b) makes the opposite predictons: D, D^*, and F^* will tend to be positive, Var$[S]$ will be relatively larger, and linkage disequilibrium will be decreased compared to the standard model.

In scenarios (c) and (d), the lineages are not exchangeable. In both cases, coalescent events in the recent past are most likely to be between lineages 1 and 2 or between lineages 3 and 4

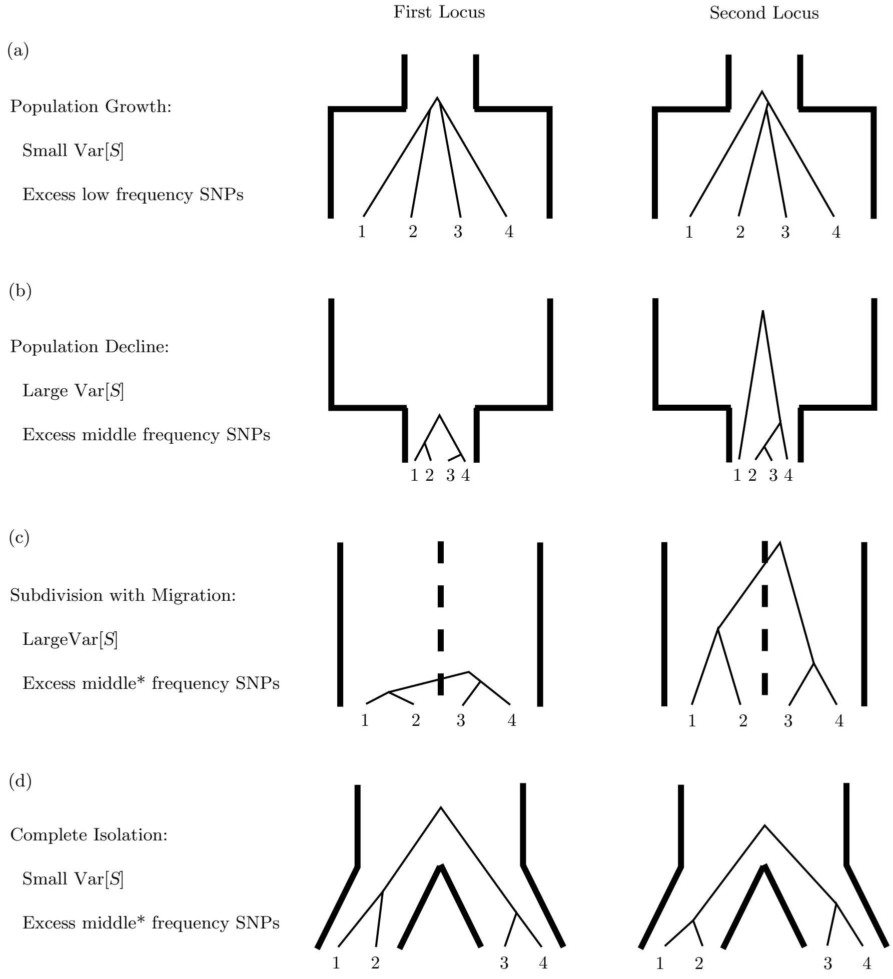


Figure 4.8: The effects of different neutral demographic histories on patterns of DNA sequence polymorphism. The asterisks emphasize that which frequency classes will be overrepresented depends on the sample configuration (see text).

than between any other pair of lineages. In the case of isolation without gene flow, at first there is no possibility of an interpopulation coalescent event, then the lineages become exchangeable when they reach the ancestral population. In the case of equilibrium migration, the lineages can move between populations so that interpopulation coalescent events are possible, but restricted migration makes them unlikely in the recent past. In both cases, samples like those shown in figure 4.8, in which a similar number of sequences is taken from each population, will show an increased proportion of middle-frequency polymorphisms and, thus, positive values of D, D^*, and F^* when these are computed from the entire sample. Note that which frequency classes will be overrepresented depends on the number of sequences sampled from the two populations. If the sample is very unbalanced, for example if a single sequence is sampled from one population and $n-1$ are sampled from the other, then ξ_1 and ξ_{n-1} will be inflated, and D, D^*, and F^* will become negative. In addition, the sign of Tajima's D will depend on both sample sizes. For instance, a positive D in an initially balanced sample will become negative eventually if more and more sequences are sampled from one of the populations (see figure 4.7).

