# Molecular Signatures of Natural Selection

Rasmus Nielsen

Center for Bioinformatics and Department of Evolutionary Biology, University of Copenhagen, 2100 Copenhagen Ø, Denmark; email: rasmus@binf.ku.dk

Annu. Rev. Genet.
2005. 39:197–218

First published online as a Review in Advance on August 31, 2005

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

doi: 10.1146/annurev.genet.39.073003.112420



0066-4197/05/1215-0197$20.00

## Key Words

Darwinian selection, neutrality tests, genome scans, positive selection, phylogenetic footprinting

## Abstract

There is an increasing interest in detecting genes, or genomic regions, that have been targeted by natural selection. The interest stems from a basic desire to learn more about evolutionary processes in humans and other organisms, and from the realization that inferences regarding selection may provide important functional information. This review provides a nonmathematical description of the issues involved in detecting selection from DNA sequences and SNP data and is intended for readers who are not familiar with population genetic theory. Particular attention is placed on issues relating to the analysis of large-scale genomic data sets.

## Contents



## INTRODUCTION

**SNP:** single nucleotide polymorphism

Population geneticists have for decades been occupied with the problem of quantifying the relative contribution of natural selection in shaping the genetic variation observed among living organisms. In one school of thought, known as the neutral theory, most of the variation within and between species is selectively neutral, i.e., it does not affect the fitness of the organisms (58, 59). New mutations that arise may increase in frequency in the population due to random factors, even though they do not provide a fitness advantage to the organisms carrying them. The process by which allele frequencies change in populations due to random factors is known as genetic drift.

A second school of thought maintains that a large proportion of the variation observed does affect the fitness of the organisms and is subject to Darwinian selection (39). These issues have not been settled with the availability of large-scale genomic data, but the debate has shifted from a focus on general laws or patterns of molecular evolution to the description of particular instances where natural selection has shaped the pattern of variation. This type of analysis is increasingly being done because it has become apparent that inferences regarding the patterns and distribution of selection in genes and genomes may provide important functional information. For example, in the human genome, the areas where disease genes are segregating should be under selection (assuming that the disease phenotype leads to a reduction in fitness). Even very small fitness effects may, on an evolutionary time scale, leave a very strong pattern. Therefore, in theory it may be possible to identify putative genetic disease factors by identifying regions of the human genome that currently are under selection (7). In general, positions in the genome that are under selection must be of functional importance. Inferences regarding selection have therefore been used extensively to identify functional regions or protein residues (12, 91). The purpose of this paper is to review the current knowledge regarding the effect of selection on a genome and to discuss methods for detecting selection using molecular data, especially genomic DNA sequence and single nucleotide polymorphism (SNP) data.

### The Nomenclature of Selection Models

Much confusion exists in the literature regarding how various types of selection are defined, in particular because some of the terminology is used slightly differently within different scientific communities. At the risk of contributing further to this confusion, I propose here

some simple definitions for some of the common terms used in the discussion of selection models before moving on to the main topics of this review.

The basic population genetic terms are well-defined. The classical population genetic models that students of biology will first encounter are models with two alleles, typically denoted $A$ and $a$. Selection then occurs if the fitnesses of the three possible genotypes ($w_{AA}$, $w_{Aa}$, and $w_{aa}$) are not all equal. There is directional selection if the fitnesses of the three genotypes are not all equal and if $w_{AA} > w_{Aa} > w_{aa}$ or $w_{AA} < w_{Aa} < w_{aa}$. Directional selection tends to eliminate variation within populations and either increase or decrease variation between species depending on whether $A$ or $a$ is the new mutant. Overdominance occurs if the heterozygote has the highest fitness if $w_{AA} < w_{Aa} > w_{aa}$. Overdominance is a case of balancing selection where variability is maintained in the population due to selection. In haploid organisms, selection occurs if $w_A \neq w_a$ and overdominance is not possible. The difference in fitness between alleles is the selection coefficient, i.e., for the haploid model the selection coefficient could be defined as $s_A = w_A - w_a$.

In the molecular evolution literature, it has been common to use the terminology of positive selection, negative selection, purifying selection, and diversifying selection. Here we define negative selection as any type of selection where new mutations are selected against. Likewise, we define positive selection as any type of selection where new mutations are advantageous (have positive selection coefficients). In the context of the simple two-allele models, both directional selection and overdominance can be cases of positive selection. Purifying selection is identical to negative selection in that it describes selection against new variants. Diversifying selection has in the population genetics literature been synonymous with disruptive selection, a type of selection where two or more extreme phenotypic values are favoured simultaneously. This type of selection will often increase variability, and diversifying selection has, therefore, in the molecular evolution literature recently been used more generically to describe any type of selection that increases variability. However, as disruptive selection may reduce genetic variability when one of the extreme types becomes fixed in the population, and since there are many other forms of selection that can increase levels of genetic variability, the more generic use of the term "diversifying selection" should probably be avoided.

When a new mutant does not affect the fitness of the individual in which it arises (i.e., $w_{AA} = w_{Aa} = w_a$), it is said to be neutral. In general, neutrality describes the condition where the loci under consideration are not affected by selection. A statistical method aimed at rejecting a model of neutral evolution is called a neutrality test.

**Balancing selection:** selection that increases variability within a population

**Positive selection:** selection acting upon new advantageous mutations

**Negative selection:** selection acting upon new deleterious mutation

**Neutrality test:** a statistical test of a model which assumes all mutations are either neutral or strongly deleterious

**Neutral mutation:** a mutation that does not affect the fitness of individuals who carry it in either heterozygous or homozygous condition

## POPULATION GENETIC PREDICTIONS

One of the main interests in molecular population genetics is to distinguish molecular variation that is neutral (only affected by random genetic drift) from variation that is subject to selection, particularly positive selection. An important point is that neutral models usually allow for the presence of strongly deleterious mutations that have such strong negative fitness consequences that they are immediately eliminated from the population (58). If selection only involves such mutations of very strong effect, the only mutations that will actually segregate in the population are the neutral mutations. Therefore, neutral models include the possible existence of pervasive strong negative selection. Although negative or purifying selection may be of great interest because it may help detect regions or residues of functional importance, much interest in the evolutionary literature focuses on positive selection because it is associated with adaptation and the evolution of new form or function. One of the main points of contention in population genetics has been the degree to which positive selection is important

**Selective sweep:** the process by which a new advantageous mutation eliminates or reduces variation in linked neutral sites as it increases in frequency in the population

in explaining the pattern of variability within and between species (39, 59).

