

Natural selection

Gil McVean

Department of Statistics, University of Oxford

February 18, 2002

Contents

| | | |
|----------|---|-----------|
| 1 | Model testing in population genetics | 2 |
| 1.1 | Goodness-of-fit tests | 3 |
| 2 | Frequency-distribution tests | 5 |
| 2.1 | Ewens-Watterson Homozygosity test | 5 |
| 2.2 | Tajima D statistic | 7 |
| 2.3 | Fu and Li D statistic | 8 |
| 2.4 | Fay and Wu H statistic | 8 |
| 2.5 | Factors affecting test power | 9 |
| 3 | Haplotype based tests | 10 |
| 3.1 | Haplotype number | 10 |
| 3.2 | Haplotype diversity | 11 |
| 3.3 | Haplotype partitions | 12 |
| 3.4 | Linkage disequilibrium | 12 |
| 4 | Tests for heterogeneity | 13 |
| 4.1 | The HKA test | 13 |
| 4.2 | The McDonald-Kreitman test | 14 |
| 4.3 | The Lewontin-Krakauer test | 15 |
| 4.4 | Other tests for heterogeneity | 16 |
| 5 | Modelling natural selection | 17 |
| 5.1 | Balancing selection | 17 |
| 5.2 | Selective sweeps | 19 |
| 5.3 | Local adaptation | 21 |
| 5.4 | Distinguishing selection and demography | 21 |

1 Model testing in population genetics

The null model in population genetics describes a world in which natural selection has played no role. A world in which genetic drift, mutation, and recombination are the only forces that influence patterns of genetic diversity. How then can we use this model to find out about selection?

There are effectively only two answers to this question. First, we can try to ask how well the null model fits the data we collect (i.e. carry out a goodness-of-fit test). If we can reject the null hypothesis that our data were generated by our null model, then we are entitled to say that some element of biological reality is missing from the model, and one element we have not included is natural selection. Of course, there are many other factors of biological reality we have not included, any one of which may also have been responsible for the observed deviation. Interpretation of goodness-of-fit tests clearly requires some understanding of the patterns of data one would expect to see under different alternative models.

A natural extension to the idea that we need to model alternative hypotheses in order to understand deviations from the null model is that we can directly compare the likelihood of observing the data under models that ignore and include natural selection. If the model that includes selection is more likely than the one that ignores it, we can claim evidence for the action of natural selection. The difficulty with the likelihood approach is that explicit models of natural selection tend to introduce multiple parameters, which makes likelihood-based inference computationally daunting. In addition, the range of possible selective scenarios that one might consider is enormous, and even if one does find a model that increases the likelihood, it is quite possible that there are other, as yet unconsidered models that have a similar, or greater likelihood, that incorporate some other deviation from the null model (e.g. demographic processes). Despite these difficulties, in the last couple of years, some important steps have been taken into the development of likelihood-based methods for inferring natural selection (Slatkin, 2001; Kim and Stephan, 2000).

The difficulties raised by attempting to infer natural selection from population genetic data are considerable. This lecture aims to follow two key themes in the population genetic treatment of selection; the rejection of the null model, and the modelling of natural selection. Clearly, good tests for selection will be informed by an understanding of the patterns of genetic variability generated by different selective models (and those generated by all other alternative models too!).

1.1 Goodness-of-fit tests

Goodness-of-fit tests are identical in spirit to any other hypothesis test in statistical inference. Hypothesis testing starts out with statement about a parameter in a model, whose validity we wish to assess. The null hypothesis is usually written as H_0 . In order to carry out a test, we identify a statistic of the data, T , and identify a set of values of T (R) for which we will reject the null hypothesis if such a value is observed. By way of example, suppose our hypothesis is that our data are generated by a WF population model with $\theta = 5$. For a sample size of $n = 20$ chromosomes, 95% of observations, were the null hypothesis true, would have between 6 and 48 segregating sites. Consequently, if we actually observed 49 (or more) or 5 (or less) segregating sites, we could reject the hypothesis at the 5% significance level.

The only difference for a goodness-of-fit test, is that the null hypothesis is the statement *The assumed model is correct*. Often, the assumed model will have parameters whose values we have to estimate. Under such conditions, the null hypothesis could be written as *The assumed model is correct, including all the parameter values I have estimated*. Ideally, one would want to test the truth of the statement *The assumed model is correct, though I may not know the exact parameter values*, however, answering that question is more in the realm of Bayesian inference, and for the time being, we will stick with the easier question where we assume that the parameter values are estimated.

