
Detecting Selection at the Molecular Level

MICHAEL W. NACHMAN

Selection acts on the phenotype but can leave its signature at the molecular level. For example, if a new mutation arises at a locus and confers a fitness advantage, it may spread quickly to all individuals in a population and thereby eliminate all pre-existing variation at the locus. Because selection is a *deterministic* process, patterns of DNA variation caused by selection can often be distinguished from patterns caused by other processes, such as genetic drift or migration. This basic idea has led to a proliferation of theoretical and empirical studies aimed at detecting the effects of selection at the molecular level.

This approach for studying selection is very different from more direct approaches that focus on observations of phenotypic change over several generations (e.g., Grant & Grant 2002), and it has both advantages and disadvantages. One advantage is that selection coefficients that are small but biologically meaningful may be impossible to measure directly but will still leave an imprint in patterns of DNA sequence variation. Another advantage is that this approach gives us a picture of selection over evolutionary, as opposed to ecological, time scales. In other words, we can detect selection that has happened many generations ago. This approach also allows us to ask questions about selection, even without knowledge of the phenotype. For example, we can ask how much selection has occurred in the recent history of a species even if we are unaware of the agent of selection. Finally, this approach holds promise that we might discover genes associated with the evolution of novel traits. One of the chief disadvantages is that when we do find a signature of selection, it is very difficult to make the link to the phenotype, much less to the environment.

The literature contains many examples of genes or genomic regions that have been clearly influenced by selection, but where the specific polymorphisms under selection and their functional consequences are unknown.

The neutral theory of molecular evolution (Kimura 1983) serves as the null hypothesis for most statistical tests in this field, and so I first describe it briefly below. Next, I describe three simple models of selection and the patterns of DNA sequence variation expected under each. Many statistical tests for detecting selection at the molecular level have been developed in the last two decades. I have grouped these tests into five basic kinds. I describe them and present a few key applications of each kind of test to real data. The genomics revolution has opened up the possibility of looking for selection at virtually every gene in a genome; I discuss the promise and difficulties of these genome-wide studies. Finally, I present a few particularly compelling case studies of selection at the molecular level.

CONCEPTS

The Neutral Theory of Molecular Evolution

Proposed by both Kimura (1968) and King and Jukes (1969) and later developed in great detail by Kimura (1983), the neutral theory states that most mutations are deleterious, but of the remaining ones, a negligible proportion are advantageous and the vast majority are neutral with respect to fitness. The fate of these neutral mutations is governed by

random genetic drift. According to this theory, it is these neutral mutations that we see as polymorphisms within species or as fixed differences between species. Deleterious mutations, although far more numerous, are eliminated quickly by selection and thus contribute neither to polymorphism nor to divergence. The neutral theory accounts for many observations in population genetics and molecular evolution, and it also has tremendous heuristic value. Because it is simple, mathematically tractable, and makes several straightforward predictions, it serves as the null model in many statistical tests. These tests are usually based on one of three mutational models. The infinite alleles model (Kimura & Crow 1964) assumes that each new mutation creates a new allele in a population. The infinite sites model (Kimura 1969a) is typically used to model DNA sequence evolution and assumes that each new mutation occurs at a site that has not previously mutated. The stepwise mutation model (Ohta & Kimura 1973) was originally developed for allozyme data but is now frequently used to model microsatellite loci; it assumes that mutations arise in steps, and that alleles can only mutate to neighboring states.

One key prediction of the neutral theory concerns the amount of variation expected within and between species. Under a neutral model, the amount of genetic variation in a population represents a balance between the input of new mutations and their loss or fixation due to genetic drift. For example, under the infinite sites model, at mutation–drift equilibrium the expected heterozygosity is $4N_e\mu$, where N_e is the effective population size and μ is the neutral mutation rate (Kimura 1969a). The rate of evolution along a lineage, ν (i.e., the substitution rate or the fixation rate) is equal to the number of new mutations entering a population each generation, $2N_e\mu$, times the probability of fixation for a new mutation, $1/2N_e$. Therefore, the rate of evolution (ν) is simply equal to the mutation rate (μ) and does not depend on the population size (Kimura 1983). This simple but important result implies that the amount of divergence between orthologous copies of a gene will depend only on the mutation rate and the amount of time separating the copies. Thus, the neutral theory predicts that the ratio of variation within a species to divergence between species will be the same for different genes, since both depend on the mutation rate (Figure 7.1A).

A second important prediction concerns the distribution of allele frequencies at mutation–drift

equilibrium. At steady state, the expected heterozygosity can be specified for any mutational model, but new alleles are constantly entering the population due to mutation and are leaving due to stochastic fixation or loss. Nonetheless, there is an expected distribution of allele frequencies that can be described based on the observed heterozygosity and the sample size. This “neutral distribution” consists of many low-frequency alleles and an increasingly smaller number of increasingly higher frequency alleles (Figure 7.1B).

Both these predictions serve as the basis for some of the statistical tests described below. There are many other predictions that follow from the neutral theory that have also been developed into statistical tests, and some of these are described in detail elsewhere (e.g., Kreitman 2000; Yang & Bielawski 2000; Luikart et al. 2003).

Models of Selection

At the molecular level, selection can act to fix alleles (positive, directional selection), eliminate alleles (negative or purifying selection), or maintain two or more alleles in a population (which I will refer to as “balancing selection” to include heterosis, spatially or temporally varying selection, and any other selection that maintains variation). These different forms of selection will affect the shape of the genealogy of alleles in a population: positive and purifying selection will produce shallower genealogies and balancing selection will produce deeper genealogies (Figure 7.2). Positive and purifying selection will reduce genetic variation since shallower genealogies have less time over which mutations can arise, and balancing selection will increase genetic variation since deeper genealogies have more time over which mutations can arise. The distribution of allele frequencies is also skewed by these different genealogies: in general, positive or purifying selection can produce an excess of low-frequency alleles while balancing selection can produce an excess of intermediate-frequency alleles. Finally, positive selection will lead to increased rates of evolution at the sites under selection.

An important concept for studies of selection at the molecular level is that linked sites will have correlated evolutionary histories. Thus, selection can affect patterns of DNA sequence variation at genes near to those that are the target of selection. This suggests that we might be able to find evidence of selection in the genome even when the targets