Much of the theoretical literature in population genetics over the past 50 years has focused on developing and analyzing models that generalize the previously mentioned basic di-allelic models to models where more than two alleles may be segregating, where multiple mutations may arise and interact—possibly in the presence of recombination, where the environment may be changing through time, and where random genetic drift may be acting in populations subject to various demographic forces (25, 39). From theory alone we have gained many valuable insights, including the fact that the efficacy of selection depends not only on the selection coefficient, but primarily on the product of the selection coefficient and the effective population size. An increased effect of selection may be due to either an increased population size or a larger selection coefficient. Among other important findings is that balancing selection may occur for many reasons other than overdominance, (e.g., fluctuating environmental conditions) and could therefore, potentially, be quite common (38, 39). However, the efficacy of selection will tend to be reduced when multiple selected alleles are segregating simultaneously in the genome. The mutations will tend to interfere with each other and reduce the local effective population size (8, 29, 40, 57). Many population geneticists used to believe that the number of selective deaths required to maintain large amounts of selection would have to be so large that selection would probably play a very small role in shaping genetic variation (43, 60, 61). These types of arguments, known as genetic load arguments, were instrumental in the development of the neutral theory. However, the amount of selection that a genome can permit depends on the way mutations interact in their effect on organismal fitness and on several other critical model assumptions (25, 62, 71, 107). Population genetic theory does not exclude the possibility that selection is very pervasive and cannot alone determine the relative importance and modality of selection in the absence of data from real living organisms (25, 39).

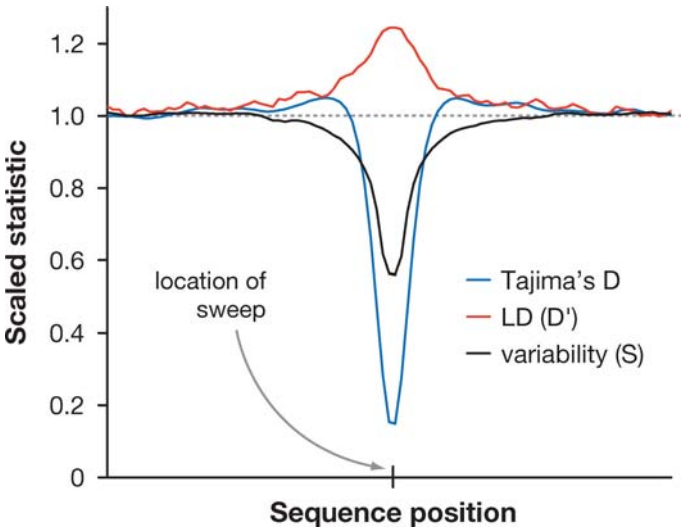


**Figure 1**

The effect of a selective sweep on genetic variation. The figure is based on averaging over 100 simulations of a strong selective sweep. It illustrates how the number of variable sites (variability) is reduced, LD is increased, and the frequency spectrum, as measured by Tajima's D, is skewed, in the region around the selective sweep. All statistics are calculated in a sliding window along the sequence right after the advantageous allele has reached frequency 1 in the population. All statistics are also scaled so that the expected value under neutrality equals one.

Much excitement currently exists in the population genetics communities over the fact that many predictions generated from the theory may now be tested in the context of the large genomic data sets. In particular, we should be able to detect the molecular signatures of new, strongly selected advantageous mutations that have recently become fixed (reached a frequency of one in the population). As these mutations increase in frequency, they tend to reduce variation in the neighboring region where neutral variants are segregating (13, 51, 52, 68). This process, by which a selected mutation reduces variability in linked sites as it goes to fixation, is known as a selective sweep (**Figure 1**). The hope is that by analysis of large comparative genomic data sets and large SNP data sets we will be able to determine how and where both positive and negative selection

has affected variation in humans and other organisms.

## POPULATION GENETIC SIGNATURES OF SELECTION

One of the main effects of selection is to modify the levels of variability within and between species (**Table 1**). A selective sweep tends to drastically reduce variation within a population, but will not lead to a reduction in species-specific differences. Conversely, negative selection acting on multiple loci will tend to reduce variability between species more drastically then variability within species. **Table 1** summarizes how various types of selection affect variability. Note that changes in the mutation rate alone will have the same effect on interspecific (between-species) and intraspecific (within-species) variability. However, selection affects intraspecific and interspecific variability differently. Many of the common population genetic methods for detecting selection are therefore based on comparing variation with and between species, most famously the HKA test (48). In this test, the rate of polymorphisms to divergence is compared for multiple genes. If the ratio varies more among genes than expected on a neutral model, neutrality is rejected.

### Population Differentiation

Selection may in many cases increase the degree of differentiation among populations. In particular, recent theory shows that a selective sweep can have a dramatic impact on the level of population subdivision, particularly when the sweep has not yet spread to all populations within a species (20, 65, 97). When a locus shows extraordinary levels of genetic population differentiation, compared with other loci, this may then be interpreted as evidence for positive selection.

One of the first neutrality tests proposed, the Lewontin-Krakauer (63) test, takes advantage of this fact. This test rejects the neutral model for a locus if the level of genetic differentiation among populations is larger than predicted by a specific neutral model. It has recently been resurrected in various forms (1, 9, 10, 53, 92, 114), primarily driven by the availability of large-scale genomic data. For example, Akey et al. (1) looked at variation in $F_{ST}$ (the most common

**Table 1 The effect of selection and mutation on variability within and between species**

| Evolutionary factor | Intraspecific variability[a] | Interspecific variability | Ratio of interspecific to intraspecific variability | Frequency spectrum |
|---|---|---|---|---|
| Increased mutation rate | Increases | Increases | No effect | No effect |
| Negative directional selection | Reduced | Reduced | Reduced if selection is not too strong | Increases the proportion of low frequency variants |
| Positive directional selection | May increase or decrease | Increased | Increased | Increases the proportion of high frequency variants |
| Balancing selection | Increases | May increase or decrease | Reduced | Increases the proportion of intermediate frequency variants |
| Selective sweep (linked neutral sites) | Decreased | No effect on mean rate of substitution, but the variance increases | Increased | Mostly increases the proportion of low frequency variants |

[a]Note that selection also affects other features of the data not mentioned here, such as levels of LD, haplotype structure, and levels of population subdivision.

measure of population differentiation) among human populations genome-wide. Beaumont & Balding (9) developed a sophisticated statistical method for identifying loci that may be outliers in terms of levels of population subdivision.

**Frequency spectrum:** the allelic sample distribution in independent nucleotide sites

**LD:** linkage disequilibrium

## The Frequency Spectrum

Selection also affects the distribution of alleles within populations. For DNA sequence or SNP data, some of the most commonly applied tests are based on summarizing information regarding the so-called frequency spectrum. The frequency spectrum is a count of the number of mutations that exist in a frequency of $x_i = i/n$ for $i = 1, 2, \ldots, n-1$, in a sample of size $n$. In other words, it represents a summary of the allele frequencies of the various mutations in the sample. In a standard neutral model (i.e., a model with random mating, constant population size, no population subdivision, etc), the expected value of $x_i$ is proportional to $1/i$. Selection against deleterious mutations will increase the fraction of mutations segregating at low frequencies in the sample. A selective sweep has roughly the same effect on the frequency spectrum (13). Conversely, positive selection will tend to increase the frequency in a sample of mutations segregating at high frequencies. The effect of selection on the frequency spectrum is summarized in **Figure 2**.

