J Mol Evol (2006) 63:635–653
DOI: 10.1007/s00239-005-0233-x

JOURNAL OF MOLECULAR EVOLUTION

Codon Usage and Selection on Proteins

Joshua B. Plotkin,[1] Jonathan Dushoff,[2] Michael M. Desai,[3] Hunter B. Fraser[4]

[1] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[2] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA
[3] Department of Molecular and Cellular Biology and Department of Physics, Harvard University, Cambridge, MA 02138, USA
[4] Department of Molecular and Cellular Biology, University of California, Berkeley, Berkeley, CA 94707, USA

Received: 3 October 2005 / Accepted: 12 June 2006 [*Reviewing Editor:* Dr. Lauren Meyers]

Abstract. Selection pressures on proteins are usually measured by comparing homologous nucleotide sequences (Zuckerkandl and Pauling 1965). Recently we introduced a novel method, termed *volatility*, to estimate selection pressures on proteins on the basis of their synonymous codon usage (Plotkin and Dushoff 2003; Plotkin et al. 2004). Here we provide a theoretical foundation for this approach. Under the Fisher-Wright model, we derive the expected frequencies of synonymous codons as a function of the strength of selection on amino acids, the mutation rate, and the effective population size. We analyze the conditions under which we can expect to draw inferences from biased codon usage, and we estimate the time scales required to establish and maintain such a signal. We find that synonymous codon usage can reliably distinguish between negative selection and neutrality only for organisms, such as some microbes, that experience large effective population sizes or periods of elevated mutation rates. The power of volatility to detect positive selection is also modest—requiring approximately 100 selected sites—but it depends less strongly on population size. We show that phenomena such as transient hyper-mutators can improve the power of volatility to detect selection, even when the neutral site heterozygosity is low. We also discuss several confounding factors, neglected by the Fisher-Wright model, that may limit the applicability of volatility in practice.

Key words: Codon usage — Selection — Fisher-Wright — Diffusion — Volatility — Protein evolution

Introduction

Background

Nucleotide coding sequences of many organisms exhibit significant codon bias—that is, unequal usage of synonymous codons. Codon bias has been attributed both to neutral processes, such as asymmetric mutation rates, and to selection acting on the synonymous codons themselves. The most common selective explanation of codon bias posits that synonymous codons differ in their fitness according to the relative abundances of iso-accepting tRNAs; a codon corresponding to a more abundant tRNA would be used preferentially so as increase translational efficiency (Ikemura 1981; Debry and Marzluff 1994; Sorensen et al. 1989). To a large extent, this hypothesis has successfully explained interspecific variation in codon usage for organisms ranging from *Escherichia coli* to *Drosophila melanogaster* (Akashi 2001).

Recently, we have noted that codon bias in a protein sequence can also result from selection at the amino acid level, even in the absence of direct selection on synonymous codons themselves (Plotkin and Dushoff 2003; Plotkin et al. 2004). Codon bias arises from selection at the amino acid level because of asymmetries in the structure of the standard genetic

Correspondence to: Joshua B. Plotkin; *email:* jplotkin@sas.upenn.edu

code. Proteins that experience different selective regimes should exhibit different synonymous codon usage. Following from this observation, we have introduced methods to screen a single genome sequence for estimates of the relative selection pressures on its proteins by comparing their synonymous codon usage, controlling for their amino acid content (Plotkin et al. 2004).

In this paper, we provide a theoretical analysis of codon usage biases that result from selection at the amino acid level. Our analysis, based on the Fisher-Wright model of population genetics, provides a theoretical grounding for techniques of estimating selection pressures on proteins using signals gathered from their synonymous codon usage. Throughout most of this paper, we will ignore any source of direct selection on synonymous codons and focus, instead, on codon biases that result from selection at the amino acid level. To the extent that any other sources of codon bias apply equally across the genome, we have devised a randomization method to control for these external sources of codon bias when estimating relative selection pressures on proteins (Plotkin et al. (2004). In the discussion, we describe a range of confounding factors that may vary across the genome and limit the applicability of methods to detect selection from synonymous codon usage. In Supplementary Information we respond to criticisms of volatility.

Codon Volatility

Codon usage biases can arise from selection on proteins because synonymous codons may differ in their *volatility*—defined, loosely, as the proportion of a codon's point mutations that result in an amino acid substitution. Although there are several possible definitions of volatility, which can all be informative, we have recently used the following formal definition (Plotkin et al. 2004).

We index the 61 sense codons in an arbitrary order $i = 1,...,61$. We use the notation aa(i) to denote the amino acid encoded by codon i. For each codon i, let $B(i)$ denote the set of sense codons that differ from codon i by a single point mutation. We define the volatility of codon i by

$$v(i) = \frac{1}{\#B(i)} \sum_{j \in B(i)} D[\mathrm{aa}(i),\, \mathrm{aa}(i)] \qquad (1)$$

where D denotes the Hamming metric, which is zero if two amino acids are identical, and one otherwise. The definition in Eq. 1 applies when all nucleotide mutations occur at the same rate. When differential nucleotide mutation rates are known (e.g., a transition/transversion bias), these rates can be incorporated into the definition of codon volatility by appropriately weighting the ancestor codons (Plotkin et al. 2004).

Minor variants of Eq. 1 yield related definitions of codon volatility. For some applications, one may want to allow termination codons in the definition of $B(i)$. It is also natural to consider alternatives to the Hamming metric, D, that weight substitutions between amino acids depending upon the differences in their stereochemical properties (Miyata et al. 1979; Plotkin and Dushoff 2003). A variety of other metrics (Tang et al. 2004; Yampolsky and Stoltzfus 2004) that reflect the effects of different amino acid substitutions on protein structure may likewise be incorporated into the definition of codon volatility. In this paper, however, we will focus on the most basic definition of codon volatility (Eq. 1, using the Hamming metric), because variant definitions are based on the same underlying principle and produce similar results in practice (Plotkin and Dushoff 2003).

The volatility of a codon is strongly influenced by the amino acid which it encodes. Therefore, in order to estimate the relative selection pressures on proteins across a genome, we have introduced a simple randomization technique to produce "volatility p-values" that control for the amino acid sequence of each gene (Plotkin et al. 2004). Under the most basic definition of volatility, there are four amino acids (glycine, leucine, arginine, and serine) whose synonymous codons differ in their volatility. As a result, when controlling for amino acid content, we will obtain a volatility signal from only those sites that contain one of these four amino acids—which amounts to roughly 30% of the sites in a typical gene. (If one uses stereochemical metrics for D in the definition of volatility ~75% of the sites in a gene contain a volatility signal). Although 30% may seem like a small proportion of sites from which to obtain a signal of selective pressures, it is larger than the proportion of sites often used to detect selection via sequence comparison of recently diverged species (Fleischmann et al. 2002; Clark 2003). For example, fewer than 4% of neutral sites exhibit substitutions when comparing human and chimpanzee sequences (Clark et al. 2003).

In the following sections we analyze the consequences of selection on proteins for codon usage in general, as well as for the volatility measure in particular. Under the Fisher-Wright model, we demonstrate that the expected codon usage at a site, as well as its temporal dynamics, depends upon the strength of positive or negative selection on the amino acid sequence. We first analyze the dynamics of negative selection in infinite and finite populations, and then we analyze positive selection. Our analysis is usually confined to the patterns of codon usage at a single site under selection at the amino acid level. However, we also discuss codon usage over many sites within a gene or genome, and we analyze how many sites are required in principle to detect a reliable signal of selection by inspecting synonymous codon usage.

Results

Negative Selection in an Infinite Population

Most nonsynonymous mutations in a protein coding sequence presumably reduce the fitness of an organism. For a large proportion of sites, therefore, natural selection opposes any change in the amino acid. We refer to this type of selection as *negative selection*. For the purposes of exploring the effect of negative selection on codon usage, we assume that selection cannot directly discriminate between the synonymous codons for the favored amino acid at a site. However, mutations are more likely to be nonsynonymous, and hence deleterious, if the codon at that site has high volatility. As we will show, this fact results in an effective preference for the less volatile codons, among those codons that code for the favored amino acid at the site. We emphasize that this preference for a codon of low volatility at a site under negative selection is *not* caused by a direct fitness difference between synonyms. Rather, more volatile codons will occur less frequently as a second-order consequence of negative selection at the amino acid level, and the structure of the genetic code.

