10

Using Molecular Data to Detect Selection: Signatures From Multiple Historical Events

Model selection is a process of seeking the least inadequate model from a predefined set, all of which may be grossly inadequate as a representation of reality — Welch (2006)

Draft version 11 April 2014

Chapter 9 reviewed tests for detecting an ongoing, or very recently completed, single *episode* of positive selection. Here we examine a complementary issue, the cumulative signature left by *multiple* historical selection events. In contrast to the large variety of test for detecting ongoing/recent selection, there are only two basic approaches that use divergence data to detect the signal from multiple episodes of positive selection. The first is to contrast the levels of polymorphism within a population with the level of divergence between populations, using either different classes of sites within the same gene (the **McDonald-Kreitman**, or **MK**, **test**) or different genes (the **HKA** test). Since these require a population sample to determine the amount of polymorphism, we refer to them as **population-based divergence tests**. The second approach contrasts the rates of evolution at different sites within a gene over a number of species within a phylogeny. Such **phylogeny-based divergence tests** do not require a population sample, as the signal comes entirely from the pattern of divergence, not polymorphism. Specifically, in the absense of positive selection, the rate of replacement substitutions is generally expected to be less than the rate of silent substitutions. Sites where the replacement rate *exceeds* the silent substitution rate provides a very robust signal of positive selection (e.g., Example 10.2). Hybrid population genetics-phylogenetics tests are starting to also appear (e.g., Wilson et al. 2011), but we will not examine these here.

While population-based divergence tests look for an excess of replacement substitutions *across* an entire gene when comparing two populations, divergence-based tests look for an excess of substitutions at *single codons* when examined over an *entire phylogeny*. As a result, phylogeny-based divergence tests likely detect the smallest percentage of selection, as they require a very special class of events: repeated positive selection on the same codon over a number of species. By contrast, the HKA and MK tests are less stringent, simply requiring multiple substitutions over the *entire gene*. Since the power of divergence-based methods is a function of the number of substitutions, their significant limitation is that one or even a few very important adaptive substitutions would leave little signal.

Tests from the previous chapter and the approaches examined here are complementary, with each picking up signals of selection that would be missed by the others. Tests for ongoing events cast a wide net in that many ongoing/recent events leave some signal, albeit perhaps very weak. However, this signal decays very quickly, so that most sites experiencing positive selection over some modest amount of evolutionary time in the past will leave no signal for these tests. Conversely, HKA and MK tests entirely miss genes with a single or a very few relatively recent adaptive substitutions, but can detect genes where multiple adaptive substitutions have occurred across the gene during the divergence of two populations. Phylogeny-based tests are even more restrictive, showing a signal only in very special cases: the same *codons* are repeated targets of selection over the species in a phylogeny. The time scales over which events can be detected also vary over approaches, with only

very recent single events having the potential to be detected using methods from Chapter 9. HKA and MK tests require enough time for a sufficient number of adaptive substitutions to accumulate in the divergence of a gene between two populations/species, while phylogeny-based tests (generally) require even longer time, allowing multiple substitutions to occur at the same codon across a number of species. An exception is the analysis of rapid-evolving viruses (such as HIV), whose high mutation rate allows the signatures of multiple rounds of selection to be found within a phylogeny generated over a very modest amount of time.

BRIEF OVERVIEW OF THE KEY CONCEPTS

We start with a short overview of population- versus phylogeny-based approaches before considering each in considerable detail. As was done in Chapter 9, this section introduces the key ideas without all of the technical burdens.

A History of Selection Alters the Ratio of Polymorphic to Divergent Sites

Population-based tests contrast the patterns of within-species polymorphism and between-species divergence to see if they are in concordance with their neutral expectations. Under the equilibrium neutral model, the standing heterozygosity (H, or the nucleotide diveristy π, see Chapter 2) and between-population divergence (d) for the ith gene being considered are

$$H_i = 4N_e\mu_i, \qquad d_i = 2t\mu_i \tag{10.1a}$$

Hence,

$$\frac{H_i}{d_i} = \frac{4N_e\mu_i}{2t\mu_i} = \frac{2N_e}{t} \tag{10.1b}$$

Since the gene-specific mutation rates cancel, under the equilibrium neutral model the H/d ratio at any locus should be roughly the same, namely $2N_e/t$ (subject to random sampling).

Example 10.1. McDonald and Kreitman (1991a) examined the *Adh* (Alcohol dehydrogenase) locus in the sibling species *Drosophila melanogaster* and *D. simulans*, as well as an outgroup species, *D. yakuba*. Within this gene, they contrasted **replacement** (**nonsynonymous**) and **silent** (**synonymous**) sites. The DNA change for a replacement mutation results in a change in an amino acid, while a silent mutation still codes for the ancestral animo acid. Equation 10.1b indicates that the ratio of number of polymorphisms to number of fixed differences should be the same for both categories. This is a simple association test, and significance can be assessed using either a χ^2 approximation or (much better) Fisher's exact test which accommodates small numbers in the observed table entries. Of the 24 fixed differences, 7 were replacement and 17 synonymous. The total number of polymorphic sites segregating in either species was 44, 2 of which were replacement and 42 synonymous. The resulting association table becomes

	Fixed	Polymorphic
Synonymous	17	42
Replacement	7	2

Fisher's exact tests gives a $p = 0.0073$, a highly significant lack of fit to the neutral equilibrium model. Based on the ratio of 42/2 syn/repl polymorphisms, the expected number x of replacement fixations is 17/ x = 42/2 , or x = 0.81. Since one replacement polymorphism is expected under drift, while 7 were seen, this suggests roughly 6 adaptive substitutions, or that 86% of the *Adh* substitutions between these species are adaptive.

A History of (Positive) Selection Alters the Silent to Replacement Substitution Rates

Phylogeny-based divergence tests do not require polymorphism data, but rather simply contrast the divergence rates at silent versus replacement sites. Silent sites are treated as proxies for neutral sites, although we have seen that they may be under weak selection (Chapter 8). Mutations at replacement sites are generally viewed as being under much stronger selection, much of it purifying. Indeed, the signal of *negative* selection (removal of new deleterious mutations) is widespread in just about every protein-coding gene, with the substitution rate of silent sites usually being much higher than that for replacement sites when averaged over the entire gene. This pattern is expected under the neutral theory if a higher fraction of mutations in replacement sites are deleterious relative to silent sites. However, there are cases where, for a limited region within a gene, the replacement substitution rate can exceed that of the silent sites (Example 10.2), suggesting adaptive fixation (i.e., positive selection).

Example 10.2. One of the classic examples of using sequence data to detect signatures of positive selection is the work of Hughes and Nei (1988, 1989). They examined mice and human major histocompatibility complex (MHC) Class I and Class II loci, highly polymorphic genes involved in antigen-recognition. Hughes and Nei compared the ratio of synonymous to nonsynonymous nucleotide substitution rates in the putative antigen-recognition sites versus the rest of these genes. For both classes of loci, they found a significant excess of nonsynonymous substitutions in the recognition sites and a significant deficiency of such substitutions elsewhere. If both types of substitutions were neutral, the per-site rates are expected to be roughly equal. If negative selection is acting, the expectation is that the synonymous substitution rate would be significantly higher (reflecting removal of deleterious nonsynonymous mutations). However, if positive selection is sufficiently common for new mutations, one expects an excess of nonsynonymous substitutions. The observed patterns for both Class I and II loci were consistent with positive selection within that part of the gene coding for the antigen recognition site and purifying selection on the rest of the gene.

A large number of studies prior to Hughes and Nei found that an excess of synonymous substitutions is by far the norm for almost all genes, implying that most nonsynonymous changes are selected against. Indeed, when one looks over an entire Class I (or II) MHC gene, this pattern is also seen. The insight of Hughes and Nei was to use data on protein structure to specifically focus on the putative antigen-binding site, and compare this region with the rest of the gene as an internal control. Further, there has to be a consistent pattern of new mutations being favored at the same few sites for such a signature to appear. A single favorable new mutation here and there through the evolution of a gene, when set against the background of most nonsynonymous mutants being deleterious, will still leave an overall signature of a vast excess of synonymous substitutions.

While there are several variant notations in the literature, we use K_s to denote the synonymous (silent) substitution rate and K_a the nonsynonymous rate (a denoting a change in an animo acid; K_{ns} and K_n for nonsynonymous is also used in the literature). A value of $K_a/K_s > 1$ indicates a long-term pattern of positive selection at replacement sites. As Example 10.2 illustrates, even if this is occurring at *specific regions* within a gene, when averaged over the *entire* gene K_a/K_s is usually less than one. Hughes and Nei had the two advantages of (i) focusing on potential sites of positive selection (so the other sites could be excluded in the analysis) and (ii) a situation in which numerous replacements driven by positive selection are expected. Thus, while an observation of $K_a/K_s > 1$ is almost universally accepted as a selection signature (more correctly, a signature of a long-term

pattern of multiple episodes of positive selection), it is almost never seen if the entire gene is taken as the unit of analysis. Phylogeny-based methods accommodate this concern by taking the codon as the unit of analysis, first placing genes within a phylogeny and then using codon-evolution models to test whether, for some subset of codons, $K_a/K_s > 1$.

POPULATION-BASED DIVERGENCE TESTS

We now turn to detailed discusses of these two different approaches, starting with methods that require polymorphism data, and hence a sample of many individuals from a single population.

The Hudson-Kreitman-Aguadé (HKA) Test

Hudson, Kreitman, and Aguadé (1987) proposed the first test to jointly use information on the amount of polymorphism within populations and the amount of divergence between populations/species. The result was their **HKA test**, which is formulated as follows. Consider two species (or distant populations) A and B both at mutation-drift equilibrium with effective population sizes $N_A = N_e$ and $N_B = \delta N_e$. Further assume they separated $\tau = t/(2N_e)$ generations ago from a common population of size $N_e^* = (N_A + N_B)/2 = N_e(1+\delta)/2$, the average of the two current population sizes. Suppose $i = 1, \cdots, L$ unlinked loci are examined in both species. We allow the neutral mutation rate μ_i to vary over loci, but assume (for a given locus) it has been the same in both species, and hence also unchanged during the divergence. The amount of polymorphism for locus i is a function of $\theta_i = 4N_e\mu_i$ in species A, and $4N_B\mu_i = 4(\delta N_e)\mu_i = \delta\theta_i$ in species B. The divergence between A and B is $2t\mu_i$, which we can express as

$$2t\mu_i = 2\,\frac{t}{2N_e}\,2N_e\mu_i = \tau\theta_i$$

For L loci, there are $L+2$ unknowns: L gene-specific values of θ_i and the two common demographic parameters δ and τ. To estimate these we have up to $3L$ observations: the numbers S_i^A and S_i^B of segregating sites at each of the L loci in each species/population, and the number D_i of substitutions between each pair of L loci. One uses the data to first estimate the model parameters, and then performs a goodness-of-fit test. If the model provides a sufficiently poor fit, the equilibrium neutral model is rejected.

More formally, the HKA test statistic X^2 is given by

$$X^2 = \sum_{i=1}^{L} X_i^2 \tag{10.2a}$$

where

$$X_i^2 = \frac{\left(S_i^A - \widehat{E}(S_i^A)\right)^2}{\widehat{Var}(S_i^A)} + \frac{\left(S_i^B - \widehat{E}(S_i^B)\right)^2}{\widehat{Var}(S_i^B)} + \frac{\left(D_i - \widehat{E}(D_i)\right)^2}{\widehat{Var}(D_i)} \tag{10.2b}$$

is the contribution to overall lack-of-fit from gene i. For n_A samples from species A and n_B samples from species B,

$$\widehat{E}(S_i^A) = \widehat{\theta}_i a_{n_A}, \quad \widehat{E}(S_i^A) = \widehat{\delta}\,\widehat{\theta}_i a_{n_B} \tag{10.3a}$$

$$\widehat{Var}(S_i^A) = \widehat{\theta}_i a_{n_A} + \widehat{\theta}_i^{\,2} b_{n_A}, \quad \widehat{Var}(S_i^B) = \widehat{\delta}\,\widehat{\theta}_i a_{n_A} + \widehat{\delta}^{\,2}\,\widehat{\theta}_i^{\,2} b_{n_B} \tag{10.3b}$$

$$\widehat{E}(D_i) = \widehat{\theta}_i \left(\widehat{\tau} + \frac{1+\widehat{\delta}}{2}\right) \tag{10.3c}$$

$$\widehat{Var}(D_i) = \widehat{\theta}_i \left(\widehat{\tau} + \frac{1+\widehat{\delta}}{2} \right) + \left(\frac{\widehat{\theta}_i(1+\widehat{\delta})}{2} \right)^2 \tag{10.3d}$$

where $a_n = \sum_{i=1}^{n-1}(1/i)$ and $b_n = \sum_{i=1}^{n-1}(1/i^2)$. Equations 10.3a and 10.3b follow from the infinite-sites model (Equation 9.21a). Equation 10.3c follows by re-writing

$$\theta_i \left(\tau + \frac{1+\delta}{2} \right) = 4N_e\mu_i \left(\frac{t}{2N_e} + \frac{1+\delta}{2} \right) = 2\mu_i t + 4\mu_i \frac{N_e(1+\delta)}{2},$$

where the first term is the between-population divergence due to new mutations and the second term the divergence from partitioning of the initial polymorphism $4N_e^*\mu_i$ present in the ancestral population. The HKA test statistic X^2 is approximately χ^2-distributed with $3L - (L+2) = 2L - 2$ degrees of freedom, given $3L$ observations and $L+2$ parameters to estimate. Hudson et al. suggest the following system of equations for the estimating the unknowns given the observed values S_i^A, S_i^B, D_i,

$$\begin{aligned}
\sum_{i=1}^{L} S_i^A &= a_{n_A} \sum_{i=1}^{L} \widehat{\theta}_i, \qquad \sum_{i=1}^{L} S_i^B = \widehat{\delta}\, a_{n_B} \sum_{i=1}^{L} \widehat{\theta}_i \\
\sum_{i=1}^{L} D_i &= \left(\widehat{\tau} + \frac{1+\widehat{\delta}}{2} \right) \sum_{i=1}^{L} \widehat{\theta}_i \\
S_i^A + S_i^B + D_i &= \widehat{\theta}_i \left(\widehat{\tau} + \frac{1+\widehat{\delta}}{2} + a_{n_A} + \widehat{\delta} \cdot a_{n_B} \right) \quad \text{for } 1 = 1, \cdots, L-1
\end{aligned} \tag{10.4}$$

These can be solved numerically for the $L\,\widehat{\theta}_i$ values unique to each locus and the two common $\widehat{\delta}$ and $\widehat{\tau}$ values, generating the estimated values for the X^2 statistic. The HKA model assumes no recombination within a gene but free recombination between genes, treating distinct genes as independent. If a significant HKA value is found, the gene-specific X_i values (Equation 10.2b) indicate which contributed the most to the lack of fit.

Example 10.3. Hudson et al. (1987) partitioned the *Adh* gene into two regions, silent sites and 4-kb of the 5′ flanking region, corresponding to a test using $L = 2$ loci. (The careful reader might be concerned that these loci are linked, while the HKA tests assumes independence across loci. For the high recombination rates in *Drosophila*, this independence assumption is not unreasonable.) A sample of 81 *Drosophila melanogaster* alleles were sequenced, along with a single allele from its sibling species *D. sechellia*. Based on sequencing data, the divergence was 210 differences in the 4052 bp flanking region and 18 differences in the 324 silent sites, for roughly equal levels of divergence per base pair between the two loci. Based on restriction enzyme data, within *melanogaster*, 9 of the 414 5′ flanking sites were variable, while 8 of 79 *Adh* silent sites were variable. Thus, while the divergence was roughly equal, there was a four-fold difference in polymorphism. Hudson et al. modified their test to accommodate having polymorphism data from only a single population. In this setting δ cannot be estimated, so the authors assume $\delta = 1$ (both species have the same effective population size, an alternative approach would be to use the value of δ giving the smallest X^2 value). Given that there is a difference in the number of sites between the polymorphism and divergence data, let θ_i be the per-nucleotide mutation rate (for locus i), so that we have to weight the θ_i value for each

term by the number of sites compared, giving Equation 10.4 as

$$S_1^A + S_2^A = 9 + 8 = a_{81}(414 \cdot \widehat{\theta_1} + 79 \cdot \widehat{\theta_2})$$
$$D_1 + D_2 = 210 + 18 = (\widehat{\tau} + 1)\left(4052 \cdot \widehat{\theta_1} + 324 \cdot \widehat{\theta_2}\right)$$
$$D_1 + S_1^A = 210 + 9 = 4052 \cdot \widehat{\theta_1}\,(\widehat{\tau} + 1 + a_{81})$$

where $a_{81} = \sum_{i=1}^{80} 1/i = 4.965$. The solutions to this system were found to be

$$\widehat{\tau} = 6.73, \quad \widehat{\theta_1} = 6.6 \times 10^{-3}, \quad \text{and} \quad \widehat{\theta_2} = 9.0 \times 10^{-3}$$

giving the resulting X^2 statistic as 6.09. There are four observations (S_1^A, S_2^A, D_1, D_2) and three parameters to fit, giving a test with one degree of freedom. Since $\Pr(\chi_1^2 > 6.09) = 0.014$, the test indicates a significant departure from the equilibrium neutral model.

Equations 10.3 and 10.4 assume that all L loci are autosomal. If all loci are X-linked, they still apply. However, if loci are a mixture of autosomal and sex-linked, the θ_i terms for sex-linked loci are multiplied by (3/4), as their expected levels of neutral polymorphism are $3N_e\mu_i$ (Begun and Aquadro 1991). If organelle sequences are included, these are completely linked and treated as a single locus. Further, this locus has a different effective population size from autosomal genes, also requiring a scaling of its θ value (typically by 1/2, but other values may be justified). Modifications of HKA were proposed by Wright and Charlesworth (2004), who present a maximum-likelihood version, and Innan (2006) who framed the test in terms of the polymorphism-divergence ratio r. This formulation allowed Innan to consider a joint test involving r and a site-frequency measure (such as Tajima's D) to provide more support for selection at a site (Innan's **Two-Dimensional test**).

As with site-frequency tests, the HKA test is *not* robust to demography. Further, since HKA comparisons are made *across* different regions in the genome, demographic effects can be exaggerated relative to the site-frequency tests examined in Chapter 9. An example of this are attempts to use the HKA test to detect selection on organelle genomes. Since all loci within an organelle are completely linked, one or more nuclear loci must also be to obtain an HKA statistic. Even after correcting for differences in effective population size, the test may still be biased if there is population structure. For most species, organelle genomes are only transmitted through females. In plants, pollen and seed have very different dispersal patterns. In many animals, there can be strong sex-specific differences in migration, often with the male traveling long distances relative to females. In such cases, the pattern of population structure on nuclear genes (an average of the two parents) can be significantly different from that on organelle genes (female migration only).

Example 10.4. Ingvarsson (2004) examined chloroplast (cpDNA) diversity in two plants in the genus *Silene* (family Caryophyllaceae). A standard HKA test contrasting four noncoding regions of the chloroplast (treated as a single locus) and two unlinked autosomal genes between *S. vulgaris* and *S. latifolira* gave a highly significant value, with most of the signal coming from the cpDNA region. However, the estimated F_{ST} value (Chapter 2) for cpDNA was 0.546 versus 0.056 for nuclear genes, showing strong population structure on the organelle genes but only modest structure on nuclear genes. Assuming an approximate island model of migration (Chapter 2), Ingvarsson attempted to correct for these differences in the amount of structure. To a first approximation, the author found that population structure increases

the amount of segregating sites and decreases the divergence, both by a factor of $1 - F_{ST}$. Hence, Ingvarsson corrected the number of segregating sites by using $S_c = (1 - F_{ST})S$ and the divergence by $D_c = D/(1 - F_{ST})$. Applying these corrections to both the cpDNA and nuclear genes and using the corrected S_c and D_c values in the HKA test gave a nonsignificant result. The apparent strong signal of selection appears to be an artifact generated by nuclear and organelle genes having different population structures.

The McDonald-Kreitman (MK) Test: Basics

One of the most straightforward, and widely used, tests of selection was proposed by McDonald and Kreitman (1991a), which contrasts the amounts of polymorphism and divergence between two categories of sites within a single gene (Example 10.1). Typically, these are the synonymous versus replacement (nonsynonymous) sites, but the basic logic can be extended to other comparisons as well. Under the neutral theory, deleterious mutations are assumed to occur, but be quickly removed by selection, not contributing to either polymorphism or divergence. In the standard neutral-theory expressions for the amount of polymorphism $4N_e\mu$ and divergence $2t\mu$, μ is the *effectively neutral* mutation rate, the rate at which mutations arise that are effectively neutral ($4N_e|s| \ll 1$). While most mutations at synonymous sites are likely effectively neutral, a much smaller fraction f of new mutations at replacement sites are neutral, resulting in a lower effectively neutral mutation rate, $f\mu$. Given that f is the reduction in successful mutations in replacement sites, $1 - f$ is a measure of *functional constraints*, with f values near one implying that most new mutations are not effectively neutral (i.e., they are deleterious). One minor bookkeeping detail is that the silent and replacement mutations rates in the MK test refer to the sum total over all sites, so that $\mu_s = \mu n_s$ and $\mu_a = \mu f n_a$ are the total mutation rates over the collection of n_s silent and n_a replacement sites in the gene of interest.