Many of the classic neutrality tests, therefore, focus on capturing information regarding the frequency spectrum. The most famous example is the Tajima's D test (112). In this test, the average number of nucleotide differences between pairs of sequences is compared with the total number of segregating sites (SNPs). If the difference between these two measures of variability is larger than what is expected on the standard neutral model, this model is rejected. The effect of a selective sweep on Tajima's D is shown in **Figure 1**. Fu & Li (34) extended this test to take information regarding the polarity of the information into account by the use of an evolutionary outgroup (e.g., a chimpanzee in the analysis of human genetic variation), and more refinements were introduced by Fu (32, 33). Fay & Wu (28) suggested a test that weights information from high-frequency derived mutations higher. These tests are probably the most commonly applied neutrality tests to date.

## Models of Selective Sweeps

The pattern of variability left by a selective sweep is a rather complicated spatial pattern (**Figure 1**). By taking information regarding this pattern into account, the power of the neutrality tests can be improved, and it may even be possible to pinpoint the location of a selective sweep. Kim & Stephan (56) developed a method based on an explicit population genetic model of a selective sweep. Using this model, they could calculate the expected frequency spectrum in a site as a function of its distance to an advantageous mutation. By fitting the data to this model, they could estimate the location of the selective sweep and the strength of the selective sweep, and perform hypothesis tests regarding the presence of a sweep. This method is particularly useful in that it takes advantage of the spatial pattern left by the sweep along the sequence.

## LD and Haplotype Structure

Levels of linkage disequilibrium (LD), the correlation among alleles from different loci, will increase in selected regions. Regions containing a polymorphism under balancing selection will tend to reduce LD if the polymorphism is old, but may increase LD in a transient phase. Selective sweeps also increase levels of LD in a transient phase (**Figure 1**), although this phase may be relatively short (82). Recently, there has been increased awareness that an incomplete sweep (when the adaptive mutation has not yet been fixed in the population) leaves a distinct pattern in the haplotype structure (87). This has led to the development of many statistical methods for detecting selection based on LD. Hudson et al. (47)

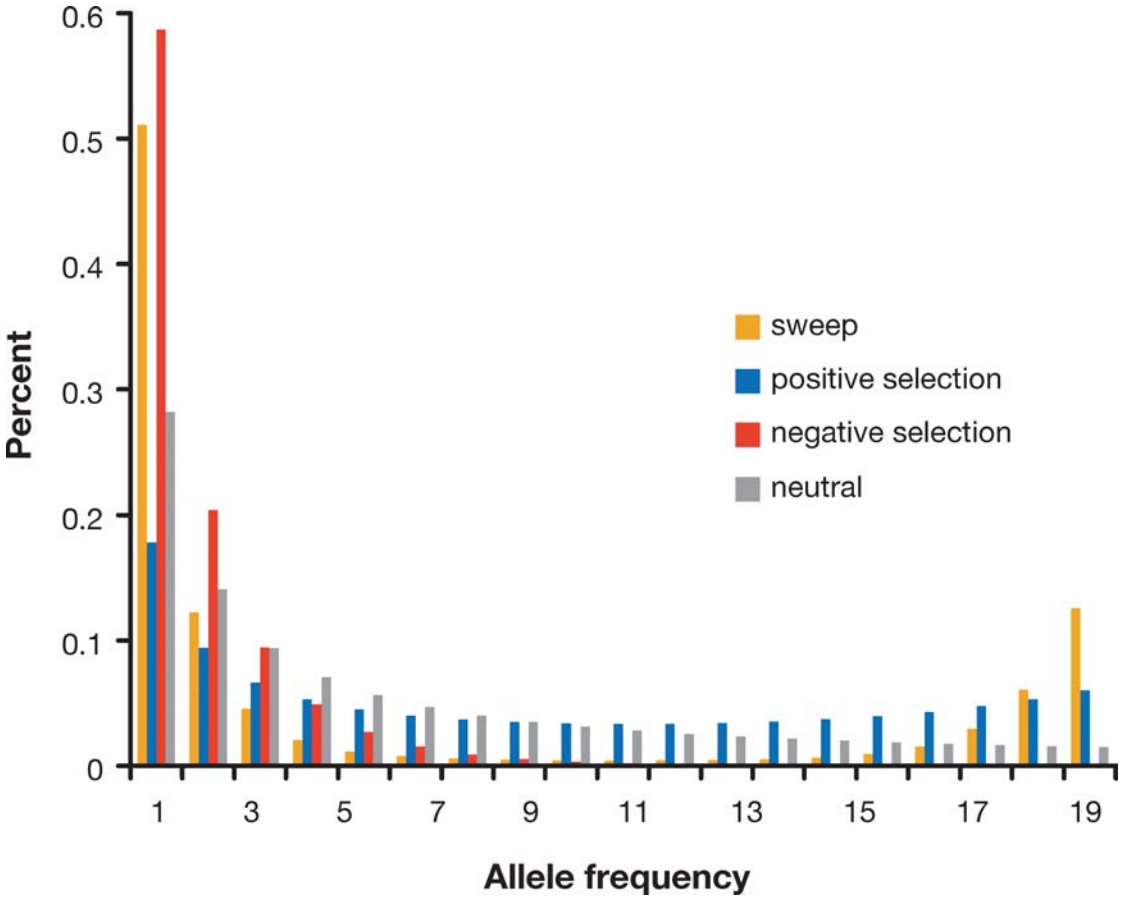


**Figure 2**

The frequency spectrum under a selective sweep, negative selection, neutrality, and positive selection. The frequency spectra under negative and positive selection are calculated using the PRF model by Sawyer & Hartl (88) for mutations with $2Ns = -5$ and 5, respectively, where $N$ is the population size and $s$ is the selection coefficient. For the selective sweep, the frequency spectrum is calculated in a window around the location of the adaptive mutation immediately after it has reached fixation in the population. In all cases, a demographic model of a population of constant size with no population subdivision is assumed.

developed a test based on the number of alleles occurring in a sample. Andolfatto et al. (4) developed a related test to determine whether any subset of consecutive variable sites contains fewer haplotypes than expected under a neutral model. A similar test was also proposed by Depaulis & Veuille (23). A variation on this theme was proposed by Sabeti et al. (87) who considered the increase in the number of distinct haplotypes away from the location of a putative selective sweep. Kelly (54) considered the level of association between pairs of loci. Kim & Nielsen (55) extended the method of Kim & Stephan (56) to include pairs of sites to incorporate information regarding linkage disequilibrium.