	Data set 1			Data set 2	
	Locus 1	Locus 2		Locus 1	Locus 2
Seq 1	...A...C...	...G...A...	Seq 1	...A...T...	...G...A...
Seq 2	...A...C...	...G...A...	Seq 2	...A...T...	...T...G...
Seq 3	...G...T...	...T...G...	Seq 3	...G...C...	...G...A...
Seq 4	...G...T...	...T...G...	Seq 4	...G...C...	...T...G...

Figure 4.9: Two hypothetical data sets which have the same, positive, value of Tajima's D and Fu and Li's D^*, and F^*.

Scenarios (c) and (d), and population subdivision in general, will not only increase the numbers of sites segregating in particular frequencies in the sample, they will also cause particular patterns of polymorphism to be repeated at different loci. For the samples of four sequences shown in figure 4.8, sequences from every locus in the genome will be expected to show an excess of polymorphic sites at which samples 1 and 2 show one base and samples 3 and 4 show the other base. Not only will the site-frequency count ξ_2 that will be inflated, one specific bi-partition ((1,2),(3,4)) of the sample, out of three possible patterns that contribute to ξ_2, will be overrepresented. Thus, linkage disequilibrium will be greater than under the standard neutral model. This also illustrates how grossly simple statistics like D, D^*, and F^* summarize the information contained in a sample of DNA sequences. Figure 4.9 shows two alternative two-locus data sets which clearly differ in their patterns of polymorphism, but which cannot be distinguished using D, D^*, and F^*. Data set 1 in the figure might be obtained under scenario (c) or (d) while data set 2 could have resulted from scenario (b), in which an increase in ξ_2 is expected but the lineages are exchangeable.

Finally, there is a difference between isolation without gene flow (d) and equilibrium migration (c) in the terms of the variability of the genealogical process. In the case of isolation without gene flow, all interpopulation coalescent events must wait until the lineages reach the ancestral population. Prior to this time, the history follows the standard coalescent. Under equilibrium migration, interpopulation coalescent events can occur at any time if a migration even occurs. It is also possible for the time to the most recent common ancestor of the sample to be very long, because migration is a random process. One effect of this is to make $\mathrm{Var}[S]$ relatively larger under scenario (c) than under scenario (d) in figure 4.8. As will be seen in Chapter 5, the difference between variances of measures of polymorphism in these two demographic scenarios becomes greater when the time of separation is longer under (d) or the migration rate is smaller under (c). In relation to this, we note in closing this section that the predictions about the values of D, D^*, and F^* made above are about average values only. Deviations from the standard neutral model will also change the shape of the distribution of these test statistics, so that the critical values calculated assuming the standard neutral model will not be valid under alternative scenarios even if the expected values of D, D^*, and F^* are equal to zero.