What statistic should be used as the basis of a goodness-of-fit test? Unfortunately, this is not an easy question to answer. For cases where alternative models

can be specified, the theory of likelihood provides a simple rationale for identifying best (uniformly most powerful) tests; specifically no test is more powerful than a likelihood ratio test, and sometimes (as in the case of Watterson's homozygosity test), single sufficient statistics can be identified as providing all possible information for a likelihood ratio test. But usually alternative models cannot be specified, or likelihoods cannot be calculated under the different models. In such cases, tests have to be formulated from some knowledge of which properties of the data are likely to be sensitive to departures from the assumed model, and which summary statistics are informative about such properties. Watterson (1977) wrote "*There is no single statistic which is best for testing the most general departures from neutrality*". The practical implication of this is that biological intuition is an essential part of using and constructing informative goodness-of-fit tests in population genetics.

Although there are no single most powerful tests in population genetics, there are very many tests that one might use. The diversity of tests can be confusing, because for the reasons outlined above different tests have been designed to detect different departures from the assumed null model, and consequently use information from different parts of the data. A test designed to detect balancing selection is unlikely to be good for detecting selective sweeps and vice versa. The again, some tests may be informative about both balancing selection and selective sweeps, and a number of other deviations from the null model (e.g. demographic complexities). Rather than classify tests on the basis of what deviations from the null model they are designed to detect, I shall classify by the information used in the tests. On this basis, there are three major classes; frequency distribution tests, haplotype-based tests and heterogeneity tests.

2 Frequency-distribution tests

Many of the best known tests of the null model in population genetics use information on the frequency spectrum of alleles or segregating sites. Tests such as those of Watterson (1977) and Tajima (1989) typically use information from a single locus, summarised in a way that is informative about the frequency distribution. Because the frequency distributions can be derived analytically, it is possible to derive test statistics that have the desirable properties of zero expectation and known variance.

2.1 Ewens-Watterson Homozygosity test

Historically, the first test of the null model (often referred to as neutrality tests, though not to be confused with tests of the neutral theory), was that proposed by Ewens (1972). In this famous paper, Ewens derived the expected number of alleles in a sample for a given θ under the infinite-alleles model. In addition, he showed that conditional on k alleles being observed in a sample of n chromosomes, the probability distribution of the numbers of alleles of each type is given by

$$P(n_1, n_2, \dots | n, k) = \frac{n!}{k! l_k n_1 n_2 \dots n_k} \quad (1)$$

Where n_i is the number of alleles of type i and l_k is a Stirling's number of the first kind (effectively a normalising constant).

Given this result, it is possible to calculate the probability of observing any configuration of allele frequencies. Ewens suggested that one way of summarising the frequency spectrum is to use the information

$$B = - \sum_i x_i \ln x_i$$

Where x_i is the frequency of allele i . When all alleles are at equal frequency, the value of the statistic is large. In contrast, when there is a single high frequency mutation, and all the rest are at low frequency, the information is small. The information in any observed data set can be compared to the distribution of information expected under the WF model. If the observed information is higher than the 97.5

percentile, or lower than the 2.5 percentile, the WF model can be rejected at the 5% level.

Although the information statistic seems a sensible choice as a means of summarising the variance in allele frequency, the choice is essentially *ad hoc*. As mentioned above, when alternative models can be specified explicitly, the theory of likelihood provides a way of identifying the most powerful test possible. Watterson (1977) showed that if the alternative hypothesis is that alleles are maintained by heterozygote advantage (with an equal fitness for all heterozygotes), the effect of such selection is to make allele frequencies more even than expected under neutrality. The likelihood ratio test for balancing selection can be reduced to a function of the population homozygosity (the probability that two alleles picked at random from a population are identical)

$$H = \sum_i x_i^2$$

In other words, homozygosity is a sufficient statistic for testing the hypothesis of symmetric heterozygote advantage (which decreases homozygosity). Although it is unlikely that selection is so even across loci, and the situation is complicated by the fact that allele frequencies have to be estimated from samples, by comparing sample homozygosity to the distribution from Ewen's sampling formula, Watterson's homozygosity test is a powerful way of detecting unexpectedly even allele frequencies, such as those seen at some HLA genes.

With the advent of DNA sequencing studies, neutrality tests based on the infinite-alleles model have largely been superseded by tests based on the infinite-sites model. However, as mentioned last week, when there is no recombination, the number of distinct haplotypes in a sample can be analysed within the framework of the infinite-alleles model. Consequently, it is possible to test for over- or under-dispersion of haplotype frequencies using the Watterson homozygosity test. However, it is worth noting that there is no guarantee that the homozygosity test will be the most powerful test for deviations in the direction of an excess of rare alleles

(increased homozygosity), as may be expected under a model of recurrent mutation to deleterious alleles.