Proteins containing a larger number of sites under negative selection will exhibit a statistical bias towards less volatile codons, after controlling for their amino acid content. In this section we calculate the expected magnitude of the codon bias as a function of the mutation rate, the strength of negative selection, and the population size. We also analyze the conditions under which we can expect to detect and draw inferences from this bias, and we estimate the time scales needed to establish and maintain such a signal.

A simplified genetic code

In an infinite population, we can describe the dynamics of codon usage at an individual site by using the standard multi-allele model, based on the work of Haldane and used throughout the literature (e.g., Nagylaki 1992, Eq. 2.25; Higgs 1994). This model describes a single site which can assume any of K states. In order to investigate codon usage, we consider $K = 64$ states, corresponding to each of the 64 possible codons. In continuous time, the proportion x_i of individuals with codon i evolves according to

$$\frac{dx_i}{dt} = \sum_{j=1}^{K} x_j(t) w_j M_{ij} - x_i W(t) \qquad (2)$$

where w_j is the Malthusian fitness of codon j, $W(t) \equiv \sum_j w_j x_j(t)$ is the mean fitness of the population, and M_{ij} is probability that codon j will produce codon i upon replication, with $\sum_i M_{ij} = 1$. Although Eq. 2 is non-linear, the equilibrium frequencies of the codons $i = 1,2,...K$ are given by the principal right eigenvector of the matrix $w_j\ M_{ij}$ (Thompson and McBride 1973). These frequencies determine the expected equilibrium codon usage at a site. A very similar model, and method of solution, has been used to study the evolution of mutational robustness (van Nimwegen et al. 1999; Wilke 2001). For the purposes of this paper, alternative formulations of the K-allele model that treat the processes of selection and mutation separately (e.g, Crow and Kimura 1970, Eq. 6.4.1) yield the exact same results. Our modeling framework is similar to that of Golding and Strobeck (1982), who also formulated a mutation-selection model for the full genetic code and pointed out that mutations to deleterious acids will have an effect on synonymous codon usage.

The equilibrium solution to Eq. 2 for the real genetic code does not lend itself to intuitive understanding. Transient dynamics are also difficult to calculate in this high-dimensional system. Therefore, in order to highlight the essential points of our analysis, we first consider a "toy" genetic code that retains those features of the true genetic code relevant to the study of synonymous codon usage under negative selection. As we will demonstrate, the solution for the simplified genetic code yields a complete understanding for the full genetic code as well.

We imagine a simplified genetic system with only three possible codons, a_1, a_2, and b. Codons a_1 and a_2 code for amino acid A, which is favored, and codon b encodes amino acid B, which has selective disadvantage σ. We assume that mutations occur at rate u between these codons according to the structure $a_1 \leftrightarrows a_2 \leftrightarrows b$ so that of the two synonymous codons, a_2 is more volatile.

According to the standard multi-allele model (Eq. 2), the relative frequencies of codons a_1, a_2 and b are described by the equation

$$\frac{d}{dt}\begin{pmatrix} a_1(t) \\ a_2(t) \\ b(t) \end{pmatrix} = \begin{pmatrix} 1-u & u & 0 \\ u & 1-2u & u(1-\sigma) \\ 0 & u & (1-u)(1-\sigma) \end{pmatrix} \times \begin{pmatrix} a_1(t) \\ a_2(t) \\ b(t) \end{pmatrix} - W(t)\begin{pmatrix} a_1(t) \\ a_2(t) \\ b(t) \end{pmatrix} \qquad (3)$$

where $W(t) = a_1(\mathrm{t}) + a_2(\mathrm{t}) + (1-\sigma)b(t)$.

The equilibrium frequencies of codons are given by the principal right eigenvector of the matrix in Eq. 3. A simple perturbation analysis of this eigenvector shows that the equilibrium proportion of a_1 depends monotonically on σ, and it exhibits a sharp transition between two regimes: the weak selection regime $\sigma \ll u$ and the strong selection regime $\sigma \gg u$. In the weak selection regime, the equilibrium relative frequencies of synonyms are given by the expansion

$$\frac{\hat{a}_1}{\hat{a}_1+\hat{a}_2}=\frac{1}{2}+\frac{1}{12}\frac{\sigma}{u}+O\left(\frac{\sigma^2}{u^2}\right) \tag{4}$$

And in the strong selection regime, the equilibrium relative frequencies are given by

$$\frac{\hat{a}_1}{\hat{a}_1+\hat{a}_2}=\frac{\sqrt{5}-1}{2}-\frac{(5-2\sqrt{5})(1-\sigma)}{5}\frac{u}{\sigma}+O\left(\frac{u^2}{\sigma^2}\right) \tag{5}$$

In the absence of selection ($\sigma = 0$) all three codons occur with equal frequency, as we would expect. In particular, the relative proportion of the two synonymous codons a_1 and a_2 equals $\frac{1}{2}$, regardless of the mutation rate. For weak selection ($\sigma \ll u$), this result is still approximately true, according to the perturbation expansion above. In the case of strong negative selection ($\sigma \gg u$), the relative proportion of the two synonymous codons is given approximately by the reciprocal of the golden mean $(\sqrt{5}-1)/2 \approx 0.62$.

We note that the results above, in the limits of strong and weak selection, are fairly straightforward when formulated in the language of mutational robustness: codons coding for the favored amino acid (under strong selection) or for any amino acid (under weak selection) form a neutral network, and the steady state concentrations are given, to first order, by the principal eigenvector of the connection matrix of this network (van Nimwegen et al. 1999; Wilke 2001) whose perturbation expansions we have given above.

The sharp transition between the weak and the strong selection regimes defines $\sigma = u$ as a critical value for negative selection. For $\sigma \ll u$ negative selection is ineffective at favoring the less volatile codon, and the site is effectively neutral. But when $\sigma \gg u$, negative selection favors the less volatile codon, and the magnitude of this effect depends only weakly on the value of σ. This is an essential point. In the strong selection regime, the magnitude of negative selection is relatively unimportant; volatile codons are disfavored at all sites where $\sigma \gg u$. The transition between the weak and strong selection regimes is shown in Fig. 1.

The effective disadvantage of a volatile codon

The critical value of σ discussed above can be understood intuitively by considering the effective selective disadvantage of the more volatile codon a_2 that results indirectly from its volatility. We will use the notion of an effective selective disadvantage to aid in our analysis of codon usage at a site under negative selection. But we emphasize that our model (Eq. 2) does not assume any direct fitness difference between synonymous codons.

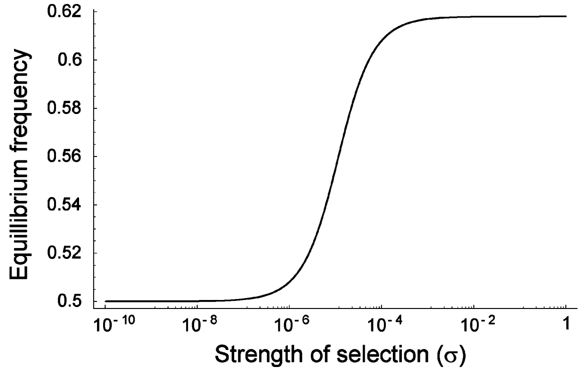

Fig. 1. The relationship between selection at the amino acid level and resulting synonymous codon usage. The graph shows relative equilibrium frequency of synonymous codons, $\hat{a}_1/(\hat{a}_1+\hat{a}_2)$ as a function of the strength of negative selection, σ. The relative frequency of codon a_1 is approximately ½ in the weak selection regime ($\sigma \ll u$) and approximately$(\sqrt{5}-1)/2$ in the strong selection regime ($\sigma \gg u$). In this figure $u = 10^{-5}$.

When the disfavored amino acid B is lethal to the organism, then the effective selective disadvantage of codon a_2 is particularly simple to understand. In this case, individuals with codon a_2 are removed from the population at rate u because they mutate to the lethal codon b, but receive no back-mutations. Hence the effective selective disadvantage, denoted s, of codon a_2 versus codon a_1 is given by $s = u$. The effective selective disadvantage of a_2 does not arise from a fitness difference between synonyms but, rather, from selection at the level of amino acids and the structure of the genetic code.