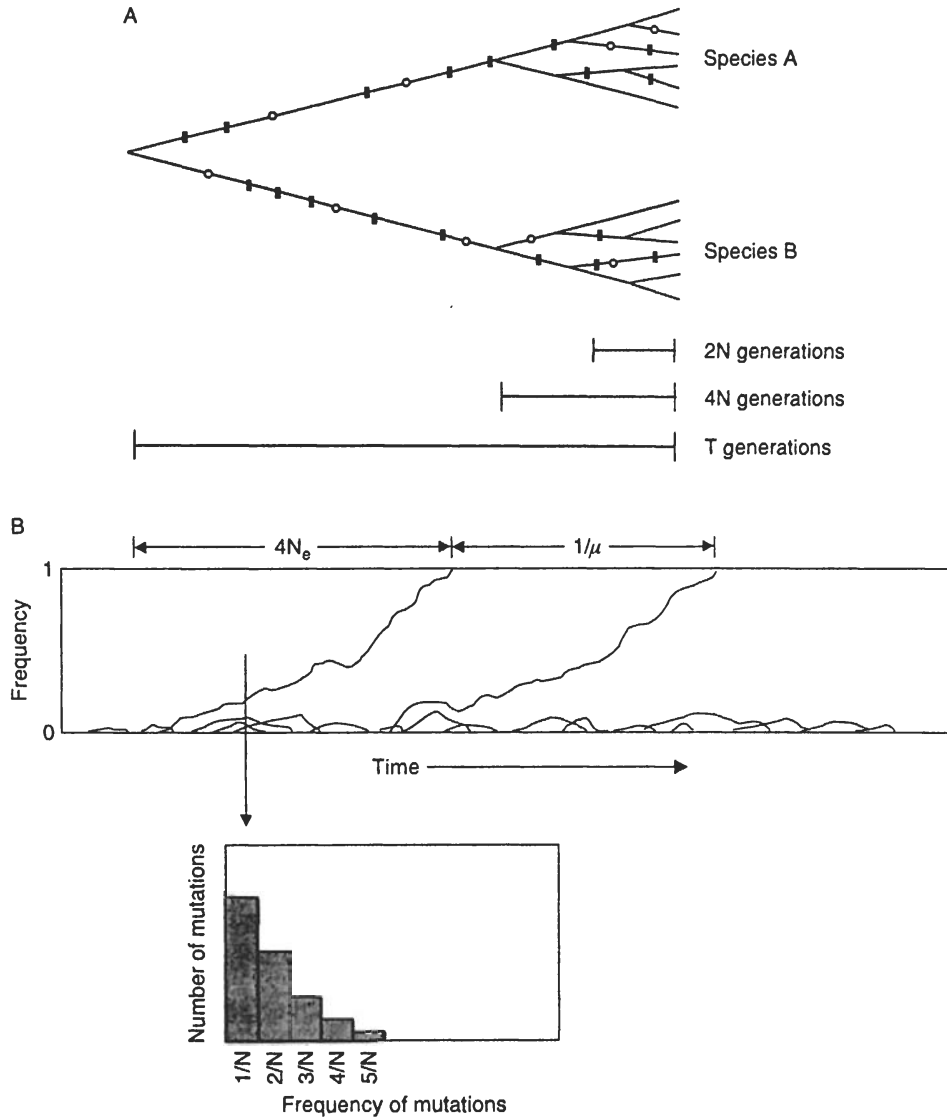


FIGURE 7.1. Two key predictions of the neutral theory. (A) Hypothetical gene genealogy for two species under the neutral model. The average time separating one allele from each species is T generations, and therefore the average sequence divergence between each species is $2\mu T$. The average coalescence time for two randomly chosen alleles within a species is $2N$ generations, and therefore the average sequence divergence between two randomly chosen alleles within a species is $4N\mu$. The average coalescence time for all alleles within a species is $4N$ generations. These expectations all have large variances. The mutation rate, μ , may vary among loci. For example, the open circles represent mutations at a locus with a low mutation rate, and the ratio of polymorphism within species A to fixed differences between A and B is 2:5 for this locus. The filled rectangles represent mutations at a locus with a high mutation rate, and the ratio of polymorphism within species A to fixed differences between A and B is 4:10 for this locus. Under the neutral theory, the ratio of polymorphism to divergence is expected to be the same for different genes (HKA test) or different classes of sites within a gene (MK test). (B) Hypothetical distribution of allele frequencies under the neutral model. Most new neutral mutations are lost due to drift. A small fraction are fixed by drift, and this takes $4N$ generations, on average. The mutation rate, μ , per gamete per generation is equal to the substitution rate, ν , per lineage per generation. Thus, the time between successive fixations is $1/\mu$. At mutation–drift equilibrium, there are many low-frequency polymorphisms and an increasingly smaller number of higher frequency polymorphisms.

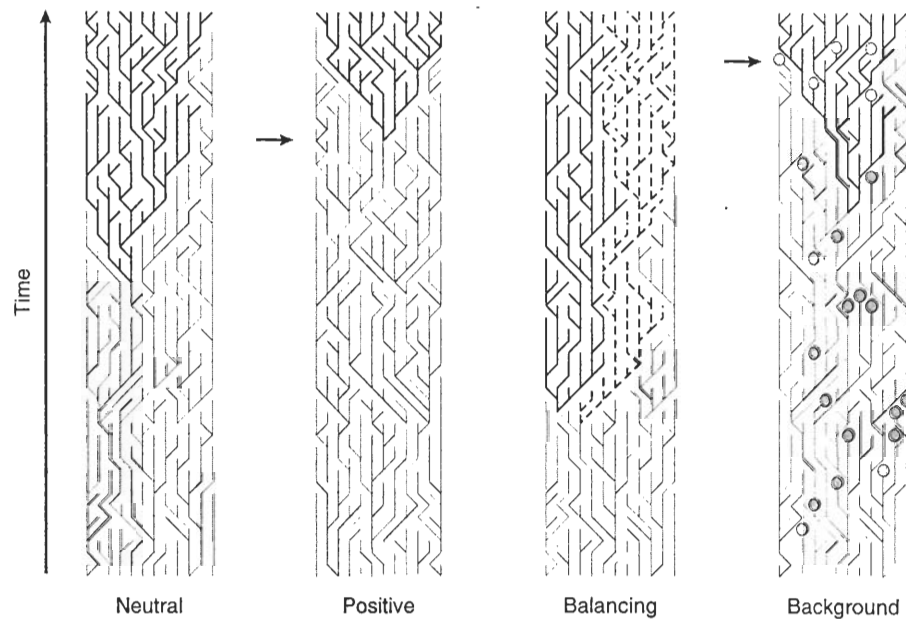


FIGURE 7.2. Hypothetical gene genealogies for a population of 12 haploid individuals under different models of selection (from Bamshad & Wooding 2003). A neutral model is given first, followed by a model of positive directional selection (which results in a shortened coalescence time for all alleles in a sample, thus reducing heterozygosity; the arrow indicates the onset of positive selection), a model of balancing selection (which results in a lengthened coalescence time for all alleles in a sample, thus increasing heterozygosity), and a model of background selection (which results in a shortened coalescence time due to the pruning of branches by deleterious mutations, shown with circles). The arrow in the last panel indicates a branch that has been pruned by background selection, relative to the neutral model in the first panel. Reprinted from Bamshad & Wooding (2003) with the permission of Nature Publishing Group.

are unknown. Conversely, when we do find evidence of selection, the target of selection may still lie at a considerable genomic distance.

Genetic hitchhiking refers to the adaptive fixation of an advantageous mutant and the associated fixation of linked, neutral variants (Maynard-Smith & Haigh 1974). In the aftermath of a selective sweep, variation at the gene under selection and at linked sites will be reduced or eliminated (Figure 7.3). The strength of this hitchhiking effect (i.e., the amount of reduction) will depend on the strength of selection and the rate of recombination in the region (Kaplan et al. 1989). In regions of low recombination, there will be little opportunity for linked neutral sites to become decoupled from the selected site during the typically short sojourn time of an adaptive fixation. In regions of high recombination, linked sites may become decoupled from selected sites.