2.2 Tajima D statistic

The first test aimed specifically at testing neutrality in the context of infinite-sites models of sequence evolution was that of Tajima (1989). As was discussed last week, two possible estimators of the population mutation rate, θ , are the average pairwise differences in a sample, π , and the number of segregating sites divided by the Watterson constant S/a_n , where $a_n = \sum_{i=1}^{n-1} 1/i$. Tajima suggested that the difference between these estimators could be used as the basis of a neutrality test

$$D = \frac{\pi - S/a_n}{\sqrt{\text{Var}(\pi - S/a_n)}} \quad (2)$$

Where the difference is normalised by the expected standard deviation of the difference. The test has two particularly desirable properties; the expectation of the numerator is zero, and (because the difference is scaled by the variance) the critical region of the test is little affected by the sample size or number of segregating sites. Tajima (1989) derived the variance analytically, under the assumption of no recombination, and showed how it could be estimated from the data.

What does the Tajima D statistic measure? Watterson's estimator of θ , is only influenced by the number of segregating sites. In contrast, π is sensitive to allele frequencies at segregating sites, such that alleles at intermediate frequencies contribute more than alleles at low frequencies. Consequently, if a sample has an excess of rare variants, π will be less than Watterson's estimator and D will be negative. In contrast, if there is an excess of intermediate frequency variants, π will be greater than Watterson's estimator and D will be positive.

Tajima's D statistic is a very general way of comparing the allele frequency spectrum against the expectations of the null model. It was not designed to pick up any particular deviation from the null model, but it will tend to be negative under selective sweeps (and population growth) and positive under balancing selection

(or population structure with sampling from many populations). It is almost certainly the most widely used neutrality test.

2.3 Fu and Li D statistic

The idea of using different estimators of the population mutation rate as the basis of neutrality tests has been developed in several different ways. A second test was derived by Fu and Li (1993) who showed that the expected number of derived mutations that are present only once in a sample, η_e , is equal to θ . Consequently, it is possible to construct a test statistic in a similar manner to Tajima's

$$D = \frac{S - a_n \eta_e}{\sqrt{\text{Var}(S - a_n \eta_e)}} \quad (3)$$

and a similar statistic (called D^*) if the direction of mutations is not known (in which case the statistic is based on the number of segregating sites at which the rare allele is only represented once - often called singletons - the expectation of which is $\theta n / (n - 1)$).

What does Fu and Li's D statistic measure? In many ways it shares much information with Tajima's D statistic, a negative value indicates an excess of singletons (which would also give a negative Tajima D), and a positive value indicates a lack of singletons (which would typically, though not necessarily, give a positive Tajima D). However, certain population genetic scenarios, particularly selective sweeps, tend to generate an excess of singletons, to which this test is more sensitive than Tajima's D .

2.4 Fay and Wu H statistic

Fu (1995) showed that the expected number of mutations at which the derived allele is represented i times in a sample, ξ_i , is given by

$$E[\xi_i] = \theta / i$$

Clearly, this result provides a whole host of possible estimators of θ , and consequently, an inexhaustible battery of neutrality tests (Fu, 1996b). Notable among

these, is the test proposed by Fay and Wu (2000) which uses an estimator of θ which is heavily influenced by high frequency derived mutations

$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i \quad (4)$$

By comparing π with the new estimator, the test statistic H provides a way of identifying samples in which there is an excess or dearth of high-frequency derived mutations. As discussed later, an excess of high frequency derived mutations may be a signal of a recent selective sweep and a nearby locus, though there are certain demographic situations (strong population structure with uneven sampling from populations) that can also give the same pattern. Unlike the previous tests, the variance of the test statistic has to be estimated by stochastic simulation.

2.5 Factors affecting test power

The rationale behind the three statistics just presented is that by comparing summary statistics of the data that use different sources of information, unusual discrepancies in the patterns observed can be detected (and hopefully associated with different departures from the null model). What makes these statistics particularly attractive is that they focus on specific aspects of data that are believed *a priori* to be sensitive to deviations from the null model in a predictable and intuitive manner.

However, there is no guarantee that such tests will have sufficient power to detect the desired deviations from the null model. Indeed, simulation studies (e.g. Fu, 1996b; Wall, 1999; Przeworski, 2002) have shown that tests often have little power to detect the deviations they were designed to pick up. The single greatest factor affecting test power is the number of mutations in the sample - more segregating sites, more power. So it is worth noting that an increase in the region sequenced, rather than the number of chromosomes, is more likely to identify new mutations (Pluzhnikov and Donnelly, 1996). An accurate modelling of recombination also usually increases test power. If there is recombination, different regions of a gene have different histories, sampled from the underlying distribution of possible histo-

ries. So the effects of evolutionary stochasticity are averaged out across a gene by recombination, which consequently means that the variance of statistics is reduced relative to the case of no recombination.