When amino acid B is not lethal the situation is slightly more complicated. Nevertheless, for $\sigma \gg u$, mutations from a_2 to b typically die due to negative selection before they mutate back from b to a_2. As a result, the effective selective disadvantage will still be $s = u$ in the regime of strong selection. We can make this argument concrete by considering the mutation-selection balance between codon b and codon a_2. According to the standard mutation-selection balance, the equilibrium proportion of codon b relative to codon a_2 equals u/σ in the regime $\sigma \gg u$. Thus for each mutant from a_2 to b, there are at most of order u/σ mutations from b to a_2. The net mutation rate from a_2 to b is therefore $u(1 - (u/\sigma))$. This is the rate at which individuals of type a_2 are lost from the population due to the fact that a_2 is more volatile than a_1. Thus the effective selective disadvantage of codon a_2 relative to a_1 is given by $s = u\ (1 - (u/\sigma))$. In the strong selection regime we neglect u/σ compared to 1, and the effective selective disadvantage of codon a_2 is simply $s = u$.

A similar argument holds for the real genetic code. In this case, the favored amino acid may correspond to several synonymous codons, each with a potentially different volatility. However, the effective selective disadvantage, s, of a more volatile codon relative to a less volatile synonym is simply the difference in the number of mutations leading to a disfavored amino acid (($\sigma \gg u$) times $u/3$, where u is the nucleotide mutation rate. (Note that $u/3$ is the rate of mutation between any two particular nucleotides.) For example, when considering the relative frequencies of codons AGA and CGG at a site under negative selection for arginine, AGA has selective disadvantage $s = 2u/3$ compared to CGG, since AGA has two more disfavored neighbors than CGG.

An analogous argument can be used to calculate the effective selective disadvantage of codon a_2 in the regime of weak selection ($\sigma \ll u$). In this regime, the relative equilibrium proportion of codon b versus codon a_2 equals $\frac{1}{2} - \sigma/(8u)$. As a result, the effective selective disadvantage of a_2 versus a_1 is approximately $s = 0$, plus a small correction of order σ. In other words, when $\sigma \ll u$ selection between a_1 and a_2 is effectively neutral; it cannot generate codon bias. We therefore refer to the regime $\sigma \ll u$ as the "almost-neutral regime." This result holds both for the simplified three-codon model and for the real genetic code.

It is also important to calculate the amount of time required to reach equilibrium codon usage in the presence of strong negative selection. Explicit solution of Eq. 3, assuming $\sigma \gg u$, indicates that the e-fold relaxation time is of order $1/u$ (the selection coefficient is s $\sim u$, and so the time scale for population sizes to change under selection is of order $1/s \sim 1/u$). In other words, starting from any initial frequencies $a_1(0)$ and $a_2(0)$, these frequencies will become e-fold closer to their equilibrium values after a duration of order $1/u$ generations. The same time scale holds for almost-neutral sites ($\sigma \ll u$) and for the real genetic code. In practice, u will be quite small, and equilibrium volatility is approached very slowly. We will revisit this point when we discuss finite populations and, again, when we discuss positive selection.

A specific example of negative selection

In this section we consider a simple example that demonstrates how our analysis applies to the real genetic code. We use Eq. 2 to model the dynamics of $K = 64$ alleles corresponding to the 64 codons, indexed in an arbitrary order. For our example, we consider a single site under negative selection for an arginine codon. In this case we define

$$M_{ij} = \begin{cases} 1 - 3u & \text{if } i = j \\ u/3 & \text{if } i \text{ and } j \text{ differ by a point mutation} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where u is the nucleotide mutation rate. We define

$$w_i = \begin{cases} 1 & \text{if } i \text{ encodes arginine} \\ 1 - \sigma & \text{if } i \text{ encodes a non-arginine amino acid} \\ 1 - \gamma & \text{if } i \text{ encodes stop} \end{cases} \tag{7}$$

so that a codon encoding an amino acid other than arginine has fitness $1 - \sigma$, and a termination codon has fitness $1 - \gamma$. We analyze this model numerically by calculating the principal right eigenvector of the matrix $w_j M_{ij}$, which yields the equilibrium proportions of all 64 codons.

In the case of no selection ($\sigma = \gamma = 0$), we find that all codons occur with the same equilibrium frequency, independent of mutation rate, as we would expect. For almost-neutral selection ($\sigma \sim \gamma \ll u$), codon usage is still approximately uniform. In the opposite case when arginine is favored and all other amino acids (or termination codons) are strongly disfavored (i.e., $\sigma \sim \gamma \gg u$), the arginine codons CGA, CGG, CGC, CGT, AGA, and AGG occur with equilibrium relative proportions $\approx$ 0.214:0.214:0.191:0.191:0.095:0.095. As expected, under negative selection the more volatile arginine codons occur with lower relative frequency in equilibrium.

The equilibrium frequencies of arginine codons determine the expected volatility at a single arginine site under negative selection. Assuming free recombination (Sawyer and Hartl 1992), an individual gene consists of many such sites randomly assembled; the mean and standard deviation in the volatility (per site) of a randomly sampled gene are shown in Fig. 2, as a function of the strength of a negative selection σ. Note that the stronger the negative selection, the lower the expected equilibrium volatility. The expected volatility exhibits a sharp transition from high to low values when the strength of negative selection σ reaches the mutation rate u, as discussed above. On either side of this transition, the volatility is insensitive to σ. The standard deviations plotted in Fig. 2 correspond to a gene comprised of $L = 200$ arginine sites, each modeled independently by the multi-allele equation.

According to Fig. 2, $L = 200$ independent arginine sites that each experience neutrality ($\sigma \ll u$) can be distinguished on the basis of their volatility from $L = 200$ sites that experience negative selection ($\sigma \gg u$). The difference in the expected volatility between these two regimes is greater than four standard deviations of the volatility within either regime. Similar results hold for serine and leucine, but glycine is less informative (Table 1).

In reality, the selective constraint σ will vary greatly across the sites of a given protein. In this case,

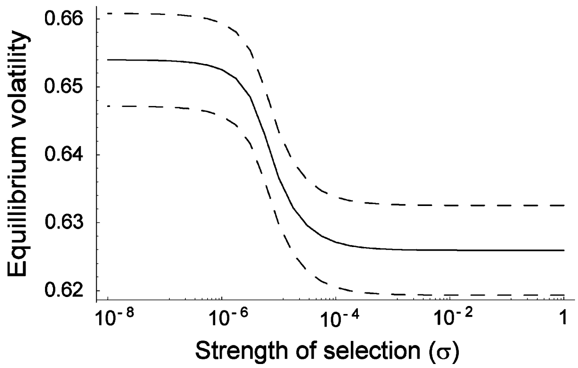

Fig. 2. The relationship between selective strength and volatility for a gene comprised of $L = 200$ freely recombining sites under selection for arginine. The graph shows expected volatility per site in the gene (±1 standard deviation, dashed) as a function of the strength of negative selection, σ. The nucleotide mutation rate is $u = 10^{-5}$. The expected volatility is significantly depressed in the regime of strong negative selection, $\sigma \gg u$. (For this figure we assume $\gamma = 1$; virtually identical results hold for $\gamma = \sigma$).