Under the equilibrium neutral model, the expected number of substitutions D_i in site class i is $2t\mu_i$, while the expected number S_i of segregrating sites in a sample of n sequences is $a_n\theta_i$ (Equation 9.21a), where θ_i is the product of $4N_e$ and the total mutation rate μ_i for site class i. Thus,

$$\frac{D_a}{D_s} = \frac{2t\mu_a}{2t\mu_s} = \frac{2t\mu f n_a}{2t\mu n_s} = f\,\frac{n_a}{n_s}, \qquad \frac{S_a}{S_s} = \frac{a_n\theta_a}{a_n\theta_s} = \frac{4N_e\mu f n_a}{4N_e\mu n_s} = f\,\frac{n_a}{n_s} \tag{10.5}$$

where the subscript a denotes replacement (amino-acid changing) sites, while s denotes silent sites. Since S_i is a measure of the amount of polymorphism, we also denote it by P_i to conform to the standard notation for MK tests. Equation 10.5 is the foundation of the MK test, as the ratio of the number of replacement to silent polymorphic sites should equal the ratio of the number of replacement to silent substitutions. If some replacement sites are under positive selection, these contribute very little to within-species polymorphism (Kimura 1969; Smith and Eyre-Walker 2002), but result in an excess of replacement substitutions, so that $D_a/D_s > P_a/P_s$. It is worth noting that a very similar approach proposed by Templeton (1987, 1996), based on contrasting patterns in the tips verses interiors of estimated gene tree topologies, predates the MK test.

McDonald and Kreitman provided a more general derivation of the polymorphism ratio in Equation 10.5, replacing $4N_e$ (the equilibrium value) by T_{tot}, the total time on all of the within-species coalescent branches (Chapter 2). By considering the ratio of the number of polymorphic sites in the two categories, the common term T_{tot} cancels, so that any effects of demography also cancel. Hence, *provided mutation rates remain unchanged*, the MK test is *not affected by population demography* (Hudson 1993; Nielsen 2001). Because the coalescent structure that determines the amount of polymorphism is explicitly removed by taking the

P_a/P_s ratio, there is no assumption that the allele frequencies are in mutation-drift equilibrium nor any assumption about constant population size. This is a very robust feature not shared by most other tests of selection. However, as we will see shortly, this test is by no means fool-proof, as changes in the effective population size can influence the *effectively neutral* mutation rates, and hence the test. Likewise, mildly deleterious alleles can contribute to within-species polymorphisms, but not between-species divergence. Their presence inflates the polymorphism ratio over the divergence ratio, reducing the power to detect positive selection.

Given the expected equality of these two ratios under neutrality, the MK test is performed by contrasting polymorphism versus divergence data at synonymous and replacement sites for the gene in question through a simple 2×2 contingency table (Example 10.1). The presentation of the data required for the MK test is often referred to as either a **MK table** or a **DPRS table**, the later based on the clockwise order of categories: **D**ivergence (number of substitutions), **P**olymorphism (number of segregating sites), **R**eplacement, and **S**ynonymous,

	Divergence	Polymorphism
Silent	D_s	P_s
Replacement	D_a	P_a

Example 10.1 presented the original data used by McDonald and Kreitman, while Example 10.5 shows how to modify this basic idea to examine different regions within the same gene.

Example 10.5. LeCore et al. (2002) examined the *FRIGIDA* (*FRI*) gene in *Arabidopsis thaliana*. This gene is a key regulator of flowering time, and was the focus of study as European populations show significant variation in flowering time, with potentially strong selection for earlier flowering following the end of the ice age. For the data below, fixed differences (divergence) were examined by comparing *thaliana* with *A. lyrata*, while data on number of segregating sites uses populations of *thaliana*.

Entire coding region	Fixed	Polymorphic	
Synonymous	59	7	
Replacement	68	21	Fisher test $p = 0.056$
Exon 1	Fixed	Polymorphic	
Synonymous	30	2	
Replacement	38	16	Fisher test $p = 0.013$
Exons 2 and 3	Fixed	Polymorphic	
Synonymous	29	5	
Replacement	30	5	Fisher test $p = 1.000$

Note the excess of replacement polymorphisms in exon one relative to exons two and three. These data could be interpreted simply as a reduction on functional constraints in Exon 1 (for example, by a recent reduction in the effective population size, increasing the effectively neutral mutation rate). However, there is a nice internal control in that exons 2 and 3 don't show this pattern, which seems to rule out a reduction in effective population size in *thaliana* accounting for the reduction in constraints. The authors note that roughly half of the replacement polymorphisms in Exon 1 are lost-of-function mutations, which result in early flowering. Hence, it appears that the excessive number of replacement polymorphisms likely result from selection for early flowering in some populations. Further, since a non-functional copy of *FRI*

results in early flowering, there are a large number of mutational targets to achieve this phenotype (and hence a high mutation rate), which likely explains the large number of replacement polymorphisms.

The *FRI* clearly shows a fairly sharp heterogeneity in patterns of selection when contrasting exon 1 with the remaining exons, and detecting such within-gene heterogeneity may provide important functional clues for a putative region under selection. The more general issue of how to detect any such heterogeneity based on a scan of a region has been examined by McDonald (1996, 1998) and Goss and Lewontin (1996).

Example 10.5 raises two important statistical issues. First, one should always use Fisher's exact test for the goodness-of-fit (which can be found in standard statistical packages, such as R). The χ^2 and G tests for contingency tables are large-sample approximations, and tend to perform poorly when any table entry has an expected value less than five.

Second, often multiple tests are performed, and the thorny issue of multiple comparisons arises (Appendix 6). If one wishes a false positive rate of q over the collection of *all* independent tests, then the Bonferroni correction requires a critical value of $p = q/n$ for each of the n tests (Equation A4.4). For Example 10.5 (with $n = 3$ tests), $p = 0.0033$ for each test for a collective false positive rate over all tests of $q = 0.01$, and $p = 0.017$ for an collective probability of $q = 0.05$. By this criteria, the experiment-wide significance is closer to 5 percent than the 1.3 percent reported for Exon 1. As detailed in Appendix 6, Bonferroni corrections are rather strict, and can be improved upon by sequential Bonferroni methods, or (where appropriate) using control of the false discovery rate.

While initially presented as a contrast between silent versus replacement sites within a single gene, the basic logic of the MK test is not limited to either two categories nor to the specific comparison of silent versus replacement sites (e.g., Hudson 1993; Templeton 1996; Podlaha et al. 2005; Chen et al. 2009).

Example 10.6. Andolfatto (2005) examined 35 coding and 153 non-coding fragments from a Zimbabwe sample of 12 *D. melanogaster* X chromosomes, with a single *D. simulans* X as an outgroup. The number of observed polymorphic and divergent sites were the lumped into various subcategories as follows:

		Polymorphisms		Fisher Test p value	
Mutational Class	Fixed	All sites	Minus singletons	All Poly	Poly(-S_1)
Synonymous	604	502	323		
Replacement	260	115	52	$4.7 \cdot 10^{-7}$	$4.3 \cdot 10^{-10}$
Non-coding	3168	2386	1295	0.0144	0.00052
5′ UTRs	328	160	71	$2.7 \cdot 10^{-6}$	$1.7 \cdot 10^{-10}$
3′ UTRs	143	86	36	0.033	$8.2 \cdot 10^{-5}$

Given the small sample size ($n = 12$ chromosomes), polymorphism data is reported both as the total number of segregating sites (all sites) and the total number of segregating sites minus the singletons. The logic for removing singletons is the concern that slightly deleterious alleles can contribute to segregating sites (although they will be rare) but will not become fixed, and thus the polymorphism ratio overpredicts the number of fixed sites. Using the synonymous class as a reference, McDonald-Krietman tests were performed on each of the four different categories (replacement, non-coding, 5′ UTR, 3′ UTR), and computed separately using either all polymorphisms or only polymorphisms that were not singletons. Exclusion of singletons decreased the p values (increased significance) in all cases. Even after correcting for multiple tests, all of the comparisons based on polymorphisms minus singletons were highly significant.

Andolfatto observed that the average nucleotide diversity π was higher for synonymous sites than for any of the other categories displayed above. This suggests stronger constraints on non-coding regions relative to synonymous sites, and hence stronger purifying selection on these sites. Conversely, the above tests values all show excessive substitutions relative to the amount of within-population variation, suggesting that many of the differences were likely fixed by positive selection. Both of these results (stronger purifying selection on polymorphisms and stronger positive selection for substitutions) for non-coding DNA over synonymous sites were very surprising, and suggested that much of what is called non-coding DNA may have some functional role (the same appears to be at least partly true for humans, ENCODE Project Consortium 2012). A similar study using polymorphism data from *D. simulans*, which has a larger effective population size than *D. melanogaster*, found an even stronger signature of purifying selection on non-coding DNA (Haddrill et al. 2008).

One issue of concern when dealing with non-coding DNA is obtaining the correct alignment to ensure that homologous sites are being compared. This is problematic for even moderately-divergent species, as insertions and deletions run rampant, making correct alignment nearly impossible. Conversely, with coding regions, strong selection to keep the sequence in frame usually allows for a very easily alignment. Care must then be taken, as with even moderately-diverged DNA, one may throw out much of the noncoding sequence because of alignment issues, which could enrich the remaining sequences used in the analysis for those sites under stronger functional constraints (which are more conversed and thus more easily aligned).

A significant McDonald-Kreitman test occurs when P_a/D_a is significantly different from P_s/D_s. Since the assumption is that the silent site ratio is unchanged by selection, a significant MK test can occur either through an excess of replacement polymorphisms (P_a too large relative to D_a and P_s/D_s) or though an excess of replacement substitutions (D_a too large relative to P_a and P_s/D_s). The **neutrality index** of Rand and Kann (1996),

$$\mathrm{NI} = \frac{P_a/D_a}{P_s/D_s} = \frac{P_a\, D_s}{P_s\, D_a} \tag{10.6a}$$

indicates which of these two scenarios has happened. Note that NI is just the odds ratio for the MK contingency table (Jewell 1986). A value greater than one indicates more polymorphic nonsynonymous (replacement) sites than expected, while a value less than one indicates an excess of replacement substitutions. Hence, values less than one suggest some of the substitutions are adaptive, while values greater than one are suggestive of weakly deleterious alleles over-inflating the effectively neutral mutation rate estimate at replacement sites (as P_s now includes both effectively neutral and slightly deleterious alleles).

Example 10.7. Consider LeCore et al.'s data on the *FRI* gene (Example 10.5). For exon one, the neutrality index is

$$\mathrm{NI} = \frac{16/38}{2/30} = 6.42,$$

showing that the significant result is due to an excess of segregating replacement sites. Conversely, for exons two and three

$$\mathrm{NI} = \frac{5/30}{5/29} = 0.97,$$

suggesting a good fit to the neutral model, with neither an excess of polymorphic nor fixed replacement sites.

Note that NI is not defined if either P_s or D_a are zero and is biased if either are small (Stoletizki and Eyre-Walker 2011). Hence, its use is problematic when the gene being considered slows little divergence. When the observed cell numbers in any MK table are small (less than 5), a number of corrections have been suggested, which basically start by adding an extra count to D_a and P_s (Haldane 1956; Jewel 1986). Stoletizki and Eyre-Walker (2011) note that these corrections are still biased, and proposed a **direction of selection** (**DoS**) statistic,

$$DoS = \frac{D_a}{D_a + D_s} - \frac{P_a}{P_a + P_s} \tag{10.6b}$$

Positive values indicate an excess of nonsynonouous substitutions (suggesting adaptive evolution), while negative values imply an excess of replacement polymorphisms (suggesting slightly deleterious alleles are segregating).

While the DoS statistic is appropriate when comparing divergence/polymorphism features of genes as a function of some other variable (such as recombination rate or GC content), other approaches have been used when the aim is to return a single summary statistic for the entire genome. A simple average of the NI values over all sampled genes is biased, as genes for which NI is not defined (P_s or D_a are zero) are excluded, and those genes with small values for either of these return biased estimates. Example 10.8 illustrates one commonly used approach to avoid these issues, namely summing over all sites to create a grand MK table for the entire collection of sampled genes.

Example 10.8. Bustamante et al. (2005) sequenced 39 humans for roughly 11,600 genes, contrasting the results with human-chimp divergence at these same genes. Summing over all sites, the resulting DPRS table (where SNPs denotes polymorphic sites) is

	Divergence	SNPs
Silent	34,099	15,750
Replacement	20,467	14,311

This is a *different analysis* from a standard MK test, as the values for a large number of loci are aggregated into a single table. The resulting p value is highly significant ($p < 10^{-16}$), so that the neutral model is rejected. What is the source of the discrepancy? The neutrality index is

$$\text{NI} = \frac{14,311/20,467}{15,750/34,099} = 1.514,$$

showing that the lack-of-fit to the neutral model is driven by an excess of replacement polymorphisms (SNPs). The authors suggest that these polymorphisms are mainly deleterious, a view echoed by Hughes et al. (2003). Consistent with this, an analysis of 47,576 replacement SNPs in a sample of 35 humans by Boyko et al. (2008) estimated that 27-29% of these SNPs were effectively neutral, 30-42% moderately deleterious, and nearly all of the rest highly deleterious (we will discuss how such values are obtained shortly). This large fraction of segregating deleterious alleles significantly lowers the power of the MK tests and related approaches. Indeed, Charlesworth and Eyre-Walker (2008) note that because of excessive replacement polymorphisms, MK tests in humans are very underpowered.

While commonly used, the above approach of summing the MK tables for single genes to create a single grand MK table for the entire genome is potentially problematic because of the **Yule-Simpson effect** (Yule 1903; Simpson 1951), also know as **Simpson's paradox** (Blyth 1972). This is a well known phenomena wherein the results of two individual 2×2 contingency tables suggest a trend in one direction, whereas their amalgamated table suggests a trend in the opposite direct (reviewed by Good and Mittal 1987). More generally, it implies that the (unweighted) average of odds ratio over individual tables is different from the odds ratio in the amalgamated table. This commonly arises when there are large disparities in the sample sizes over tables (i.e., large differences in D_a values, as would be expected). To avoid this issue, Stoletzki and Eyer-Walker (2011) suggest using a weighted approach proposed by Tarone (1981) and Greenland (1982) for combining the odds ratio over general 2×2 contingency tables,

$$NI_{TG} = \frac{\sum D_{si} P_{ai} / (P_{si} + D_{si})}{\sum P_{si} D_{ai} / (P_{si} + D_{si})} \tag{10.6c}$$

where i denotes the ith gene. This is defined for all genes that show any silent site variation (either P_s or D_s is nonzero), and also weights each gene by its total silent site sample variation ($P_s + D_s$).

The McDonald-Kreitman Test: Caveats

One of initial criticisms against the McDonald-Kreitman test was that estimates of the number of segregating sites were rather sensitive to sampling, especially when the number of samples is small (Graur and Li 1991; Whittam and Nei 1991). McDonald and Kreitman (1991b) countered that this was not serious, as these effects would influence estimates of number of polymorphic sites in both silent and replacement sites equally. While largely correct, this is not strictly true, as there are generally two-fold more potential replacement than silent sites, giving them a slightly smaller sampling error. Still, this issue has more to do with power, and is unlikely to give false positives. The serious concerns with this test are more subtle, and as such, took longer to be fully appreciated.

As mentioned, the MK test does not require constant population size nor that mutation-drift equilibrium has been reached, and hence is rather robust to many demographic concerns that plague other tests. Balancing this strength are two subtle caveats, both relating to the distribution of fitness effects in observed variants (polymorphisms and substitutions). First, the MK framework assumes that deleterious mutations are strongly deleterious and make essentially no contribution to either the number of segregating or fixed sites. However, if present, weakly deleterious mutations (i.e., $-10 < 4N_e s < -1$) can contribution to segregating polymorphisms (especially since MK uses number of polymorphic sites, not their frequencies) but are highly unlikely to become fixed. Such mutations are over-represented in polymorphic sites relative to fixed sites, reducing the power of the MK test to detect an excess of replacement substitutions (and hence a signature of positive selection). Since we assume most mutations at silent sites (our neutral proxy) are either neutral or under very weak deleterious selection, this (generally) has little impact on silent sites, but a significant impact on replacement sites. One proposed correction for this is to drop "rare" polymorphisms, but this is rather subjective. Dropping singletons (Templeton 1996) as we did in Example 10.5 provides one simple correction, while other authors (e.g., Fay et al. 2002, Smith and Eyre-Walker 2002, Gojobori et al. 2007) have suggested only including "common" polymorphisms in the analysis, such as those with minor allele frequencies above ten percent.

The other concern is much more problematic. At the heart of the test is Equation 10.5 — the ratio of polymorphic sites and the ratio of substitutions both estimate the same quantity, the ratio of neutral mutation rates for the two categories. The caveat is that the *effectively*

neutral mutation rate changes with N_e. Recall that any mutation for which $4N_e|s| \ll 1$ behaves as if it is effectively neutral (Chapter 7). Under the equilibrium neutral model, the ratio of D_a/D_s has expected value f (assuming $n_a = n_s$), the reduction in the neutral mutation rate at replacement sites. Figure 10.1 shows that f decreases as effective population size increases, so that the amount of constraint $1-f$ increases with N_e. For the same distribution of selection coefficients, one can raise, or lower, the effectively neutral mutation rate by decreasing, or increasing, the effective population size. If the effective population size was significantly different during the divergence phase (where substitutions where fixed) than it is at the current phase (which generated the observed number of polymorphisms), then these two phases can have different effectively neutral mutation rates.

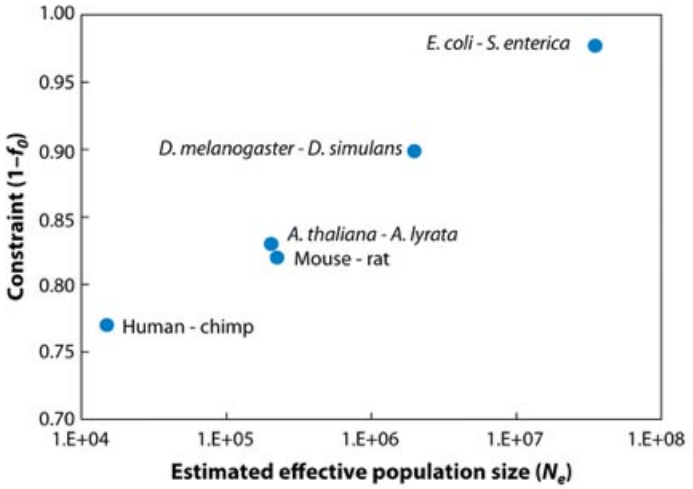

Figure 10.1 The estimated constraint $1-f$ on replacement sites as a function of effective population size, where f is ratio of effectively neutral mutation rates at replacement versus silent sites. As N_e increases, more deleterious mutations move from the effectively neutral class into the strongly deleterious class (f decreases), reducing the effectively neutral mutation rate and increasing the amount of constraint on a gene. After Wright and Andolfatto (2008).

McDonald and Kreitman (1991a) were aware than an increase in the effective population size when slightly deleterious alleles are present could create a situation where these were fixed during divergence under the smaller population size, but do not even contribute to within-species polymorphisms if N_e significantly increases. This would given an inflated D_a/P_a ratio, and hence a false signal of positive selection. Eyre-Walker (2002) shows that even a modest increase in N_e can generate such false signals, and that the problem is exacerbated by culling rare polymorphisms, which (as discussed above) is common practice. In the words of Hughes (2007), this feature implies that the MK test "cannot distinguish between positive Darwinian selection and any factor that causes purifying selection to become relaxed or to become less efficient". Phrased in terms of the neutrality index (Equation 10.6), a value greater than one can be generated by either segregating deleterious alleles or a relaxation in the functional constraints during the polymorphism phase. The later could occur by a change in the environment (Example 10.9) or by a change in N_e. Conversely, an index value less than one (which is normally taken as support for adaptive evolution) could similarly be generated by relaxation of functional constraint during the divergence phase, so that more mutations (relative to those currently segregating in the population) were effectively

neutral, and hence fixed. This effect can also occur between populations of the same species. For example, Lohmueller et al. (2008) observed a higher fraction of segregating deleterious mutations in human populations from European than from Africa, which they attribute to the bottleneck in the founding European population (and hence a reduction in N_e) during their migration of out Africa.