4.4 A Footprint of Positive Selection *Drosophila simulans*

All genetic variation was assumed to be neutral in the demographies of figure 4.8. Here, we will consider one possibility for the action of natural selection, namely a *selective sweep*, and will see evidence for a recent sweep in the genome of *Drosophila simulans.* The term selective sweep refers to the effect on neighboring genomic regions of the fixation of a selectively-favored allele. When a mutation arises and creates a selectively-favored allele, it necessarily occurs on a particular chromosome. It is imbedded in a single string of bases, and is thus physically linked to just one of the alternative bases at each polymorphic site. If the new allele is favored strongly enough to rise in frequency from a single copy to a frequency of one in the population, *i.e.* it becomes fixed, and this happens quickly, then all the particular bases physically associated with it also become fixed. This phenomenon is called genetic *hitchhiking* (Maynard Smith and Haigh, 1974; Kaplan *et al.*, 1989). Variation at the loci linked to the favorable mutation is "swept" from the population, hence the term selective sweep. Since the pioneering work of Kaplan *et al.* (1995), there has been a growing literature on the effects of selective sweeps on genetic variation (Braverman *et al.*, 1995; Stephan *et al.*, 1992; Barton, 1998; Fay and Wu, 2000; Kim and Stephan, 2002; Przeworski, 2002; Kim and Nielsen, 2004).

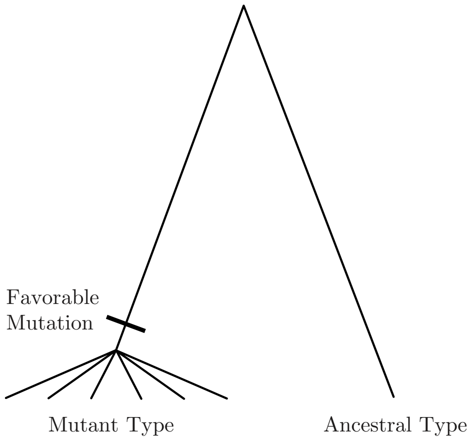
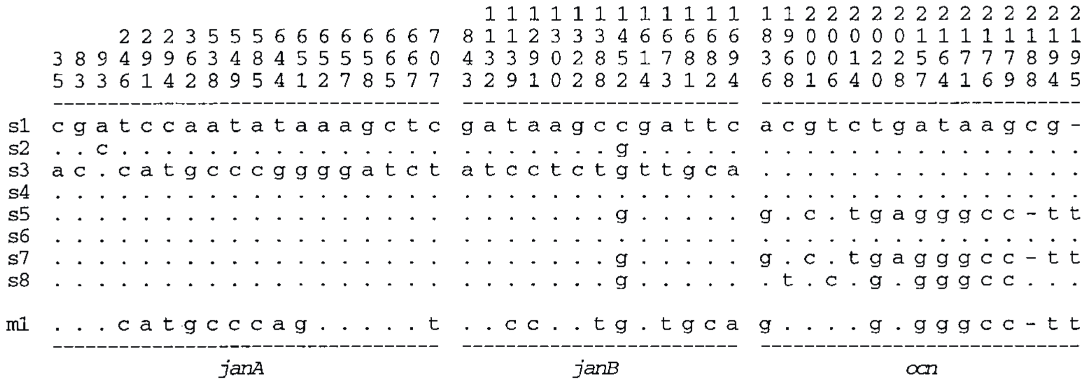


Figure 4.10: A hypothetical genealogy of a sample of size $n = 7$ sequences at a locus which is the focus of a recent selective sweep.

Figure 4.10 shows the type of genealogy one might observe at a locus that has undergone a recent selective sweep. Looking back into the history of the sample, those sequences which possess the mutant, selectively-favored type all must coalesce because the mutant alleles all trace back to a single mutation event on a single chromosome. If the sweep has not yet gone to completion, which means the mutant has not yet fixed, then it is possible, as in figure 4.10, for the sample to contain one or more copies of the ancestral type. A similar pattern of mutant and ancestral copies in the sample can be produced if there is recombination at the locus (see Chapter 6). Before the time of the mutation, in this hypothetical example of a single sweep, there were no selective differences and the remainder of the genealogy is predicted by the standard neutral model. We can think of this as a two-population model in which one of the populations is created from a single gene copie at some time, then grows in size. Eventually, the new 'population' takes over completely as the ancestral type decreases in frequency to zero. The trajectory of the allele frequency can be described by the usual, forward-time population genetic analyses (Kaplan *et al.*, 1989; Barton *et al.*, 2004).