Table 1. Equilibrium codon usage under neutrality versus negative selection

	Neutral	Neutral*	Negative	v
Leucine				
cta	0.16667	0.17300	0.21353	5/9
ctc	0.16667	0.18580	0.19098	6/9
ctg	0.16667	0.17890	0.21353	5/9
ctt	0.16667	0.18580	0.19098	6/9
tta	0.16667	0.12990	0.09549	5/7
ttg	0.16667	0.14650	0.09549	6/8
$\mathbb{E}[v]$	0.65146	0.64590	0.63172	
$\sigma[v]$	0.07362	0.07259	0.07022	
Arginine				
aga	0.16667	0.15210	0.09549	6/8
agg	0.16667	0.17050	0.09549	7/9
cga	0.16667	0.15210	0.21353	4/8
cgc	0.16667	0.17740	0.19098	6/9
cgg	0.16667	0.17050	0.21353	5/9
cgt	0.16667	0.17740	0.19098	6/9
$\mathbb{E}[v]$	0.65278	0.65400	0.62592	
$\sigma[v]$	0.09854	0.09660	0.09354	
Serine				
agc	0.16667	0.18510	0.00000	8/9
agt	0.16667	0.18510	0.00000	8/9
tca	0.16667	0.13440	0.25000	4/7
tcc	0.16667	0.17190	0.25000	6/9
tcg	0.16667	0.15162	0.25000	5/8
tct	0.16667	0.17190	0.25000	6/9
$\mathbb{E}[v]$	0.71792	0.72981	0.63243	
$\sigma[v]$	0.12504	0.12561	0.03913	
Glycine				
gga	0.25000	0.22460	0.25000	5/8
ggc	0.25000	0.26180	0.25000	6/9
ggg	0.25000	0.25170	0.25000	6/9
ggt	0.25000	0.26180	0.25000	6/9
$\mathbb{E}[v]$	0.65625	0.65724	0.65625	
$\sigma[v]$	0.01804	0.01859	0.01804	

Note. In each regime, we report the equilibrium relative abundance of codons, and the resulting mean and standard deviation in volatility per site. The second column corresponds to neutrality ($\sigma = \gamma \ll u$): the third column corresponds to neutrality but with disfavored termination codons ($\sigma \ll u$, $\gamma = 1$); the fourth column corresponds to negative selection in an infinite population ($\sigma \gg u$, $\gamma \gg u$). The final column gives the volatility of each codon, assuming no transition/transversion bias (Plotkin et al. 2004).

disregarding the possibility of positive selection, the volatility of a gene (after controlling for its amino acid sequence) essentially reflects the relative number of informative sites that experience negative selection versus neutrality. For example, the volatility of gene X that contains $L = 200$ informative sites under negative selection and an equal number of neutral sites will be significantly greater (with a Z-score of about three) than the volatility of gene Y that consists of 2L informative sites all under negative selection. A more thorough discussion of variable selection pressures across genes is described in a subsequent section (Eq. 18), below.

Table 1 shows the equilibrium relative frequencies of synonymous codons for each of the informative amino acids (G, L, R, and S) under neutrality versus negative selection. In Table 1 we assume, as we do throughout this paper, that volatility is measured using the Hamming metric and that there is no transition/transversion bias. Corresponding values for different metrics or including a mutational bias may be calculated using the same approach. As shown in Table 1, the difference in the expected volatility between selective regimes is least extreme (indeed, barely informative) for glycine sites. The volatility difference is most extreme for serine sites: the highly volatile codons AGT and AGC are not expected to occur at a site under negative selection, but they are expected occur at a site under neutrality or, especially, under positive selection (see below). This extreme case results from the fact that codons AGT and AGC are not connected by synonymous point mutations to the other serine codons. As a result, the relaxation time to equilibrium for codons AGT and AGC is longer than $1/u$ generations, and these codons may be less informative for negative selection because they take too long to be effectively selected against. This situation does not necessarily imply that codons AGT and AGC should be treated separately from the other serine codons. In fact, when treated as an entire group, the serine codons are particularly informative for discriminating between positive selection and neutrality (Table 2).

Negative Selection in a Finite Population

The models presented in the previous section describe the processes of mutation and negative selection in an infinite population. In finite populations, however, genetic drift also affects allelic frequencies. In this section, we study the combined effects of mutation,

Table 2. The mean and standard deviation of volatility at a site under various models of positive selection

	Pos[1]	Pos[2]	$\theta = 1$	$\theta = 0.1$	$\theta = 0.01$
Arg					
$\mathbb{E}\,[v]$	0.6683	0.6607	0.6684	0.6683	0.6675
$\sigma\,[v]$	0.0941	0.0927	0.0819	0.0889	0.0896
Leu					
$\mathbb{E}\,[v]$	0.6554	0.6484	0.6541	0.6532	0.6538
$\sigma\,[v]$	0.0722	0.0698	0.0636	0.0691	0.0684
Ser					
$\mathbb{E}\,[v]$	0.7468	0.7337	0.7372	0.7396	0.7422
$\sigma\,[v]$	0.1271	0.1268	0.1275	0.1301	0.1301
Gly					
$\mathbb{E}\,[v]$	0.6576	0.6576	0.6576	0.6576	0.6576
$\sigma\,[v]$	0.0172	0.0172	0.0163	0.0171	0.0171

Columns 2 and 3 summarize the expected volatility after a positively selected sweep according to heuristics 1 and 2 described in the main text. Columns 4 through 6 report simulations of the full Fisher-Wright model in a population of $N = 1000$ individuals. For each simulated selective sweep, the population is initially fixed for a random, disfavored codon. Codons for the favored amino acid are assigned fitness 1, and all others fitness $1 - \sigma$ (here $\sigma = 0.1$ but results do not depend on σ provided $\sigma \gg 1/N$ and $\sigma \gg u$). After the favored amino acid has swept the population (proportion $> 99\%$), the average volatility at the site is recorded. Results reported for each amino acid and θ value represent the average and standard deviation obtained in >2000 independent runs. As discussed in the main text, the appropriate effective population size that determines the value of θ does not necessarily equal the average neutral site heterozygosity.

negative selection, and drift, which we analyze using diffusion equations. These equations can be very complex. A full treatment of even the simplified three-codon genetic code requires a two-dimensional diffusion process, and the real genetic code involves a 63-dimensional process. To make this problem tractable, we use the notion of the "effective selective disadvantage" of more volatile codons, discussed above. This allows us to consider the dynamics only at the favored codons, thereby reducing the dimensionality of the diffusion process.

The neutral ($\sigma = 0$) or almost-neutral ($\sigma \ll u$) regimes are straightforward: here all synonymous codons for the favored amino acid have the same effective fitness. In this regime, each synonymous codon occurs with the same probability in steady state, independent of population size.

For the remainder of this section, we analyze the case of strong negative selection ($\sigma \gg u$) at a single site. We assume that $N\sigma \gg 1$, so that selection at the amino acid level is effective. (This assumption is not restrictive, because when $N\sigma \lesssim 1$, then either $\theta = 2Nu$ or σ will be too small to support a volatility signature of selection.) We consider a diffusion approximation to the process of mutation, selection, and drift operating only on the synonymous codons, to each of which we assign an effective selective coefficient. For the simplified three-codon genetic system, the more volatile codon a$_2$ has an effective selective disadvantage of $s = u$ compared to codon a_1. For the real genetic code, more volatile codons will have a selective disadvantage of this order, but the precise value of s will depend on the specific codon in question. In the following analysis, we consider the case of the simplified three-codon system. However, we do not explicitly make the substitution $s = u$, so that our results can also be applied (with a slightly different value of s) to the real genetic code.

The time-dependent proportion $f(x,t)$ of allele a_1 relative to allele a_2 can be described in the diffusion limit of the Fisher-Wright model by the Komolgorov forward equation (Kimura and Crow 1964),

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\{a(x)f(x,t)\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{b(x)f(x,t)\} \tag{8}$$

where the instantaneous mean and variance in the change of allelic frequency are given by

$$a(x) = sx(1-x) - ux + u(1-x)$$
$$b(x) = x(1-x)/N$$

The stationary distribution of allele frequencies $\hat{f}(x)$ satisfies the equation

$$\frac{\partial}{\partial x}\{b(x)\hat{f}(x)\} = 2a(x)\hat{f}(x) \tag{9}$$

which has the solution (Wright 1931)

$$\hat{f}(x) = Cx^{\theta-1}(1-x)^{\theta-1}e^{Sx} \tag{10}$$

where $\theta = 2Nu$, $S = 2Ns$, and C is chosen so that $\int_0^1 \hat{f}(x)dx = 1$. Since $s \sim u$ (and thus $S \sim \theta$), the shape of the stationary distribution $\hat{f}(x)$ falls into two categories: a bell-shaped distribution in the regime $\theta > 1$ and a U-shaped distribution in the regime $\theta < 1$. In other words, for $\theta > 1$ the steady-state population is typically polymorphic at the locus, much like the infinite population mutation-selection balance. In contrast, for $\theta < 1$ the steady-state population is usually near-monomorphic at the locus, occasionally switching between alleles a_1 and a_2, with a bias (whose strength is determined by S) toward allele a$_1$.