The practical implication of recombination is that the critical regions for a test (the rejection region) depend on the population recombination rate, $4N_e r$. Given an estimate of $4N_e r$, the critical regions of a test can be estimated by stochastic simulation. However, obtaining an accurate estimate of $4N_e r$ is difficult (Hudson, 1983; Fearnhead and Donnelly, 2001), so it is safer to take a lower bound for the parameter.

Finally, it is worth noting that stochastic simulation enables the use of more complex tests of the null model, for example tests that contrast multiple summary statistics. The only difficulty is interpreting such tests. In effect, the only bound on the design of neutrality tests is imagination.

3 Haplotype based tests

Tests that use information on the frequency spectrum of segregating sites are clearly ignoring a significant source of information; namely associations between alleles, or the haplotype structure of the sample. However, because haplotype structure is influenced by recombination, as well as selective and demographic forces, it is necessary to have an accurate estimate of the population recombination rate at the loci in question before the null model can be tested. If an estimate of $4N_e r$ is obtainable (for example, if there is no recombination, or the data are on unlinked SNPs, or $4N_e r$ can be estimated by a population genetic method) a number of tests of the null model are available.

3.1 Haplotype number

A simple test of the null model based on haplotype structure is to consider whether there are more or less haplotypes than are expected given the number of segregat-

ing sites. The two extreme cases are where the data is partitioned into two or a few distinct haplotypes, such that the number of haplotypes is much less than the number of segregating sites, and where each segregating site defines a new haplotype. The former case may be indicative of balancing selection, or population structure, while the latter situation may be expected under selective sweeps or population growth. Note the the number of haplotypes cannot exceed $S + 1$.

Tests of haplotype number have been proposed that condition on S (Depaulis and Veuille, 1998) and θ (Fu, 1996a). Given that the number of segregating sites sets the limits for the number of haplotypes, it makes sense to condition on S , however, it should be noted that the simulation scheme used by Depaulis and Veuille, following Hudson (1993), treats S as a parameter. Given that S is a statistic, and θ is the parameter, simulations conditioning on S should weight individual samples by the probability of observing S segregating sites, given an estimate of θ . In practice, there is little difference in the critical regions for statistics defined by the two simulation schemes. A factor of much greater importance is the estimate of the population recombination. There are a number of possible schemes for estimating $4N_e r$ from population genetic data (Hudson, 1987; Fearnhead and Donnelly, 2001; Hudson, 2001; McVean et al., 2002) which differ considerably in estimator variance and computational tractability (Wall, 2000).

3.2 Haplotype diversity

As indicated previously, Watterson's homozygosity test can be applied directly to haplotypes in the case of no recombination. When there is recombination, a similar test can be designed that compares haplotype diversity to the distribution expected under the null model given an estimate of θ and $4N_e r$, and conditioning on the number of segregating sites (Depaulis and Veuille, 1998). Given S , haplotype diversity and haplotype number share much information; haplotype diversity is high under balancing selection (certain demographic models) and low under selective sweeps (population growth).

3.3 Haplotype partitions

A related test was designed by Hudson *et al.* (1994) to detect subsets of the data that showed unusual patterns. Called the haplotype partition test, the idea is to ask whether given a sample of size n with S segregating sites, there is likely to exist a subset of sequences of size i which has j or fewer segregating sites.

Although the test sounds complex, the test is aimed to pick up incomplete selective sweeps, or local adaptation, the effect of which is to reduce diversity in a subset of the population, without affecting the rest. Such processes create strong haplotype structure for only a subset of the sample. A similar idea has been used to identify domains within a larger region that show unusually strong haplotype structure (Wall, 1999; Andolfatto *et al.*, 1999).

3.4 Linkage disequilibrium

Given that the key point in using haplotype data rather than just allele frequency spectrums is to use the information gained from considering associations between alleles, it makes sense to consider such associations directly. The Z_{nS} test of Kelly (1997) considers the average across all pairs of segregating sites of one measure of allelic association (or linkage disequilibrium)

$$Z_{nS} = \frac{2}{n(n-1)} \sum_{ij, i \neq j} r_{ij}^2$$

Where $r_{ij}^2 = D_{ij}^2 / (p_i q_i p_j q_j)$, the square of the correlation coefficient of the two alleles at sites i and j (Hill and Robertson, 1968). Alternative tests could be constructed from other summaries of linkage disequilibrium such as the $|D'|$ statistic of Lewontin (1964). Although the idea of directly using linkage disequilibrium is appealing, the variance of measures of linkage disequilibrium is large, and is affected by allele frequencies. In addition, linkage disequilibrium is also influenced by recombination, hence an accurate estimate of $4N_e r$ is necessary to define critical regions for test statistics.