The pattern of polymorphism at a locus subjected to a selective sweep will depend on how recently the sweep occurred, on whether or not it has gone to completion, and on recombination rate in that genomic region. We will return to these topics in Chapters 5 and 6. Consider for the moment genealogies like the one shown in figure 4.10. If the sweep is recent, then few mutations will have occurred along the lineages which descend from the favorable mutant and genetic variation will be lacking among these. Between these sequences and those possessing the ancestral base at the selected site, a more typical level of variation should be observed. Genealogies of this shape will show a site-frequency spectrum which differs greatly from the standard neutral prediction, and should look similar to the pattern generated by population structure, such as shown in figure 4.8(c) and 4.8(d). In particular, some frequency classes will be overrepresented relative the standard neutral predictions, but which clases these are will depend on the numbers of copies of each type of sequence (ancestral and mutant) in the sample. With one ancestral copy and $n-1$ mutant copies, we predict that the unfolded site-frequency spectrum will be U-shaped, with both low-frequency mutants (ξ_1) and high-frequency mutants (ξ_{n-1}) overrepresented.

Parsch *et al.* (2001) recently sequenced parts of a genomic region containing three genes on chromosome 3R in *Drosophila simulans* and found just such a pattern of polymorphism. Their data are reproduced in figure 4.11. The three genes — *janusA*, *janusB*, and *ocnus* — are all expressed in the testes of the flies, *janusB* and *ocnus* exclusively so, while *janusA* is also expressed in other tissues and in both sexes. Recent evidence has uncovered an apparently general pattern of rapid evolution in genes associated with reproduction in biparental organisms (Civetta and Singh, 1998; Swanson and Vacquier, 2002), *i.e.* that the evolution at these loci has proceeded by a series selective sweeps. The data in figure 4.11 are consistent with a recent selective sweep, probably one which is still in progress (Meiklejohn *et al.*, 2004). The first eight sequences in figure 4.11, s1 though s8, are from *Drosophila simulans* and the ninth sequence, m1, is from *Drosophila melanogaster*. The dots represent positions that are identical to those shown in s1. A total of about 1.7 kilobases were sequenced for each sample, and the numbers at the top of figure 4.11 are those of the polymorphic sites, while monomorphic sites are not shown. There is no ambiguity in obtaining a single sequence from a diploid fly in *Drosophila* studies such as this one because each 'sample' is actually the sequence of an isofemale line which was originally taken from the wild then made homozygous by generations of inbreeding.



```
                                         1 1 1 1 1 1 1 1 1 1 1 1   1 1 2 2 2 2 2 2 2 2 2 2 2 2 2
          2 2 2 3 5 5 5 6 6 6 6 6 6 6 7  8 1 1 2 3 3 3 4 6 6 6 6 6   8 9 0 0 0 0 0 1 1 1 1 1 1 1 1
    3 8 9 4 9 9 6 3 4 8 4 5 5 5 5 6 6 0  4 3 3 9 0 2 8 5 1 7 8 8 9   3 6 0 0 1 2 2 5 6 7 7 7 8 9 9
    5 3 3 6 1 4 2 8 9 5 4 1 2 7 8 5 7 7  3 2 9 1 0 2 8 2 4 3 1 2 4   6 8 1 6 4 0 8 7 4 1 6 9 8 4 5
    -----------------------------------  -------------------------   -----------------------------
s1  c g a t c c a a t a t a a a g c t c  g a t a a g c c g a t t c   a c g t c t g a t a a g c g -
s2  . . c . . . . . . . . . . . . . . .  . . . . . . . g . . . . .   . . . . . . . . . . . . . . .
s3  a c . c a t g c c c g g g g a t c t  a t c c t c t g t t g c a   . . . . . . . . . . . . . . .
s4  . . . . . . . . . . . . . . . . . .  . . . . . . . . . . . . .   . . . . . . . . . . . . . . .
s5  . . . . . . . . . . . . . . . . . .  . . . . . . . g . . . . .   g . c . t g a g g g c c - t t
s6  . . . . . . . . . . . . . . . . . .  . . . . . . . . . . . . .   . . . . . . . . . . . . . . .
s7  . . . . . . . . . . . . . . . . . .  . . . . . . . g . . . . .   g . c . t g a g g g c c - t t
s8  . . . . . . . . . . . . . . . . . .  . . . . . . . g . . . . .   . t . c . g . g g g c c . . .

m1  . . . c a t g c c c a g . . . . . t  . . c c . . t g . t g c a   g . . . . g . g g g c c - t t
    -----------------------------------  -------------------------   -----------------------------
                   janA                            janB                            ocn
```