In stationary state, the expected proportion of allele a_1 is given by

$$M(\theta, S) = \int_0^1 x\hat{f}(x)dx = \frac{1}{2} + \frac{\mathscr{B}(\theta + 1/2, S/2)}{2\mathscr{B}(\theta - 1/2, S/2)} \quad (11)$$

where $\mathscr{B}(x, y)$ is the modified Bessel function of the first kind. Similarly, the variance in the frequency of allele a_1 is given by

$$V(\theta, S) = \int_0^1 x^2\hat{f}(x)dx - M(\theta)^2 \quad (12)$$

$$\text{where, } V(\theta, S) = \frac{1}{4 + 8\theta} + \{[2\theta\mathscr{B}(\theta - 1/2, S/2)\mathscr{B}(\theta + 3/2, S/2) - (1 + 2\theta)\mathscr{B}(\theta + 1/2, S/2)^2]\} \div \{[(4 + 8\theta)\mathscr{B}(\theta - 1/2, S/2)^2]\} \quad (13)$$

We use the standard Taylor series expansion of $\mathscr{B}(x, y)$ to obtain a simple approximation for the mean stationary frequency of allele a_1:

$$M(\theta, S) = \frac{1}{2} + \frac{S}{4} + O(\theta^2) \quad (14)$$

valid for $\theta \sim S \ll 1$. Thus the difference in expected volatility at a site under neutral versus negative selection is of order S, when $\theta \ll 1$.

When $\theta = S = 1$, the mean stationary frequency of allele a_1 assumes the value 1/(e −1) ≈ 0.58. For $\theta \sim$ S $\gg 1$, the mean frequency quickly approaches the asymptotic value $\lim_{\theta \to \infty} M\ (\theta,\ \theta) = (\sqrt{5} - 1)/2$, in agreement with our earlier result for an infinite population.

The results in this section generalize our analysis of an infinite population. For an infinite population, we found that the expected relative frequency of codon a_1 equals 1/2 in the almost neutral regime, and it equals $(\sqrt{5} - 1)/2$ in the strong selection regime. In a finite population with $\theta \gg 1$, the same results hold. In a finite population with $\theta \ll 1$, the expected relative frequency of the more volatile codon equals $\frac{1}{2}$ in the neutral regime, and it equals 1/2 + (Ns/2) in the strong selection regime. For any population size, the relative frequency of codon a_1 depends monotonically on the strength of selection at the amino acid level σ (Eq. 11), and it exhibits a sharp transition at the critical value $\sigma = u$. Note that the effective selective disadvantage s of codon a_2 versus a_1 is due solely to the lower fitness of amino acid B, and thus it depends monotonically on σ.

It is worth noting that our exact expression (Eq. 11) for the mean stationary frequency of allele a_1 generalizes earlier work by Bulmer (1991) on the relative frequency of two synonymous codons that experience a direct fitness difference. In the limit of small θ, we find that

$$\lim_{\theta \to 0} M(\theta, S) = \frac{1}{2} + \frac{\mathscr{B}(1/2, S/2)}{2\mathscr{B}(-1/2, S/2)} = \frac{1}{1 + e^{-S}} \quad (15)$$

which agrees with Bulmer's result (his Eq. 6). In other words, Bulmer's approximation applies only in the limit of small mutation rates or population sizes. Our results also agree with prior work on mutational robustness: for $\theta \gg 1$, the population assumes the infinite-population solution, and for $\theta \ll 1$ all nodes are populated with equal frequencies (van Nimwegen et al. 1999; Wilke 2001). Equation 11 extends earlier work on mutational robustness by providing an analytic solution for all intermediate values of θ.

We can again use the standard Taylor expansion of the Bessel function to obtain a simple expression for the variance in the stationary frequency of allele a_1,

$$V(\theta, S) \approx \frac{(3 + 2\theta)(4 + 8\theta) - 3S^2}{16(3 + 2\theta)(1 + 2\theta)^2} \quad (16)$$

which is a highly accurate approximation for all θ, provided as usual that S is of order θ or smaller. Note that when $\theta \ll 1$ the variance is approximated by $\frac{1}{4} - (\theta/2)$, and when $\theta \gg 1$ the variance is of order $1/\theta$.

Inferring negative selection in a finite population

Our exact (Eq. 11) or approximate (Eq. 14) expressions for the stationary mean frequency of codon a_1 allow us to determine the minimum number of sites required for codon volatility to distinguish between neutrality and negative selection. When sites are modeled independently (equivalent to the assumption of linkage equilibrium [Sawyer and Hartl 1992]), under neutrality ($\sigma \ll u$; $s = 0$) the relative frequency of codon a_1 versus codon a_2 across a gene of length L is binomially distributed with mean $\frac{1}{2}$ and variance 1/4L. If, on the other hand, the gene experiences negative selection ($\sigma \gg u$; $s = u$), then the relative frequency of codon a_1 is binomially distributed, with mean $M(\theta, S)$ and variance $M(\theta, S)[1 - M(\theta, S)]/L$. Therefore, in order to reject neutrality at the 95% confidence level with 50% power, we require

$$M(\theta, S) - \frac{1}{2} > 1.96\sqrt{\frac{1/4 + M(\theta, S)[1 - M(\theta, S)]}{L}} \quad (17)$$

Using this equation, Fig. 3 shows the minimum number of sites required to distinguish between negative selection and neutrality on the basis of codon volatility, under our simplified "genetic code" consisting of three codons.

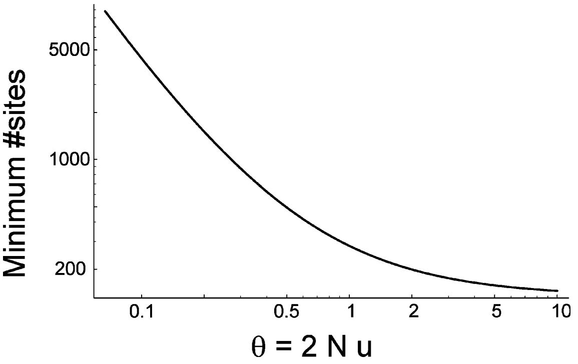

Fig. 3. The relationship between the scaled population size, $\theta = 2Nu$, and the minimum number of sites required to distinguish negative selection from neutrality, at the 95% confidence level. Sites are assumed to be unlinked. It is important to note that the appropriate effective population size that determines the value of θ does not necessarily equal the average neutral site heterozygosity (see text).

Under the real genetic code, about one-third of the sites in a typical gene are informative, to varying degrees. Thus, the y-axis in Fig. 3 would typically be increased by roughly a factor of three, depending upon the amino acid content. Therefore, within an order of magnitude, Fig. 3 reflects the power of the volatility technique for detecting negative selection. Figure 3 indicates that, according to the Fisher-Wright model, synonymous codon usage will not provide a robust signal of negative selection on individual genes except for micro-organisms that experience relatively large effective population sizes (or mutator episodes, see below).

Equation 17 applies when comparing a collection of neutral sites against a collection of sites under negative selection. In most situations, however, the selective constraint σ will vary across the sites of a protein. For example, consider gene X, with $L + J$ sites under negative selection, compared to gene Y, with L neutral sites and J sites under negative selection. In this case, the expected frequency of codon a_1 in gene Y is $(L/2 + JM(\theta, S))/(L + J)$. Therefore, in order to infer that gene X experiences more negative selection than gene Y, at the 95% confidence level we require

$$M(\theta, S) - \frac{L/2 + JM(\theta, S)}{(L + J)} > 1.96\sqrt{\frac{L/4 + (L + 2J)M(\theta, S)[1 - M(\theta, S)]}{(L + J)^2}} \tag{18}$$

As Eq. 18 indicates, the power to discriminate between two genes is decreased when both genes contain many sites under the same selective regimes and only a few sites under different selective regimes. Nevertheless, provided $J \sim$ L, the power to discriminate between gene X and gene Y is decreased by a factor of two at most (compared to $J = 0$), and so the minimum number of sites required to detect negative selection (Fig. 3) would change by a factor of two at most.