Example 10.9. An example of some of the potential difficulties in interpreting the results of a McDonald-Kreitman test is seen in Harding et al. (2000), who examined the human Melanocortin 1 receptor (*MC1R*), a key regulatory gene in pigmentation. Comparing the canonical *MC1R* haplotype in humans with a sequence from Chimp found 10 nonsynonymous (replacement) and 6 synonymous (silent) substitutions. An African population sample found zero nonsynonymous and 4 synonymous polymorphisms, giving

	Fixed (Human-Chimp)	Polymorphic (African)
Silent	6	4
Replacement	10	0

Fisher's exact test gives a p value of 0.087, close to significance. Taken on face value, one might assume that this data implies that the majority of the replacement substitutions between human and chimp were selectively-driven. However, the authors also had data from populations in Europe and East Asia, which showed ten nonsynonymous and three synonymous polymorphisms, giving the DPRS table as

	Fixed (Human-Chimp)	Polymorphic (Europe/East Asia)
Silent	6	3
Replacement	10	10

with a corresponding p value of 0.453. The authors suggest that the correct interpretation of these data is very stringent purifying selection due to increased functional constraints in African populations, with a release of constraints in Europe and East Asian. Asians in Papua New Guinea and India also showed very strong functional constraints, again consistent with a model of selection for protection against high levels of UV.

The key point is that the population chosen as the reference standard for the polymorphism ratio is critical. The two tests above both used the same divergence data, but the significance (or lack thereof) of the MK test critically depended on whether the population sample was African or European/East Asia.

Example 10.10. The effect of slightly deleterious alleles on the expected value of the neutrality index was examined by Welch et al. (2008), who assumed that new mutations have their fitness values drawn from some probability distribution. Assuming that s values are drawn from a gamma distribution (Appendix 2) over $-\infty < s < 0$ with shape parameter $\beta > 0$ (the coefficient of variation for s is given by $1/\sqrt{\beta}$, with $\beta = 1$ corresponding to the exponential distribution), Welch et al. showed that the expected value of the neutrality index is approximately

$$NI \approx 1 + \beta K$$

where $K > 0$ is a function of the sample sizes. Thus, the presence of deleterious mutations inflates the neutrality index above one. Conversely, a neutrality index less than one implies an excess of replacement substitutions, which is usually taken as support for positive selection. Welch et al. caution that this need not be the case. Assume the same model as above, with

new mutations only being deleterious, but now suppose that the population size has changed over time. In particular, suppose that the population had a constant size N_e for some fraction q of the total divergence time, after which is increases by a factor $\delta > 1$ to $\delta N_e > N_e$. In this case, the expected value of the neutrality index becomes

$$NI \approx \frac{1 + \beta K}{1 + q(\delta^{\beta} - 1)}$$

Welch et al. note that if the population expansion is recent and/or substantial (q near one and/or δ large), NI can easily be less than one, giving a false signature of selection. This expression quantifies our concern that a smaller population size during divergence results in a higher effectively neutral mutation rate, and hence more substitutions, than expected given the number of segregating replacement sites in the current population with a much larger size and hence a smaller effectively neutral mutation rate.

Finally, silent sites may be a rather poor proxy for neutral sites, especially in species with large effective population sizes. Chapter 8 reviewed codon usage bias, wherein some synonymous codons are preferentially used over others, with the discrepancy simply not being a function of nucleotide frequencies. Selection is thought to the weak on such sites, but can still have an impact (Hartl et al. 1994; Akashi 1995). For example, DuMont et al. (2004) found that preferred synonymous codons are substituted significantly faster than unpreferred synonymous changes at the *Notch* locus in *D. simulans*, while *melanogaster* (with a smaller N_e) has a significantly higher substitution rate for unpreferred changes. The consensus on codon bias is that the strength of selection is weak ($s \ll 0.001$), making synonymous changes are effectively neutral in small populations, but subjected to both purifying and positive selection in larger populations where $4N_e|s|$ is sufficiently large that these alleles are no longer effectively neutral. However, there may be a bit of silver lining to selection on the neutral proxy sites. As mentioned, one concern for false positives under an MK test is an increase in the effective population size. Eyre-Walker (2002) shows that selection on the neutral proxy sites (synonymous codons), restricts the conditions under which a false positive signal can arise via a change in N_e. Presumably this occurs because changes in N_e now influences both the test and neutral proxy sites, creating a bit of an internal control.

Dominance in Fitness and the MK Test

One might be concerned that dominance alters the polymorphism to divergence ratio at replacement sites, as both the frequency spectrum and the probability of fixation for a selected site is influenced by dominance. While Weinreich and Rand (2000) and Williamson et al. (2004) show that most types of dominance have little impact on this ratio, an important exception is weak to moderate overdominance (i.e., weak to moderate balancing selection). Williamson et al. showed that overdominance can increase the substitution rate relative to that predicted from the amount of polymorphism, giving a signal of positive direction selection in an MK test (a neutrality index less than one). The reason for this behavior follows from Robertson's (1962) classic result examined in Chapter 7, wherein overdominance can *increase*, rather than retard, the rate of fixation when the equilibrium values are extreme (minor allele equilibrium frequency 0.2 or less). The idea is that selection rapidly moves allele frequencies to these equilibrium values, at which point drift can cause alleles to become substituted if selection is relatively weak.

Fluctuating Selection Coefficients and MK Tests

While we have been assuming that the selection coefficient s on an allele remains constant, this is likely not the case. The impact of such fluctuating selection on MK tests has been

examined by Huerta-Sanchez et al. (2008) and Gossmann et al. (2014). Both these papers assumed selection coefficients were randomly sampled over time from a distribution with mean value zero. Huerta-Sanchez et al. found that fluctuating values of s results in an increase in the probability of fixation (relative to a neutral allele) and a decrease in the amount of polymorphism. This can generate false positives for selection in an MK test. Gossmann et al., however, noted that the results are more subtle. Since selection coefficients are randomly sampled over time, some alleles will, by chance, end up with a net positive value of s over their sojourn, and such mutations contribute disproportionately to levels of polymorphism and divergence. They conclude that MK signals under fluctuating selection are therefore genuine as fixed mutations are those that, by chance, end up with a net positive s value over their sojourn. Further, they find that the real impact of fluctuating selection is that MK methods tend to *underestimate* the fraction of adaptive sites.

Recombinational Bias in Extended MK Tests

The standard MK test, contrasting silent and replacement sites within a single gene, is very robust to recombination. As succinctly stated by Andolfatto (2008), this occurs because the comparison sites are fully interdigitated. Denoting the classes by a and b, within-gene comparisons are of the form $abababab$. Adjacent sites thus share the same coalescent structure and recombination (or lack thereof) has little effect. Conversely, extensions of the MK test may compare sites that are *not* interdigitated (but still closely linked), such as contrasting the silent sites in a gene with 3′ or 5′ UTR sites adjacent to that gene. This comparison has the form $aaaabbbb$. Even more extreme, one may compare silent sites in one region with other sites in different regions. In both these settings, Andolfatto (2008) found that recombination can indeed bias the MK test, generating an increased number of false-positives. The bias is most severe when the ratio of recombination to mutation rates is around one. For very small values (no recombination), there is little bias, and likewise for very large values (unlinked sites), there is little bias as well. Heterogeneity in recombination and/or mutation rates among less that fully interdigitated comparisons can also generate false-positives.

ESTIMATING PARAMETERS OF ADAPTIVE EVOLUTION

As shown in Example 10.1, DPRS tables lead to a simple prediction about the expected number of replacement substitutions given the ratio of silent to replacement polymorphisms. This allows us to directly ask how many (if any) excess substitutions at replacement sites have occurred within our target gene. While straightforward, one issue is power: at any particular gene the true excess has to be fairly substantial in order for the MK test to be significant. However, when we sum up such excesses over a large number of genes, we have the power to detect even a small average increase. This ability to look at the cumulative evidence over a large number of genes to detect a small individual effect is one of the advantages of genome-wide studies. A second approach to estimating the number of adaptive substitutions places this idea into a more formal statistical framework, called the Poisson random field model, which allows us to estimate the average selection coefficients of sites under positive selection. We consider these two approaches in turn.

Estimating the Fraction α of Substitutions that are Adaptive

It was realized fairly quickly that DPRS tables offer much more than simply an opportunity to test selection (Sawyer and Hartl 1992; Charlesworth 1994; Fay et al 2001, 2002; Smith and Eyre-Walker 2002). A neutrality index less than one indicates that the observed number of replacement substitutions is greater than expected from the ratio of the number of silent to

replacement polymorphic sites. Assuming that the P_a/P_s ratio does indeed reflect the ratio of effectively neutral mutation rates at these two classes of sites, then (when coupled with the observed number of silent substitutions) it predicts the expected number of effectively neutral replacement substitutions. Any statistically-significant excess over this predicted value are either sites fixed under positive selection or the result of changes in the effectively neutral mutation rate between the populations that generated the observed polymorphism and divergence data. As mentioned, the later can happen if the effective population size was much smaller during the divergence phase, allowing more slightly deleterious mutations to escape selection and become fixed.

As above, let μ and $f\mu$ denote the per-site mutation rate for silent and replacement sites, so that $\mu_a = f\mu n_a$ and $\mu_s = \mu n_s$ are the total mutation rates for the replacement and silent sites in our sample. Under neutrality, the expected number of substitutions for each class is $D_s = 2\mu_s t$ and $D_{a,n} = 2\mu_a t$. Now suppose there are a additional replacement substitution that were fixed by positive selection, giving the total number of replacement substitutions as $D_a = D_{a,n} + a = 2\mu_a t + a$. Ideally, we would like to estimate both the number a and the fraction $\alpha = a/D_a$ of replacement substitutions that are adaptive. To estimate a, note that the expected number of segregating sites are given by $\theta_i a_n$, or $P_s = 4\mu_s N_e a_n$ and $P_a = 4\mu_a N_e a_n$, where the later assumes that the vast bulk of segregating sites are neutral (adaptive mutations are assumed to be both rare and also fixed quickly, and hence make little contribution to P_a). First note that

$$D_s \frac{P_a}{P_s} = 2\mu_s t \, \frac{\mu_a}{\mu_s} = 2\mu_a t \tag{10.7a}$$

From above, this last expression is just the expected number of neutral replacement substitutons, so that $D_{a,n} = D_s(P_a/P_s)$. Since $a = D_a - D_{a,n}$, our estimate becomes

$$\widehat{a} = D_a - D_s \frac{P_a}{P_s} \tag{10.7b}$$

as obtained by Charlesworth (1994), Fay et al. (2001, 2002), and Smith and Eyer-Walker (2002). This directly suggests an estimator for the fraction α of replacement substitutions that are adaptive,

$$\widehat{\alpha} = \frac{\widehat{a}}{D_a} = 1 - \frac{D_s P_a}{D_a P_s} = 1 - \text{NI} \tag{10.7c}$$

Note that a positive estimate of α requires a neutrality index of less than one. Using the data from Example 10.6 on for non-coding regions on the X in *D. melanogaster*, $\widehat{\alpha} = 1 - 0.906 = 0.094$ (using all polymorphic sites) and $\widehat{\alpha} = 1 - 0.764 = 0.236$ (singletons ignored). Hence, between roughly 10 and 25 percent of all substitution in these non-coding regions might be adaptive. Kousathanas et al. (2010) also obtained estimates of around 10% adaptive substitutions in the immediate up- and down-stream regions around protein-coding genes in the house mouse (*Mus musculus castaneus*).

Finally, note that we can estimate the fraction f of replacement mutations that are effectively neutral by noting that

$$\frac{P_a}{P_s} = \frac{4\mu_a N_e a_n}{4\mu_s N_e a_n} = \frac{\mu_a}{\mu_s} = \frac{f\mu\, n_a}{\mu\, n_s} = f\,\frac{n_a}{n_s}, \tag{10.7d}$$

giving an estimate of

$$\widehat{f} = \frac{P_a}{P_s}\,\frac{n_s}{n_a} = \frac{P_a/n_a}{P_s/n_s} \tag{10.7e}$$

This is just the ratio of the fraction of replacement sites that are polymorphic divided by the fraction of silent sites that are polymorphic. Recall that $1 - f$ is a measure of the amount of constraint relative to a silent site, as f is the relative fraction of replacement site mutations that (relative to silent sites) are effectively neutral. For *Drosophila*, estimated $1 - f$ values are 0.94 for replacement sites, 0.81 for UTRs, 0.61 for intergenic regions, and 0.56 for intron sequences (summarized by Sella et al. 2009). Further, Halligan and Keightley (2006) showed that synonymous sites are not the fastest evolving sequences in *Drosophila*. In comparison with these **FEI sites** (**fastest evolving intronic sites**), the constraint in synonymous sites is 0.09, suggesting that nine percent of new silent mutations are deleterious.

While Equations 10.7b/c can be applied to single genes, individual-gene estimates of α are expected to have large sampling variance and low power. If the actual number a of adaptive substitutions is modest to small, this may not sufficiently inflate D_a for estimates of α to be significantly different from zero for most single genes. For example, if five substitutions are expected given the silent/replacement polymorphism ratio, an observed value of eight is unlikely to be significantly different. However, if $3/8$ were indeed driven to fixation by positive selection then $\alpha = 0.375$, which is quite substantial. Despite lower power for any *single* gene, considerable power can be obtained by estimating α over a *number of genes*, using the accumulation of any small deviations to detect even small values of α. The question is how best to do so. Fay et al. (2001, 2002) suggested the estimator

$$\widehat{\overline{\alpha}}_{Fay} = 1 - \frac{\overline{D}_s}{\overline{D}_a}\left(\frac{\overline{P_a}}{\overline{P_s}}\right) \tag{10.8a}$$

In here (and what follows), a bar over a quantity denotes its average for observed values or its expected value for parameters over the sample of genes considered. In particular, we use α when referring to estimation issues on a single gene and $\overline{\alpha}$ for the average of the α values over a set of genes. The estimator given by Equation 10.8a has two potential sources of bias (Smith and Eyre-Walker 2002; Welch 2006), both of which lead to overestimation of $\overline{\alpha}$. Let μ and $f\mu$ denote the effectively neutral per-site mutation rates for silent and replacement sites within a gene, where f is allowed to vary over genes. Following Welch (2006), one can show that

$$E\left(\frac{\overline{D}_s}{\overline{D}_a}\right) = \frac{\overline{n}_s}{\overline{n}_a}\frac{1}{\overline{f}}\overline{\left(\frac{1}{1-\alpha}\right)}^{-1} \simeq \frac{\overline{n}_s}{\overline{n}_a}\frac{1}{\overline{f}}\left[1 - \overline{\alpha} - \sigma^2(\alpha)\right] \tag{10.8b}$$

When there is between-locus variation in α (so that $\sigma^2(\alpha) > 0$), $\overline{\alpha}$ is over-estimated by Equation 10.8a. A more subtle bias occurs if f and $4N_e\mu$ are *negatively-correlated* over genes (Smith and Eyre-Walker 2002; Welch 2006), as

$$E\left(\frac{\overline{P}_a}{\overline{P}_s}\right) = \frac{\overline{n}_a}{\overline{n}_s}\left(\overline{f} + \frac{\sigma(4N_e\mu, f)}{4N_e\overline{\mu}}\right) \tag{10.8c}$$

Equation 10.7e *underestimates* f and therefore results in an overestimation of $\overline{\alpha}$ if $4N_eu$ and f are negatively correlated over genes (and underestimates $\overline{\alpha}$ if they are positively correlated). Smith and Eyre-Walker (2002) note that a negative correlation is biologically reasonable, as the effective population size can vary over the genome (Chapters 3, 8), and regions with smaller N_e likely have higher f values, as more mutations become effectively neutral.

To reduce bias from correlations between f and N_e, Smith and Eyre-Walker (2002) suggest the estimator

$$\widehat{\overline{\alpha}}_{SEW} = 1 - \frac{\overline{D_s}}{\overline{D_a}}\overline{\left(\frac{P_a}{P_s + 1}\right)} \tag{10.9a}$$

Provided that the number of polymorphic silent sites in the sample is modest (five or greater), this adjusted polymorphism ratio is unbiased by correlations between f and N_e (Smith and Eyre-Walker 2002; Welch 2006), giving

$$E\left(\widehat{\overline{\alpha}}_{SEW}\right) \simeq \overline{\alpha} + \sigma^2(\alpha) \tag{10.9b}$$

While this correction removes concern over correlations between f and N_e, it still results in an overestimation of $\overline{\alpha}$ when between-locus variation in α is present.

Example 10.11. A simple model provides some insight into the amount of bias possible when using Equation 10.9b. Imagine there are just two types of genes: a fraction q have $\alpha_* > 0$, while the rest have only neutral substitutions. Under this model $\overline{\alpha} = q\alpha_*$, while

$$\sigma^2(\alpha) = E[\alpha^2] - \overline{\alpha}^2 = q\alpha_*^2 - q^2\alpha_*^2 = q\alpha_*^2(1-q)$$

Suppose that $\alpha_* = 0.2$ and $q = 0.5$, so that $\overline{\alpha} = 0.1$. The expected value from the Smith-Eyre-Walker estimate is

$$\overline{\alpha} + \sigma^2(\alpha) = 0.1 + 0.5 \cdot 0.2^2(1 - 0.5) = 0.11$$

or a ten percent overestimation. Conversely, consider the extreme case where at ten percent of the genes, all substitutions are adaptive, so that $\alpha_* = 1$ and $q = 0.1$. Again, $\overline{\alpha} = 0.1$ while the expected value from the Smith-Eyre-Walker estimate is $0.1 + 0.1 \cdot 1^2(1 - 0.1) = 0.19$, so that even in this extreme case $\overline{\alpha}$ is only a two-fold overestimate.

A potential concern with both Equation 10.8a and 10.9a is bias due to the Yule-Simpson effect. Recalling Equations 10.7c and 10.6c suggests the estimator

$$\widehat{\overline{\alpha}}_{TG} = 1 - NI_{TG} = 1 - \frac{\sum D_{si}P_{ai}/(P_{si} + D_{si})}{\sum P_{si}D_{ai}/(P_{si} + D_{si})} \tag{10.9c}$$

is perhaps the most robust. While Stoletzki and Eyre-Walker (2011) found very close argument between $\widehat{\overline{\alpha}}_{TG}$ and $\widehat{\overline{\alpha}}_{Fay}$ over the data sets they examined, all of the above considerations suggest that the most prudent estimator is given by Equation 10.9c.

Confidence intervals for $\overline{\alpha}$ using any of the above estimators can be obtained using bootstrap resampling. One generates a sample of genes by drawing *with replacement* from the original list of all genes and estimates $\overline{\alpha}$ for this sample. This process is repeated a large number of times to generate a distribution for the estimate under resampling. Taking the lower 2.5% and upper 97.5% in this distribution gives the 95% bootstrap confidence interval. Collectively, we will refer to estimators (such as Equations 10.7-10.9) that use departures from the expectation under neutrality in a DPRS table as **MK estimators**.

While the above sources of bias are generally modest and in a predictable direction (overestimation of $\overline{\alpha}$), the presence of mildly deleterious alleles provides a major bias, which can be either positive or negative (Eyre-Walker 2002; Bieren and Eyre-Walker 2004; Welch 2006; Charlesworth and Eyre-Waker 2008; Eyre-Walker and Keightley 2009; Halligan et al. 2010; Schneider et al. 2011; Keightley and Eyre-Waker 2012). Estimates are downwardly biased by the presence of low-frequency deleterious alleles that contribute to P_a but not D_a,

inflating the polymorphism ratio relative to the divergence ratio (Eyre-Walker 2006, Eyre-Walker and Keightley 2009). As with MK tests, one approach is to count only "common" polymorphisms for P_a and P_s. However, Charlesworth and Eyre-Walker (2008) note that while this approach is "better than doing nothing", estimates of α still tend to be downwardly biased even after this correction unless the true α is fairly substantial. Further, the bias is a function of the complex distribution of fitness effects (Charlesworth and Eyre-Walker 2008; Welch et al. 2008; Eyre-Walker and Keightley 2009; Schneider et al. 2011; Keightley and Eyre-Waker 2012).