Figure 4.11: The pattern of polymorphism at 44 segregating sites in the *janus-ocnus* region in *Drosophila simulans*. The first eight sequences are from *D. simulans* and the ninth is an outgroup sequence, from *D. melanogaster*. Reproduced from Figure 2 of Parsch *et al.* (2001).

We will focus our attention on the polymorphic sites in *janusA* and *janusB* only, since these can reasonably be assumed to share a single genealogy whereas between these sites and those in *ocnus* some recombination must have occurred to produce different genealogies (Parsch *et al.*, 2001). The latter can be inferred by a variant of the four-gamete test (see Section 1.1), one that takes the outgroup sequence into account. There are a total of $S = 31$ polymorphic sites among these eight sequences from *janusA* and *janusB*, and these show a pattern which is remarkably consistent with the example genealogy shown in figure 4.10. Sequence s3 differs from sequences s1, s2, and s4-s8 at 29 out of these 31 sites. Using the *melanogaster* sequence to estimate the ancestral states at every site, which can be done without error if mutations occurred according to the infinite-sites model, we can see that s3 possesses the derived type at 12 of these 29 sites and retains the ancestral state at the other 17. Again under the infinite-sites model, these would be due to 12 and 17 mutations which occurred, respectively, on the branch leading from s3 to the most recent common ancestor of all eight sequences and on the branch leading from the common ancestor of sequences s1, s2, and s4-s8 to the most recent common ancestor of all eight sequences. The other two polymorphic sites, 93 and 1452, would be the result of mutations which occurred on the branches belonging to the subtree which relates sequences s1, s2, and s4-s8. In all, the 31 sites fall into just three site-requency classes: $\xi_1 = 13$, $\xi_2 = 1$, and $\xi_7 = 17$.

Tajima's test for the *janusA*-*janusB* data in figure 4.11 gives $D = -1.74$ and is significant at the 5% level. Other tests, which unlike Tajima's D (see figure 4.7) distinguish between singletons (ξ_1) and mutations found in $n-1$ copies in the sample, are more powerfull against the alternative hypothesis of a selective sweep (Fu and Li, 1993; Fay and Wu, 2000) and reject the standard neutral model at the 1% level (Parsch *et al.*, 2001). However, as with D, D^*, and F^*, there are still just two ways in which these more powerfull statistics can deviate from the neutral prediction of zero, and a significant result can be due to any of a number of causes (Wakeley and Aliacar, 2001; Przeworski, 2002). Based only on the *janusA*-*janusB* data in figure 4.11, we can question whether the null model was rejected due to selection or because *Drosophila simulans* does not conform to some other assumption of the standard neutral model.