## MacDonald-Kreitman Tests

Finally, the MacDonald-Kreitman test (69) explores the fact that mutations in coding regions come in two different flavors:

nonsynonymous mutations and synonymous mutations. It summarizes the data in what has become known as a MacDonald-Kreitman table, which contains counts of the number of nonsynonymous and synonymous mutations within and between species. If selection only affects the nonsynonymous mutations, negative selection will reduce the number of nonsynonymous mutations and positive selection will increase the number of nonsynonymous mutations, relative to the number of synonymous mutations. However, the effect will be stronger in divergence data than in polymorphism data. A test similar to the HKA test can therefore be constructed comparing the ratios of nonsynonymous to synonymous mutations within and between species. If these ratios differ significantly, this provides evidence for selection.

## STATISTICAL CONCERNS

The neutrality tests are all tests of complicated population genetic models that make specific assumptions about the demography of the populations, in particular a constant population size and no population structure. In addition, in some of the tests there may be other implicit assumptions regarding distributions of recombination rates and mutation rates. Many of these tests have long been known to be highly sensitive to the demographic assumptions. For example, Simonsen et al. (96) showed that Tajima's D test (112) would reject a neutral model very frequently in the presence of population growth. The molecular signature of population growth is in many ways similar to the local effect of a selective sweep, and neutrality tests are often used as a method to detect population growth (85). Nielsen (73), Przeworski (82), and Ingvarsson (50) also argued that simple models of population subdivision can lead the commonly used neutrality tests to reject the neutral model with high probability, even in the absence of selection. In addition, even if the presence of selection can be established, in many cases it can be difficult to distinguish between the pattern left by selective sweeps and selection on slightly deleterious mutations (so-called background selection) (18, 19).

Tests based on patterns of LD may be particularly sensitive to the underlying model assumptions, because they (in addition to assumptions regarding demography) contain strong assumptions regarding the underlying recombination rates. Recent studies suggest that recombination rates are highly variable among regions (70) and among closely related species (83, 117). If that is true, it may not be advisable to focus attention toward patterns of LD when attempting to detect selection. Nonetheless, haplotype structure can be highly informative, particularly in detecting incomplete selective sweeps (87). Further research into how haplotype patterns can be used robustly to infer selection may be warranted.

Because of the effect of demographic assumptions on the population genetic neutrality tests, the results of these tests have often been contentious and often have not led to firm conclusions regarding the action of selection. One exception is the MacDonald-Kreitman (69) test. This test has increased robustness because the sites in which synonymous and nonsynonymous mutations occur are interspersed among each other and therefore similarly affected by demography and genetic drift. In fact, the MacDonald-Kreitman (69) test is robust to any demographic assumption (73). Unfortunately, it may not be very suitable for detecting recent selective sweeps because both nonsynonymous and synonymous mutations, linked to the beneficial mutation, will be similarly affected by the selective sweep. Also, the MacDonald-Kreitman (69) test cannot distinguish between past and present selection. Reducing the information in the data simply to the number of nonsynonymous mutations and synonymous mutations leads to a significant loss of information.

One possible way to circumvent the problem of demographic confounding effects is to compare multiple loci. For example, Galtier et al. (35) have implemented a statistical

method, applicable to microsatellite loci, to test whether the signature of population growth is constant among loci or varies among loci. If the effect varies significantly among loci, beyond what can be explained by the demographic model, this may be interpreted as evidence for a selective sweep. In general, one can assume that if strong departures from the neutral model are seen only on one or a few outlier loci, this may be interpreted as evidence for selection on these loci. However, certain demographic factors, such as population subdivision, may increase the variance among loci (73). Certain demographic models may be more likely than others to produce outlier loci even in the absence of selection.

The application of population genetic tests other than the MacDonald-Kreitman test requires careful consideration of the possible range of demographic factors that may affect the results (2, 73). It is not very meaningful in itself to reject the standard neutral model using these methods without paying careful attention to the underlying demographics. Even the interpretation of significant results of the MacDonald-Kreitman test requires attention to demography if the directionality (positive versus negative) of selection is to be inferred (26). Fortunately, many recent studies go to great lengths in trying to exclude the possibility that rejections of a neutral model may be caused by demographic effects (3, 116).

## SIGNATURES OF SELECTION IN COMPARATIVE DATA

While population genetic approaches aim at detecting ongoing selection in a population, comparative approaches, involving data from multiple different species, are suitable for detecting past selection. The major tool used to detect selection from comparative data is to compare the ratio of nonsynonymous mutations per nonsynonymous site to the number of synonymous mutations per nonsynonymous site ($d_N/d_S$). If there is no selection, not even strongly deleterious mutations, synonymous and nonsynonymous substitutions should occur at the same rate and we would expect $d_N/d_S = 1$. If there is negative selection, $d_N/d_S < 1$ and if there is positive selection, $d_N/d_S > 1$. The $d_N/d_S$ ratio is therefore a proxy for the effect of selection that helps to identify not only selection, but also the directionality of selection. It is therefore a very commonly used tool for detection of positive selection and has been used in a variety of cases, for example, to demonstrate the presence of positive selection on HIV sequences (78) and on the human major histocompatibility locus (MHC) (49). However, as negative selection will tend to dominate in evolution, comparing the average rate of synonymous and nonsynonymous substitution in aligned sequences is a very conservative tool. If the gene is functional so that many or most mutations will disrupt function, the amount of positive selection needed to elevate the $d_N/d_S$ above one is enormous. To overcome this problem, methods have been devised for detecting positive selection that takes variation in the $d_N/d_S$ ratio into account (78, 127). The basic idea is to allow the $d_N/d_S$ ratio to follow a statistical distribution among sites. If a distribution that allows values of $d_N/d_S > 1$ fits the data significantly better than a model that does not allow for such values, this is interpreted as evidence for positive selection. The methodology has been widely used and has led to a sharp increase in the number of loci where researchers have detected the presence of positive selection (31, 100, 125). This has also led to some skepticism toward this methodology (105, 106), although it has been found to perform well in simulation studies and is based on well-established statistical principles (5, 120, 124).

Several different statistical methods allow site-specific inferences regarding positive selection (30, 78, 104). The objective of these methods is to determine if specific sites have been targeted by positive (or negative) selection. In several cases, these methods have been used to make functional prediction regarding particular protein residues (91).

**$d_N$:** number of nonsynonymous mutations per nonsynonymous site

**$d_s$:** number of synonymous mutations per synonymous site

**$d_N/d_S$ ratio:** the rate ratio of nonsynonymous to synonymous substitutions

The same type of methodology used to model variation in the $d_N/d_S$ ratio among sites has also been used to model estimates of $d_N/d_S$ along particular lineages of a phylogeny (123, 126, 128). This allows the testing of hypotheses regarding selective pressures on particular evolutionary lineages. Models have also been developed that allow site-specific inferences on a particular group of lineages on a phylogeny (128). Several excellent recent reviews describe the statistical methods used to detect selection from comparative data in more detail (124, 125). A summary of the different tests of neutrality is given in **Table 2**.