The presence of mildly deleterious alleles also biases estimates of α if the population size differed during the divergence and polymorphism phases. If the population has recently undergone an expansion, this can upwardly bias estimates of α. In such cases, slightly deleterious alleles may become fixed contributing to divergence, but are quickly removed in the new larger population, not contributing to P_a. Conversely, if the population has recently undergone a contraction, this inflates P_a as more deleterious alleles are segregating, downwardly biasing estimates of α. Eyre-Walker and Keightley (2009) and Halligan et al. (2010) obtained a simple expression for the bias in α when the recent population size N_P generating the polymorphism data differs from the ancestral size N_D generating the divergence data. Assuming beneficial mutations are sufficiently strong that α is invariant under the two population sizes, while deleterious new mutations have their fitness effects drawn from a gamma distribution with shape parameter $\beta > 0$, then the connection between the expected value α_{est} of an estimated α and its true value is

$$\alpha_{true} = 1 + (\alpha_{est} - 1)\left(\frac{N_P}{N_D}\right)^{\beta} \tag{10.10}$$

A contraction in N_e ($N_P < N_D$) leads to an underestimation of α, while an increase in N_e results in an overestimation. The same approach leading to Equation 10.10 was used in Example 10.10 to examine the behavior of the neutrality index (which is closely related to α, see Equation 10.7c) under changes in N_e.

Maximum-likelihood (ML) estimators of α have been proposed in an attempt to account for segregating deleterious mutations (Bierne and Eyre-Walker 2004; Welch 2006; Boyko et al. 2008;, Eyre-Walker and Keightley 2009; Schneider, et al. 2011; Keightley and Eyre-Waker 2012). This is done by assuming a standard form (such as a gamma) for the distribution of deleterious fitness effects, and then using the site-frequency spectrum data to estimate the parameter(s) of this distribution. We sketch the basic outline of this approach in the next section (in the context of Poisson random field models). Corrections for the effects from changes in N_e have also been suggested (Eyre-Walker and Keightley 2009). While elegant and powerful when the model assumptions are correct, the concern is that all of these approaches are highly-dependent on the assumed functional form (e.g., gamma, normal, or other) of the unknown distribution of fitness effects for the slightly deleterious mutations. Indeed, Kousanthanas and Keightley (2013) found that these models perform poorly when the distribution of fitness effects is multimodal, suggesting some nonparametric approaches for such cases.

Another potential source of bias, first noted by Akashi (1995), is codon usage. Strong usage bias results in the synonymous substitution rate underestimating the neutral divergence rate, which in turn inflates estimates of α. If strong bias occurs on just a few genes, this is not likely a global effect, and will have minor impact on $\overline{\alpha}$. However, recall that the *D. melanogaster* X chromosome has a higher usage bias than its autosomes. Since this appears to be due to stronger selection on X-specific genes, it can result in different biases in α between X and autosomal loci (Campos et al. 2012).

Given these competing sources of bias (overestimation of $\overline{\alpha}$ when $\sigma^2(\alpha) > 0$ and underestimation of α when deleterious alleles are segregating), are MK estimators (e.g., Equations

10.8a, 10.9a, 10.9c) more likely to over- or under-estimate the true $\overline{\alpha}$? As Example 10.11 highlights, the overestimation of $\overline{\alpha}$ when α varies over genes, while not trivial, is often modest, especially if $\overline{\alpha}$ is modest to large. Conversely, segregating deleterious alleles inflate the polymorphism ratio P_a/P_s, underestimating the actual excess number of substitutions. This effect can be quite dramatic. In particular, if deleterious alleles are not uncommon, a neutrality index value greater than one can occur, which results in a negative estimate of α (Equation 10.7c). Putting these two sources of bias together, $\overline{\alpha}$ is generally likely to be underestimated unless the population has undergone a recent size expansion. A final complication, noted by Fay (2011), is the assumption that each site evolves independently (also see Messer and Petrov 2012). Two possible sources of overestimation of α are possible when this assumption fails. First, slightly deleterious mutations may be fixed by hitchhiking to a favorable substitution, potentially inflating D_a and hence estimates of α. Second, epistasis in fitness between sites may occur such that the fixation of one site changes the constraints on other sites within the gene, which again can potentially result in an inflation in D_a.

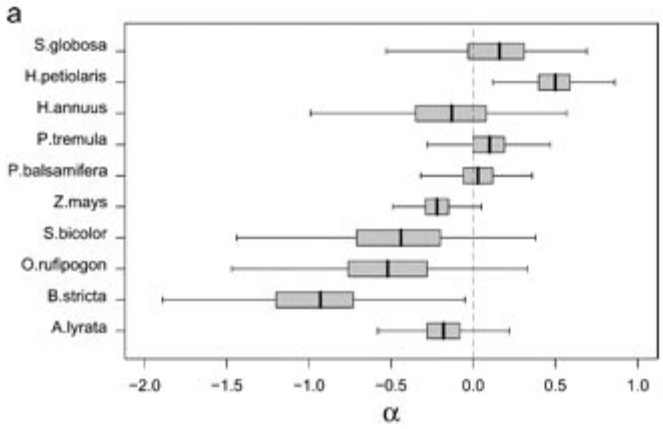

Figure 10.2. Estimated $\overline{\alpha}$ values for ten plant species. Boxes and whiskers indicate, respectively, the 50% and 95% confidence intervals for the estimates, which were obtained using Eyre-Walker and Keightley's (2009) ML method. This allows for a distribution of deleterious fitness effects and potentially different effective population sizes in the divergence and polymorphism phases. Only the comparison involving sunflowers (polymorphism data from *Helianthus petiolaris*, divergence between *petiolaris* and *annuus*) had an estimated average α that was significantly positive. Surprisingly, the comparison using polymorphism data from *H. annuus* and the same divergence (*petiolaris* versus *annuus*) gave a negative estimate of average α (but not significantly different from zero). After Gossmann et al. (2010).

How Common are Adaptive Substitutions?

There has been an explosion of genome-wide estimates of $\overline{\alpha}$ (Eyre-Walker 2006) that will likely continue, as the required data (divergence between a set of genes in two species, polymorphism data for the same genes from one, or both, species) is becoming increasingly easy to obtain. Table 10.1 summarizes some of these studies, and Figure 10.2 shows a recent analysis from ten species pairs in plants. The quest for $\overline{\alpha}$ values is very reminiscent of the mad "find them and grind them" dash in the 1970's to estimate levels of protein variation for just about any species one could get their hands on (e.g., Lewontin 1974).

The general observation for *Drosophila* is that estimates of $\overline{\alpha}$ for amino acid substitutions are high, close to 50%, with estimates of the fraction of adaptive changes in noncoding regions

approaching 30% in some cases. High $\overline{\alpha}$ values for replacement sites are also seem for the mouse, bacteria, and three plants (*Populus, Helianthus*, and *Capsella*), while very low levels are seen in other plants (Table 10.1 and Figure 10.2). Low levels on *Arabidopsis thaliana* were originally attributed to the high levels of selfing in this species (Bustamante et al. 2002), but an outcrossing close relative (*A. lyrata*) showed similar very low levels of $\overline{\alpha}$ (Foxe et al. 2008). The case receiving the most interest is humans, where an initially rather high estimate of 0.35 by Fay et al. (2001) for a small set of genes was followed by several studies showing much lower values. One trend that has been suggested is that $\overline{\alpha}$ increases with effective population size. While intriguing, there are also apparent counter-examples. For example, Bachtrog (2008) found that *D. miranda*, thought to have a low effective population size, has a similar value of $\overline{\alpha}$ to *Drosophila* species thought to have significantly larger sizes.

Table 10.1. Partial list of estimates of the fraction $\overline{\alpha}$ of replacement substitutions that are adaptive. The organism listed is the species that provided the polymorphism data. MK refers to a MacDonald-Kreitman estimator (Equations 10.8 or 10.9), ML to maximum-likelihood extensions of MK estimators (Bierne and Eyre-Walker 2004; Welch 2006; Eyre-Walker and Keightley 2009; Schneider et al. 2011), and PRF to Poisson Random Field estimators (examined in detail in the next section). Estimates of zero indicate a neutrality index score exceeding one (and hence a negative estimate of $\overline{\alpha}$).

Organism	$\overline{\alpha}$	Method	Reference
Mus musculus castaneus (Mouse)	0.57	ML	Halligan et al. 2010
Oryctolagus cuniculus (Rabbit)	0.6	MK, ML	Carneiro et al. 2012
Gallus gallus(Chicken)	0.20	MK	Axelsson and Ellegren 2009
Drosophila simulans	0.45	MK	Smith and Eyre-Walker 2002
	0.43	ML	Bierne and Eyre-Walker 2004
	0.41	ML	Welch 2006
D. melanogaster	0.44	ML	Bierne and Eyre-Walker 2004
	0.95	PRF	Sawyer et al. 2007
	0.85	ML	Schneider et al. 2011
D. miranda	0.48	ML	Bachtrog 2008
X chromosome	0.33	MK	Haddrill et. al. 2010
	0.14	ML	
autosomal	0	MK	
	0	ML	
D. pseudoobscura (X)	0.44	MK	Haddrill et. al. 2010
	0.70	ML	
autosomal	0.59	MK	
	0.87	ML	
Escherichia coli	0.56	MK	Charlesworth and Eyre-Walker 2006
Arabidopsis thaliana	0	PRF	Bustamante et al. 2002
A. lyrata	0	PRF	Foxe et al. 2008
Capsella grandiflora (crucifer)	0.40	ML	Slotte et al. 2010
Populus tremula (Aspen)	0.30	ML	Ingvarsson 2010
Helianthus annuus (sunflower)	0.75	MK	Strasburg et al. 2009
Humans	0.35	MK	Fay et al. 2001
	0	MK	Zhang and Li 2005
	0.06	PRF	Bustamante et al. 2005
	0.12	MK	Gojobori et al. 2007

Drawing a clear conclusion for these initial data is problematic for several reasons. First, even in the same species, different genes may be used and/or different populations chosen

as the polymorphism benchmark. The effect of the later is especially prominent in Figure 10.2, with the same divergence data between sunflower species (*Helianthus annuus* versus *H. petiolaris*) showing either a significantly positive estimate of mean α when using *Helianthus petiolaris* as the polymorphism reference population but a negative (but not significant) estimate when using *H. annuus* as the reference population. Clearly, differences in the current N_e values between the two species being considered can inflate, or deflate, estimates of α (Equation 10.10). Second, different studies used different methods, ranging from simple MK-type estimators (Equations 10.8, 10.9) to much more sophisticated ML-based estimators that attempt to account for both changes in N_e and the presence of segregating deleterious alleles (Bierne and Eyre-Walker 2004; Welch 2006; Eyre-Walker and Keightley 2009). While certainly powerful when the modeling assumptions are correct, the robustness of these ML approaches to model misspecification is unclear.

Despite these potential misgivings, the pattern of estimates of $\overline{\alpha}$ over species, and even within genomes, is of fundamental importance to evolutionary biologists. If indeed some species have very low $\overline{\alpha}$ values, does this automatically imply that they have lower rates of adaptation? One surprisingly group that showed a very low estimated $\overline{\alpha}$ was the Hawaiian silverswords *Schiedea* (family Caryophyllaceae), a plant group with rapid (and dramatic) morphological evolution over a very recent time window (Gossmann et al. 2010). One possible resolution is that most current studies have focused on estimation of α in coding sequences, whereas perhaps most adaptation (especially over short time scales) occurs at the level of gene regulation. Based upon the estimated α values in noncoding regions, Andolfatto (2005, Wright and Andolfatto 2008) suggests that the number of adaptive substitutions in noncoding regions in *Drosophila* could easily be far greater than the number of adaptive replacement substitutions. Given that *Drosphila* has a compact genome relative to humans and many other species, the bulk of adaptive variation may not be where current scans of α have looked.

One observation consistent with this was the work of Torgerson et al. (2009), who compared polymorphism and divergence levels in roughly 15,000 conserved non-coding (CNCs) regions flanking human genes. CNCs are operationally defined as noncoding sequences at least 100 nucleotide in length that are at least 70% conserved between mouse and humans. The idea is that these are putative regulatory control regions, and hence under purifying selection. Comparing human-chimp divergence, the authors estimated an overall $\overline{\alpha} \sim 0.05$ for all CNC, ~ 0.15 and 0.23 for 5′ and 3′ UTRs, and ~ 0.12 for upstream and downstream regions. Their most interesting finding was a disconnect between the estimated α values for the CNC regions flanking a gene versus that for nonsynonymous substitutions within the gene. Namely, an apparent disconnect between regulatory (CNC) and structure (amino acid substitution) substitutions. In particular, some of the strongest signals came from genes expressed in the fetal brain, but the protein sequences for these same genes showed no such signatures.

Estimating the Rate λ of Adaptive Substitutions

A quantity that prominently appeared in expressions in Chapter 8 on the effects of periodic sweeps was λ, the per generation rate at which adaptive substitutions are fixed. While it might seem that estimates of λ would be very difficult to obtain, fortunately this is not the case, and they follow almost directly from estimates of α (Smith and Eyre-Walker 2002; Andolfatto 2007), being simply the number of adaptive substitutions divided by the total time of divergence $2t$ (as each of the two branches of length t can fix adaptive mutations). If $d_a = D_a/n_a$ denotes the per-site number of replacement substitutions between two species, then an upper bound is simply $\lambda \leq d_a/(2t)$. (The use of D_a to compute d_a makes the assumption that all substitutions have been observed, so that no corrections for multiple substitutions at the same site are needed. This is not unreasonable when comparing two closely-related

species.) With an estimate of α, the number of adaptive replacement substitutions is just αD_a, giving Andolfatto's (2007) estimator,

$$\widehat{\lambda} = \frac{\alpha\, d_a}{2t} \tag{10.11a}$$

for the per-site, per-generation rate of adaptive substitutions. Noting that K_a, the per-site *rate* of replacement substitutions, is just $K_a = d_a/(2t)$, we can also write Equation 10.11a as $\lambda = \alpha K_a$. If an estimate of t is not available, an estimator (scaled as $\tau = t/[2N_e]$) can be obtained from the ratio of D_s/P_s. From Equations 10.12a/b,

$$\frac{E[D_s]}{E[P_s]} = \frac{1}{a_m + a_n}\left(\tau + \frac{1}{m} + \frac{1}{n}\right) \tag{10.11b}$$

where m and n are the sample sizes for the two populations and a_x given by Equation 4.3b. Substituting the observed values of D_s and P_s for their expected values and rearranging provides a simple method-of-moments estimator for τ,

$$\widehat{\tau} = (a_m + a_n)\ \frac{D_s}{P_s} - \left(\frac{1}{m} + \frac{1}{n}\right) \tag{10.11c}$$

Example 10.12. The preliminary estimate of the percent amino acid divergence between human and chimp proteins is 0.8, giving $d_a = 0.008$ (Chimpanzee Sequencing and Analysis Consortium 2005), with a divergence time of roughly 7 million years. If we take $\alpha = 0.10$ (ten percent of replacement substitutions are adaptive), then our estimate of the rate of adaptive replacement substitutions per-site per-generation is

$$\lambda = \frac{0.10 \cdot 0.008}{14 \times 10^6} = 5.7 \times 10^{-11} \text{ per-site per-year}$$

Assuming a generation time of 25 years, this corresponds to a rate of 2.3 $\times 10^{-12}$ per-site per-generation. As a point of comparison, Andolfatto (2007) contrasted X chromosome genes in *Drosophila melanogaster* (for polymorphism data) and *D. simulans* (as the outgroup for divergence). The estimated α was 0.5, while $d_a = 0.028$ (roughly three percent amino acid divergence), and $t = 10^7$ generations, giving

$$\lambda = \frac{0.50 \cdot 0.028}{2 \times 10^7} = 7.0 \times 10^{-10} \text{ per-site per-generation}$$

The Sawyer-Hartl Poisson Random Field Model: Basics

A second approach for extracting information from DPRS tables on the nature and amount of selection is the **Poisson random field model** (**PRF**) of Sawyer and Hartl (1992). Their initial version assumed that all sites within a region evolve independently and that the strength of selection on all replacement sites is the same. Strongly deleterious mutations are allowed to occur, but the assumption is that these do not contribute to either polymorphism (observed segregating sites) or divergence, and are accounted for by simply reducing the mutation rate to exclude such mutations. Under this model, the observed counts (P_s, D_s, P_a, and

D_a) in a DPRS table follow independent Poisson distributions, whose expected values are functions of four parameters $(\theta_a, \theta_s, \tau, \gamma)$. With four observations (the DPRS entries) and four unknowns, we can estimate the parameters, but cannot assess how well the model fits the data. Two of the parameters are the scaled total mutation rates $\theta_a = 4N_e\mu_a$ and $\theta_s = 4N_e\mu_s$, while the third parameter is the scaled divergence time $\tau = t/(2N_e)$. Of most interest is the fourth parameter, the scaled strength of selection $\gamma = 2N_e s$. Sawyer and Hartl assumed additive fitness, so that a new mutation has fitness $1 + s$ as a heterozygote and $1 + 2s$ as a homozygote. In contrast to MK approaches, the PRF model does not estimate the fraction α of adaptive substitutions directly, but knowledge of γ can allow one to do so indirectly (Example 10.13).

The PRF model assumes each site evolves independently and hence there are no effects from selection at linked sites — the assumption is that selection can only influence a site by directly acting on it. To obtain the expected values for entries in a DPRS table, Sawyer and Hartl used diffusion theory (Appendix 1) to obtained the expected equilibrium distributions (under mutation-selection-drift balance) for polymorphisms at neutral and selected sites as well as the expected divergence between sites. The PRF model is an infinite sites model (Chapter 2), with each new mutation being unique and at a different site than previous ones. For a sample of m and n sequences from the two species, the expected values for the DPRS entries are

$$E[D_s] = \theta_s \left(\tau + \frac{1}{m} + \frac{1}{n} \right) \tag{10.12a}$$

$$E[P_s] = \theta_s \left(\sum_{j=1}^{m-1} \frac{1}{j} + \sum_{j=1}^{n-1} \frac{1}{j} \right) = \theta_s \left(a_m + a_n \right) \tag{10.12b}$$

$$E[D_a] = \theta_a \left(\frac{2\gamma}{1 - \exp(-2\gamma)} \right) \Big(\tau + G(m, \gamma) + G(n, \gamma) \Big) \tag{10.12c}$$

$$E[P_a] = \theta_a \left(\frac{2\gamma}{1 - \exp(-2\gamma)} \right) \Big(F(m, \gamma) + F(n, \gamma) \Big) \tag{10.12b}$$

where

$$F(n, \gamma) = \int_0^1 \left(\frac{1 - x^n - (1 - x)^n}{1 - x} \right) \left(\frac{1 - \exp^{-2\gamma x}}{2\gamma x} \right) dx \tag{10.13a}$$

$$G(n, \gamma) = \int_0^1 (1 - x)^{n-1} \left(\frac{1 - \exp^{-2\gamma x}}{2\gamma x} \right) dx \tag{10.13b}$$

The full derivation is given by Sawyer and Hartl, but a brief sketch of the underlying ideas is as follows. First, a classic result (Wright 1938) is that the amount of time a new mutation (with selection coefficient s) spends in the interval $(x, x + dx)$ is

$$\phi(x \mid N_e, s) = \frac{1 - \exp^{-2\gamma(1-x)}}{1 - \exp^{-2\gamma}} \, \frac{1}{x(1 - x)} \, dx \tag{10.14a}$$

In the limit as $\gamma \to 0$, this reduces to dx/x, recovering Watterson's expression for the site frequency spectrum for neutral alleles (Equation 2.34a). Equation 10.14a is the expected equilibrium frequency spectrum for sites under selection, and is valid for both positive and negative values of s.

As a brief aside, we mentioned above that certain maximum likelihood versions of the basic MK test use a distribution of fitness effects (often denoted by **DFE** in the literature),

$\varphi(s \mid \boldsymbol{\Delta})$, where $\boldsymbol{\Delta}$ are the distribution parameters (Bierne and Eyre-Walker 2004; Welch 2006; Eyre-Walker et al. 2006; Boyko et al. 2008; Eyre-Walker and Keightley 2009; Keightley and Eyre-Walker 2012). The expected site frequency spectrum becomes

$$\phi(x \mid N_e, \boldsymbol{\Delta}) = \int \phi(x \mid N_e, s)\, \varphi(s \mid \boldsymbol{\Delta})\, ds \tag{10.14b}$$

which is then used to obtain a maximum likelihood estimate of the distribution parameters $\boldsymbol{\Delta}$, with the resulting DFE used to adjust for the effects of segregating deleterious alleles.