The dependency of haplotype statistics on allele frequencies and recombination rates clearly makes the construction of good haplotype-based tests of the null model problematic. Ideally, we would hope to consider the effects of the frequency distribution and allelic associations separately. By analogy with the tests which compare estimators of θ , one approach would be to construct tests that compare different estimators of $4N_e r$, or the joint distribution of estimators of θ and $4N_e r$. However, because of the technical difficulties of estimating $4N_e r$, such tests are not yet available.

4 Tests for heterogeneity

All the tests of the null model discussed so far have considered patterns of diversity at a single locus, and assumed that all sites at the locus are equivalent and subject only to neutral mutations. However, there is an extremely important class of tests of the null model which use contrasts between two or more loci (or classes of mutations) as a source of information. Deviations from the null model are identified by greater variation between loci (or classes of mutations) than expected. An important element in several of such tests is the contrast between patterns of polymorphism and divergence.

4.1 The HKA test

Perhaps the best known heterogeneity test in population genetics is the Hudson-Kreitman-Aguadé test (1987). The test considers patterns of polymorphism and divergence at two loci, that have different mutation rates θ_1 and θ_2 . Suppose we have collected n_1 and n_2 chromosomes at each locus, and an outgroup sequence for both. Under the null model, the expected number of fixed differences is

$$E[d_i] = \theta_i \left(\tau - \frac{n_i - 1}{n_i} \right)$$

Where τ is the divergence time between the outgroup and sample species scaled in units of $2N_e$ generations; the effective population size of the sample species. The

expected number of segregating sites is given by the Watterson formula $E[S_i] = \theta_i a_n$. Given information on the number of polymorphic and fixed differences at each locus, the three parameters (θ for each locus and τ) can be estimated by a least-squares fit (maximum likelihood is another possibility). Hudson *et al.* suggested that the sum of the square of differences between observed and expected values should be approximately χ^2 distributed under the null model (the exact distribution can be estimated by stochastic simulation). By contrasting a locus of interest against a reference null locus, the HKA test provides a potentially powerful method for detecting loci that demonstrate unusual patterns of polymorphism.

The HKA test can easily be extended to comparing more than two loci, and patterns of polymorphism in two species. The test was originally formulated to test the null model at the *Adh* locus of *Drosophila melanogaster* (Hudson *et al.*, 1987). This locus has two different electrophoretic variants (called Fast and Slow alleles) which are associated with a two-fold difference in enzyme activity (the result of variation in the regulatory sequence with which the electrophoretic variants are in strong association). The polymorphism exists in a longitudinal cline in North America, with the fast allele being found further north and at higher altitudes (Berry and Kreitman, 1993). It is most likely that the polymorphism is maintained by balancing selection (local selection balanced by migration). In line with this hypothesis, the locus shows an excess of polymorphisms relative to reference loci.

4.2 The McDonald-Kreitman test

Another test originally formulated to analyse the *Adh* locus of *D. melanogaster* is the McDonald-Kreitman or MK test. Rather than contrast patterns of polymorphism and divergence at different loci, the MK test contrasts patterns of polymorphism and divergence for different types of mutations (for example, synonymous and nonsynonymous) at a single locus. A contingency table of counts of polymorphisms and fixed differences is obtained from the data and non-independence between the rows and columns is assessed by means of a Fisher's exact test (or a

permutation test).

The MK test is actually a much more general test than the HKA test, as it is insensitive to almost all deviations from the null model, except natural selection. The central idea is that if the two classes of mutations are equivalent, then patterns of polymorphism and divergence for the two will be the same (as long as they are interspersed). Even if there has been population growth, or geographical structuring, the relative proportions of polymorphisms and fixed differences should be the same under the assumption of homogeneity (though allele frequency distributions and haplotype patterns may well indicate a deviation from the null model).

How should deviations from the assumption of homogeneity be interpreted? The key is that one class of mutations must be deemed neutral (or less strongly selected) on an *a priori* basis; for example synonymous mutations. If such a class can be identified, an excess of fixed differences in the putatively selected class is typically indicative of adaptive evolution (because advantageous mutations rapidly fix in populations). In contrast, an excess of segregating mutations can be interpreted as evidence of purifying selection. Deleterious mutations (as most amino acid changing mutations are expected to be) may contribute to polymorphism, but are very unlikely to become fixed in a population.