## Targets of Positive Selection

Using analyses of comparative data, a clear picture emerges of the systems that most often are involved in positive selection of the kind that leads to increases in the $d_N/d_S$ ratio (75). Typically, it involves an interaction between two organisms, or two different genetic components within the same organism, that compete or interact in such a way that an equilibrium is never reached. The best known examples are host-pathogen interactions that lead to positive selection of genes in pathogens (27, 30, 45, 78, 100) or in host immune and defense systems (49, 75, 90, 100). Other examples include genes involved in gametogenesis or expressed on the surface of gametes (75, 109, 110, 122). The forces creating positive selection in these genes may include sperm competition (122) and genetic conflicts between sperm and egg-cell (108). Positive selection also seems to be common in cases where selfish genes have the opportunity to create segregation distortion, potentially

**Table 2 A very incomplete list of methods for detecting selection from DNA sequence and SNP data**

| Test | Data | Pattern | Requires multiple loci | Robust to demographic factors? | References |
|---|---|---|---|---|---|
| Tajima's D and related | Population genetic data | Frequency spectrum | No | No | (28, 32–34, 112) |
| Modeling of selective sweep—spatial pattern | Population genetic data | Frequency spectrum/spatial pattern | No | No | (55, 56) |
| Tests based on LD | Population genetic data | LD and/or haplotype structure | No | No | (4, 23, 47, 54, 87) |
| $F_{ST}$ based and related tests | Population genetic data | Amount of population subdivision | Yes | No[a] | (1, 9, 10, 53, 92, 114) |
| HKA test | Population genetic and comparative data | Number of polymorphisms/substitutions | Yes | No | (48) |
| Macdonald-Kreitman-type tests | Population genetic and comparative data | Number of nonsynonymous and synonymous polymorphisms | No | Yes | (16, 69) |
| $d_N/d_S$ ratio tests | Comparative data or population genetic data without recombination (6) | Nonsynonymous and synonymous substitutions | No | Yes | (49, 78, 104, 123, 128, 129) |

[a]The degree to which these tests are robust to the underlying demographic assumptions is controversial and has not been fully explored.

reducing the fitness of the organism (46, 75). This type of genomic conflict may, for example, occur in loci associated with centromeres (46, 66, 67) or involved in apoptosis during spermatogenesis (75). Positive selection in terms of elevated $d_N/d_S$ ratios tend to detect selection situations where repeated selective fixations have occurred in the same gene or in the same site, due to a continued dynamic interaction. In contrast, population genetic methods have the ability to detect selection on a single adaptive mutation that recently has swept through the population.

So far, very little research has been done to detect positive selection in noncoding regions based on comparative data. Although methods similar to those used to detect elevated $d_N/d_S$ ratios can be devised for noncoding regions (119), sites in noncoding regions cannot easily be divided into possible selected sites and nonselected sites, similarly to nonsynonymous and synonymous sites in coding regions. Nonetheless, the presence of highly variable sites in noncoding regions may be signs of positive selection, and methods to identify such sites may find good use in the analysis of comparative genomic data. A serious practical problem that may arise in the application of such methods is the possibility of confounding misalignments with hypervariable regions.

Most of the literature on statistical methods for detecting selection from comparative data (e.g., from $d_N/d_S$ ratios) and from population genetic data has been poorly connected. Although the comparative approaches have provided the most unambiguous evidence for positive selection, results have rarely been interpreted in terms of population genetic theory. One probable reason is that multiple population genetic models could generate the same pattern of observed $d_N/d_S$ ratios, and that any detailed inferences of population genetic processes using comparative data would be based on a very strong assumptions regarding the way fitnesses are assigned to mutations (79). Comparative data in themselves are, therefore, unlikely to provide more detailed information regarding population genetic processes but relatively vague assertions of positive and negative selection and their distribution in the genome. Inferences regarding the type of negative or positive selection operating (e.g., balancing versus positive directional selection) must involve population genetic data. Moreover, comparative approaches cannot alone determine if selection is currently acting in a population. For such inferences population genetic data are also needed.

## GENOMIC APPROACHES

The availability of large-scale genomic data has created new challenges and opportunities, especially in allowing for more nonparametric outlier analyses. Genes with increased levels of LD, reduced or enhanced levels of variability, increased levels of population differentiation, or skewed allele frequency spectra may be good candidates for selected loci. Recently, there has been heightened interest particularly in using increased population subdivision among populations as a method for detecting selection (1, 9, 44, 53, 64, 92, 93, 101, 102, 114). For example, Akey et al. (1) used variation in $F_{ST}$ (a common measure of population subdivision) in the human genome to identify regions of increased population subdivision.

However, the availability of genomic data does not solve the fundamental problem that population-level demographic processes and selection are confounded. Many demographic processes, such as certain types of population subdivision, may increase the variance in the statistics used to detect selection. Certain demographic models are, therefore, more likely than other models to produce outliers. The outlier approach in population genetics does not solve the problem that a postulated signature of selection, inferred from population genetic data, may instead be the product of complicated demographics. Nonetheless, certain approaches based on detecting extreme levels of population subdivision seem to have some robustness to the model assumptions (9, 114).

## PRF Models

**PRF:** Poisson random field

The simultaneous analysis of multiple genomic loci allows the estimation of parameters that are common among loci, potentially leading to increased power and robustness. For example, Bustamante et al. (16) analyzed MacDonald-Kreitman tables from *Arabidopsis* and *Drosophila* in a statistical framework that allows the divergence time between species to be a shared parameter among all loci, leading to increased statistical power. Similar approaches can be used to increase the robustness of the statistical methods by explicitly estimating demographic parameters, thereby taking the uncertainty introduced by the unknown demographic processes into account. This is particularly convenient in the framework of Poisson random field (PRF) models introduced by Sawyer & Hartl (88). These models assume that all loci (individual SNP sites) are independent, i.e., effectively unlinked. This implies that they may provide a good approximation in the analysis of SNP data from multiple locations throughout the genome, but less so in the analysis of DNA sequence data from a single or a few loci. In these models, the expected frequency spectrum (or the entries of a MacDonald-Kreitman table) can be calculated directly using mathematical models. This means that selection coefficients for particular classes of mutations can be estimated directly, and various hypotheses regarding selection can be tested in a rigorous statistical framework (15–17, 89). For example, it is possible to estimate which types of amino acid-changing mutations have the largest effect on fitness (15, 118). Such methods may eventually be very useful when designing statistical methods for predicting which mutations are most likely to cause disease. However, inferences based on PRF models differ fundamentally from most other methods for identifying selection, because the effect of selection on linked neutral sites is not incorporated into the models. Whereas most methods for detecting positive selection in terms of selective sweeps consider the effect of a positively selected mutation on the nearby neutral variation, PRF models provide predictions regarding the selected mutation itself. In most applications, estimates based on PRF models will, therefore, be biased (17). Nonetheless, the PRF models provide a convenient computationally tractable statistical framework for examining the effect of selection on different classes of mutations.