Returning to the PRF model, we do not use the site-frequency spectrum, but rather just the four cell counts in the DPRS table. If x is the frequency of a segregating allele, the probability we score it as a polymorphic site in a sample of size n is $1 - x^n - (1-x)^n$, where the last two terms account for either all n draws being the derived allele or all n draws being the ancestral allele (Eqution 2.36b). Hence, the probability we score a truly segregating site as polymorphic becomes

$$\int_0^1 (1 - x^n - (1-x)^n)\ \phi(x \mid N_e, s) \tag{10.14c}$$

Using this expression, the function $F(n, \gamma)$ given by Equation 10.13a follows upon substitution into Equation 10.14a and some simplification (a similar approach was used in Chapter 9 for ML-based detection of hard sweeps, see Equation 9.16a). The Sawyer-Hartl model also correctly accounts for the possibility that segregating mutations are scored as substitutions because the sample size was insufficient to contain both alleles. If the derived allele frequency is x, the probability that we score a polymorphic site as a substitution event (for the derived allele) is x^n, giving the additional increment to the probability of an observed substitution as

$$\int_0^1 x^n\, \phi(x \mid N_e, s) \tag{10.14d}$$

This term is added to the probability of a true substitution to give a full accounting of the number of sites in the sample scored as substitutions, and $G(n, \gamma)$ follows from Equation 10.14d by a change of variables.

The basic similarities, and fundamental differences, between MK estimators (e.g., Equations 10.7-10.9) and the PRF approach can be easily obscured by the impressive nature of the PRF equations. The similarity is that both approaches use the same the data, the four values in a DPRS table. However, the two approaches estimate different quantities and have different underlying model assumptions. MK estimators make no assumption about the nature or strength of selection on replacement sites, but instead estimate f, the reduction in the effectively neutral mutation rate at replacement sites, and α, the fraction of replacement substitutions at a gene that are adaptive. The effects of purifying selection enters only through f, while the effects of positive selection only thorough α. In contrast, the PRF equations estimate θ_a and θ_s, the scaled total mutation rates over all sites of that category within the gene. The ratio of θ_a/θ_s (suitably corrected for number of sites within each category, see Equation 10.7e) is *not* an estimate of f, as the PRF model *does* allow for slightly deleterious alleles to be segregating. It also allows for advantageous alleles to be present, where γ is (very roughly) the scaled average selection coefficient over all replacement mutations. Thus, it does *not* estimate α directly, but given estimates of γ we can compute the expected fraction of substitutions fixed by positive selection (Example 10.13, Equation 10.16). The original Sawyer-Hartl model was very restrictive, with only a single fitness class for replacement sites (which is approximately treated as an average selection coefficients over mutations). Extensions discussed shortly remove this restriction, allowing for neutral, deleterious, and advantageous classes, with separate estimates of γ for each class.

The original Sawyer-Hartl analysis equated the observed entries in a DPRS table with their corresponding expected values (Equation 10.12) and then solved for the unknowns of interest (the ratio θ_a/θ_s, the scaled average strength of selection γ, and the scaled time of divergence τ). A value of γ significantly different from zero implies selection on replacement sites, with $\gamma > 0$ implying positive selection and $\gamma < 0$ negative selection. This original model only assumed a single selective class, with silent sites being neutral. This base model can be placed in a likelihood framework by recalling that each observed entry is an independent Poisson random variable. The resulting probability that the count in a specific category is k given its expected value κ is

$$\text{Prob}(X = k \,|\, \kappa) = \kappa^k \exp(-\kappa)/k!, \quad \text{where} \quad \kappa = E(X)$$

The likelihood of the data in the DPRS table for gene i is thus given by

$$L_i = \prod_{j=1}^{4} \left(\frac{\kappa_{i,j}^{x_{i,j}} \exp(-\kappa_{i,j})}{(x_{i,j})!} \right) \tag{10.15}$$

where $x_{i,j}$ denotes the observed values for category j in gene i, with

$$x_{i,1} = P_{s,i}, \quad x_{i,2} = P_{a,i}, \quad x_{i,3} = D_{s,i}, \quad x_{i,4} = D_{s,i}$$

and $\kappa_{i,j}$ are the corresponding gene-specific expected values,

$$\kappa_{i,1} = E[P_{s,i}], \quad \kappa_{i,2} = E[P_{a,i}], \quad \kappa_{i,3} = E[D_{s,i}], \quad \kappa_{i,4} = E[D_{a,i}]$$

Note from Equation 10.12 that these are functions of the unknown parameters $(\theta_{a,i}, \theta_{S,i}, \gamma_i, \tau)$, so a numerical search over these to maximize Equation 10.15 given the data obtains the likelihood solutions. Under the assumption of independence across genes, the combined likelihood over k genes becomes

$$L = \prod_{i=1}^{k} L_i$$

where $\theta_a, \theta_S, \gamma$ can potentially vary over the genes, while the divergence time τ is shared by all.

As might be expected, this basic model has been expanded by considering more realistic fitness models. Nielsen et al. (2005) allowed three fitness classes for replacement sites: neutral, deleterious, and advantageous. While fitness is assumed to be the same within each class, this is a significant improvement over the basic Sawyer-Hartl model. The resulting likelihood now has four parameters for selection (as opposed to one, γ). These are $p_a,\ p_0,\ p_d$, the frequencies of advantageous, neutral, and deleterious mutations (where $p_a = 1 - p_0 - p_d$), and γ_a and γ_d, the scaled selection coefficients for the favored and deleterious alleles. These values were assumed to be the same over all genes. Nielsen et al. applied their method to a set of 50 human genes that presented other evidence for possible positive selection. The resulting ML estimates were $p_d = 0.748, p_0 = 0.172$, and $p_a = 0.080$ as the fraction of deleterious, neutral, and advantageous mutations, and $\gamma_d = -34.96$ and $\gamma_a = 267.11$ as the scaled strength of selection of deleterious and advantageous mutations. Note that even in this case where genes were ascertained as likely to be under positive selection, most mutations are still deleterious. A similar analysis of two *Drosophila melanogaster* data sets by Schneider et al. (2011) found that about 1.5% of all nonsynonymous mutations were adaptive, but with a much smaller scaled strength of selection, $\gamma_a \sim 10$.

While the PRF model does not directly estimate α, it can be obtained from the estimates of γ and the fraction p_a of advantageous mutations. The expected rate of effectively neutral

substitutions is μp_0, the neutral mutation rate. The expected rate of adaptive substitutions, λ is obtained as follows. The expected number of favorable mutations arising each generation is $2N\mu p_a$, where μp_a is the favorable mutation rate. For large γ, each of these have fixation probability $2sN_e/N$, for an expected per generation substitution rate of favorable alleles of

$$\lambda \simeq (2N\mu p_a)(2sN_e/N) = \mu p_a(2\gamma) \tag{10.16a}$$

The fraction of adaptive substitutions is the rate of adaptive substitutions divided by the total rate of substitutions (adaptive plus neutral),

$$\alpha = \frac{\lambda}{\lambda + \mu p_0} \tag{10.16b}$$

Substituting Equation 10.16a gives

$$\alpha = \frac{2\gamma\mu p_a}{2\gamma\mu p_a + \mu p_0} = \frac{2\gamma p_a}{2\gamma p_a + p_0} = \frac{2\gamma}{2\gamma + p_0/p_a} \tag{10.16c}$$

Equation 10.16c relates the selection estimates p_a and γ from a PRF model with the selection estimate α from an MK approach. Inspection shows that small p_a (or more precisely a small value of p_a/p_0) does not mean that α is small, as $\alpha > 0.5$ when $2\gamma > p_0/p_a$. One final result emerges from Equation 10.16a. Since μp_a is the rate of benefical mutations, which (in keeping with our notation from Chapter 8) we denote by μ_b, giving

$$\lambda = 2\gamma\mu_b \tag{10.16d}$$

which immediately suggests the estimator of Bachtrog (2008),

$$\mu_b = \frac{\lambda}{2\gamma} \tag{10.16e}$$

Example 10.13: What is the estimate of α for the subset of genes considered by Nielsen et al. (2005)? Here $p_a = 0.08, p_0 = 0.172$ and $\gamma_a = 267.11$. While only eight percent of all new replacement mutations were advantageous, α is considerably larger that 0.08, with Equation 10.16c giving

$$\alpha = \frac{2 \cdot 267.11 \cdot 0.08}{2 \cdot 267.11 \cdot 0.08 + 0.172} = 0.996$$

The reason for this high value is that the estimated advantageous mutation rate (0.08μ) is slightly below half of the estimated neutral rate (0.172μ), while the fixation probabilities for advantageous mutations are over five hundred times greater. If we lumped the neutral and deleterious mutations rates together and assumed these were all effectively neutral (i.e., replacing 0.172 by 0.920), our estimate of α is still very high, 0.980. However, it is also important to recall that this was a highly ascertained set of genes, chosen to be enriched for positive selection. Now consider the Schneider et al. (2011) values for their population of *Drosophila melanogaster* ($p_a \sim 0.015$, $\gamma_a \sim 10$). If we assume all of the remaining mutations are neutral,

$$\alpha = \frac{20}{20 + (1 - 0.015)/0.015} = 0.23$$

If we assume that 50% of all new mutation are deleterious ($p_0 = 0.485$), then $\alpha = 0.38$.

The robustness of estimates from the PRF when the model assumptions are relaxed has been examined by several authors. While the model assumes additive selection, estimates of γ are relatively insensitive to dominance (Williamson et al. 2004). Wakeley (2003) examined the effects of population structure (assuming an island model, Chapter 2). While estimates of the divergence time τ where significantly affected, estimates of γ were only weakly affected, and tend to be conservative (closer to neutrality). Desai and Plotkin (2008) note that the infinite-sites assumption (mutations never reoccur at the same site) breaks down under high mutation rates ($\theta > 0.05$), as might be found for viruses and microbes. In such cases, recurring mutations at the same site can result in genes under weak negative selection giving a signal of strong positive selection.

One critical difference between PFR and MK analyses is the contribution of information from silent sites (e.g., P_s, D_s), a point stressed by Li et al. (2008). Estimates of selection under an MK analysis are in the form of estimates of α, which are critically dependent upon P_s and D_s (e.g., Equations 10.8a, 10.9a), in addition to D_a and P_a. Conversely, under the PRF model, positive selection is estimated only through γ. Examination of Equations 10.12c and d shows that estimates of γ depend *only* on D_a and P_a, and that information from silent sites (P_s and D_s) does not enter. As a consequence, the control for demographic effects on P_a provided by P_s does not enter, and over- or under-inflated estimates of P_a from population structure can significantly bias estimates of γ. Further, Equation 10.14a (from which the PRF equations follow) is an *equilibrium* model, namely that the population size has been stable for sufficient time to reach the mutation/selection/drift equilibrium. Chapter 9 was littered with the bodies of tests that critically depend on this same assumption. In contrast, since MK estimates involve the ratio of P_a/P_s, recent demographic effects influencing polymorphism levels are accounted for, and there is no assumption about the population being at an equilibrium value for the current amount of genetic variation. Thus, while both MK and PRF approaches face bias from differences in population size between the divergence and polymorphism phases, PRF approaches have additional bias introduced by any nonequilibrium patterns in the polymorphism data. As noted by Li et al. (2008), tests of selection using PRF theory (γ significantly greater than zero) are closer to an HKA than a MK test, as the former compares the P/D ratio over different genes, lacking the internal control of comparing polymorphism levels from two different classes within the same gene.

Finally, while we have framed the PRF approach in terms of analysis of simple DPRS data, it can also be modified to directly estimate γ from the site-frequency spectrum from a single population (Hartl et al. 1994; Bustamante et al. 2001; Williamson et al. 2004; Huerta-Sanchez et al. 2008). DPRS data is very granular, collapsing all of the polymorphism and divergence information into just four data points. In contrast, the site-frequency spectrum is a very rich source of additional information on the structure of the polymorphism data (Chapter 9). Using the PRF model to estimate γ directly from the frequency spectrum is done in analogous fashion to estimating sweep parameters using the frequency spectrum discussed in Chapter 9. In particular, Equation 10.14a is substituted into Equation 9.16a to form the likelihood, from which a MLE for γ can be obtained by standard approaches (LW Appendix 4). While very elegant, this approach is not generally recommended due to the very delicate dependence of the frequency spectrum on demographic structure, which is not accounted for by the current models. Likewise, Equation 10.14a assumes additive fitnesses, whereas even small amounts of dominance can alter the the site-frequency spectrum (Williamson et al. 2004).

The Sawyer-Hartl Poisson Random Field Model: Bayesian Extensions

More fined-grain variation in fitness was allowed by Bustamante et al. (2002) and Sawyer et al. (2003) in the form of Bayesian models (an approach discussed more fully in Chapter 19 and in great detail Appendices 2 and 3). Instead of returning a point estimate $\widehat{\theta}$ for an unknown

parameter θ (or vector of parameters $\boldsymbol{\Theta}$), a Bayesian analysis return the full distribution (the **posterior**) $\varphi(\theta \mid \mathbf{x})$ for that parameter given any previous information (a **prior**) and the likelihood given the data $\mathbf{x}$. An especially powerful feature of a Bayesian analysis is the notion of a **marginal posterior**. Partition the vector of parameters as $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_i, \boldsymbol{\Theta}_n)$, where the vector $\boldsymbol{\Theta}_i$ contains parameters that are of interest to us (for example γ), while $\boldsymbol{\Theta}_n$ is a vector of **nuisance parameters**, quantities needed to specify the model, but are often of no interest (for example, θ_a, θ_s, τ). The marginal posterior for $\boldsymbol{\Theta}_i$ is given by integrating the full posterior over the nuisance parameters,

$$\varphi(\boldsymbol{\Theta}_i \mid \mathbf{x}) = \int_{\boldsymbol{\Theta}_n} \varphi(\boldsymbol{\Theta}_i, \boldsymbol{\Theta}_n \mid \mathbf{x})\, d\boldsymbol{\Theta}_n$$

This can be done directly through the use of MCMC sampling (Appendix 3). The importance of marginal posteriors is that they capture how the uncertainty in estimating *all* of the parameters in a model influences the uncertainty in those particular parameter(s) of interest (such as γ).

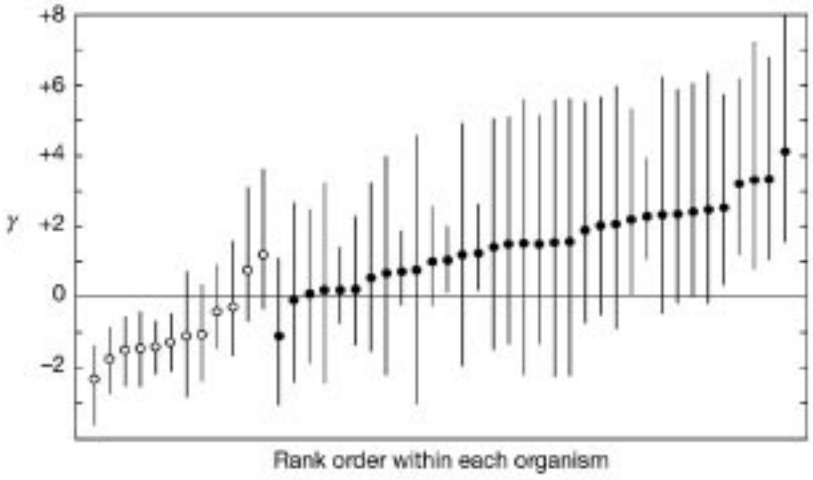

Figure 10.3 Bustamante et al. (2002) examined 12 genes from *Arabidopsis thaliana* (using a single allele from *A. lyrta* to compute divergence) and 34 genes from *D. melanogaster* (with a single allele for *D. simulans*). This figure plots the resulting posterior distribution for γ for each gene. The circle represents the mean, and the vertical lines denote the 95% creditable intervals (the shortest span of the posterior containing 95% of the probability). These are plotted by rank order within the two species, with *Arabidopsis* plotted first as open circles and *melanogaster* second as filled circles. If the vertical line is entirely below zero, selection on this locus is significantly negative. For lines entirely above zero, selection on that gene is significantly positive. Half (6 of 12) of the *Arabidopsis* genes are significantly negative, while none are significantly positive. Conversely, no *melanogaster* genes are significantly negative, while 9/34 are significantly positive.

Bayesian analysis of PRF data typically uses a **hierarchical model**. The motivation for this approach comes from random-effects models (Chapter 19). Suppose we have p parameters of interest. Treating these as fixed effects requires p degrees of freedom, but often $p \gg n$, the number of observations. In some settings, we can treat these p quantities as random effects, draws from some unknown distribution, such as a normal, with unknown mean and variance. Since all draws (realizations) are assumed to come from this common distribution,

we can borrow information across observations to estimate the distribution parameters, using (for the case of a normal) only two degrees of freedom. Bayesian hierarchical models take this idea a step further. Consider data structured into a number of categories (say genes), with multiple observations (draws) from each category (say mutations at the gene). Assuming that the draws from a given category are all from the same distribution (say a normal with category-specific mean and variance), then when the number of categories is large, so to is the parameter set (all of the category-specific means and variances). A hierarchical model reduces the number of parameters to estimate by assuming that the mean (and/or variance) for each category-specific distribution was itself a draw from a second distribution. Once that draw is made, those parameter values are fixed for that category. This reduces the estimation problem to one of just estimating the hyperparameters in the second distribution.

An example of this approach is Bustamante et al. (2002), who assumed a constant value γ_i for gene i, but allowed these gene-specific values to vary. This was done by assuming each γ_i is a random variable drawn from a normal distribution with mean μ_γ and variance σ^2_γ, which are estimated from the data. This model allows selection to vary over loci, but as a function of just two parameters $(\mu_\gamma, \sigma^2_\gamma)$. Since τ is a common factor over all genes, this allows information to be borrowed across loci, improving power. Figure 10.3 shows an example of the output from such an analysis. Because the analysis in done over each set of genes, only loci which sufficient information, namely $P_a + D_a \geq 4$, are likely to be informative.

Sawyer et al. (2003) extended the Bustamante et al. approach by allowing each new mutation at gene i to have different fitness, which are drawn from a normal distribution,

$$\gamma \sim N(\mu_{\gamma,i}, \sigma^2_w), \quad \text{where} \quad \mu_{\gamma,i} \sim N(\mu_\gamma, \sigma^2_\gamma) \tag{10.17}$$

The mean (scaled) selection coefficient $\mu_{\gamma,i}$ was itself allowed to vary over loci, but the variance about this mean σ^2_w was assumed common over all loci (allowing us to again share information over genes). As in the Bustamante et al. model, the gene-specific mean $\mu_{\gamma,i}$ is itself drawn from a normal with mean μ_γ and variance σ^2_γ. Thus, there are only three basic fitness parameters in this model, $\mu_\gamma, \sigma^2_\gamma$, and σ^2_w.

Example 10.14: Sawyer et al. (2007) applied their 2003 model to a sample of 91 genes from an African population of *D. melanogaster*, using a *simulans* sequence to assess divergence. Ignoring very strong deleterious mutations that are unlikely to contribute to polymorphisms, they found that approximately 95% of all new replacement mutations are deleterious, with 70% of all replacement polymorphisms observed in a sample being deleterious. Conversely, they estimated that over 95% of the fixed differences at replacement sites are due to positive selection, albeit fairly weak. Approximately 46% of replacement substitutions are estimated to have $N_e s < 2$, 85% a value of less than four, and 99% a value less than seven.

While Bayesian models allowing fitness to vary over new mutations are powerful and potentially offer a solution to the problem of segregating deleterious mutations that plagued MK tests and estimates of α, just how robust they are remains unclear. Current versions all assume normal distributions of fitness effects, but this is clearly not a realistic model (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2007; Welch et al. 2008; Boyko et al. 2008). While it is conceptually straightforward (but computationally much more demanding) to replace a normal by other candidate distributions, a very reasonable concern is just how robust the results are to alternative distributions. The normal has symmetry about the mean, while asymmetric or more heavy-tailed distributions might be a better reflection of biology.

A second concern was noted by Li et al. (2008), who found a very strong effect of the prior. Specifically, the number of genes with γ values declared to be significantly different from zero increased with the assumed variance σ_γ^2 in the prior for fitness effects. This makes intuitive sense, in that restricting this variance to be small constrains most realized values of γ to be close to the mean value, while increasing it allows estimates to deviate substantially from the mean (and hence have their credible intervals avoid overlapping zero). Strong dependency of the posterior on the prior is always problematic in a Bayesian analysis, and good practice is to run the model over several rather different sets of prior hyperparameters (such as σ_γ^2) to assess the stability of the posterior under these different models. Li et al. (2008) noted that a plot of number of positively selected sites (genes with γ values whose credible intervals are all greater than zero) increases with the assumed variance in γ, but appeared to show signs of approaching an asymptote in humans and *Drosophila simulans* over the values for σ_γ^2 used in the analysis. However, the same curve for yeast (*Saccharomyces cerevisiae*) showed no signs of approaching an asymptote over this range.