The results in this section were derived for a simplified genetic code, whereas the real genetic code would require analysis of multi-dimensional diffusion equations. The scaling behavior of the one-dimensional case is expected to hold in higher dimensions—i.e., for the real genetic code when comparing neutrality to negative selection, for $\theta \ll 1$ the expected difference in volatility per site will be of order θ, and for $\theta \gg 1$ the expected difference in volatility can be calculated from the infinite population model (Eq. 2 and Table 1).

Relaxation toward steady state

Although Eq. 11 predicts the steady-state relative frequencies of codons a_1 and a_2 in the selected regime ($\sigma \gg u$), we have not yet discussed how long it takes, on average, to reach this steady state. In the case of a very large population, $\theta \gg 1$, we know from the infinite population model (Eq. 2) that the e-fold relaxation time to equilibrium is of order $1/u$ generations. In this section, we demonstrate that the same result applies to the time scale of relaxation toward steady state in the regime $\theta \ll 1$.

Again, we consider a single site under negative selection. In the regime $\theta \ll 1$, we have seen that the steady-state population will spend most of the time in a nearly monomorphic state, with a preference (of order S) for the less volatile codon, a_1. Therefore, in order to calculate the time scale of relaxation toward steady state, we may simply calculate the amount of time required such that, starting with a population fixed for allele a_2, the probability of the population remaining fixed for allele a_2 has been reduced e-fold.

Given a population initially fixed for codon a_2, there are Nu mutations to codon a_1 generated per generation. Each of these mutations has an effective selective advantage $s = u$ over allele a_2 and will therefore fix with probability $2s/(1 - e^{-2\mathrm{Ns}})$ (Crow and Kimura 1970). Hence the rate of production of a mutation that will eventually fix is given by

$$P_{fix} = \frac{2Nus}{1 - e^{-2Ns}} \approx u, \tag{19}$$

assuming $\theta \ll 1$. According to this calculation, the mean time until fixation of codon a_1 is of order $1/u$ generations, which gives the time scale of relaxation to the steady-state codon usage in a finite population under negative selection.

About Population Sizes

As discussed above, the strength of the signal of negative selection depends upon the parameter θ. θ is twice the product of the effective population size and the (per-site) mutation rate: $\theta = 2Nu$. What is the appropriate value of θ in practice?

Unfortunately, this question is far easier asked than answered. Population geneticists have struggled to reconcile estimates of θ deduced from polymorphism data with direct measurements of N and u across broad taxonomic ranges. The effective population sizes of micro-organisms, in particular, are topics of active debate. Estimates of θ are usually obtained by comparing polymorphism data at neutral (or presumably neutral) sites against the expected site diversity or the expected number of segregating sites under a neutral model (Ewens 2004). The assumption of synonymous site neutrality will typically lead to underestimates of θ. In one of the few cases where neutrality was not assumed, θ was estimated as 0.18 per site in *E. coli* (Hartl and Sawyer 1994). In a recent survey where neutrality was assumed (Lynch and Conery 2003) the authors reported an average value of $\theta \approx 0.15$ among the prokaryotes studied. But estimates of θ for a microbial species can vary by four orders of magnitude, and they depend strongly on assumptions about population structure (Berg 1996). To complicate matters further, heterogeneity in mutation rates leads to substantial underestimates of θ (Tajima 1996).

Aside from uncertainty in its estimation, the value of θ deduced from neutral SNP data (Lynch and Conery 2003) may not be relevant to questions of selection and volatility. Monomorphism observed at neutral sites may result from non-neutral processes, such as background selection (Charlesworth et al. 1993) or hitchhiking on periodically sweeping sites (Maynard Smith and Haigh 1974). As a result, the variance effective population size estimated from SNP data may not be relevant to other aspects of evolution, such as substitutions at linked weakly selected sites (Gillespie 2001).

One particularly striking example of a discrepancy in the appropriate effective population sizes arises from the consideration of mutator phenotypes. Populations of microbial species periodically experience a transient increase in the mutation rate, usually 100–1000 times greater than that of a non-mutator strain (Bjedov et al. 2003). Between 2 and 36% of bacterial populations isolated in the wild at any given time exhibit a mutator phenotype (Giraud et al. 2001; Oliver et al. 2000; LeClerc et al. 1996). The mutator phase can be induced in several ways. A defective DNA repair gene (e.g., mismatch repair gene) may arise and sweep to fixation by hitchhiking on a positively selected mutation. The entire population then experiences an elevated mutation rate until a non-mutator allele sweeps and replaces the mutator (Notley-McRobb et al. 2001; Denamur et al. 2000). A second, perhaps more common, mechanism is stress-induced mutagenesis; natural isolates of *E. coli* often experience an increase in their mutation rate in response to stress (Bjedov et al. 2003). As a result of these and other observations, researchers have argued that bacterial populations evolve primarily by periodic acquisition of mutator phenotypes followed by adaptive sweeps and subsequent loss of the mutator (Giraud et al. 2001; Denamur et al. 2000; Notley-McRobb et al. 2001). For example, Denamur et al. (2000) concluded that "a significant fraction of genomic diversity of *E. coli* natural isolates was generated during the MMR-deficient intermezzos of their evolutionary past"; and Tenallion et al. (2004) concluded that "mutations produced by stress-induced mutator alleles can represent a large fraction of the mutations produced by a clone during its evolution." As we shall see, the effect of mutators on synonymous codon usage is dramatic: the expected site diversity is driven by the value of θ in the wild type regime ($\theta_w = 2Nu_w$) but the pattern of synonymous codon usage at a site under negative selection is driven by the value of θ in the mutator regime ($\theta_m = 2Nu_m \gg \theta_w$).

As a simple example of this phenomenon, we have simulated a Fisher-Wright model of a single locus in a population of constant size $N = 1000$. The simulated site is subject to recurrent mutation between "alleles" a_1 and a_2 at wildtype rate $u_w = 10^{-5}$. As above, the alleles a_1 and a_2 differ in fitness by s, where s equals the mutation rate. In each generation, we record the frequency, x, of allele a_1. Periodically, we model the fixation of a mutator allele (or, equivalently, the stress-induced mutagenesis across the entire population) by exogenously increasing the mutation rate to $u_{\mathrm{m}} = 10^3 \times u_w$ for 100 generations; thereafter we enforce a selective sweep at the site, followed by reversion to the wildtype mutation rate. Overall, the population experiences the mutator regime for 5% of the time, consistent with observed frequencies of mutator phenotypes in the wild (Giraud et al. 2001; Oliver et al. 2000; LeClerc et al. 1996). According to our simulations, the average site diversity, $2x(1 - x)$, at a randomly chosen time equals 0.028, which is close to its expected value assuming that θ is given by θ_w: $\mathrm{E}\ [2x(1 - x)] \approx \theta_w = 0.02$. But the average frequency of allele a_1 equals 0.611, which is close to its expectation assuming that θ is given by θ_m: $\mathrm{E}[x] = M(\theta_m, \theta_m) = 0.616$ (Eq. 11). In other words, the average frequency of the less volatile codon a_1 is dominated by the mutator periods, but the average site heterozygosity (and any estimate of θ based on it) is dominated by the non-mutator periods.

There is a simple, intuitive explanation for this result. The average heterozygosity at the site is low at virtually all times (except during the brief mutator periods) because selective sweeps cause monomorphism, followed by long periods of low θ. Therefore, the effective θ for SNP diversity is small, i.e., close to Nu_w. But the site converges quickly toward the less volatile codon during the mutator periods, since the rate of convergence is determined by $s = u_m$. And the site is essentially frozen during the non-mutator periods, since the decay rate of volatility is only u_w. Therefore the expected frequency of a_1 at a random time is primarily determined by the frequency reached during the mutator regime. As is clear from this explanation, the expected frequency of codon a_1 will, in general, depend on the stochastic scheduling of mutator periods. For example, the site will converge toward $M(\theta_m, \theta_m)$ provided the population experiences at least one mutator phase of duration of order $1/u_m$ generations, within every $1/u_w$ generations. In fact, even if the mutator phases are very brief and infrequent, the average frequency of allele a_1 can greatly exceed the value predicted by θ estimated from the neutral site heterozygosity. This result indicates that, for organisms that experience mutator periods, the value of θ relevant to volatility may be 100- to 1000-fold greater than the value of θ generally reported in the literature, because the latter value is usually estimated from neutral site heterozygosity.