Thus, if genetic hitchhiking is common, we might expect to see a general correlation between levels of neutral nucleotide diversity and recombination rate for different genomic regions. First demonstrated in *Drosophila melanogaster* (Begun & Aquadro 1992), this pattern has now been seen in many organisms, including humans (see Case Studies). One trivial explanation for this pattern might be that recombination is mutagenic; i.e., that high nucleotide diversity results from a greater input of new mutations in some genomic regions. A simple test of this idea is provided by comparing recombination rate with interspecific divergence for different regions of the genome. If recombination is mutagenic, then regions of high recombination should show higher rates of evolution between species. This pattern is not seen in *Drosophila* (Begun & Aquadro 1992), but a weak positive correlation

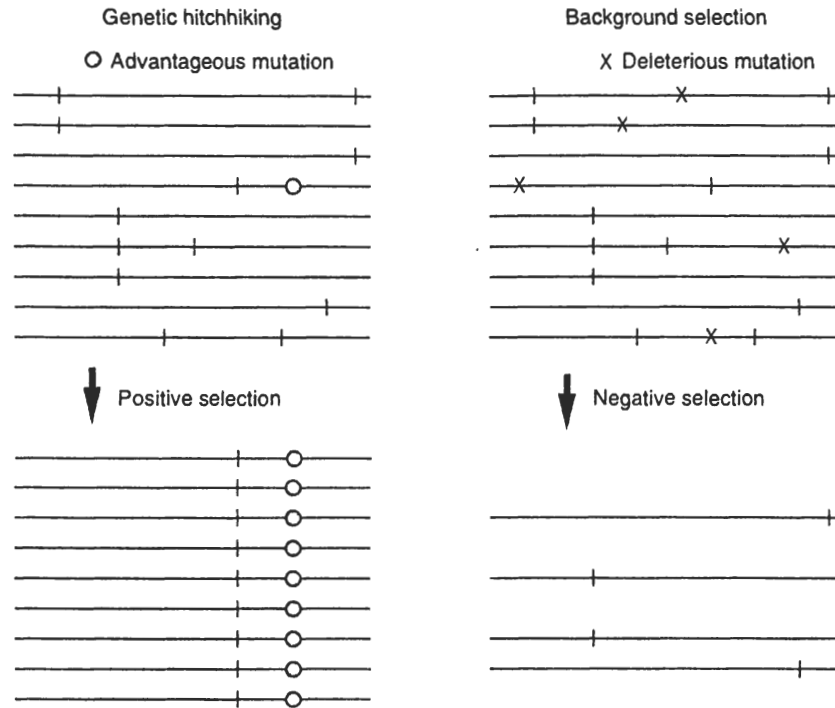


FIGURE 7.3. Schematic model of genetic hitchhiking and background selection without recombination (from Nachman 2001). Horizontal lines depict haplotypes in a population, and vertical marks depict neutral mutations. Under genetic hitchhiking, an advantageous mutation arises and is fixed by positive selection, dragging linked neutral variants with it. In the aftermath of a complete selective sweep without recombination, all individuals possess the same haplotype. If recombination occurs during the selective sweep (not shown), some variation may remain in the population. Under background selection, deleterious mutations arise and are eliminated by selection, eliminating linked neutral variants with them. In the presence of recombination (not shown), neutral variants may escape elimination. Formally, background selection is equivalent to a reduction in the effective population size by a fraction f_{B} , the equilibrium frequency of chromosomes free of deleterious mutations.

between recombination rate and interspecific divergence is seen in mammals (Hellman et al. 2003). Although positive selection will lead to increased rates of evolution for sites under selection, it does not increase the rate of evolution at linked sites (Birky & Walsh 1988). Thus, the association between recombination and divergence cannot be explained by genetic hitchhiking and instead suggests that recombination is associated, perhaps causally, with higher mutation rates in some species.

Another model of selection at linked sites involves purifying selection. Background selection (Charlesworth et al. 1993) is the removal of deleterious mutations by selection and the associated

removal of linked neutral variants (Figure 7.3). Like genetic hitchhiking, this process can reduce genetic variation, particularly in regions of low recombination. Background selection thus represents an alternative hypothesis that might explain the correlation seen between nucleotide diversity and recombination rate (Begun & Aquadro 1992).

Statistical Tests of the Neutral Model

In the standard neutral model, genetic variation is assumed to be selectively neutral, and populations are assumed to be panmictic (randomly mating),

constant in size, and at equilibrium with respect to mutation and drift. Point mutations in DNA sequences are typically modeled using the infinite sites model (Kimura 1969a) in which each new mutation occurs at a site that has not previously mutated. Deviations from these assumptions may lead to rejection of the null hypothesis. For example, many statistical tests of neutrality will return a significant result if there are changes in population size, even in the absence of selection. Many such "population-level" processes can lead to rejection of the null hypothesis, and thus an important goal of molecular studies of selection is to distinguish the effects of demographic processes (e.g., migration, nonrandom mating, population growth, population bottlenecks) from selection. One way to do this is to compare patterns of variation at multiple loci; in general, selection will affect only specific genes, while population processes will affect all loci in the genome.

Another important consideration is that different statistical tests differ tremendously in their power to detect particular non-neutral events. Some tests seem to be effective in detecting more recent selection while others seem to be more effective in detecting older selection. Below, I describe different statistical tests, and where such information exists, I describe the power and utility of the tests for detecting particular non-neutral patterns.

I have grouped statistical tests of neutrality into five general categories. This classification is based mostly on the kinds of data that are used, but it is somewhat arbitrary and other classifications are possible.

Tests Based on the Distribution of Allele Frequencies

The first test of this kind (Watterson 1978) was based on the infinite alleles model of mutation (Kimura & Crow 1964) rather than the infinite sites model, and it predated population-level DNA sequence data by several years. Watterson's test compares the observed number of alleles and the observed heterozygosity in a sample. An excess of heterozygosity, given the observed number of alleles at a locus, is consistent with balancing selection or a population contraction, while a deficit of heterozygosity is consistent with positive directional selection, a population expansion, or the presence of weakly deleterious alleles in the sample.

Tajima (1989) introduced a test that is conceptually similar but is based on the infinite sites model; it compares the observed number of polymorphic sites corrected for sample size (θ_w) with the observed nucleotide heterozygosity (θ_π) in a sample. Formally, both θ_w and θ_π are estimators of the neutral mutation parameter, $4N_e\mu$. At equilibrium $\theta_w = \theta_\pi = 4N_e\mu$. The test statistic, D , is constructed as the difference between these two estimators [$D = (\theta_\pi - \theta_w) / \text{Var}(D)$] and therefore takes on positive values when $\theta_\pi > \theta_w$ and negative values when $\theta_\pi < \theta_w$. Since low-frequency mutations contribute proportionally more to θ_w than to θ_π negative values of D occur when there is an excess of low-frequency polymorphisms and positive values of D occur when there is an excess of intermediate-frequency polymorphisms. An excess of low-frequency polymorphisms is consistent with positive directional selection, since in the aftermath of a selective sweep variation is eliminated and new mutations arise at low frequencies. It is also consistent with the presence of mildly deleterious alleles (which remain at low frequency because of selection against them) or with a recent population expansion (since rare alleles are preserved in expanding populations). Conversely, an excess of intermediate-frequency polymorphisms may be indicative of balancing selection or a population contraction (since rare alleles are lost during population contractions). Recently, several authors have also shown that sampling from subdivided populations may create a skew in the distribution of allele frequencies (e.g., Hammer et al. 2003). Tajima (1989) showed that for small insertion-deletion polymorphisms (indels) in *D. melanogaster*, the value of D was not significantly different from 0, while for large indels, D was significantly negative, leading him to conclude that large indels are more harmful than short indels.