PHYLOGENY-BASED DIVERGENCE TESTS

Finally, we briefly consider divergence tests that examine the pattern of substitutions over a known *phylogeny*. These tests are designed to detect a rather different pattern of selection than was assumed in Chapter 9 (single events) or earlier in this chapter (multiple substitutions *across* a gene between two populations/species). While multiple substitutions are also required for a signal in phylogeny-based divergence tests, these must at the same *site* (typically a codon) within a gene. Single substitutions at even a large number of different codons across a gene leaves very little signal for these tests. As such, phylogenic tests are biased towards detecting sites that undergo repeated evolution (Hughes 2007), and are likely to miss many, indeed perhaps *most*, adaptive substitutions.

The required data for phylogeny-based tests is a set of aligned sequences and a predetermined phylogenetic tree for the sampled species. The assumption is that all sequence differences are the result of fixation events. If a site is segregating in one (or more) of the taxa from which a single sequence is drawn, one may incorrectly infer it as a substitution event. The taxa must also have the right amount of divergence, as too little, or too much, results in very low power. With too little divergence, there are not many changes, and hence low power to detect small percentage differences in silent versus replacement changes at particular sites. Further, if little true divergence has occurred, even a few segregating sites incorrectly called as substitutions can significantly inflate the divergence. With too much divergence, multiple substitutions at a single site between two lineages may occur, and improper adjustments for such multiple hits can introduce substantial bias if the statistical model used to account for these is incorrect.

A few comments are in order on the phylogeny for the sampled taxa, as this determines the covariance structure of the data. We not only require the topology (the pattern of common ancestry), but also the *branch lengths*, the distances (time) between the taxa. Errors in either obviously compromise these tests. For example, one route for repeated selection is the independent evolution of the same key amino acid at a particular critical site (e.g., Example 10.15). The topology of a phylogeny can inform us as to whether a cluster of taxa sharing this key amino acid are all descendants from a single fixation event or comprise a collection of independent events. Likewise, if some of the branch lengths are taken to be too short (or long) relative to the rest of the phylogeny, this can also bias rate estimates.

There is a very rich literature on molecular evolution, and our purpose here is to only provide a brief overview of divergence-based tests at the phylogenetic level. Readers seeking a fuller treatment of many of the important side issues (such as tree construction) not

addressed here should consist any number of excellent texts on the subject (Kimura 1983; Page and Holmes 1998; Hughes 1999; Graur and Li 2000; Nei and Kumar 2000; Felsenstein 2004; Yang 2006; Li 2006).

The K_a to K_s ratio, ω

The basis for divergence-based tests is $\omega = K_a/K_s$, the (per-site) ratio of nonsynonymous (replacement)to synonymous (silent) substitution rates, which Miyata and Yasunaga (1980) refer to as the **acceptance rate**. For sites under the standard neutral model (deleterious mutations can arise, but are quickly removed), the expected value of ω at a site (or gene) is $\omega = \mu f/\mu = f < 1$, where f is the ratio of the effectively neutral mutation rates. Thus, in the absence of positive selection, we expect $\omega < 1$. Moreover, if adaptive mutations are absent (or very rare), then $1 - \omega$ is a direct measure of the amount of constraint $(1 - f)$ on a site. Conversely, $\omega > 1$ is usually taken as an unmistakable signature of selection (Kimura 1983). Even if a demographic change results in a lowering of the effective population size (increasing the effectively neutral mutation rate at replacement sites), such a change only brings K_a closer to, but still smaller than, K_s.

There are cases where $\omega > 1$ is *not* a signal for positive selection. Ratnakumar et al. (2010) note that resolution of heteroduplex DNA during gene conversion events often results in a bias towards G and C bases (also see Webster and Smith 2004). Given that nonsynonymous codon positions often have lower GC content than synonymous sites, biased GC-gene conversion can inflate the K_a/K_s ratio even in the absence of selection. Ratnakumar et al. analyzed a dataset of roughly 18,000 human genes compared against their orthologues in at least two other mammalian genomes. They found genes giving divergence-based signals of selection had a significant tendency to also display genomic signals of GC conversion bias. They estimated that over twenty percent of elevated ω values in this data set could be the result of biased gene conversion.

A second factor that can upwardly bias estimates of ω is the presence of strong selection constraints on *silent* sites. Chamary et al. (2006) review some of the evidence that silent sites may still be subjected to constraints (beyond any weak ones from codon usage bias, Chapter 8) because they affect mRNA stability, splicing, or microRNA binding. A cautionary tale is offered by Hurst and Pál (2001), who examined constraints on the breast cancer *BRCA1* gene. A sliding window of roughly 300 nucleotides scanned across this gene in two pairs of comparisons: human-dog and mouse-rat (the use of a window allowed for an average regional estimate of K_a and K_s based on comparing the two species). The window around position 200-300 showed a relatively normal level of K_a (relative to the rest of the gene), while K_s plummeted dramatically, especially in the human-dog comparison. The result was an ω value significantly greater than one, not due to an elevation in the replacement substitution rate, but rather a decrease in the silent substitution rate. Wolf et al. (2009) found that an upward bias in ω from genes from reduced K_s values can be especially problematic when using closely related taxa as D_s is small. Pond and Muse (2005) note that if variation in K_s occurs over the gene, failure to include this heterogeneity in the model can easily result in false positives (estimated $\omega > 1$ for particular codons). Thus, while $\omega > 1$ is usually taken as a gold standard for positive selection, a little more humility in its use may be in order.

While conceptually straightforward, the operational problem in using ω is that while one or a few sites may be under *repeated* strong directional selection ($\omega > 1$ at these residues), most sites in a protein are expected to be under some selective constraints ($\omega < 1$), so that the average over all sites gives $\omega < 1$. Indeed, a meta-analysis by Endo et al. (1996) found that only 17 out of 3595 proteins showed $\omega > 1$. There were, however, a few early success stories. Example 10.2 discussed Hughes and Nei (1988), who used the 3-D protein structure of the major histocompatibility complex to suggest potential sites to examine (those amino acids on the surface in critical positions). Within this set of residues, $\omega > 1$, while ω was

less than one when averaged over the entire gene. Unfortunately, most proteins lack this amount of biological detail to draw upon. Because amino acid residues in close proximity on the three-dimensional structure of a protein can be scattered all over the primary (i.e., linear) sequence, grouping sites for analysis by their position in the primary sequence is very ineffective. The key is to base tests of ω values on a *codon-by-codon* basis, so that codons, rather than genes, become the unit of analysis.

Two general approaches have been suggested to estimate ω. Both require a phylogeny, and issues such as the correct multiple sequence alignment as well as errors in the assumed tree potentially loom in the background. **Parsimony-based** approaches reconstruct the sequence at each node in the tree, and then use these to count up the number of synonymous and nonsynonymous substitutions for each codon. **Likelihood approaches** (LW Appendix 4) are on a more firm statistical footing, but are computationally intense and can be rather model-specific. Both approaches allow for tests of whether a protein is under selection and (more excitingly) tests for selection at specific *sites* in that protein. More recently, tests are being built around **Bayesian approaches** (Appendix 2), which allow for the management of uncertainly in very complex statistical models.

Parsimony-Based Ancestral Reconstruction Tests

Fitch et al. (1997) and Suzuki and Gojobori (1999) proposed similar parsimony-based approaches for detecting selection at single sites. Both start with a phylogeny and then use parsimony (choose the solution requiring the fewest number of changes) to reconstruct the ancestral sequences at all of the nodes in the tree. With these estimated sequences in hand, one can them simply count the number of synonymous and nonsynonymous substitutions on the tree. Fitch et al. compute an average ω rate for the entire gene and then look for excessive variations at particular codons, while Suzuki and Gojobori perform the analysis considering each codon separately. The false-positive rate of these methods is generally small (Suzuki and Gojobori 1999; Suzuki and Nei 2002), but they suffer from low power (Wong et al. 2004). Further, they have to address several rather delicate issues of sequence evolution that, if not correctly accounted for, can provide rather significant artifacts. First, it is well known that **transitions** (A $\leftrightarrow$ G, C $\leftrightarrow$ T) can occur at different rates than **transversions** (e.g., A $\leftrightarrow$ T, etc.), and (at third base positions) transitions are more likely to given synonymous changes. Failure to incorporate these rate differences can result in an overestimation of the number of nonsynoymous substitutions (Yang and Nielsen 2002). Second, any codon usage bias (Chapter 8) must be accommodated. Third, when divergence times are modest to large, one must correct for the possibility of multiple substitutions between two lineages at a site, otherwise one undercounts the number, and nature, of the actual substitution events. All of these issues can have a highly significant effect on estimates of ω (Yang and Bielawski 2000). Finally, given that the ancestral states are likely estimated with error, parsimony analysis has no formal procedure to take this uncertainty into account. Bayesian posterior distributions can account for these errors, but this requires moving from a parsimony to a likelihood framework. For all of these reasons, most analysis now turn to likelihood-based approaches (and their Bayesian counterparts), wherein one explicitly allows the model to account for transitions vs. transversions, codon usage bias, and multiple substitutions.

Maximum-Likelihood-Based Codon Tests

Maximum-likelihood (ML) methods following the evolution of a codon over a phylogenetic tree were introduced by Goldman and Yang (1994) and Muse and Gaut (1994). While conceptually straightforward, they involve a fair bit of bookkeeping. They assume that each site is evolving independently, which can be compromised by two rather different factors. First, a substitution at one site can change the nature of selection at other sites. Second, high levels

of recombination can lead to false signals of selection (Anisimova et al. 2003). This can be especially problematic with viral sequences, which rapidly evolve over a short time span.

ML methods require a specific probablity model for the transition between codon types over a tree. They start with a vector representing the 61 different codons (stop codons are excluded). At any point in time, a codon can mutate to one of nine other codons following a single base change (Figure 10.4). The base model given by Goldman and Yang (1994) defines the following transition probabilities between codons i and j,

$$Q_{ij} = \begin{cases} 0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega\pi_j & \text{for a nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{for a nonsynonymous transition} \end{cases} \quad \text{for } 1 \leq i, j \leq 61 \qquad (10.18)$$

The 61×61 $\mathbf{Q}$ matrix is specified by Equation 10.18. The π_j are the equilibrium frequencies of codon j (often calculated from the nucleotide frequencies at the three codon positions), while κ and ω are estimated parameters to account for biases in codon changes. Potential differences in transition versus transversion rates are accounted for by κ. One takes the current observed codons over the phylogeny, and then runs the model by considering all possible ancestral codons at each of the internal nodes (ancestors) in the tree. The model thus corrects for multiple hits. The key parameter of interest is ω, the strength of selection on replacement substitutions. In the initial models, ω was a fixed constant over all genes, but subsequent models (nicely paralleling the development of extensions of the Poisson random field model to allow γ to vary over genes/alleles) increasingly allowed ω to vary over sites, a point we expand on shortly. Figure 10.4 shows the basic structure of these transitions for a particular codon.

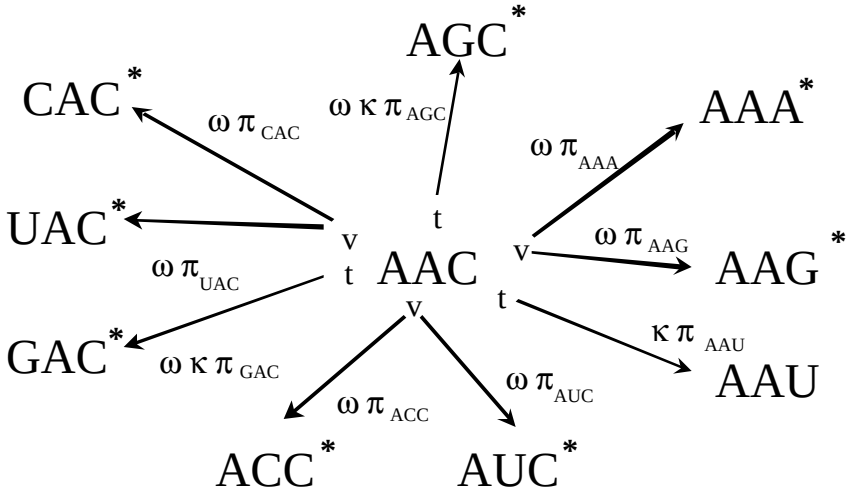

Figure 10.4. The various transmission probabilities under the codon evolution model (Equation 10.18) for the nine new codons that are within a single nucleotide change from the target codon (here AAC). Asterisks denote a replacement change, where the rate is a function of selection and hence ω. Since transitions (t) and transversions (v) may occur at different rates, setting the transversion rate as the baseline, κ denotes any transition rate correction (with $\kappa = 1$ if the two rates are equal). All changes are a function of π_j, the equilibrium frequency of codon j. Performing these same calculations over all 60 other non-stop codons generates the full transition matrix.

Tests for directional selection on a gene are accomplished by using this codon model superimposed on the phylogenetic tree, running likelihood calculation (over all codons) to find the ML solutions for $\mathbf{Q}$ matrix parameters. This allows for a direct test that $\omega > 1$ using the Likelihood Ratio approach (LW Appendix 4). The key to these likelihood calculations is that $\mathbf{P}(t)$, the codon transition matrix at time t, is related to the instantaneous rate matrix $\mathbf{Q}$ by

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) \tag{10.19}$$

Here,

$$P_{ij}(t) = \Pr(\text{codon } = i \text{ at time } t \,|\, \text{codon is } j \text{ at time } t = 0) \tag{10.20}$$

Diagonalize the matrix $\mathbf{Q}$ as $\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ where $\boldsymbol{\Lambda}$ is a diagonal matrix, with ith diagonal elements being the eigenvalues λ_i of $\mathbf{Q}$ (Appendix 4). Then

$$\exp(\mathbf{Q}t) = \mathbf{U}\exp(t\boldsymbol{\Lambda})\mathbf{U}^T$$

where

$$\exp(t\boldsymbol{\Lambda}) = \text{diag}(e^{t\lambda_1}, e^{t\lambda_2}, \cdots, e^{t\lambda_n}) = \begin{pmatrix} e^{t\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{t\lambda_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & e^{t\lambda_n} \end{pmatrix} \tag{10.21}$$

A variety of likelihood models based on Equation 10.21 are tested (much in the same way that one tests subsets of complex segregation analysis models, see LW Chapter 13), adding additional factors (i.e., nonzero κ, etc.) if they improved model fit (i.e., give a significant likelihood ratio). Evidence for selection on a gene is indicated if the likelihood ratio test for $\omega > 1$ is significant.

The base model (Equation 10.18) assumes all codons have the same ω value, which is not only unreasonable but also destroys most of the power of this approach, as our assumption is that $\omega < 1$ for most codons, which would mask those codons where $\omega > 1$. Nielsen and Yang (1998) and Yang et al. (2000) extend the base model by assuming a mixture-model, with the codons in a sequence being drawn from one of several categories, each with different ω values. For codons from class k, Equation 10.18 becomes

$$Q_{ij}^{(k)} = \begin{cases} 0 & \text{If } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{for a synonymous transversion} \\ \kappa\pi_j & \text{for a synonymous transition} \\ \omega^{(k)}\pi_j & \text{for a nonsynonymous transversion} \\ \omega^{(k)}\kappa\pi_j & \text{for a nonsynonymous transition} \end{cases} \tag{10.22a}$$

The simplest version has three classes, with codons either being neutral (with probability p_0), deleterious (with probability p_d), or advantageous (with probability $p_a = 1 - p_n - p_d$), with

$$\omega^{(k)} = \begin{cases} 0 & \text{deleterious class} \\ 1 & \text{neutral class} \\ \omega > 1 & \text{positively-selected class} \end{cases} \tag{10.22b}$$

The parameters p_0, p_d, and ω are estimated from the data by maximum likelihood (LW Chapter 13 examines ML on mixtures models). The idea is that one fits a base model (allowing only neutral and deleterious classes), and then fits the full model (Equation 10.22b or other extensions), using a likelihood-ratio test to see if the fit is significantly improved. If so, this

is taken as support for a history of repeated positive selection on a subset of codons in the gene of interest.

While Equation 10.22 is clearly an improvement, assigning all codons in the deleterious class an $\omega = 0$ (i.e., no substitutions) is clearly restrictive, as is assigning all codons in the advantageous class the same ω value. Nielsen and Yang (1998) and Yang et al. (2000) further expand Equation 10.22b by taking

$$\omega^{(k)} = \begin{cases} 0 < \omega < 1 & \text{deleterious class} \\ 1 & \text{neutral class} \\ 1 < \omega < \infty & \text{positively-selected class} \end{cases} \tag{10.23}$$

where now the fitness values ω for any particular codon in class k are random draws from some distribution whose parameters are again estimated by maximum-likelihood. This is exactly the approach used previously to allow γ to vary over genes in the PRF model (e.g., Equation 10.17). A number of candidate distributions for ω are possible, depending on whether we wish to restrict values to between (0,1) or to (1, ∞), for codons in the deleterious and positively-selected classes (respectively). For example, these authors use either a Beta or truncated Gamma distribution (restricted to returning values of $0 < \omega < 1$) for the deleterious class and a truncated Gamma (restricted to returning values of $\omega > 1$) for the positively-selected class (Appendix 2 reviews the Beta and Gamma distributions). Again, a model-fitting approach is used where one first fits a lower-order model, and then adds in the next set of parameters to see if the fit is significantly better. One unfortunate aspect of having so many potential models is nomenclature, with specific names (or model numbers) often being assigned to a particular model, making the literature a bit daunting to the uninitiated.

The power of the basic ML approach has been examined by Anisimova et al. (2001, 2002, 2003) and Wong et al. (2004), and is a function of two different sample sizes: the number of codons in the sequence and the number of actual sequences (number of taxa in the phylogeny). The more codons in a gene, the better, although 100 seems to give reasonable power. Power is more efficiently increased by adding more sequences (taxa), as opposed to looking at more codons. For moderately long sequences with a modest phylogeny (10 – 20 species), power can be quite reasonable, at least under the parameters simulated (typically 5 – 10% adaptive codons, each with ω around 5). They also found that using the χ^2 test to compute significance of likelihood ratios was conservative, and hence can be safely used, albeit suffering some reduction in power.

As might be expected, this basic framework can be modified in a number of additional ways, for example by letting some branches be under selection and others not (Yang and Nielsen 2002; Zhang et al. 2005). **Branch models** assume the same value of ω over all sites, but allow it to vary over branches, **site models** (our main focus here) allow ω to vary over sites, but not branches, while **branch-site** models allow ω to vary over both (e.g., Kosakovsky Pond and Frost 2005; Kosiol et al. 2008; Kosakovsky Pond et al. 2011). Anisimova and Yang (2007) discuss multiple-test corrections with branch-site models.

These methods are not fool-proof, and their robustness to the underlying distributional assumptions remains unclear (Suzuki and Nei 2004; Nozawa et al. 2009). For example, Zhang (2004) found 20-70% false positives in a branch-site model by Yang and Nielsen (2002) that allowed selection to operate on some branches, but not others. Relaxation of purifying selection on otherwise neutral branches can generate these spurious results. Zhang et al. (2005) simply replaced the assumption of $\omega = 0$ for the deleterious class with ω being an unknown to be estimated that lies within the interval (0,1), and obtained much better behavior (also see Yang and dos Reis 2011). These results point out how fragile some of these models may be, with essentially no internal controls to check for model consistency.

Finally, while our discussion of phylogeny-based divergence tests has focused exclusively on coding sequences, this need not be the case. Wong and Nielsen (2004) extend the

logic of codon-based models to noncoding regions. Here the test is the substitution rate in noncoding regions versus the rate at nearby silent sites. Using this approach, Wong and Nielsen found little signal of selection on noncoding regions of the sequences from 13 viral data sets, but strong signals of positive selection in protein-coding regions in five of these data sets. The major complication with using noncoding sequences is alignment, as homologous positions need to be compared over a phylogeny. Given that insertions and deletions are common, the time window for unambiguous alignments tends to be rather short.

Bayesian Estimators of Sites Under Positive Selection

Provided that one has the correct model, likelihood can be used to infer which actual *sites* have likely been under repeated positive selection. This powerful idea, due to Nielsen and Yang (1998), first tests the data to see if a model allowing for a subset of codons to be positively selection significantly improves the fit. If so, this provides evidence of positive selection *somewhere* in the gene of interest, but does not specify which particular codon(s) are the actual targets. To find these, Nielsen and Yang used Bayes' theorem (Equation A2.2), which is done as follows. Equations 10.18, 10.22, and 10.23 can be used to generate the conditional probability $\Pr(\text{data} \mid \omega_i)$ of the data (the pattern of observed states at a particular codon over the sampled tree of taxa), given that the codon came for fitness class i (typically three classes: neutral, deleterious, and advantageous). However, what we would really like is to "flip" this condition, and obtain $\Pr(\omega_i \mid \text{data})$, i.e., $\Pr(\text{ in class } i \mid \text{data})$, and in particular obtain the posterior probability of a codon being in the advantageous class given the observed data. Bayes' theorem allows us to this. Suppose there are k classes, with each class having a different associated ω. The posterior probability that a specific codon is in fitness class i is

$$\Pr(\text{class } i \mid D) = \frac{\Pr(D \mid \omega_i)\Pr(\text{class } i)}{\Pr(D)} = \frac{\Pr(D \mid \omega_i)\Pr(\text{class } i)}{\sum_{i=1}^{k}\Pr(D \mid \omega_i)\Pr(\text{class } i)} \tag{10.24}$$

where D is the pattern of codons for that site in the tree and the prior, $\Pr(\text{class } i)$, is estimated by maximum likelihood (i.e., the p_0, p_a, p_d). The case of interest is whether the codon belongs to the class of advantageous sites, $\Pr(\omega > 1 \mid D)$. This approach allows us to directly assign probabilities of selection to any particular site. Anisimova et al. (2002) found that large ω values and a modest to large number of sequences are required for this approach to have reasonable power.