4.3 The Lewontin-Krakauer test

The MK test can potentially be used to identify genes that have undergone recurrent adaptive evolution (through an excess of fixed differences). However, there are many instances where a single instance of local adaptive evolution is a more realistic model (for example as may apply to the null *Duffy* allele - which confers resistance to the *Plasmodium vivax* agent of malaria - which probably underwent a selective sweep around the time of the human transition to agriculture, and the spread of malaria). What type of test may be used to identify such events?

The obvious consequence of local adaptation is that a subset of globally sampled sequences will show a markedly different pattern of polymorphism than the

rest of the sample, and this subset will be geographically restricted. Such loci may therefore be identifiable by the demonstration of an unusually high level of geographic structuring (in addition to the haplotype-based tests considered previously).

The Lewontin-Krakauer test was originally formulated to detect such outliers (Lewontin and Krakauer, 1973). Geographic structuring is assessed by Wright's fixation index F_{ST} , which will be discussed in the next lecture, and loci with unusually high levels of the statistic can potentially be identified. Actually carrying out the test is difficult, because ideally we wish to test the hypothesis *Is the variation in F_{ST} between loci greater than that expected under a model of geographical structure alone*. However, there is an essentially infinite number of possible models of geographical structure, and statistical inference under any but the simplest is currently intractable. For this reason, the Lewontin-Krakauer test is impossible to use, however one way round the problem was suggested by Taylor *et al.* (1995) who used the *a priori* hypothesis that the locus they were interested in (one that conferred insecticide resistance) should show greater geographic subdivision than a set of reference loci. That such a pattern was observed was taken as statistical evidence for the hypothesis of local adaptation.

4.4 Other tests for heterogeneity

Just as there is a host of possible tests of the null model at single loci, there is a potentially inexhaustible supply of tests for heterogeneity between loci. A simple approach is to use the variance between loci in some summary statistic as a test statistic in its own right. For example, Frisse *et al.* (2001) showed unexpectedly high levels of variation in Tajima's D statistic between anonymous non-coding loci in a human population sample from China. Tests based on an ensemble of summary statistics are potentially powerful for detecting subtle deviations from the null model that could not be detected on a locus-by-locus basis.

Finally, it is worth noting that as the scale of available sequence information

increases, it will be possible to compare patterns of variation at the locus of interest to the empirical distribution. Although there is no guarantee that the tails of the empirical distribution are in any way unusual, the *a priori* expectation is that if there is a small subset of loci subject to recent selective forces, these are likely to be at the extreme of the empirical distribution.

5 Modelling natural selection

Throughout the lecture, an implicit understanding of the patterns of polymorphism expected under different selective regimes has been assumed. Intuitively, it is quite easy to describe the types of pattern one might expect to find, but it is only through explicit modelling of the selective process that such intuition can be tested. Furthermore, explicit modelling, enables us to identify important parameters in natural selection, and potentially derive ways of estimating such parameters from empirical data. In this last section, I will outline simple population genetic models for balancing selection, selective sweeps, and local adaptation, and describe some of the patterns of polymorphism they are likely to generate.

5.1 Balancing selection

Historically, balancing selection is the form of natural selection that has received the most attention. Balancing selection can come about through heterozygote advantage (thought to be the case for HLA class I alleles, and mutations causing sickle-cell anaemia), frequency-dependent selection (as for self-incompatibility loci in plants) or a balance between local directional selection and migration (for example the *Adh* cline in *Drosophila melanogaster*). Whatever the cause, balancing selection causes multiple alleles to be maintained in a population, often at fairly constant frequencies.

There have been two major classes of models for loci under balancing selection. Wright (1931) and Watterson (1977) considered the case of multiple alleles, which

can drift in frequency (and even be lost). Recently, models of balancing selection with two alleles maintained at a constant frequency over time have been used to caricature the selective process.

Using diffusion theory, Watterson and Wright derived several properties of the frequency distribution of multiple alleles subject to balancing selection. However, in terms of interpreting patterns of putatively neutral diversity at sites linked to the position subject to selection, the two allele models have been more useful. Under the assumption that there is a single selected site held at constant frequency, the genealogical process at linked sites can be treated as a structured coalescent process (Hudson, 1990) in which chromosomes carrying the same selected allele can coalesce, but those carrying different selected alleles cannot. Chromosomes can only move from one background to the other by recombination at rate rx_i per generation, where x_i is the frequency of the alternative background.