Many other tests based on the distribution of allele frequencies have subsequently been developed. These different tests focus on different aspects of the data. For example, Fu and Li's (1993) test focuses on the number of singletons in a sample, and subsequent tests by Fu (1997) utilize other aspects of the data. Fay and Wu's (2000) test focuses on the number of derived, high-frequency mutations. For most of these tests, the significance is assessed through coalescent simulations to generate a distribution for the test statistic under the null hypothesis. Since population-level processes can lead to deviations, a key problem is choosing the appropriate

null model. For example, in humans, most loci show an excess of rare alleles. Since this pattern is seen at many loci, it is probably a consequence of population expansion in the history of our species. Detecting a signature of positive selection against this backdrop of a genome-wide skew toward rare alleles is difficult. One approach is to simulate a null distribution with a model that includes the appropriate changes in population size. Recently, this approach was used by Wooding et al. (2004) to conclude that the values of Tajima's D seen at the PTC locus in humans are indicative of selection acting at this locus. This gene is polymorphic in most human populations and is involved in the ability to taste bitter substances.

Several researchers have conducted power analyses of these tests under different models of selection or changes in population size (e.g., Simonsen et al. 1995; Fu 1997). These studies generally show that fairly large samples are needed to achieve reasonable power, and that even with these samples, the ability to detect selection (or changes in population size) is restricted to fairly short windows of time following the event, but not necessarily immediately after the event. For example, following a complete selective sweep, variation is entirely eliminated and these tests consequently have no power. Once new mutations arise, the tests have some power to detect a skew in the allele frequency spectrum, but power is again lost once a neutral distribution is reached ($< 4N$ generations).

These tests have been used to detect both positive directional selection and balancing selection. In most cases, demographic explanations can be ruled out only when it is shown that the distribution of allele frequencies at other loci is distinct.

Tests Based on Intraspecific Polymorphism and Interspecific Divergence

A second class of tests utilizes data from variation within and between species. These tests are based on the basic idea that the ratio of polymorphism to divergence is expected to be the same for different genes under neutrality, even if the actual amount of variation is different among genes. For example, histones are extremely conserved proteins and therefore have a low neutral mutation rate (i.e., the fraction of sites at which changes do not affect function is very small). Most intergenic, noncoding DNA, on

the other hand, is generally not conserved and therefore has a higher neutral mutation rate (i.e., changes at most sites do not affect function). Histone coding sequences are therefore expected to show less variation than intergenic noncoding sequences, both within and between species. However, the ratio of polymorphism to divergence should be the same for both loci provided that they are evolving neutrally (i.e., all mutations are either deleterious or neutral).

The first formulation of this basic idea into a statistical test was by Hudson et al. (1987), and this has become known as the HKA test. The test incorporates any number of loci and utilizes polymorphism data from one or two species and divergence data from interspecific comparisons. The test calculates expected values of polymorphism and divergence using least-squares estimators, and compares these with observed values to generate a test statistic that is approximately χ^2 distributed. Loci are assumed to be independent (i.e., unlinked), but the test is conservative if they are linked. By comparing multiple loci, the HKA test is able to disentangle locus-specific selection from population-level effects. For example, if a population has undergone a severe contraction in one species, genetic variation will be reduced at all loci, and the test will not be significant. If selection has reduced variation at just one locus, however, the ratio of polymorphism to divergence will be lower at that locus, and the test may be significant. In principle, rejection of the null model might be due to either elevated or reduced polymorphism or elevated or reduced divergence. However, as described above, selection at linked sites will affect levels of polymorphism (Maynard-Smith & Haigh 1974) but not levels of divergence (Birky & Walsh 1988), so unless the sites surveyed are themselves the target of selection, deviations in polymorphism levels are more likely to be the cause of any non-neutral pattern. This raises an important point: significant rejection of the null model in an HKA test may be due to selection at the sites surveyed, or it may be due to selection at linked sites, potentially at a considerable distance from the surveyed loci. For example, in *D. melanogaster* nucleotide variability is reduced over a considerable distance near the tip of the X chromosome (Begun & Aquadro 1991) and along the fourth chromosome (Jensen et al. 2002), presumably as a result of either positive or negative selection, but the targets of selection are unknown. If two loci which are both experiencing

similar selection are compared in an HKA framework, the null model may not be rejected.

A related test compares the number nonsynonymous and synonymous mutations within and between species (McDonald & Kreitman 1991) and has become known as the MK test. The ratio of polymorphism to divergence for these two kinds of mutations is expected to be the same under neutrality, following the same logic as for the HKA test. Deviations can be tested using a simple 2×2 test of independence, such as a χ^2 test or a Fisher's Exact Test. In principle, deviations could be due to either an excess or deficiency of counts in any of the four cells (synonymous polymorphisms, synonymous divergence, nonsynonymous polymorphisms, nonsynonymous divergence), although it is typically assumed that selection on amino acids is more likely than selection on silent mutations. Thus, in practice, it is often assumed that silent changes are neutral and that deviations reflect selection on nonsynonymous mutations. This assumption may not be valid in situations where codon bias due to selection occurs (Chen & Stephan, Ch. 9 of this volume).

There are some important and often unappreciated differences between the HKA and MK tests. The HKA test compares two loci that are freely recombining with respect to each other while the MK test compares two classes of sites that are interspersed. The loci being compared in an HKA test have independent evolutionary histories, and may have shorter or longer gene genealogies (and therefore more or less polymorphism) due to chance. The two classes of sites in an MK test are completely interspersed and are therefore expected to share the same gene genealogy. The HKA test therefore has two sources of variance (the sampling variance and the evolutionary variance) while the MK test has only one (the sampling variance). This difference is reflected in the statistical treatment of each test: the MK test can be performed with any conventional test of independence while the HKA test utilizes a framework that incorporates evolutionary variance. A second consequence of the interspersed sites in the MK but not the HKA test is that selection at linked sites must be considered in the latter but not the former. In other words, a significant rejection of the MK test implies that the sites surveyed are themselves the target of selection. The tests are thus complementary: the HKA test can be used to detect selection at a distance but it typically does not provide direct information about the genes under selection, while the MK test will not

detect selection at a distance but it will provide information about specific genes under selection. The MK test has been used to provide evidence of strong positive selection at several genes in *Drosophila* (e.g., McDonald & Kreitman 1991; Eanes et al. 1993) and other organisms (Ford 2002).

Tests Based on Population Differentiation

Cavalli-Sforza (1966) first suggested that differences among loci in measures of population differentiation, such as F_{ST} , could be used to infer the action of selection, and this was later developed into a statistical test by Lewontin and Krakauer (1973). The idea behind this test is that gene flow among populations will generate some average value of differentiation for most loci, against which outliers can be identified. In principle, these outliers could lie in either direction: local adaptation might produce unusually high levels of differentiation at some loci, while balancing selection acting similarly in different populations might produce below-average levels of differentiation at some loci. The test developed by Lewontin and Krakauer (1973) was criticized by Nei and Maruyama (1975) and Robertson (1975) who argued that the expected variance in F_{ST} for the test was not valid under many models of population structure. However, the fundamental logic of comparing levels of population differentiation among loci has served as the basis for several newer tests. For example, Beaumont and Nichols (1996) and Beaumont and Balding (2004) have used coalescent simulations to develop a null distribution of F_{ST} values, conditioned on observed heterozygosity under various models of population structure. Stephan et al. (1998) also used coalescent simulations to generate expected distributions of F_{ST} under neutral and selection models.

Another approach for detecting selection based on comparisons among populations is to look at levels of heterozygosity for different loci in different populations. Schlötterer et al. (1997) and Schlötterer (2002) pointed out that in comparisons of heterozygosity at multiple loci from two or more populations, it is possible to disentangle the effects of selection from the effects of demography. For example, if two populations differ in size, the expected level of variation in the smaller population will be lower for all loci. However, individual loci that have been under selection may have even less variation than this genome-wide difference.