Example 10.15. Bishop et al. (2000) examined the class I chitinase genes from 13 species of mainly North American *Arabis* (tower mustards), crucifers closely related to *Arabidopsis*. Chitinase genes are thought to be involved in pathogen defense, as they destroy the chitin in cell walls of fungi. Many fungi have evolved resistance to certain chitinases, so these genes are excellent targets for repeated cycles of selection. Codon evolution models estimated that between 64 and 77 percent of replacement substitutions were deleterious, with 5-14% advantageous (analysis using phylogenies estimated by different methods all yielded similar results). These favored sites had an estimated value of $\omega = 6.8$. Using the criteria of a posterior probability of membership in the advantageous class in excess of 0.95 (i.e. $\Pr(\text{selective class} \mid D) > 0.95$), 15 putative sites were located (using Equation 10.24). Seven of these sites involved only one alternative substitution, which evolved multiple times over the phylogeny. The authors had access to the three dimensional structure of chitinase, which shows a distinctive cleft, thought to be the active site. Mapping putative sites of positive selection showed a significant excess of sites cluster at the cleft (28% of cleft sites versus 19% elsewhere).

Balancing this apparently successful application of these methods to detect selected sites is the work of Yokoyama et al (2008). These authors examined the evolution of dim-light vision in

vertebrates, which is determined by the wavelength of maximal absorption of rhodopsin. This can be directly measured in the lab, which allowed the authors to specifically determine the role of particular substitutions in dim-light adaptation using 11 engineered ancestral rhodopsin sequences. They found that most of the change in maximal absorption can be accounted for by 12 sites. In contrast, Bayesian methods predicted a total of eight positively selected sites, none of which corresponded to sites show by mutagenesis to have adaptive roles.

A few final technical comments are in order. There are several different approaches based on Equation 10.24 that assign a posterior probability that a particular site is in a particular fitness class. Let $\eta_i = \Pr(\text{class} = i)$ denote the probability that a random site is in fitness class i (with $\omega = \omega_i$). This is the prior, and when applying Bayes' theorem, the assumption is that both η_i and ω_i are fully known without error. However, when applying Equation 10.24 this is not the case, as the η_i and ω_i values are estimated from the data, an approach referred to as **Empirical Bayes** (parameters of the prior estimated from the data). Further, Nielsen and Yang used the *point* estimates (the MLEs) for η_i and ω_i associated with each class, which does not incorporate any measure of the uncertainty of these estimates. Such an approach is often called **Naive Empirical Bayes**. A more powerful approach, **Bayes Empirical Bayes**, instead computes Equation 10.24 by integrating it over the posterior of the joint distribution of the η_i, ω_i. Such approaches have been developed by Huelsenbeck and Dyer (2004), Yang et al. (2005), Scheffler and Seoighe (2005), and Aris-Brosou (2006), and are reviewed by Anisimova and Liberles (2007).

Finally, control of false positives is a critical issue in scans for codons under positive selection. A typical gene has several hundred codons, resulting in several hundred tests. With a false positive rate of five percent at any particular site, we expect 10 and 20 false positives for genes with 200 and 400 codons. While standard or sequential Bonferroni methods (Appendix 6) could be used to control the gene-wide error rate, these result in very stringent tests and hence additional loss of power in a setting that may already have power issues. Use of the false discovery rate (FDR, Appendix 6) provides one solution. A false discovery rate of five percent means that *among the tests declared to be significant*, only five percent are false positives. Guindon et al. (2006) proposed two different approaches to implement the FDR in codon models. Their direct probability approach (following from Newton et al. 2004) constructs a list L containing all codons being declared as being significant. Using Equation 10.24 to assign a posterior probability of being in the selected class ($\omega > 1$), codon j is added to the list when $\beta_j = 1 - \Pr(\omega > 1 \mid D) \leq \delta$, where δ is some predefined threshold (note the small δ implies high posterior probability). Under this criteria, the expected number of false discoveries is just

$$FD(\delta) = \sum_{\beta_j \leq \delta} \beta_j \tag{10.25a}$$

This follows since β_j is the expected number of false positives at site j and we sum over all included sites to get the total number. A false discover rate of q is obtained by finding the largest value δ such that

$$\frac{FD(\delta)}{L_n} \leq q \tag{10.25b}$$

where $L_n > 0$ is the number of members in the list.

Example 10.16. Suppose Equation 10.24 is applied over a large number of codons within a gene, and we rank the values of $\Pr(\omega > 1 \mid D)$ from largest to smallest. In descending order,

the 15 largest values are 0.97, 0.97, 0.97, 0.96, 0.96, 0.96, 0.95, 0.95, 0.94, 0.94, 0.93, 0.93, 0.92, 0.915, and 0.91, for corresponding β values (in ascending order of) 0.03, 0.03, 0.03, 0.04, 0.04, 0.04, 0.05, 0.05, 0.06, 0.06, 0.07, 0.07, 0.08, 0.085, and 0.09. If we take our threshold as $\delta = 0.06$, what is the FDR? This list has $L_n = 10$ members, and the sum of their δ values is FD(0.06) = 0.43, giving the FDR as $\alpha = 0.43/10 = 0.043$, so that just slightly over four percent of the codons on this list are expected to be false-positives. If we use $\delta = 0.08$ as our threshold for list inclusion, $L_n = 13$, FD(0.08) = 0.65, giving the FDR as 0.65/13 = 0.05.

The second approach considered by Guindon et al. is a **parametric bootstrap**, which uses the estimated model parameters (the ω_i and probabilities of class membership) to generate a large number of simulated data sets D^*, where sites truly under selection are known. Using these simulated data, Equation 10.24 is used to compute $\Pr(\omega > 1 \mid D)$. Again, some threshold value is chosen, and codons declared to be included when they exceed this value. Since the sites truly under selection are known for the simulated data, the number of false discoveries for this threshold level is easily obtained. As above, the threshold value for declaring significance is varied until the FDR value reaches its desired level. Guindon et al. examined the performance of both approaches for controlling the false positive rate (PB for parametric bootstrapping and DP, direct probabilty, for Equation 10.25) as well as the two approaches for computing Equation 10.24 (Naive empirical Bayes, with Equation 10.24 using point estimates for model parameters; and BEB, Bayes empirical Bayes, integrating Equation 10.24 over the full posterior distribution of model parameters). They found that the combination of DP with BEB works best with weak selection (ω between 1.5 and 2 at positively sited sites), while PB works best under strong selection (selected sites have $\omega > 4$).

CONNECTING THE PARAMETERS OF ADAPTIVE EVOLUTION

As summarized in Table 10.2, a number of different parameters of adaptive evolution have been introduced in this chapter (as well as in Chapter 8), along with various machinery to estimate them. We have examined the connections between some these parameters (e.g., Equation 10.16). However, we have yet to develop a connection between the two key parameters in the different approaches to using divergence data: Poisson random field models estimate the scaled strength of selection γ at a site, while codon models estimate $\omega = K_a/K_s$ over a phylogeny.

We can connect these parameters as follows. Assume that synonymous sites are taken as the neutral benchmark, so that (as a first approximation) the per-site mutation rate μ_s is also the neutral mutation rate. Two types of mutations contribute to the rate of replacement substitutions: a fraction f (which is interchangeable with p_0) that are effectively neutral and a much smaller (perhaps zero) fraction p_a that are favored. Effectively neutral substitutions accrue at a rate of $f\mu_s$, while (Equation 8.22a) advantageous substitutions accrue at rate $\lambda = (2N\mu_b)(2sNe/N) = 2(2N_es)\mu_b = 2\gamma\mu_b = 2\gamma p_a\mu_s$. Hence

$$\omega = \frac{K_a}{K_s} = \frac{f\mu_s + 2\gamma p_a\mu_s}{\mu_s} = f + 2\gamma p_a = p_0 + 2\gamma p_a \tag{10.26a}$$

so that very strong selection ($\gamma\, p_a > 1$) is required for $\omega > 1$. Likewise,

$$\gamma = \frac{\omega - f}{2p_a} = \frac{\omega - p_0}{2p_a} \tag{10.26b}$$

If $f = 0.5$ and $p_a = 0.01$, so that half of the mutations are effectively neutral and one percent are favored, $\gamma = 25$ is required for $\omega = 1$, while $\omega = 3$ requires $\gamma = 125$. If p_a is now 0.001, a

value of $\gamma = 400$ only gives $\omega = 1.3$, which is a sufficiently small deviation to avoid detection in many cases. Finally, to connect α and ω, from Equations 10.16b and 10.26a

$$\alpha = \frac{2\gamma p_a}{2\gamma p_a + p_0} = \frac{2\gamma p_a}{\omega} \tag{10.26c}$$

which can alternately be expressed as

$$\alpha\,\omega = 2\gamma p_a, \qquad \text{and} \qquad \omega = \frac{2\gamma p_a}{\alpha} \tag{10.26d}$$

Table 10.2. Summary of the key parameters of adaptive evolution and their connections. Chapter 8 first introduced several of these (α, γ, μ_b), while ω and f were introduced in this chapter, which details with the estimation of all these parameters.

α	The fraction of substitutions that are adaptive
γ	The scaled strength of selection, $2N_e s$
μ_b	The adaptive mutation rate
p_a	The fraction of new mutations at a site that are advantageous
λ	The rate of adaptive fixations, $\lambda = 2\gamma\mu_b$
f	The fraction of neutral mutations (relative to some standard, typically silent sites)
p_0	Probability a new mutation is effective neutral ($p_0 = f$) when standard is neutral sites
$1 - f$	The amount of constraint on a site (relative to some standard, typically silent sites)
ω	The ratio of the replacement to silent substitution rates

$$\omega = f + 2\gamma p_a = \frac{2\gamma p_a}{\alpha} \quad \text{(Equations 10.26a,d)}$$

$$\gamma = \frac{\omega - f}{2p_a} = \frac{\omega - p_0}{2p_a} \quad \text{(Equation 10.26 b)}$$

$$\alpha = \frac{\lambda}{\lambda + \mu p_0} = \frac{2\gamma}{2\gamma + p_0/p_a} = \frac{2\gamma p_a}{\omega} \quad \text{(Equations 10.16b,c, 10.26c)}$$

THE SEARCH FOR SELECTION: CLOSING COMMENTS

Detecting selection using molecular data is a major growth industry and will continue to accelerate as whole-genome sequencing becomes increasingly faster and cheaper. As the last two chapters indicate, there is an enormous amount of statistical machinery proposed to carry out this task, but every method has major limitations. Ecologists and evolutionary biologists search for selection using complementary trait-based approaches, which require specifying potential traits under selection, and measuring the association between these and individual fitness (Chapters 28 and 29). While such trait-based approaches allow us to consider particular traits of interest, molecular data has several advantages. Two are fairly obvious in that traits need not be specified and measurement of individual fitness is not needed. Their greatest advantage, however, is that molecular data are a time machine. We cannot go back in the past to measure traits and fitness, but past selection *may* leave a number of different signals in the genome. Very recent events *may* leave sweep-like signatures (Chapter 8), and Chapter 9 reviews the myriad of tests for detecting these. If polygenic adaptation is the rule, major changes in trait values can occur through only minor changes in a number of loci, each of small effect. In this case, very little molecular signal is expected, but ongoing selection can easily be detected using trait-based methods (if one knows the correct traits!). Over a longer time scale, repeated selection events *may* leave molecular patterns. Population-based divergence tests (HKA, MK) can detect patterns of repeated positive selection over an entire gene

during the divergence of two populations/species, while phylogeny-based divergence tests (codon models) can detect repeated positive selection at the same *codon* over a phylogeny.

Caution is in Order When Declaring Positive Selection!

Since just about every test can give a false-positive for reasons other than positive selection, any detected region should *always* be viewed as no more that a candidate to be followed up by the hard work of accessing whether it has any functional impact and, if so, what the nature of selection might be. In particular, investigators should be extremely careful of "just-so" stories, wherein once a region is detected, some clever story is proposed as to the cause of selection in this region. One must resist the notion that functional differences can automatically be equated to adaptive changes (Gould and Lewontin 1979; Storz and Wheat 2010; Barrett and Hoekstra 2011). In the words of Nielsen (2009), "evidence of selection, and knowledge of the function of a gene, does not constitute evidence for adaptation", as the following cautionary tale illustrates.

Example 10.17. Humans show dramatic expansion of brain size with respect to most mammals, with this increase in (relative) size usually assumed to be corrected with increased cognitive abilities. Primary microcephaly is a condition in humans resulting in small heads, but other normal features. Nonfunctional alleles at the genes *microcephalin* and *ASPM* (abnormal spindle-like microcephaly associated) both display the microcephaly phenotypes, with a typical individual having a brain size of around 400 cm^3 (versus the normal 1400 cm^3), comparable to that in early hominids. Not surprising, several studies have looked for selection on these genes within the primate lineage. Zhang (2003) inferred a K_a/K_s ratio of 1.03 for *ASPM* on the branch from the human-chimp common ancestor to humans, but a ratio of 0.66 on the branch from this ancestor to chimps. Values of 0.43 to 0.29 were found along other branches in mammals, suggested positive selection along the human lineage. Evans et al. (2004a) also examined *ASPM* over a larger phylogeny ranging from new world monkeys through humans. Accelerated ($K_a/K_s > 1$) rates of evolution were seen between gibbons and the ancestor the great apes, and a large acceleration ($K_a/K_s = 1.44$) was seen on the lineage from the human/chimp ancestor to humans. Evans et al. also performed a McDonald-Kreitman test, comparing the polymorphisms within humans to the divergence since the human-chimp common ancestor, finding

	Fixed	Polymorphic
Synonymous	7	10
Replacement	19	6

Fisher's exact test gives a p value of 0.01, with an excess of around 15 replacement substitutions over what is expected from the replacement/synonymous ratio seen in the polymorphism data ($\alpha \sim 80\%$.). Similar results were seen for *microcephalin*. Evans et al. (2004b) found K_a/K_s = 1.05 in the simian lineages leading to humans, and ratios of 0.4 to 0.6 along other mammalian lineages. A further breakdown showed that most of the excess in K_a/K_s occurred from prosimians to the branching of the great apes, with values less than one within the great apes. They also found a significant McDonald-Kreitman result, with an estimated 45 adaptive (replacement) substitutions occurring between prosimians and humans. Thus, *microcephalin* seems associated with expansion of brain size leading to the great apes, while *ASPM* is further associated with the increase in brain size specifically along the lineage leading to humans.

Building on these strong observations of selection leading to the human lineage, Mekel-Bobrov et al. (2005) and Evans et al. (2005) searched for *ongoing* selection in these two genes, and found strong signals in each. Evans et al. (2005) found that the *microcephalin* gene had one haplotype (associated with a replacement substitution) at much higher frequencies than the

others, with extended linkage disequilibrium and small intra-allelic variation. Indeed, using intra-allelic variation, the age of this haplotype was estimated at 37 thousand years (with a range of 14 to 60 thousand). Young alleles at high frequencies are hallmark indicators of positive selection (Chapter 9). Extensive coalescent simulations using a variety of population structures all gave high levels of significance to these results. This exact pattern was seen by Mekel-Bobrov et al. (2005) with *ASPM*: a common haplotype with long LD and a very recent estimated origin (5,800 years). Again, coalescent simulations of neutral drift under a variety of proposed models of human population growth and expansion showed these results to be highly significant. Together, these studies strongly suggested on-going selection in these two genes. They gathered a significant amount of attention, not the least of which was do to the finding that the putative adaptive haplotypes were in higher frequencies in Europe and Asia relative to Africa, and the connection that is often drawn between cognition and brain size.

Although Evans et al. (2005) cautioned that "it remains formally possible that an unrecognized function of *microcephalin* outside the brain is actually the substrate of selection", many interpreted the above data as an adaptive response in intelligence. After all, two functional genes that both influence brain size, a presumed correlate of intelligence, coupled with a history of past, and ongoing, selection does indeed suggest a case for selection on intelligence. This view, however, was quickly dispelled. Timpson et al (2007) and Mekel-Bobrov et al. (2007) showed in large sample sizes (900 and 2400, respectively) that there was no correlation between the putative adaptive haplotypes and increased intelligence. Any on-going selection on these genes does not appear to correlate with any selection for increased cognition. Currant et al. (2006) further noted that *spatial* models of population growth were not considered, and here it is possible to see the above patterns for neutral mutations that arise along the leading lead of a recent population expansion (through allelic surfing, see Chapter 9). If not for the concern among many geneticists at drawing incorrect social implications from the initial selection findings, this saga might have become a textbook standard in the search for selection. It is unlikely that most loci with strong signatures of selection are likely to receive this leave of scrutiny.

Curbing Enthusiasm

We started this set of chapters with a plea for caution and will do so again to end them. Just like the great electrophoresis hunt in the 1970's (grinding up every species/population in sight to measure segregating protein variation) and the great QTL hunt in the 1990's (trying to find QTLs for just about every trait in your favorite organism), we are now experiencing the great selection hunt phase of evolutionary genetics. The obvious excitement of detecting either ongoing selection or targets with a history of repeated past selection must also be tempered with caution. There are a huge variety of different tests, but no one best test even for a particular situation (much less over all settings). Simple methods may lack power, but very sophisticated highly parametric tests may not be very robust to modeling assumptions. As mentioned multiple times, issues of demography (changes in population size) and population structure can cripple most tests. More sophisticated versions developed to circumvent some of these issues are not yet fully vetted, so must be used with caution. Finally, there is the **Beavis effect** (LW Chapter 15), also know as the **winner's curse** (Kraft 2008), in which a parameter declared significant is often overestimated. This problem is especially acute when the power for detection is low. When a selection signal is detected (likely out of a sea of candidates, with each test having moderate to low power), the actual effect is likely overestimated, and potentially by a very large amount. These comments are not meant to discourage the use of these powerful methods, but rather to ensure that the enthusiasm with which they are applied is somewhat tempered by the cold reality of their limitations.

Finally, as stressed throughout the last few chapters, even when successful, these tests

give us an insight into *just a tiny fraction* of all selective substitutions. How representative this subsample is of adaptive selection in general is unclear, but it is certainly biased, so significant caution is in order in extrapolating these results to general statements about adaptation. However, it is also clear that multiple selection events (be they recurrent sweeps or background selection) clearly leave an impact on linked neutral sites, and most genomes show ample signals that this a very common phenomena (Chapter 8).