Williamson et al. (118) used PRF models to estimate the average selection coefficient acting on different classes of mutations in the human genome. The novelty of their approach (118) was that a demographic model was fitted to the data from synonymous mutations, while selection coefficients were estimated for the same demographic model applied to nonsynonymous mutations. The resulting test was shown to be robust to many different assumptions regarding demographic processes. By explicitly incorporating demography into the model, a high degree of robustness was achieved. Unfortunately, there are no similar approaches for detecting selection from individual loci containing multiple linked mutations. The current methods for taking demographic processes into account when analyzing data from loci with linked mutations involve extensive simulations of data under various demographic models (3, 75, 116).

## SNP Data

With the availability of large-scale SNP data sets, it should, in principle, be possible to provide detailed selection maps in humans and other organisms. Standard methods for detecting selection from population genetics can, in principle, be applied to provide a detailed picture of the regions of the genome that may have been targeted by selection. However, most SNP data have been obtained through a complicated SNP discovery process that minimally involves the discovery (or ascertainment) of SNPs in a small sample followed by genotyping in a larger sample. The process by which the SNPs have been selected affects levels of LD observed in the data (77),

the frequency spectrum (77), and levels of population subdivision (74, 115). It also affects the variance in these statistics, complicating genomic methods based on outlier detection. The solution to this problem is to explicitly take the ascertainment process into account. Most statistical methods can be corrected relatively easily (76, 77), leading to new valid methods for detecting selection that take the SNP ascertainment process into account. Unfortunately, most current SNP databases and large-scale SNP genotyping efforts (37) are not associated with sufficiently detailed information regarding the ascertainment process necessary for appropriate ascertainment bias corrections. At present, it is difficult or impossible to make valid inferences regarding selection from most large-scale SNP data sources. It is to be hoped that this will change in the future as researchers become more aware of the importance in maintaining detailed records regarding SNP ascertainment processes.

### Comparative Genomic Data

As more and more genomes are sequenced, comparative approaches for detecting positive selection at a genome-wide scale are becoming increasingly common (22, 75). The standard methods for detecting positive (or negative) selection using $d_N/d_S$ ratios can be applied directly in studies on a genomic scale. However, current methods can be improved by establishing models that take advantage of the fact that (ignoring within-species variability) all genes in a phylogeny share the same evolutionary tree.

## FUNCTIONAL INFERENCES

In the field of bioinformatics there has been a long tradition of using conserved sites in comparative data to infer function. The implicit assumption is that high levels of conservation are caused by negative selection against new deleterious selection, i.e., functional constraints. In the absence of site-specific suppression of the biological mutation rate, highly reduced levels of variability must be caused by negative selection.

### Phylogenetic Footprinting

Although there exist many methods for quantifying how conserved a site, or a set of sites, is, the most statistically solid methods for identifying conserved sites are known as phylogenetic footprinting. In these methods, the rate of substitution in a particular site (or collection of sites) is estimated by considering the pattern of mutation along the underlying phylogeny. This is typically done by mapping mutations onto the phylogenetic tree using parsimony (12) and is complicated by the fact that the alignment may be ambiguous in noncoding regions for divergent species. These methods have been used for a variety of purposes and have been particularly successful in identifying regulatory elements in noncoding DNA (24, 111). The advantage of these methods is that they explicitly take the underlying evolutionary correlations (the phylogeny) into account, leading to increased statistical power and accuracy over methods that do not consider the phylogeny.

One of the most exciting recent discoveries in the field of genomics is the presence of extremely conserved regions, with no known function, in mammalian genomes (11). Such regions may be regulatory regions, containing conserved structural features or unannotated protein-coding genes or RNA genes. To determine if these regions are truly under selection, neutrality tests comparing intraspecific and interspecific variability could be used. There is even the possibility of positive selection in noncoding regions. More research is needed to develop appropriate statistical methods for identifying selection outside coding regions from genomic scale comparative and population genetic data.

### Disease Genetics

In disease genetics, there is an increased awareness that regions of the human genome

that have been targeted by positive selection may be disease associated (7). Disease-causing mutations should affect organismal fitness, except if the age of onset of the disease is very late. There is, therefore, an intimate relationship between disease and selection that potentially can be exploited in identifying candidate disease loci and candidate SNPs.

A very promising application is in the identification of putative disease-causing SNPs. Evolutionary inferences from comparative and population genetic data, in combination with functional and structural information, can be used to predict which mutations most likely have negative fitness consequences. The mutations with the most severe fitness consequences are obviously the mutations that are most likely to be disease causing. Several different methods have already been described that allow predicting of potential disease-causing mutation (72, 84). These methods may potentially be improved by using explicit population genetic models. This seems to be a particularly promising application of PRF models as these models can describe explicitly the selection coefficients acting on particular classes of mutations (15).

### Positive Selection

While there has long been a focus on the use of conservation (negative selection) to find functional elements, increased attention has recently been directed toward the possibility of using inferences regarding positive selection to elucidate functional relationships. In human genetics, several cases are known where recessive disease-causing mutations were thought to be carried to high frequencies in the populations, because they confer a fitness advantage in the heterozygote condition. Diseases that have been hypothesized to have been targeted by this type of overdominant selection include sickle-cell anemia (42), glucose-6-phosphate dehydrogenase deficiency (86), Tay-Sachs disease (99), cystic fibrosis (94), and Phenylketnonuria (121). Not known is how many of the common disease factors have been influenced by overdominant selection, but these observations do suggest that regions of the human genome that have been targeted by balancing selection may contain disease-causing variants worth exploring.

In virology, site-specific inferences regarding positive selection have been used in several cases to identify functionally important sites. In the HIV virus, site-specific inferences of $d_N/d_S$ ratios have been used to identify positions that may be involved in drug resistance (21). In HIV and other viruses, sites that may interact with the host immune system have been identified by detecting site-specific selective pressures, and it has been proposed that such methods may assist in the development of vaccines (36, 95). It has also been proposed that site-specific inferences of $d_N/d_S$ ratios may help predict the evolution of virulent strains of influenza (14). Recently, site-specific inferences of $d_N/d_S$ ratios from different primate species were used to identify a new species-specific retroviral restriction domain (91).