Thus by looking at locus and population combinations, it is possible to identify genomic regions that have been under selection.

Tests Based on Linkage Disequilibrium

Linkage disequilibrium (LD) is the nonrandom association of alleles (nucleotides) at different genes (sites). Positive, directional selection will decrease genetic variation, but can also lead to an increase in LD, especially if the selected allele has not yet been fixed. This idea has motivated several tests that look for an excess of LD (Kelly 1997; Toomajian & Kreitman 2002; Sabeti et al. 2002). In general, these tests are likely to be most sensitive at identifying recently selected alleles, since complete selective sweeps will eliminate variation. In humans, it appears that selection can create LD over considerable genomic distances (e.g., Saunders et al. 2002).

Tests Based on Fixation Rates and Patterns

All the tests described above require data on genetic variation within species. A different class of tests is typically based on patterns of molecular evolution between species, and these tests can therefore detect selection that has occurred in the more distant past. These tests are based on the partitioning of DNA sequences into sites at which mutations will change the amino acid (nonsynonymous sites) and sites at which mutations will not change an amino acid (synonymous sites). For a given gene, the numbers of synonymous and nonsynonymous sites are counted. Sequences from two species are then aligned, and the observed numbers of synonymous and nonsynonymous mutations are counted. If no selection is operating, the observed numbers of synonymous and nonsynonymous mutations should occur in proportion to the numbers of each kind of site. Expressed on a per site basis, the ratio of nonsynonymous substitutions per nonsynonymous site (K_A or d_N) to synonymous substitutions per synonymous site (K_S or d_S) should be 1. Under purifying selection K_A/K_S (or d_N/d_S) will be less than 1. If $K_A/K_S > 1$, this suggests that positive selection has driven the fixation of amino acids. This test can also be applied to variation within a species, but unless the number of observed differences is large, the test has little power. For many genes, K_A/K_S in interspecific comparisons is on the

order of 0.1–0.2, reflecting the fact that most genes are under considerable selective constraint. In general, selection has to be quite strong to drive K_A/K_S above 1. Thus, this test has little power to detect weak selection. For example, genes with high K_A/K_S values that are still less than 1 (e.g., 0.8) might reflect weak positive selection or relaxed constraint. Some of the first methods were proposed by Miyata and Yasunaga (1980) and Li et al. (1985). The first applications of these methods to demonstrate positive selection were by Hill and Hastie (1987) who studied serine protease inhibitors, and Hughes and Nei (1988) who showed that $d_N/d_S > 1$ for the antigen recognition sites of the major histocompatibility complex class I loci in both humans and mice.

Subsequent studies have modified this basic idea to incorporate a maximum likelihood framework for estimating d_N/d_S and for hypothesis testing. These newer models have two main advantages: they incorporate lineage-specific effects and codon-specific effects. For example, models developed by Yang and Nielsen (2002) allow one to test the hypothesis that all sites have a d_N/d_S ratio that is drawn from a single distribution against the alternative that some sites have $d_N/d_S < 1$ while other sites have $d_N/d_S > 1$. This is biologically sensible since it is unlikely that positive selection will affect all codons within a gene. Even in genes subject to positive selection, many codons are probably still subject to selective constraints. There are many variations of these models, and they are summarized along with empirical examples of positive selection from the literature by Yang and Bielawski (2000). A major result of molecular evolutionary studies of the last two decades is that many of the genes that are under positive selection are involved in either immunity or reproduction (Ford 2002). This is probably a consequence of the fact that both classes of genes are involved in co-evolutionary processes in which selection pressures are constantly changing.

Genomics and Selection

The recent completion of whole-genome sequences from many organisms is making it possible to conduct tests for selection on a genome-wide scale. In principle, with enough markers at sufficient density, one might be able to “scan the genome” to identify many or most of the regions that have recently been under selection. This approach has been used with tests based on linkage disequilibrium

(Huttley et al. 1999), the distribution of allele frequencies (Payseur et al. 2002), patterns of population differentiation (Akey et al. 2002), relative levels of heterozygosity (Kayser et al. 2003; Storz et al. 2004) and patterns of interspecific evolution (Clark et al. 2003), among others (reviewed in Luikart et al. 2003). These studies often include thousands of loci and therefore thousands of tests. This presents a formidable statistical challenge. With 1000 tests, for example, we expect 50 to be significant at the 0.05 level under the null model. Separating these false positives from a true signature of selection is difficult. At present there is no clear solution, but two approaches have been used. One is to use a very conservative P value, thus minimizing the likelihood that low probability test results will be due to chance. A second is to simply identify outliers in the observed distribution and to treat these outliers as candidates for genes under selection. Some authors have looked for confirmatory evidence that a particular locus may be under selection by looking at adjacent markers (e.g., Kayser et al. 2003), although this neglects the fact that linked sites will have correlated evolutionary histories even under the null model. One approach for confirming that a particular locus is a truly under selection is to use different tests that rely on independent data. For example, if a locus shows significantly elevated K_A/K_S between species and significantly reduced variation within species in an HKA test, we have greater confidence that selection has operated (Karn & Nachman 1999). Ultimately, however, a statement about selection is a hypothesis of function. The best evidence that a gene is currently under selection comes from functional studies in which allelic variants are shown to be associated with functional differences that affect fitness.

CASE STUDIES

There are many good examples of selection at the molecular level. Here I restrict discussion to recent studies that incorporate DNA sequence data. Thus, I do not provide examples from the earlier allozyme literature, although these are some of the best examples linking molecular variation to fitness differences in a known ecological context (see Ford 2002). I also restrict discussion to examples that rely heavily or exclusively on population-level data, thus excluding the many interesting molecular evolutionary studies that document selection over longer

evolutionary time scales between species (reviewed in Yang & Bielawski 2000; Ford 2002).

Reduced Nucleotide Variation in Low-Recombination Regions

One very general result to come out of studies of selection at the molecular level is that genomic regions with reduced rates of recombination show reduced levels of neutral variation. The was first demonstrated in an HKA framework by Begun and Aquadro (1991) for the tip of the X chromosome and by Berry et al. (1991) for the fourth chromosome in *D. melanogaster*. Begun and Aquadro (1992) later provided evidence of a general genome-wide correlation between levels of genetic variation and local rates of recombination (Figure 7.4A), a result that was initially interpreted as evidence of recurrent selective sweeps until Charlesworth et al. (1993) presented a model of background selection as an alternative explanation. The correlation between recombination and nucleotide heterozygosity now appears to be quite general, having been demonstrated in humans (Nachman et al. 1998; Nachman 2001) and many other organisms. Background selection and genetic hitchhiking are not mutually exclusive, and both are likely operating to some degree (Kim & Stephan 2000). Distinguishing between these models is difficult and has been the subject of considerable recent work. These models make different predictions concerning the relative levels of genetic variation on the X chromosome and autosomes (Begun & Whitley 2000), the distribution of allele frequencies (Jensen et al. 2002), patterns of population differentiation (Stephan et al. 1998), and levels of variation at markers with different mutation rates (Payseur & Nachman 2000).

The observation of reduced variation in low-recombination regions is instructive in several respects. First, it seems likely that some form of selection at linked sites is responsible for the pattern in most cases (but see Hellmann et al. 2003), although it is unclear whether the pattern is caused predominantly by positive or negative selection (Kreitman 2000). The biological implications are obviously very different for these two models of selection, and thus resolving the relative contributions of each is important for understanding the underlying biology. Second, it seems that we are detecting the effects of selection at linked sites, but the actual targets of selection (i.e., the sites affecting function, either beneficially or detrimentally)

are unknown and may lie at a considerable genomic distance from the sites that have been surveyed. Thus, although there is now a fairly large empirical and theoretical literature on genetic hitchhiking and background selection, and despite the fact that recombination and heterozygosity appear to be correlated in many organisms, the full evolutionary significance of this pattern remains elusive.