Literature Cited

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophla* DNA. *Genetics* 139: 1067–1076. [10]

Andolfatto, P.. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437: 1149-1152. [10]

Andolfatto, P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* 17: 1755–1762. [10]

Andolfatto, P. 2008. Controlling type-1 error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* 180: 1767–1771. [10]

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18: 1585–1592. [10]

Anisimova, M., J. P. Bielawski, and Z. Yang. 2002. Accuracy and power of Bayes prediction of amino acids sites under positive selection. *Mol. Biol. Evol.* 19: 950–958. [10]

Anisimova, M., and D. A. Liberles. 2007. The quest for natural selection in the age of comparative genomics. *Heredity* 99: 567–579. [10]

Anisimova, M., R.. Nielsen, and Z. Yang. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164: 1229–1236. [10]

Anisimova, M., and Z. Yang. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol. Biol. Evol.* 24: 1219–1228. [10]

Aris-Brosou, S. 2006. Identifying sites under positive selection with uncertain parameter estimates. *Genome* 49: 767–776. [10]

Axelsson, E., and H. Ellegren. 2009. Quantification of adaptive evolution of genes expressed in avian brain and population size effect on the efficacy of selection. *Mol. Biol. Evol.* 26: 1073–1079. [10]

Bachtrog, D. 2008. Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes. *BMC Evol. Biol.* 8: 334. [10]

Barrett, R. D. H., and H. E. Hoekstra. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nat. Reviews Genet.* 12: 767–780. [10]

Begun, D. J., and C. F. Aquadro. 1991. Molecular population genetics of the distal portion of the *X* chromosome in *Drosophila*: Evidence for genetic hitchhiking of the *yellow-achaete* region. *Genetics* 129: 1147–1158. [10]

Bierne, N., and A. Eyre-Walker. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* 21: 1350–1360. [10]

Bishop, J. F., A. M. Dean, and T. Mitchell-Olds. 2000. Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. *Proc. Natl. Acad. Sci. USA* 97: 5322–5327. [10]

Blyth, C. R. 1972. On Simpson's paradox and the sure-thing principle. *J. Amer. Stat. Assoc.* 67: 364-–366. [10]

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, T. J. White, R. Nielsen, A. G. Clark, and C. D. Bustamante. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics* 4: e1000083. [10]

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez, D.Civello, M. D. Adams, M. Cargill, and A. G. Clark. 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437: 1153–1157. [10]

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416: 531–534. [10]

Bustamante, C. D., J. Wakeley, S. A. Sawyer, and D. L. Hartl. 2001. Directional selection and the site-frequency spectrum. *Genetics* 19: 1779–1788. [10]

Campos, J. L., K. Zeng, D. J. Pakrer, B. Charlesowrth, and P. R. Haddrill. 2012. Codon usage bias and effective population sizes on the X chromosome versus autosomes in *Drosophila melanogaster. Mol. Biol. Evol.* 30: 811–823. [8, 10]

Carneiro, M., F. W. Albert, J. Melo-Ferreira, N. Galtier, P. Gayral, J. A. Blanco-Aguiar, R. Villafuerte, M. W. Nachman, and N. Ferrand. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. *Mol. Bio. Evol.* 29: 1837–1849. [8, 10]

Chamary, J. V., J. L. Parmley, and L. D. Hurst. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7: 98–108. [10]

Charlesworth, B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* 63: 213–227. [10]

Charlesworth, J., and A. Eyre-Walker. 2006. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* 23: 1348–1356. [10]

Charlesworth, J., and A. Eyre-Waker. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* 25: 1007–1015. [10]

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69–87. [10]

Chen, C.-H., T.-J. Chuang, B.-Y. Liao, and F.-C. Chen. 2009. Scanning for signatures of positive selection for human-specific insertions and deletions. *Genome Biol. Evol.* 1: 415–419. [10]

Currat, M., L. Excoffier, W. Maddison, S. P. Otto, N. Ray, M. C. Whitlock, and S. Yeaman. 2006. Comment on "Ongoing adaptive evolution of *ASPM*, a brain size determinant in *Homo sapiens*" and "*Microcephalin*, a gene regulating brain size, continues to evolve adaptively in humans". *Science* 313: 172. [10]

Desai, M. M., and J. B. Plotkin. 2008. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics* 180: 2175–2191. [10]

DuMont, V. B., J. C. Fay, P. P. Calabrese, and C. F. Aquadro. 2004. DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans:* A case of accelerated synonymous site divergence. *Genetics* 167: 171–185. [10]

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74. [8, 10]

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* 13: 685–690. [10]

Evans, P. D., J. R. Anderson, E. J. Vallender, S. L. Gilbert, C. M. Malcom, S. Dorus, and B. T. Lahn. 2004a. Adaptive evolution of *ASPM*, a major determinant of cerebral cortical size in humans. *Hum. Mol. Genet.* 13: 489–494. [10]

Evans, P. D., J. R. Anderson, E. J. Vallender, S. S. Choi, and B. T. Lahn. 2004b. Reconstructing the evolutionary history of *microcephalin*, a gene controlling human brain size. *Hum. Mol. Genet.* 13: 1139–1145. [10]

Evans, P. D., S. L. Gilbert, N. Mekel-Bobrov, E. J. Vallender, J. R. Anderson, L. M. Vaez-Azizi, S. A. Tishkoff, R. R. Hudson, and B. T. Lahn. 2005. *Microcephalin*, a gene regulating brain size, continues to evolve adaptively in humans. *Science* 309: 1717–1720. [10]

Eyre-Walker, A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* 162: 217–2024. [10]

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends Ecol. Evol.* 10: 569–575. [10]

Eyre-Walker, A., and P. D. Keightley. 2007. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8: 610–618. [10]

Eyre-Walker, A., and P. D. Keightley. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* 26: 2097–2108. [10]

Eyre-Walker, A., M. Woolfit, and T. Phelps. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173: 891–900. [10]

Fay, J. C. 2011. Weighing the evidence for adaptation at the molecular level. *Trends Genet.* 27: 343–349. [10]

Fay, J. C., G. J. Wyckoff, and C.-I. Wu. 2001. Positive and negative selection on the human genome. *Genetics* 158: 1227–1234. [10]

Fay, J. C., G. J. Wyckoff and C.-I Wu. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila. Nature* 415: 10024–1026. [10]

Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer, Sunderland, MA [10]

Fitch, W. M., R. M. Bush, C. A. Vender, and N. J. Cox. 1997. Long-term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. USA* 94: 7712–7718. [10]

Foxe, J. P., V. Dar, H. Zheng, M. Nordborg, B. S. Gaut, and S. I. Wright. 2008. Selection on amino acid substitutions in *Arabidopsis. Mol. Biol. Evol.* 25: 1375–1383. [10]

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol* 11: 725–736. [10]

Gojobori, J., H. Tang, J. M. Akey, and C.-I. Wu. 2007. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proc. Natl. Acad. Sci. USA* 104: 3907–3912. [10]

Good, I. J., and Y. Mittal. 1987. The amalgamation and geometry of two-by-two contingency tables. *Ann. Stat.* 15: 694–711. [10]

Gould, S. J., and R. C. Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond.* B. 205: 581–598. [10]

Goss, P. J. E., and R. C. Lewontin. 1996. Detecting heterogeneity of substitution along DNA and protein sequences. *Genetics* 143: 589–602. [10]

Gossmann, T. I., B.-H. Song, A. J. Windsor, T. Mitchell-Olds, C. J. Dixon, M. V. Kapralove, D. A. Filatove, and A. Eyre-Walker. 2010. Genome wide analysis reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.* 27: 1822–1832. [10]

Gossmann, T. I., D. Waxman, and A. Eyre-Walker. 2014. Fluctuating selection models and McDonald-Kreitman type analyses. *PloS One* 9: e84540. [10]

Graur, D., and W.-H. Li. 1991. Neutral mutation hypothesis test. *Nature* 354: 114-115. [10]

Graur, D., and W.-H. Li. 2000. *Fundamentals of molecular evolution.* Sinauer, Sunderland, MA [10]

Greenland, S. 1982. Interpretation and estimation of summary ratios under heterogeneity. *Stat. Med.* 1: 217–227. [10]

Guindon, S., M. Black, and A. Rodrigo. 2006. Control of the false discovery rate applied to the detection of positively selected amino acid sites. *Mol. Biol. Evol.* 23: 919–926. [10]

Haddrill, P. R., D. Bachtrog, and P. Andolfatto. 2008. Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol. Bio. Evol.* 25: 1825–1834. [10]

Haddrill, P. R., L. Loewe, and B. Charlesworth. 2010. Estimating the parameters of selection on non-synonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. *Genet.* 185: 1381–1396. [10]

Haldane, J. B. S. 1956. The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Hum. Genet.* 20: 309–311. [10]

Halligan, D. L., F. Oliver, A. Eyer-Walker, B. Harr, and P. D. Keightley. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics* 6: e1000825. [10]

Halligan, D. L., and P. D. Keightley. 2006. Ubiquitous selective constraints in the *Drosphila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16: 875–884. [10]

Harding, R. M., E. Healy, A. J. Ray, N. S. Ellis, N. Flanagan, C. Todd, C. Dixon, A. Sajantile, I. J. Jackson, M. A. Brich-Machin, and J. L. Rees. 2000. Evidence for variable selective pressures at MC1R. *Am. J. Hum. Genet.* 66: 1351–1361. [10]

Hartl, D. L., E. N. Moriyama, and S. A. Sawyer. 1994. Selection intensity for codon bias. *Genetics* 138: 227–234. [10]

Hudson, R. R. 1993. Levels of DNA polymorphism and divergence yield important insights into evolutionary processes. *Proc. Natl. Acad. Sci. USA* 90: 7425–7426. [10]

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159. [10]

Huelsenbeck, J. P., and K. A. Dyer. 2004. Bayesian estimation of positively selected sites. *J. Mol. Evol.* 58: 661–672. [10]

Huerta-Sanchez, E., R. Durrett, and C. D. Bustamante. 2008. Population genetics of polymorphism and divergence under fluctuating selection. *Genetics* 178: 325–337. [10]

Hughes, A. L. 1999. *Adaptive evolution of genes and genomes.* Oxford University Press, Oxford. [10]

Hughes, A. L. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99: 364–373. [10]

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335: 167–170. [10]

Hughes, A. L., and M. Nei. 1989. Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. *Proc. Natl. Acad. Sci. USA* 86: 958–962. [10]

Hughes, A. L., B. Packer, R. Welch, A. W. Bergan, S. J. Chanock, and M. Yeager. 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. *Prod. Natl. Acad. Sci. USA* 100: 15754–15757. [10]

Hurst, L. D., and C. Pál. 2001. Evidence for purifying selection acting on silent sites in *BRCA1*. *Trends Genet.* 17: 62–65. [10]

Ingvarsson, P. K. 2004. Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Gene. Res.* 83: 31–39. [10]

Ingvarsson, P. K. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula. Mol. Biol. Evol.* 27: 650–660. [10]

Innan, H. 2006. Modified Hudson-Kreitman-Aguadé test and two-dimensional evaluation of neutrality tests. *Genetics* 173: 1725–1733. [10]

Jewell, N. P. 1986. On the bias of commonly used measures of association for 2 x 2 tables. *Biometrics* 42: 351–358. [10]

Keightley, P. D., and A. Eyre-Walker. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J. Mol. Evol.* 74: 61–68. [10]

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61: 893–903. [10]

Kimura, M. 1983. *The neutral theory of molecular evolution*, Cambridge Univ. Press, Cambridge. [10]

Kosakovsky Pond, S. L, and S. D. W. Forst. 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol. Biol. Evol.* 22: 478–485. [10]

Kosakovsky Pond, S. L, B. Murrell, M. Fourment, S. D. W. Forst, W. Delport, and K. Scheffler. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.* 28: 3033–3043. [10]

Kosiol, C., T. Vinar, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen, and A. Siepel. 2008. Patterns of positive selection in six mammalian genomes. *PLos Genetics* 4: e1000144. [10]

Kousathanas, A., and P. D. Keightly. 2013. A comparison of methods to infer the distribution of fitness effects of new mutations. *Genet.* 193: 1197–1208. [10]

Kousathanas, A., F. Oliver, D. L. Halligan, and P. D. Keightley. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol. Biol. Evol.* 28: 1183–1191. [10]

Kraft, P. 2008. Curses — Winner's and otherwise — in genetic epidemiology. *Epidemiology* 19: 649–651. [10]

Le Core, V., F. Roux, and X. Rebound. 2002. DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana:* Extensive nonsynonymous variation is consistent with local selection for flowering time. *Mol. Biol. Evol.* 19: 1261–1271. [10]

Lewontin, R. C. 1974. *The genetic basis of evolutionary change.* Columbia University Press, New York. [10]

Li, W.-H. 2006. *Molecular evolution.* Sinauer, Sunderland, MA [10]

Li, Y. F., J. C. Costello, A. K. Holloway, and M. W. Hahn. 2008. "Reverse ecology" and the power of population genomics. *Evolution* 62: 2984–2994. [9, 11]

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clarck, and C. D. Bustamante. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451: 994–997. [10]

McDonald, J. H. 1996. Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphisms to divergence. *Mol. Biol. Evo.* 13: 263–260. [10]

McDonald, J. H. 1998. Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphisms to divergence. *Mol. Biol. Evo.* 15: 377–384. [10]

McDonald, J. H., and M. Kreitman. 1991a. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652–654. [10]

McDonald, J. H., and M. Kreitman. 1991b. Neutral mutation hypothesis test. *Nature* 354: 116. [10]

Mekel-Bobrov, N., S. L. Gilbert, P. D. Evans, E. J. Vallender, J. R. Anderson, R. R. Hudson, S. A. Tishkoff, and B. T. Lahn. 2005. Ongoing adaptive evolution of *ASPM*, a brain size determinant in *Homo sapiens*. *Science* 309: 1720–1722. [10]

Mekel-Bobrov, N. D. Posthuma, S. L. Gilbert, P. Lind, M. F. Gosso, M. Luciano, S. E. Harris, T. C. Bates, T. J. C. Polderman, L. J. Whalley, H. Fox, J. M. Starr, P. D. Evans, G. W. Montgomery, C. Fernandes, P. Heutink, N. G. Martin, D. I. Boomsma, I. J. Deary, M. J. Wright, E. J. C. de Geus, and B. T. Lahn. 2007. The ongoing adaptive evolution of *ASPM* and *microcephalin* is not explained by increased intelligence. *Hum. Mol. Genet.* 16: 600–608. [10]

Miyata, T., and T. Yasunaga. 1980. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* 16: 12–36. [10]

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11: 715–724. [10]

Nei, M., and S. Kumar. 2000. *Molecular evolution and phylogenetics.* Oxford University Press, Oxford. [10]

Newton, M., A. Noueiry, D. Sarkar, and P. Ahlquist. 2004. Detecting differential expression with a semi-parametric hierarchical mixture method. *Biostats* 5: 155–176. [10]

Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* 86: 641–647. [10]

Nielsen, R. 2009. Adaptionism – 30 years after Gould and Lewontin. *Evolution* 63: 2487–2490. [10]

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and application to the HIV-1 envelope gene. *Genetics* 148: 929–936. [10]

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Bio* 3: e170. [10]

Nozawa, M., Y. Suzuki, and M. Nei. 2009. Reliabilities of identifying positive selection by branch-site and the site-prediction methods. *Proc. Natl. Acad. Sci. USA* 106: 6700–6705. [10]

Page, R. D. M. and E. C. Holmes. 1998. *Molecular evolution: a phylogenetic approach.* Blackwell. [10]

Podlaha, O., D. M. Webb, P. K. Tucker, and J. Zhang. 2005. Positive selection for indel substitutions in the rodent sperm protein *catsper1*. *Mol. Biol. Evol.* 22: 1845–1852. [10]

Pond, S. K., and S. V. Muse. 2005. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* 22: 2375–2358. [10]

Rand, D. M., and L. M. Kann. 1996. Excessive amino acid polymorphism in mitochondiral DNA: Contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* 13: 735–748. [10]

Ratnakumar, A., S. Mousseet, S. Glémon, J. Berglund, N. Galtier, L. Duret, and M. T. Webster. 2010. Detecting positive selection within genomes: the problem of baised gene conversion. *Phil. Trans. R. Soc. B* 365: 2571–2580. [10]

Robertson, A. 1962. Selection for heterozygotes in small populations. *Genetics* 47: 1291–1300. [10]

Sawyer, S. A., and D. L. Hartl. 1992. Population genetics of polymorphism and divergence. *Genetics* 132: 1161–1176. [10]

Sawyer, S. A., R. J. Kulathinal, C. D. Bustamante, and D. L. Hartl. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* 57: S154–S164. [10]

Sawyer, S., J. Parsch, Z. Zhang, and D. L. Hartl. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Prod. Natl. Acad. Sci.* 104: 6504–6510. [10]

Scheffler, K., and C. Seoighe. 2005. A Bayesian model comparison approach to inferring positive selection. *Mol. Biol. Evol.* 12: 2513–2540. [10]

Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, P. D. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189: 1427–1437. [8, 10]

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto. 2009. Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics* 5: e100049. [10]

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *J. Roy. Stat. Soc. B* 13: 238–241. [10]

Slotte, T., J. P. Foxe, K. M. Hazzouri, and S. I. Wright. 2010. Genome-wide evidence for efficient positive and purifying seleciton in *Capesella grandiflora*, a plant species with a large effective population size. *Mol. Biol. Evol.* 27: 1813–1821. [10]

Smith, N. G. C., and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415: 1022–1024. [10]

Strasburg, J. L., C. Scotti-Saintagne, I. Scotti, Z. Lai, and L. H. Rieseberg. 2009. Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species. *Mol. Biol. Evol.* 26: 1341–1355. [10]

Stoletzki, N., and A. Eyre-Walker. 2011. Estimation of the neutrality index. *Mol. Biol. Evol.* 28: 63–70. [10]

Storz, Jay F., and C. W. Wheat. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* 64: 2489–2509. [10]

Suzuki, Y., and T. Gojobori. 1999. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol* 16: 1315–1328. [10]

Suzuki, Y., and M. Nei. 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: 1865–1869. [10]

Suzuki, Y., and M. Nei. 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: 914–921. [10]

Tarone, R. E. 1981. On summary estimators of relative risk. *J. Chron. Dis.* 34: 463–468. [10]

Templeton, A. R. 1987. Genetic systems and evolutionary rates. *In* K. S. W. Campbell and M. F. Day (eds), *Rates of evolution*, pp. 218–234. Allen and Unwin, London. [10]

Templeton, A. R. 1996. Contingency tests of neutrality using intra/interspecific gene trees: The rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the Hominoid primates. *Genetics* 144: 1263–1270. [10]

Timpson, N., J. Heron, G. D. Smith, and W. Enard. 2007. Comment on papers by Evans *et al.* and Mekel-Bobrov *et al.* on evidence for positive selection of *MCPH1* and *ASPM*. *Science* 317: 1036. [10]

Torgerson, D. G., A. R. Boyko, R. D. Hernandez, A. Indap, X. Hu, T. J. White, J. J. Sninsky, M. Cargill, M. D. Adams, C. D. Bustamante, and A. G. Clark. 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genetics* 5: e1000592. [10]

Wakeley, J. 2003. Polymorphism and divergence for island-model species. *Genetics* 163: 411–420. [10]

Webster, M. T., and N. G. C. Smith. 2004. Fixation biases affecting humans SNPs. *Trends Genet.* 20: 122–126. [10]

Weinreich, D. M., and D. M. Rand. 2000. Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondiral genomes. *Genetics* 156: 385–399. [10]

Welch, J. J. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* 173: 821–837. [10]

Welch, J. J., A. Eyre-Walker, and D. Waxman. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *J. Mol. Evol.* 67: 418–426. [10]

Whittam, T. S., and M. Nei. 1991. Neutral mutation hypothesis test. *Nature* 354: 115-116. [10]

Williamson, S., A. Fledel-Alon, and C. D. Bustamante. 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* 168: 463–475. [10]

Wilson, D. J., R. D. Hernandez, P. Andelfatto, and M. Przeworski. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.* 7: e1002395. [8]

Wolf, J. B. W., A. Künstner, K. Nam, M. Jakobsson, and H. Ellegren. 2009. Nonlinear dynamics of nonsynonymous (d_N) and synonymous (d_S) substitution rates affects inference of selection. *Genom. Biol. Evol.* 1: 308–319. [10]

Wong, W. S. W., and R. Nielsen. 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* 167; 949–958. [10]

Wong, W. S. W., Z. Yang, N. Goldman, and R. Nielsen. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in the protein coding sequences and for identifying positively selected sites. *Genetics* 168: 1041–1051. [10]

Wright, S. 1938. The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. USA* 24: 253–259. [10]

Wright, S. I., and P. Andolfatto. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Ann. Rev. Ecol. Evol. Syst.* 39: 193–213. [10]

Wright, S. I. , and B. Charlesworth. 2004. The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics* 168: 1071–1076. [10]

Yang, Z. 2006. *Computational molecular evolution.* Oxford University Press, Oxford. [10]

Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15: 496–503. [10]

Yang, Z., and M. dos Reis. 2011. Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.* 28: 1217–1228. [10]

Yang, Z., and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19: 908–917. [10]

Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressures at amino acid sites. *Genetics* 155: 431–449. [10]

Yang, Z., W. S. W. Wong, and R. Nielsen. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22: 1107–1118. [10]

Yokoyama, S., T. Tada, H. Zhang, and L. Britt. 2008. Elucidation of phenotypic adapations: molecular analysis of dim-light vision proteins in vertebrates. *Proc. Natl. Acad. Sci. USA* 105: 13480–13485. [10]

Yule, G. U. 1903. Notes on the theory of association of attributes in statistics. *Biometrika* 2: 121-–134. [10]

Zhang, J. 2003. Evolution of the human *ASPM* gene, a major determinant of brain size. *Genetics* 165: 2063–2070. [10]

Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol. Biol. Evol.* 21: 1332–1339. [10]

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22: 2472–2479. [10]

Zhang, L., and W.-H. Li. 2005. Human SNPs reveal no evidence of frequency positive selection. *Mol. Biol. Evol.* 22: 2504–2507. [10]