## EVIDENCE FOR SELECTION

There is an increasing amount of evidence that selection is important in shaping variation within and between species. In human SNP data, there is a clear difference in the frequency spectrum between nonsynonymous and synonymous mutations (103, 118). This observation in itself shows that a large proportion of the mutations that are segregating in humans (and presumably in other species as well) are affected by selection. In addition, there is a rapidly growing list of specific genes that show evidence for positive selection in both humans and other organisms (7, 31, 98, 113, 125). This explosion of results showing a presence of positive selection may in fact suggest that positive selection is much more common than previously believed. Positively selected mutations may just have remained hidden among all the negatively selected

mutations. In addition, ambiguity in the interpretation of classical population genetic neutrality tests, due to the presence of confounding demographic factors, may have precluded the establishment of firm conclusions regarding the pervasiveness of selection. As more large-scale data have accumulated, and methods that are robust to demographic assumptions have been applied, a clearer picture of the pervasiveness of positive selection has been established. Modern versions of the neutral theory (80, 81) allow for a substantial amount of negative selection, and even some positive selection. As the evidence for selection accumulates, the debate regarding the causes of molecular evolution should focus on whether selection is so dominating that effective population sizes and standing levels of variation are best described by the models of repeated selective sweeps favored by Gillespie (40, 41), or whether classical models of genetic drift are most appropriate. In the models that Gillespie has proposed, known as genetic draft models, mutations causing species differences are not neutral mutations increasing in frequency due to genetic drift, but primarily neutral mutations increasing in frequency due to linkage with adaptive mutations sweeping through the population. Even though only few mutations are adaptive, the population genetic dynamics is determined by the selective forces acting on the adaptive mutations, not by genetic drift. There is no mathematical or empirical evidence to suggest that this model is unrealistic, and as the evidence in favor of positive selection accumulates, the question arises whether models of draft should replace models of drift.

With the new availability of very large population genetic and comparative genomic data sets, we should soon be able to determine how many genes, and how big a proportion of mutations, have been affected by positive and negative selection. This will also lead to more evolutionary explorations into the molecular nature of adaptation, help predict which SNPs in humans may be disease associated, and lead to improved functional annotations of genomic data. Methods that combine comparative and population genetic data, and methods that have a high degree of robustness to the underlying demographic factors may be particularly useful in this endeavor.

## SUMMARY POINTS

1. Both positive and negative selection leave distinctive signatures at the molecular level that can be detected using statistical tests.
2. In population genetic data, selection may affect levels of variability, linkage disequilibrium, haplotype structure and allelic distribution in each nucleotide site (frequency spectrum). In comparative data, selection has a strong effect on the $d_N/d_S$ ratio.
3. Statistical methods for detecting selection differ in the assumptions they make and how powerful they are. Most methods applicable to population genetic data rely on strong assumptions regarding the demography of the populations, while comparative methods are free of such assumptions.
4. An increasing amount of evidence suggests that positive selection is much more pervasive than previously thought.
5. Inferences regarding selection provide a powerful tool in functional studies, for example for the prediction of possible disease-related genomic regions.

## UNRESOLVED ISSUES

1. Can robust statistical population genetic tests be developed that can help identify genomic regions targeted by positive selection?
2. Will inferences regarding selection help identify disease loci in humans and other organisms?
3. Should we focus on genetic draft instead of genetic drift?

## LITERATURE CITED


1. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–14
2. **Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.* 11:635–41**
3. Andolfatto P, Przeworski M. 2000. A genome-wide departure from the standard neutral model in natural populations of Drosophila. *Genetics* 156:257–68
4. 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* 153:1297–311
5. Anisimova M, Bielawski JP, Yang ZH. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18:1585–92
6. Anisimova M, Nielsen R, Yang ZH. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229–36
7. Bamshad M, Wooding SP. 2003. Signature of natural selection in the human genome. *Nat. Rev. Genet.* 4:99
8. Barton NH. 1995. Linkage and the limits to natural selection. *Genetics* 140:821–41
9. Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* 13:969–80
10. Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. London Ser. B* 263:1619–26
11. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–25
12. Blanchette M, Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* 12:739–48
13. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms. *Genetics* 140:783–96
14. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. 1999. Predicting the evolution of human influenza A. *Science* 286:1921–25
15. Bustamante CD, Nielsen R, Hartl DL. 2003. Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* 63:91–103
16. Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in Arabidopsis. *Nature* 416:531–34
17. Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159:1779–88


**A related review focusing on the problem of distinguishing background selection from selective sweeps, with particular focus on *Drosophila* populations.**

18. Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* 63:213–27
19. Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–303
20. Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* 70:155–74
21. Chen L, Perlina A, Lee CJ. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol.* 78:3722–32
22. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–63
23. Depaulis F, Veuille M. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol.* 15:1788–90
24. Duret L, Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* 7:399–406
25. **Ewens WJ. 2004. *Mathematical Population Genetics*. I. *Theoretical Introduction*. Berlin/Heidelberg/New York: Springer**
26. Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* 162:2017–24
27. Fares MA, Moya A, Escarmis C, Baranowski E, Domingo E, Barrio E. 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. *Mol. Biol. Evol.* 18:10
28. Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–13
29. Felsenstein J. 1974. Evolutionary advantage of recombination. *Genetics* 78:737–56
30. Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. USA* 94:7712–18
31. Ford MJ. 2002. Applications of selective neutrality tests to molecular ecology. *Mol. Ecol.* 11:1245–62
32. Fu YX. 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics* 143:557–70
33. Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915–25
34. Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709
35. Galtier N, Depaulis F, Barton NH. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* 155:981–87
36. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, et al. 2002. Diversity considerations in HIV-1 vaccine selection. *Science* 299:1515–18
37. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, et al. 2003. The International HapMap Project. *Nature* 426:789–96
38. Gillespi JH, Langley CH. 1974. General model to account for enzyme variation in natural populations. *Genetics* 76:837–84
39. Gillespie JH. 1991. *The Causes of Molecular Evolution*. New York: Oxford Univ. Press. 336 pp.
40. Gillespie JH. 2000. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155:909–19

**A great introduction to population genetic theory for the mathematically literate.**

41. Gillespie JH. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55:2161–69
42. Haldane JBS. 1949. Disease and evolution. *Ricerca Sci.* 19:3–10
43. Haldane JBS. 1957. The cost of natural selection. *Genetics* 55:511–24
44. Harr B, Kauer M, Schlotterer C. 2002. Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 99:12949–54
45. Haydon DT, Bastos AD, Knowles NJ, Samuel AR. 2001. Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics* 157:7
46. Henikoff S, Malik HS. 2002. Selfish drivers. *Nature* 417:227
47. Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. 1994. Evidence for positive selection in the superoxide-dismutase (Sod) region of *Drosophila-melanogaster*. *Genetics* 136:1329–40
48. Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–59
49. Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. *Nature* 335:167–70
50. Ingvarsson PK. 2004. Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Genet. Res.* 83:31–39
51. Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. *Genetics* 120:819–29
52. Kaplan NL, Hudson RR, Langley CH. 1989. The hitchhiking effect revisited. *Genetics* 123:887–99
53. Kayser M, Brauer S, Stoneking M. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol. Biol. Evol.* 20:893–900
54. Kelly JK. 1997. A test of neutrality based on interlocus associations. *Genetics* 146:1197–206
55. Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167:1513–24
56. Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160:765–77
57. Kim Y, Stephan W. 2003. Selective sweeps in the presence of interference among partially linked loci. *Genetics* 164:389–98
58. Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624
59. Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. New York: Cambridge Univ. Press. 367 pp.
60. Kimura M. 1995. Limitations of Darwinian selection in a finite population. *Proc. Natl. Acad. Sci. USA* 92:2343–44
61. Kimura M, Crow J. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 40:725–38
62. Kondrashov AS. 1982. Selection against harmful mutations in large sexual and asexual populations. *Genet. Res.* 40:325–32
63. Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* 74:175–95
64. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4:981–94