Although the simple model used in this section does not describe any but the most phenomenological features of mutator alleles, it does reveal an important general observation: the value of θ estimated from presumably neutral polymorphism data does not in general equal the effective value of θ that determines synonymous codon usage at a site under negative selection. This result is of utmost importance to any discussion of the relationship between θ and the power of volatility to detect selection.

Positive Selection

In the sections above, we have considered selection that opposes a change to the amino acid at a site. Under the Fisher-Wright model, we have seen that negative selection induces a bias toward the less volatile codons for the favored amino acid at a site. However, selection sometimes favors a change in the amino acid at a particular site. As we will demonstrate, after a positively selected sweep has occurred, a site is more likely to be occupied by a codon of greater than average volatility (controlling for the amino acid now present at the site).

A variety of mechanisms are known to cause positive selection, such as exogenous changes in the environment or frequency dependent fitnesses. We do not here model all of the various types of positive selection but, rather, focus on the essential aspect shared by these mechanisms: pressure for an amino acid substitution at a site. One common model of positive selection (Goldman and Yang 1994; Yang et al. 2000) stipulates continual pressure to change the amino acid at any and all sites in a gene as quickly as possible: as soon as an amino acid substitution has occurred at a site, there is immediately pressure for another substitution. Over the time scales of our interest, we view this scenario as unlikely. Instead, we will study a more straightforward conception of positive selection: we analyze the dynamics at a single site that has, for a period of time, experienced selection for amino acid A, and that subsequently experiences selection for a different amino acid, B (for whatever reason). We refer to the change in the favored amino acid at the site as a positive selection event.

Prior to the onset of positive selection, amino acid A is assigned fitness 1 and all other amino acids fitness $1 - \sigma$; subsequently, amino acid B is assigned fitness 1 and all others fitness $1 - \sigma$. We assume that $N\sigma \gg 1$ (otherwise, the site is effectively neutral at the amino acid level) and that $\sigma \gg u$ (otherwise, the expected codon frequencies are uniform). At some point after the change in selective regimes, a codon for amino acid B will eventually arise and quickly sweep close to fixation. We are interested in calculating the relative probabilities of the various B codons after the positively selected sweep has occurred. Whatever B codon arises and fixes, we can be sure that it will have arisen through a non-synonymous mutation, since we have assumed that the population was initially dominated by amino acid $A \neq B$. Intuition would therefore suggest that a highly volatile B codon (which has more non-synonymous neighbors) is more likely to occur immediately after the sweep than a less volatile B codon (which as has fewer non-synonymous neighbors). In other words, intuition would suggest that the expected volatility at the B-site is elevated immediately following the selective sweep. Subsequent to the sweep, since B is now favored, negative selection will reduce the volatility at the site. However, this process takes time. Thus, for some time after the positive selection event, we expect to find a bias toward elevated volatility at the site, which gradually decays. In this section, we analyze this process under the Fisher-Wright model.

Positive selection under a simplified genetic code

Similar to previous sections, we initially consider a simplified genetic code consisting of four codons, a_1, a_2, b_1, and b_2, the first two of which encode amino acid A, and the latter two amino acid B. Mutations can only occur between codons a_1 and a_2, a_2 and b_2 and b_2 and b_1, creating the mutation structure

$$a_1 \leftrightarrows a_2 \leftrightarrows b_2 \leftrightarrows b_1$$

In this simplified genetic code, codons a_2 and b_2 are the more volatile codons for their respective amino acids.

After the change in selection from amino acid A to B, a mutation to codon b_2 that survives stochastic drift will eventually arise. Thus, at least initially, the more volatile codon b_2 is more prevalent than the less volatile codon b_1. During this period, we can detect the signature of the positively selected sweep because of the elevated volatility at the site. However, negative selection for amino acid B will eventually favor codon b_1. Therefore, the volatility signature of the positive selection event will be present provided that the time scale of decay toward codon b_1 is longer than the interval since the positive selection event.

Fortunately, the time scale of decay toward b_1 is quite long. For $\theta \gg 1$, we can use the infinite population model to find this time scale: As discussed above, the time required to reduce the volatility e-fold is of order $1/u$. For $\theta \ll 1$, we must use a finite population size calculation. In this regime, the population is nearly monomorphic at almost all times. Following the selective sweep, the site will be monomorphic for b_2 with almost unit probability. We are interested in the duration of time required such that the probability of being monomorphic for b_2 (as opposed to b_1) has been reduced e-fold. The probability of switching between b_2 and b_1, however, is of order u per unit time (even before b_2 has finished outcompeting a_2), according to Eq. 19. Thus, the time scale of decay in a finite population is also $1/u$.

According to this analysis, a selective sweep will result in the presence of a more volatile codon for order $1/u$ generations—a very long time indeed. (In the case of *E. coli*, for example, there are estimated only to be only 100–300 generations per year [Gibbons and Kapsimalis 1967; Ochman *et al.* 1999].) Equivalently, repeated sweeps for amino acid changes at a site will result in the presence of more volatile codons at almost all times, provided that new sweeps occur more often than every $1/u$ generations. It is important to recall, however, that an episode of elevated mutation rates following a selective sweep will induce a faster decay rate of elevated volatility.

Positive selection under the real genetic code

The analysis above for a simplified genetic code may be generalized to the real genetic code, for arbitrary amino acids A and B. We do not assume that A and B are separated by a single point mutation. After a change in selective regimes from A to B, the population will eventually acquire a B codon that survives drift and fixes in the population. As above, we are interested in the relative probabilities of each of the B codons—and thus the expected volatility at the site—immediately after the selective sweep has occurred. Under the Fisher-Wright model applied to the real genetic code, the relative probabilities of the B codons following the selective sweep are difficult to calculate analytically. Instead, we will introduce two heuristic approximations, and compare these approximations to simulations of the Fisher-Wright model.

Our first heuristic assumes that the relative probability of occurrence of a particular B codon after a selective sweep is proportional to the number of its non-synonymous non-termination neighbors, compared to the other B codons. For example, codon AGG (Arg) has 7 non-synonymous non-termination neighbors; whereas codons AGA, CGA, CGC, CGG, and CGT have 6, 4, 6, 5, and 6 such neighbors, respectively. Thus, the probability that AGG will dominate the population immediately after a selective sweep for arginine is given by $7/(7 + 6 + 4 + 6 + 5 + 6)$ according to heuristic 1. The intuition behind heuristic 1 is simple: we know that the mutation giving rise to the selective sweep was non-synonymous, and there are more non-synonymous routes into certain arginine codons than into others.

We also consider a second heuristic that makes a different simplifying assumption. Under heuristic 2, we assume that each of the possible non-synonymous codons within a point mutation of amino acid B is equally likely to have given rise to the B codon that eventually sweeps the population. In addition, if a potential ancestor codon has more than one way of mutating into B, we assume that each of these mutations was equally likely. For example, there are 27 non-termination, non-arginine codons within a point mutation of arginine. According to heuristic 2, we assume that each of these possible ancestor codons was equally likely to precede the arginine sweep. Thus the probability that codon AGA (Arg) will sweep is given by $1/27 \times (1 + 1 + 1 + 1 + 1/3 + 1/3)$, where the various terms represent the probabilities of a mutation to AGA from ATA, ACA, AAA, GGA, ACT, and AGC respectively. The fractions in this expression arise because codons AnA and GGA each have only one way of mutating to arginine, whereas codons AGT and AGC each have three ways, which we have assumed to be equally likely. Heuristic 2 can likewise be applied to calculate the probability of each arginine codon after the sweep, and thus the expected volatility at the site (Table 2).