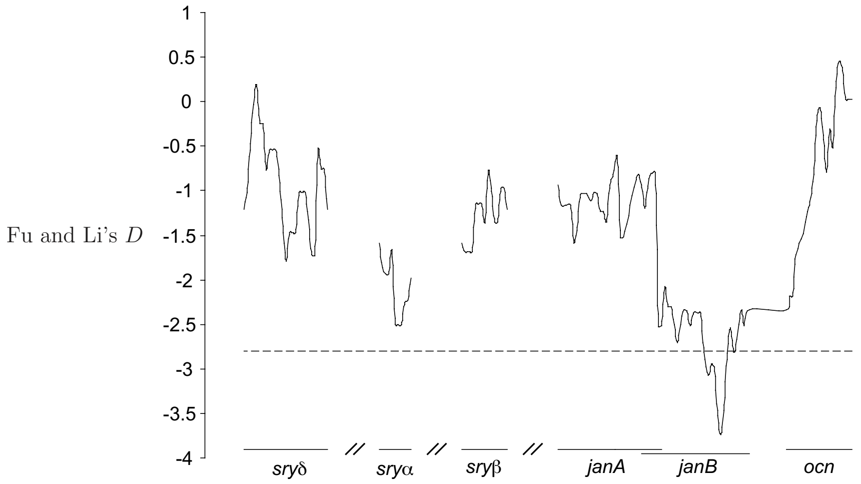


Figure 4.12: Fu and Li's (1993) statistic D in a sliding window of 200? base paris around *janusB*, reproduced from figure 3B in Meiklejohn *et al.* (2004). Fu and Li's (1993) D compares S/a_1 and ξ_1. The dashed line shows the cuttoff for 5% significance.

In fact, there is one very important difference between neutral demographic histories, *e.g.*

the various scenarios depicted in figure 4.8, and selection: selection acts on particular loci while the effects of demographic history are expected to be identical at every locus in the genome. This implies that if most loci are not strongly affected by selection and we scan a genome using Tajima's D or some other statistic, the bulk of loci should display the neutral background and loci under selection should appear as outliers. While this appears promising, it should be clear that it has not changed the basic structure of the problem of detecting selection, which was from the beginning to detect outliers. In addition, some non-trivial questions now confront us, such as how to deal with mutliple-testing problems, which of the myriad possibilities we should adopt as a new null model, and how to compute significance levels in the face possibly numerous unknown demographic parameters.

Hoping that future work will resolve some of these issues, a reasonably well-founded way to use this knowledge of the difference between selection and demography is to first have some biological indication that selection might be operating or have operated at a locus, together with a significant test result. Then, computing the same test statistic at a number of other, presumably neutral loci provides a control which guards against the possibility that the significant result obtained for the locus of interest is due to neutral demographic factors. In a subsequent study of variation in the *janus-ocnus* region, Meiklejohn *et al.* (2004) sequenced thirty-six samples across a broader genomic region, again finding significant results consistent with a selective sweep. Figure 4.12 plots the value of the test statistic across the entire region sequenced. At three flanking loci, *serendipity* δ, *serendipity* α, and *serendipity* β, as well as at *janusA* and *ocnus* in this larger sample, non-significant values were observed, indicating that *janusB* is the focus of selection.

4.5 Exercises

1. What is the probability that the first event in the history of a sample of size n is a mutation event on the lineage ancestral to sequence 1?

2. What is the probability that the first event in the history of a sample of size n is a coalescent event involving sequence 1? As in excercise 1, assume that mutations are also possible.

3. What is the covariance of the numbers of mutations on two lineages that span the interval during which there were i lineages ancestral to a sample?

4. Show that $E[\xi_{n-1}] = \theta/(n-1)$.

5. What is the expected number of number of nucleotide-sequence differences between a single sequence in a sample of size n and the most recent common ancestor of the entire sample?

6. What is the expected number of segregating sites in a sample of size n that are present on both sequence 1 and sequence 2? Note: the mutations may be present on other sequences as well.

7. What is the covariance between the numbers of mutations on two distinct lineages that span the intervals during which there were i and $i-1$ lineages ancestral to a sample

8. Compute the quantitiy $E[k_{ij}, k_{rj}]$ in table 4.1. Hint: it is easier to first compute $\mathrm{Cov}[k_{ij}, k_{rj}]$.

9. What is the joint distribution of the number of alleles and the number of segregating sites for a sample of two sequences?

10. What is the probability that there are three alleles and two segregating sites in a sample of three sequences?