65. Majewski J, Cohan FM. 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* 152:1459–74
66. Malik HS, Henikoff S. 2001. Adaptive evolution of cid, a centromere-specific histone in Drosophila. *Genetics* 157:1293–98
67. Malik HS, Henikoff S. 2002. Conflict begets complexity: the evolution of centromeres. *Curr. Opin. Genet. Dev.* 12:711–18
**68. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. *Genet. Res.* 23:23–35**
69. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652–54
70. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–84
71. Milkman RD. 1967. Heterosis as a major cause of heterozygosity in nature. *Genetics* 55:493–95
72. Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res.* 11:863–74
73. Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* 86:641–47
74. Nielsen R. 2004. Population genetic analysis of ascertained SNP data. *Human Genomics* 3:218–24
75. Nielsen R, Bustamante CD, Clark AG, Glanowski S, Sackton TB, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* In press
76. Nielsen R, Hubisz MJ, Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168:2373–82
77. Nielsen R, Signorovitch J. 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* 63:245–55
78. Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–36
79. Nielsen R, Yang ZH. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol.* 20:1231–39
80. Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23:263–86
81. Ohta T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. USA* 99:16134–37
82. Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* 160:1179–89
83. ¨Ptak SE, Roeder AD, Stephens M, Gilad Y, Paabo S, Przeworski M. 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol.* 2:849–55
84. Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30:3894–900
85. Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* 19:2092–100
86. Ruwende C, Khoo SC, Snow AW, Yates SNR, Kwiatkowski D, et al. 1995. Natural-selection of hemizygotes and heterozygotes for G6pd deficiency in Africa by resistance to severe malaria. *Nature* 376:246–49

**The classic paper introducing the idea of a selective sweep.**

87. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–37

**The first paper introducing PRF models as a statistical framework for population genetic inferences.**

**88. Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132:1161–76**

89. Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *J. Mol. Evol.* 57(Suppl.) 1:S154–64

90. Sawyer SL, Emerman M, Malik HS. 2004. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol.* 2:1278–85

**An elegant paper showing how inferences regarding selection can be used to make functional predictions that can be tested in the lab.**

**91. Sawyer SL, Wu LI, Emerman M, Malik HS. 2005. Positive selection of primate TRIM5 alpha identifies a critical species-specific retroviral restriction domain. *Proc. Natl. Acad. Sci. USA* 102:2832–37**

92. Schlotterer C. 2002. A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160:753–63

93. Schlotterer C. 2003. Hitchhiking mapping—functional genomics from the population genetics perspective. *Trends Genet.* 19:32–38

94. Schroeder SA, Gaughan DM, Swift M. 1995. Protection against bronchial-asthma by Cftr Delta-F508 mutation—a heterozygote advantage in cystic-fibrosis. *Nat. Med.* 1:703–5

95. Sheridan I, Pybus OG, Holmes EC, Klenerman P. 2004. High-resolution phylogenetic analysis of Hepatitis C virus adaptation and its relationship to disease progression. *J. Virol.* 78:3447–54

96. Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141:413–29

97. Slatkin M, Wiehe T. 1998. Genetic hitch-hiking in a subdivided population. *Genet. Res.* 71:155–60

98. Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* 415:1022–24

99. Spyropoulos B, Moens PB, Davidson J, Lowden JA. 1981. Heterozygote advantage in Tay-Sachs carriers. *Am. J. Hum. Genet.* 33:375–80

100. Stahl EA, Bishop JG. 2000. Plant-pathogen arms races at the molecular level. *Curr. Opin. Plant Biol.* 3:299

**A related review of methods for detecting selection in genomic scans, with particular focus on statistics based on population differentiation.**

**101. Storz JF. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.* 14:671–88**

102. Storz JF, Payseur BA, Nachman MW. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Mol. Biol. Evol.* 21:1800–11

103. Sunyaev SR III WCL, Ramensky VE, Bork P. 2000. SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.* 16:335–37

104. Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16:1315–28

105. Suzuki Y, Nei M. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 19:1865–69

106. Suzuki Y, Nei M. 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. *Mol. Biol. Evol.* 21:914–21

107. Sved JA, Reed TE, Bodmer WF. 1967. The number of balanced polymorphisms that can be maintained in a natural population. *Genetics* 55:469–81

108. Swanson WJ, Aquadro CF, Vacquier VD. 2001. Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive Darwinian selection of sperm lysin. *Mol. Biol. Evol.* 18:376–83
109. Swanson WJ, Nielsen R, Yang QF. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol.* 20:18–20
110. Swanson WJ, Zhang ZH, Wolfner MF, Aquadro CF. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. USA* 98:2509–14
111. Tagle D, Koop B, Goodman M, Slightom J, Hess D, Jones R. 1988. Embryonic $\varepsilon$ and $\gamma$ globin genes of a prosimian primate (*Galago crassicaudatus*); nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 203:439–55
**112. Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–95**
113. Vallender EJ, Lahn BT. 2004. Positive selection on the human genome. *Hum. Mol. Genet.* 13:R245–54
114. Vitalis R, Dawson K, Boursot P. 2001. Interpretation of variation across marker loci as evidence of selection. *Genetics* 158:1811–23
115. Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K. 2001. The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *Am. J. Hum. Genet.* 69:1332–47
**116. Wall JD, Andolfatto P, Przeworski M. 2002. Testing models of selection and demography in *Drosophila simulans*. *Genetics* 162:203–16**
117. Wall JD, Frisse LA, Hudson RR, Di Rienzo A. 2003. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *Am. J. Hum. Genet.* 73:1330–40
118. Williamson SH, Hernadez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA.* 102:7882–87
119. Wong WSW, Nielsen R. 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* 167:949–58
120. Wong WSW, Yang ZH, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–51
121. Woolf LI, McBean MS, Woolf FM, Cahalane SF. 1975. Phenylketonuria as a balanced polymorphism—nature of heterozygote advantage. *Ann. Hum. Genet.* 38:461–69
122. Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304
123. Yang ZH. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568–73
124. Yang ZH. 2002. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* 12:688–94
**125. Yang ZH, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends. Ecol. Evol.* 15:496–503**
126. Yang ZH, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46:409–18
127. Yang ZH, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17:32–43

**Tajima's classic paper introducing his well-known neutrality test.**

**A nice study showing how both selection and demography must be taken into account when interpreting genetic data.**

**A review of the statistical methodology used to detect positive selection using $d_N/d_S$ ratios.**

128. Yang ZH, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19:908–17
129. Yang ZH, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–49

## RELATED RESOURCES

Fay JC, Wu C-I. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet*. 4:213–35
Lewontin RC. 2002. Directions in evolutionary biology. *Annu Rev. Genet.* 36:1–18
Kreitman M. 2000. Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* 1:539–59