Despite the fact that the relative contributions of positive and negative selection to the correlation between heterozygosity and recombination are unknown, this correlation still implies that the

dynamics of variation at many or most sites in the genome may be governed by selection, even if these sites are not functionally important themselves (Ford 2002). This suggests that some parameters estimated in neutral models (such as N_e) from many genes may be substantial underestimates of true values.

Adh in *Drosophila melanogaster*

In contrast with the example above, an example where the target of selection has been clearly

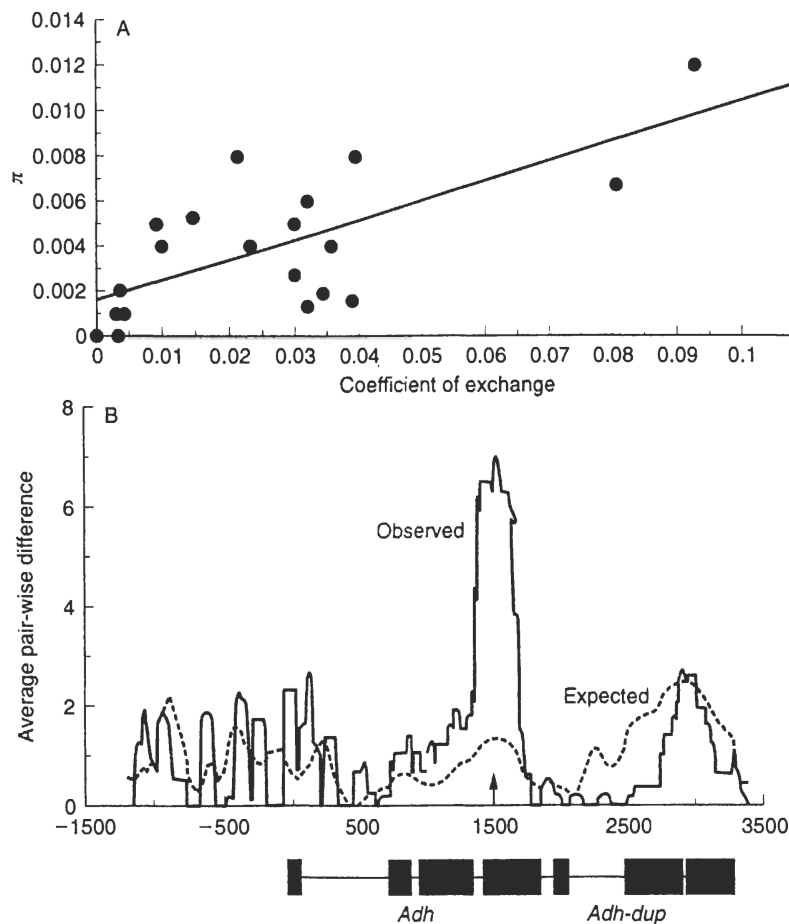


FIGURE 7.4. Three examples of selection at the molecular level in *Drosophila*. (A) Nucleotide variability is positively correlated with recombination rate in *D. melanogaster*. Begun and Aquadro (1992), reprinted with the permission, © 1992 National Academy of Sciences, USA. (B) Sliding window analysis of *Adh* in *D. melanogaster*, showing a peak of polymorphism surrounding the target of selection. Kreitman and Hudson (1991), reprinted with permission of the Genetics Society of America.

(continued)

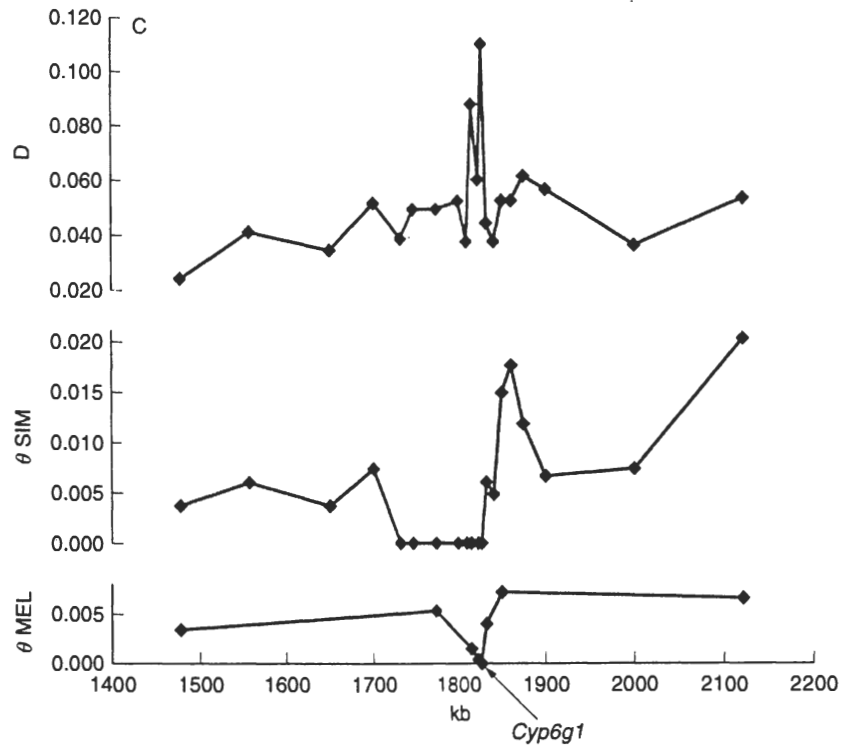


FIGURE 7.4 (cont.) (C) Patterns of polymorphism (θ) and divergence (D) at *Cyp6g1* and neighboring loci in *D. melanogaster* (MEL) and *D. simulans* (SIM). Schlenke and Begun (2004), reprinted with the permission of Nature Publishing Group. *Cyp6g1* shows reduced polymorphism in both species, but increased divergence, relative to neighboring genes. See text for discussion of these three examples.

identified is *Adh* in *D. melanogaster*. The HKA test was first applied to this locus to show that the ratio of polymorphism to divergence is higher for the coding region of the gene compared with the 5' regulatory region, an observation that was interpreted as evidence for balancing selection (Hudson et al. 1987). Kreitman and Hudson (1991) later used a sliding-window approach to show that the peak of polymorphism was centered on the known fast/slow allozyme polymorphism in the third exon (Figure 7.4B). The hypothesis that this locus is under some form of balancing selection is further strengthened by the observation of clinal variation with latitude along the eastern coast of North America (Berry & Kreitman 1993). Interestingly, the MK test comparing nonsynonymous and synonymous substitutions within and between species of *Drosophila* reveals an excess of nonsynonymous mutations between species, suggesting that positive directional selection has fixed amino acid substitutions

over evolutionary time. Thus, this locus serves as a good example of the potential complexity of selective forces. We see evidence for balancing selection in the recent history of *D. melanogaster* (based on patterns of variation within and between populations) as well as evidence for positive directional selection deeper in the history of *D. melanogaster* (based on patterns of nonsynonymous substitution between species).