This model shares many similarities to the coalescent process in geographically structured populations, where chromosomes can only coalesce within the same population, and movement between populations occurs by migration (Hudson, 1990) at a rate m per generation. Several properties of the structured coalescent are known; for example the expected coalescence time for a pair of chromosomes sampled from the same genetic background is identical to that in the null model, 2 (scaled in units of $2N_e$ generations), and the expected coalescence time for a pair sampled from different backgrounds is $2(1 + 1/R)$.

What does this result tell us about patterns of polymorphism under balancing selection? Three conclusions can be drawn. First, balancing selection effectively structures polymorphism into two subpopulations. Second, while diversity within each class should be similar to that in other parts of the genome, between genetic backgrounds there will be an excess of polymorphism. Finally, because the increase in coalescence times for chromosomes on different backgrounds is a decreasing function of R , the increase in diversity will peak at the balanced site, and decay as a function of $1/R$. Such a peak of polymorphism can be seen at the *Adh*

locus of *Drosophila melanogaster* (Hudson et al., 1987).

The key feature of genealogies at balanced loci is that coalescence typically occurs rapidly within the classes, but there is a long waiting time for coalescence events between classes. A consequence of such topologies is that the internal branches of an n -coalescent are much longer than the terminal branches which leads to an excess of mutations at intermediate frequencies (hence positive Tajima D values), and sets of mutations at different sites that have identical allele frequencies, hence strong linkage disequilibrium (mutations which are fixed between the classes). In short, balancing selection leads to high levels of diversity, and strong haplotype structure, which both decay with distance from the selected locus. An example of a locus that displays exactly these properties is β -globin (Harding et al., 1997)

5.2 Selective sweeps

When an advantageous mutation appears in a population and sweeps through to fixation, it leaves behind a trail in patterns of linked neutral diversity (Maynard Smith and Haigh, 1974). Suppose a favourable mutation has just become fixed in a population. Looking back in time, we can think of the coalescent process at sites linked to the selected mutation in a similar manner as for balancing selection. That is, there is a structured coalescent, in which chromosomes can only coalesce if they carry the same allele at the selected site, and where recombination allows linked sites to move between genetic backgrounds (Braverman et al., 1995). The only difference from balancing selection is that the population sizes of the two backgrounds are not constant, rather the selected class has undergone rapid expansion, and the non-selected background has undergone a rapid decrease in population size.

Close to the selected mutation, all chromosomes will coalesce on the selected background. This has two key effects on the genealogy; first the coalescence times are much less than under the null model, so diversity is reduced. Because the

rate of coalescence is a function of the reciprocal of the population size, the rate of coalescence accelerates as you look back in time, so the time intervals between successive coalescent events gets shorter (rather than longer). So the second feature of genealogies under hitch-hiking is that if there are mutations, they are likely to be recent, hence are likely to be at a low frequency in the sample. Consequently, hitch-hiking is associated with negative Tajima D and Fu and Li D statistics (Braverman et al., 1995; Fu, 1996a)

A third feature of genealogies under selective sweeps only occurs at sites some genetic distance from the selected site (Fay and Wu, 2000). As mentioned above, looking back in time, chromosomes can potentially move to the non-selected background through recombination. Suppose that only a single chromosome escapes in such a manner, then the ancestor of the chromosomes within the selected class and the escaped chromosome can only coalesce prior to the origin of the selected mutation. Because there is no selection operating at this point in time, the number of mutations that occur in this portion of the genealogy is simply the expected number for a pair of chromosomes under the null model.

The key point about such genealogies is that they are very unbalanced, with $n - 1$ and 1 descendants for the two most ancient lineages. Furthermore, because the majority of mutations occur in the pre-selection portion of the genealogy, there is an excess of mutations for which the derived state is at high frequency ($n - 1$ of n sequences) (Fay and Wu, 2000). Fay and Wu's H test, discussed earlier, is specifically designed to detect such high frequency derived mutations.

We noted for the case of balancing selection that in many ways it mimics population structure. Likewise, selective sweeps mimic population growth (at the selected site). Are there demographic processes that can mimic the genealogical process at sites partially linked to strong favourable mutations? The situation most likely to give a similar pattern is strong population structure with most chromosomes being sampled from a single population. Under such conditions, chromosomes within the population will rapidly coalesce, but the time to the sample

MRCA may be considerable.

5.3 Local adaptation

A third model of natural selection that we may wish to model is local adaptation, where a mutation arises that is locally favourable, so sweeps to fixation in some populations, but does not become globally fixed. An example of such a situation is the *FY*O* allele at the *Duffy* locus in humans (Hamblin et al., 2002).