Cyp6g1 in *D. simulans* and *D. melanogaster*

A remarkable case of parallel evolution at the molecular level is provided by studies of *Cyp6g1* in both *D. simulans* and *D. melanogaster*. Schlenke and Begun (2004) studied patterns of intraspecific variation and interspecific divergence at this locus and at neighboring loci. Their study was initially motivated by interest in another gene in the same

region of the *D. simulans* genome. Schlenke and Begun sequenced eight inbred *D. simulans* lines from a single population from California. They surveyed 28 primarily noncoding regions, each about 900 bp in length, spanning over 3 Mb of the genome. Surprisingly, they discovered a region of approximately 100 kb that was completely devoid of polymorphism but that did not show reduced divergence between *D. simulans* and *D. melanogaster*. This suggests that the low level of polymorphism in *D. simulans* is not caused by a lower mutation rate. Their observation is extremely unlikely under a neutral model, but is consistent with recent, strong selection somewhere within this 100 kb region. Schlenke and Begun also surveyed a subset of the 28 loci in *D. melanogaster*. One locus, *Cyp6g1*, showed reduced variation in both species, but elevated levels of divergence between species, consistent with strong positive selection in the recent history of both species (Figure 7.4C).

Previous work by Daborn et al. (2002) showed that the insertion of an *Accord* transposable element upstream of *Cyp6g1* is polymorphic in *D. melanogaster* and is associated with higher expression of this gene. Moreover, the *Accord* insertion is associated with insecticide resistance in *D. melanogaster*. Remarkably, Schlenke and Begun found that the insertion of a different transposable element (*Doc*) was polymorphic in *D. simulans*, and was also inserted upstream of *Cyp6g1*. Just as in *D. melanogaster*, the insertion of the transposable element in *D. simulans* was associated with higher *Cyp6g1* expression. Finally, the *Doc* insertion in *D. simulans* seems to be associated with slightly stronger resistance to DDT insecticide. While the link to DDT resistance for *Cyp6g1* in *D. melanogaster* is strong (Daborn et al. 2002; Schlenke & Begun 2004), the link between DDT resistance and *Cyp6g1* expression in *D. simulans* is weaker, and the story appears more complicated. For example, Schlenke and Begun found that selection associated with the *Doc* insertion upstream of *Cyp6g1* in *D. simulans* was found in a California population, with no evidence of selection in a population from Zimbabwe, despite the fact that DDT is still used in Zimbabwe but has been banned from California for more than 30 years. They suggest that selection at *Cyp6g1* in *D. simulans* may have been caused by some other insecticide or contaminant (or even a natural toxin), with the *Doc* insertion conferring some cross-resistance to DDT.

Several lessons come from these studies. First, in both species it seems that adaptation resulted from the insertion of a transposable element upstream of a gene in a way that altered expression. In a remarkable example of parallel evolution, the same gene is involved in both species, although changes in expression were caused by the insertion of a different transposable element in each case. In *D. simulans*, this genomic region was studied out of interest in a completely different gene, and the discovery of adaptation at *Cyp6g1* was fortuitous. This represents one of the best examples of a scan for selection based on patterns of DNA sequence variation where the actual target of selection seems to have been identified and where functional consequences have been demonstrated. It is important to recognize, however, that the agent of selection is still unknown in *D. simulans*. This underscores the difficulty of making the complete link from genotype to phenotype to environment.

From Phenotype to Genotype: *Tb1* in Maize and Coat Color in Mice

Two unrelated examples provide case studies of a different approach, in which investigators started with a phenotype of interest and then used mapping and statistical analyses of DNA sequence variation to identify the causative genes and to better understand the nature of selection. In both cases, the agent of selection was known in advance, and thus the link between phenotype and environment was clear.

Doebley and colleagues (Doebley et al. 1997; Wang et al. 1999) have been studying the genetic basis of morphological evolution by studying the genes underlying the domestication of maize from teosinte, its wild ancestor. One major difference between these plants is that teosinte typically has long branches with tassels at the ends while maize has short branches with ears at the ends. Through mapping and cloning, Doebley et al. (1997) identified *teosinte branched1* (*Tb1*) as the gene that largely controls this difference. Wang et al. (1999) then sequenced the *Tb1* gene and the upstream nontranscribed 5' region of this gene in a sample that included both maize and teosinte. In the transcribed region of the gene, they found that maize had 39% of the heterozygosity seen within teosinte, but in the 5' region, maize had only 3% of the heterozygosity seen within teosinte (Figure 7.5A). Wang et al. (1999) performed HKA tests comparing

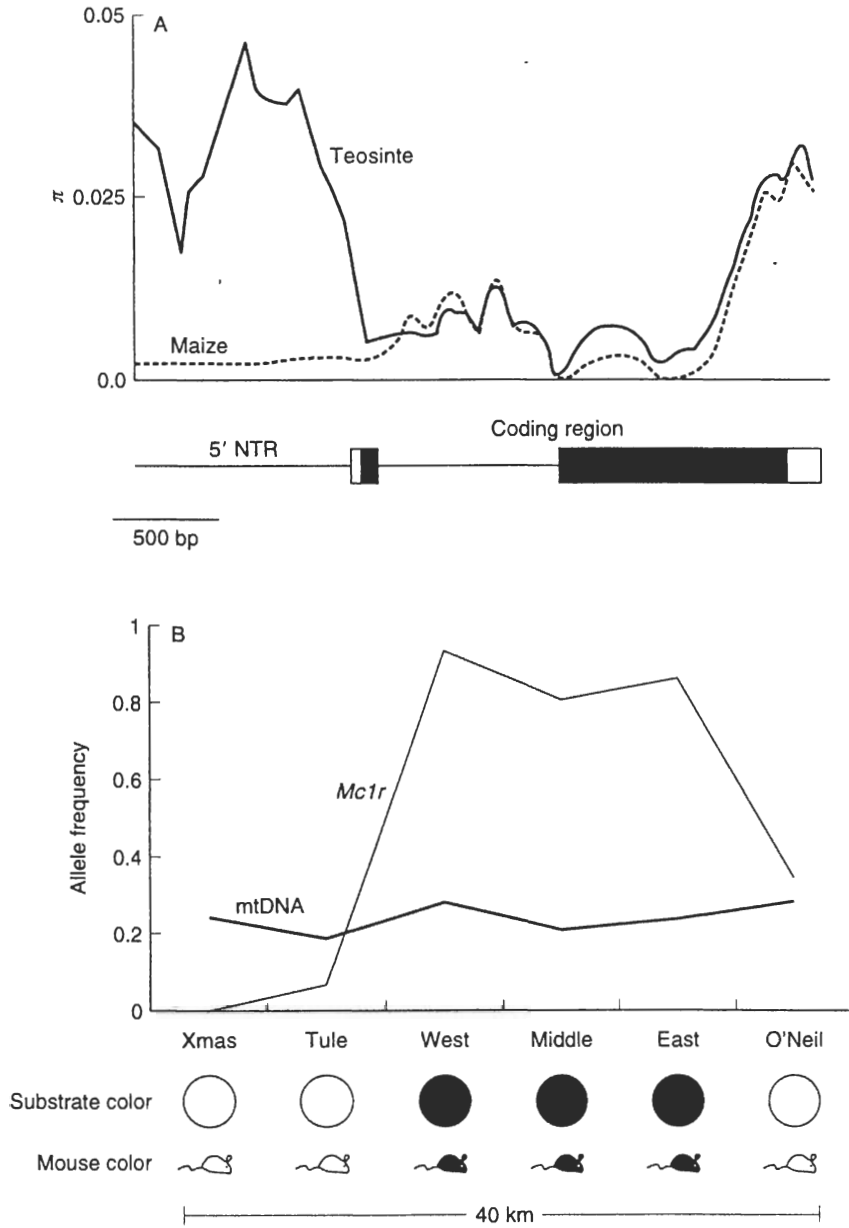


FIGURE 7.5. (A) Sliding window analysis of polymorphism in maize and in teosinte at the *Tb1* gene and at the 5' regulatory region of *Tb1*. Maize and teosinte show similar levels of variability throughout the transcribed region, but maize shows drastically reduced polymorphism in the 5' region, suggesting that an adaptive fixation occurred in the regulatory region of *Tb1* during the domestication of maize. Wang et al. (1999), reprinted with the permission of Nature Publishing Group. (B) Change in allele frequency at *Mc1r*, a gene underlying coat color in pocket mice, and for a presumably neutral locus across five populations (from Hoekstra et al. 2004). Three populations are on light-colored rocks and three populations are on dark lava, as indicated. Mitochondrial DNA(mtDNA) haplotypes show similar frequencies across these populations, while *Mc1r* alleles show dramatic changes in frequency, consistent with selection on this gene.