In many ways, the patterns of genetic diversity expected under local adaptation are a mixture of those generated by balancing selection and selective sweeps; geographically localised reductions in diversity and excesses of rare mutations, but also strong haplotype structure (at least for a subset of the chromosomes sampled). As discussed with reference to the Lewontin-Krakauer test, high levels of geographic structuring may be expected at such loci.

5.4 Distinguishing selection and demography

It should be clear from the above discussion of how selection can be modelled, that there are many strong similarities between the effects of natural selection and various demographic processes on patterns of genetic variability. How then can one be sure that the patterns observed are the result of selection?

It is, of course, impossible to be certain. If there is a strong *a priori* reason for thinking the locus in question has been subject to selection, it seems parsimonious to accept such patterns as evidence for selection. However, for loci where the selective forces are obscure, the only option is to contrast patterns of polymorphism at multiple loci. The ideal choice for such reference loci are genomic regions with no known function (Frisse et al., 2001). If the same patterns are observed at the test and reference loci, then demographic factors are the most likely cause.

References

- Andolfatto P, Wall JD, Kreitman M (1999). Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. *Genetics* 153:1297–1311.
- Berry A, Kreitman M (1993). Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. *Genetics* 134:869–893.
- Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995). The hitchhiking effect on the site-frequency spectrum of DNA polymorphisms. *Genetics* 140:783–796.
- Depaulis F, Veuille M (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol.* 15:1788–1790.
- Ewens WJ (1972). The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3:87–112.
- Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
- Fearnhead P, Donnelly PJ (2001). Estimating recombination rates from population genetic data. *Genetics* 159:1299–1318.
- Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001). Gene conversion and difference population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* 69:831–843.
- Fu YX (1995). Statistical properties of segregating sites. *Theor. Pop. Biol.* 48:172–197.
- Fu YX (1996a). New statistical tests of neutrality for DNA samples from a population. *Genetics* 143:557–570.

- Fu YX (1996b). Statistical tests of neutrality against population growth, hitchhiking and background selection. *Genetics* 147:915–925.
- Fu YX, Li WH (1993). Statistical tests of neutrality of mutations. *Genetics* 133:693–709.
- Hamblin MT, Thompson EE, Di Rienzo A (2002). Complex signatures of selection at the Duffy blood group locus. *Am. J. Hum. Genet.* 70:369–383.
- Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, et al (1997). Archaic African *and* Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* 60:772–789.
- Hill WG, Robertson AR (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226–231.
- Hudson RR (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* 23:183–201.
- Hudson RR (1987). Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* 50:245–250.
- Hudson RR (1990). Gene genealogies and the coalescent process. In D. Futuyama and J. Antonovics (Eds.), *Oxford Surveys in Evolutionary Biology*, Volume 7, pp. 1–44. Oxford University Press.
- Hudson RR (1993). The how and why of generating gene genealogies. In A. G. Clark and N. Takahata (Eds.), *Mechanisms of Molecular Evolution*, pp. 23–36. Tokyo: Japanese Scientific Societies Press.
- Hudson RR (2001). Two-locus sampling distributions and their application. *Genetics* 159:1805–1817.
- Hudson RR, Bailey K, Skarecky D, Kwiatkowski J, Ayala FJ (1994). Evidence for positive selection in the superoxidase dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* 136:1329–1340.

- Hudson RR, Kreitman M, Aguadé M (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
- Kelly JK (1997). A test of neutrality based on interlocus associations. *Genetics* 146:1197–1206.
- Kim Y, Stephan W (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* 155:1415–1427.
- Lewontin RC (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49:49–67.
- Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74:175–195.
- Maynard Smith J, Haigh J (1974). The hitch-hiking effect of a favorable gene. *Genet. Res.* 23:23–35.
- McVean G, Awadalla P, Fearnhead P (2002). A coalescent-based method for detecting and estimating recombination rates from gene sequences. *Genetics* in press.
- Pluzhnikov A, Donnelly P (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144:1247–1262.
- Przeworski M (2002). The signature of selection at randomly chosen loci. *Genetics* in press.
- Slatkin M (2001). Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.* 78:49–57.
- Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
- Taylor MFJ, Shen Y, Kreitman ME (1995). A population genetic test of selection at the molecular level. *Science* 270:1497–1499.

Wall JD (1999). Recombination and the power of statistical tests of neutrality. *Genet. Res.* 74:65–79.

Wall JD (2000). A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* 17:156–163.

Watterson GA (1977). Heterosis or neutrality? *Genetics* 85:789–814.

Wright S (1931). Evolution in Mendelian populations. *Genetics* 16:97–159.