the untranscribed 5' region with other, presumably neutral genes, and showed that the ratio of polymorphism to divergence at the 5' region of *Tb1* was significantly reduced. This suggests that a beneficial mutation was fixed in the 5' regulatory region of *Tb1* during maize domestication. This is also consistent with the observation that the maize *Tb1* allele is expressed at a higher level than the teosinte *Tb1* allele (Doebley et al. 1997). Remarkably, an HKA test between the 5' region of *Tb1* and the transcribed region of *Tb1* was significant, showing that the effects of selection were limited to a narrow genomic region (about 2 kb). Wang et al. were able to use this observation, together with estimates of the recombination rate, to provide an estimate of the strength of selection on this gene during maize domestication ($0.04 < s < 0.08$) and the time over which selection drove this allele to fixation (315–1023 years). Finally, a phylogenetic analysis of maize and teosinte alleles in the 5' region helped pinpoint the specific teosinte ancestor for maize, and suggested that maize domestication occurred in the Balsas river valley of southwestern Mexico. This example shows how statistical analysis of DNA sequences helped provide insight into the nature of selection for a trait whose genetic basis had been previously identified. It also provides a cautionary note. Wang et al. (1999) failed to find a fixed difference between teosinte and maize in the 5' regulatory region of *Tb1*, suggesting either that the functional site lies upstream of the sequenced region, or that the genetic basis is more complex and involves interactions with other sites elsewhere in the genome.

Nachman and colleagues (Nachman et al. 2003; Hoekstra et al. 2004) have been studying the genetic basis of adaptive color differences in rock pocket mice (*Chaetodipus intermedius*). These mice inhabit rocky areas of the Sonoran desert. They are typically light-colored and live on light-colored rocks. However, in several different areas, melanic mice are found on dark rocks. The close match between the color of the mice and the color of the rocks on which they live is presumably an adaptation to avoid predation from owls and other vertebrate predators. Thus, this is a situation where the ecological importance of alternative phenotypes is reasonably well understood. The light color is ancestral and is geographically widespread, while the dark color is derived and is found on several different, geographically isolated lava flows. Nachman et al. (2003) used association studies with markers in

candidate genes to identify the genetic basis of these adaptive differences. In one population in Arizona, allelic variation at the melanocortin-1-receptor gene (*Mc1r*) was found to be perfectly associated with color variation in the mice. One allele of *Mc1r* (the *D* allele) was distinguished by four amino acids from other *Mc1r* alleles and was only found in dark animals. Patterns of DNA sequence variation at *Mc1r* revealed that the *D* alleles were nearly devoid of genetic variation, suggesting that a selective sweep had recently driven this allele to its current frequency. This hypothesis of selection was supported by in vitro functional studies which showed that the *D* allele encodes a receptor with hyperactive function relative to the other *Mc1r* alleles. Increased activity of this receptor is known to be associated with dark color in laboratory mice. Thus the *Mc1r* gene appears to be responsible for adaptive melanism in one population of pocket mice in Arizona. Surprisingly, Nachman et al. (2003) found that similar melanic phenotypes in mice from populations in New Mexico were not caused by changes at *Mc1r*. Thus, population genetic studies revealed that adaptive dark color has evolved independently in this species through changes at different genes. In the Arizona population, Hoekstra et al. (2004) compared patterns of population differentiation at *Mc1r* with patterns of population differentiation at a presumably neutral mitochondrial DNA locus (Figure 7.5B) to estimate the strength of selection on the *Mc1r* *D* alleles ($s > 0.05$). This example thus represents a situation where the link from genotype to phenotype to environment is reasonably clear. Even in this situation, however, population genetic analyses cannot tell us which of the four amino acid mutations that comprise the *Mc1r* *D* allele is responsible for the phenotypic differences in color. In principle, this might be solved by introducing each mutation separately into the gene, and testing the function of these different receptors in vitro.

FUTURE DIRECTIONS

While we have made tremendous progress detecting selection at the molecular level in the last two decades, there are at least three clear directions for future work. First, additional studies that make the link between genotype and phenotype in an explicit ecological context are needed. Despite the extensive evidence for selection in different regions of the

genome and in different species, there are still relatively few examples where the functional significance of allelic variants is well understood in a particular environmental setting. Functional studies that dissect the biochemical consequences of genetic variation, combined with phenotypic studies of the fitness effects of functional differences, will help us to fully understand the evolutionary significance of genetic variation. Such studies will also help make the connection between ecological and evolutionary timescales.

A second area of needed research concerns selection on regulatory regions. Adaptation may often result from changes in gene regulation rather than changes in gene structure, yet most studies of selection at the molecular level have focused on coding regions of genes. One difficulty is that in many cases the specific regulatory elements of genes are unknown. Empirical studies to identify regulatory elements and theoretical studies to develop appropriate statistical methods for analyzing regulatory sequences will provide real insight into the genetics of adaptation.

A third area in which future work is likely to be especially rewarding are studies that take full advantage of complete genome sequences. We now have the potential to interrogate every gene in the genome to ask questions about selection. For example, it will soon be possible to perform MK tests on every gene in the *D. melanogaster* or human genome. In principle, such data should allow us to ask questions about the relative frequency of different kinds of selection, about the kinds of genes that are under selection, and about the amount of selection in the recent history of a species. New theoretical and statistical work will be needed to overcome the inherent problems associated with multiple tests, but these studies have the potential for the first

time to give us a detailed genome-wide view of the genetics of adaptation.

SUGGESTIONS FOR FURTHER READING

A good review of basic concepts is provided by Kreitman and Akashi (1995). Ford (2002) provides a comprehensive review of specific examples where selection has been detected at the molecular level. There are two recent reviews that focus specifically on humans: Bamshad and Wooding (2003) and Vallender and Lahn (2004). Luikart et al. (2003) provide a nice overview of studies that use genome-wide data to detect selection.

Bamshad M & SP Wooding 2003 Signatures of natural selection in the human genome. *Nat. Rev. Genet.* 4:99–111.

Ford MJ 2002 Applications of selective neutrality tests to molecular ecology. *Mol. Ecol.* 11:1245–1262.

Kreitman M & H Akashi 1995 Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* 26:403–422.

Luikart G, England PR, Tallmon D, Jordan S & P Taberlet 2003 The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4:981–994.

Vallender EJ & BT Lahn 2004 Positive selection on the human genome. *Hum. Mol. Genet.* 13:R245–R254.

Acknowledgments I thank David Begun for discussion and Jeff Good for comments on this chapter.