The difference between the two heuristics is subtle, but important. Heuristic 1 assumes that each possible mutational route into the new amino acid is equally probable, whereas heuristic 2 assumes that each possible ancestor codon is equally probable. Neither heuristic yields a completely accurate description of the Fisher-Wright model. In a small population, however, we might expect heuristic 2 to be a more accurate approximation, because the population will usually be almost monomorphic before the sweep, as it drifts between non-B codons. As soon as a substantial number of individuals drift to a B-adjacent codon, however, one of the adjacent B codons will fix in the population long before the population drifts into other non-B codons. Thus each ancestral codon is equally probable. If a non-B ancestor has more B neighbors, this does not make it a more likely ancestor to the B codon: it simply chooses one of these paths with equal probability. On the other hand, in a large population we expect heuristic 1 to be a more accurate approximation of the Fisher-Wright model. In this case, the population will tend to be polymorphic for non-B codons before the sweep. All mutational routes into B will be used. Thus all possible mutations yielding a B codon are equally probable. If a non-B ancestor has more B neighbors, it will produce more mutations to B codons.

Neither heuristic, however, yields a completely accurate description of the expected volatility at a site after a positively selected sweep (see Table 2). Both heuristics neglect second- and higher-order effects that determine volatility after a sweep under the full Fisher-Wright model. For instance, if amino acid B is two mutations away from amino acid A, then the relative probabilities of various B codons after the sweep will depend not only on the number of their non-B neighbors but also on the number of their neighbors' non-B neighbors, and so on. Therefore, the heuristics are at best first-order approximations to the Fisher-Wright model.

Table 2 shows the expected volatility at a positively selected site as predicted by our two heuristics, compared to simulations of the full Fisher-Wright model. According to either heuristic and according to Fisher-Wright simulations over a range of θ values, the expected volatility at a site after a positively selected sweep is always greater than the expected volatility of a neutral or negatively selected site (Table 1). As discussed above, following a positively selected sweep, the elevated volatility at the site will subsequently decay on a time scale of order of $1/u$ generations.

There is an important distinction between the volatility signature of positive selection versus that of negative selection. The depressed volatility at a site under negative selection is caused by a mutation-selection-drift balance. When the effective population size is small, a large number of sites are required to distinguish negative selection from neutrality. By contrast, the volatility signature of *positive* selection is *not* an equilibrium property, and it is not as sensitive to population size. As shown in Table 2, the volatility of a site after a positively selected sweep depends only weakly on θ.

It is worth noting that the elevated volatility after a positively selected sweep for serine will decay even more slowly than for other amino acids, because the highly volatile codons AGC and AGT are not connected by synonymous mutations to other serine codons. A mutator period subsequent to a selective sweep will accelerate the decay toward lower volatility.

Inferring positive selection

As we have seen, a gene that contains many sites under positive selection will exhibit a greater volatility (controlling for its amino acid composition) than a gene under mostly neutral or, especially, negative selection. How many positively selected sites are required in order to detect a reliable signal? In the case of arginine, for example, at $\theta = 1$ the expected volatility of a site that has recently experienced a positively selected sweep is approximately 0.6684 ± 0.0819, whereas a neutral arginine site has expected volatility 0.6528 ± 0.0985, and an arginine site under negative selection has expected volatility 0.6404 ± 0.0693 (obtained by Fisher-Wright simulation at $\theta = 1$; cf. Table 1). Therefore, the volatility of 56 positively selected arginine sites will be significantly greater than that of 56 negatively selected sites, at the 95% confidence level. Similarly, the volatility of 260 arginine sites under positive selection will be significantly greater than that of 260 neutral sites. For $\theta = 0.1$, 190 sites are required to reliably distinguish positive selection from negative selection; and 280 sites are required to distinguish positive selection from neutrality. Similar results hold for leucine and serine; glycine is less informative.

Discussion

In the preceding sections we have analyzed the consequences of selection on proteins for synonymous codon usage, and for volatility in particular. For all parameter values, we have demonstrated that negative selection at the amino acid level depresses volatility and positive selection increases volatility, compared to neutrality. We have elucidated the conditions under which we can expect to draw inferences about selection on proteins from synonymous codon usage. All of these results have been derived under the Fisher-Wright model of population genetics—the same model that has been used to quantify the power of most other methods to detect selection (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Goldman and Yang 1994; Ewens 2004).

As others have noted (Dagan and Graur 2004; Hahn et al. 2005), the idea of estimating selection on a protein using a single nucleotide sequence challenges the essentiality of the comparative method in evolutionary research—a paradigm that has dominated the field for four decades (Zuckerkandl and Pauling 1965). Although we have demonstrated the reliability of this new approach in certain parameter regimes, there is the potential for confounding factors ignored by the Fisher-Wright model. In the discussion below we compare methods of detecting selection on proteins, and we discuss several important, practical limitations of the volatility method.

Codon Volatility Versus Comparative Sequence Analysis

Selection pressures on proteins are usually estimated by comparing homologous nucleotide sequences (Zuckerkandl and Pauling 1965). Orthologous genes are identified in different organisms and sequenced; their sequences are then aligned, and the changes that have accumulated since divergence are used to infer the selection pressures that have been acting (Goldman and Yang 1994). When available, sequence variation sampled from individuals within a species can be compared with variation across species to produce an elegant test for adaptive evolution at a locus (McDonald and Kreitman 1991; Sawyer and Hartl 1992). In addition, there is a variety of statistical tests designed to detect a departure from neutrality in the site frequency spectrum sampled within a single species (see Kreitman 2000 and references therein). In many cases, the complete distribution of these statistics under the neutral null model are difficult to derive, but they have been studied through computer simulation (Simonsen et al. 1995).

Techniques for estimating selective constraints via comparison of divergent sequences are typically applied to one or several genes at a time. When extensive intra- or inter-specific sequence data are available at a locus of interest, such techniques have proven enormously useful for measuring selection, and it is unlikely that they will be significantly improved by incorporating information about synonymous codon usage. But the accurate estimation of selective constraints requires a large number (approximately six or more [Anisimova et al. 2001]) of orthologous sequences for each gene of interest. At the genome-wide scale, comparative data (i.e., orthologous gene sequences) will not be available for all genes, and methods to estimate selective constraints based on sequence comparison will often be inapplicable. Furthermore, the genes under positive selection are often of particular interest, but such genes are even less likely to have identifiable orthologs in related species due to their rapid sequence divergence (Plotkin *et al.* 2004). Even in the lineage of the *Saccharomyces* genus, which is currently the best-case scenario for comparative genomics, the genomes of four species have been fully sequenced and only two-thirds of the genes in *S. cerevisiae* have unambiguously identifiable orthologs in related species (Kellis et al. 2003). Unlike comparative techniques, the analysis of synonymous codon usage offers a computational tool to screen for selection pressures on *all* genes in a sequenced genome. Genome-wide screens based on analyzing synonymous codon usage may prove useful in identifying important classes of genes under strong selection, such as the antigens of pathogens (Plotkin et al. 2004). Nevertheless, at the scale of individual genes, screens for selection based on codon usage will probably not have enough statistical power except for some micro-organisms, which experience very large effective population sizes or periods of hyper mutation.

Unlike most statistics that test for a departure from neutrality based on comparative data, estimates of selection based on "volatility *p*-values" (Plotkin *et al.* 2004) are not estimators in a rigorous statistical sense—i.e., statistics whose sampling properties can be derived from a null model, and which can be used in likelihood ratio tests of a null hypothesis (Yang et al. 2000; Clark et al. 2003). Given the expected relative frequencies of codons that we have derived for each of the three regimes (neutral, negative, and positive selection; Tables 1 and 2), it would be possible to develop maximum-likelihood methods that estimate the number of sites of a gene in each regime. This approach will be complicated, however, by the need to control for other sources of codon bias (see below). It is critical to note that volatility *p*-values, which control for each gene's amino acid content, measure only the relative selection pressures on the proteins in a genome, and therefore cannot distinguish between positive or relaxed negative selection. This point has been emphasized previously (Plotkin et al. 2004, 2005) but misunderstood by several authors (see Supplementary Information).

Aside from the different situations in which they are applicable, estimates of selection based on codon volatility differ in a fundamental way from most estimates based on sequence comparison. Homologous sequence comparison between species is often used to assess, by either maximum likelihood (Goldman and Yang 1994) or maximum parsimony (Li 1993), the rates of synonymous and non-synonymous substitutions in a coding sequence. The ratio of these rates, dN/dS, is used as a measure of the selective constraints that have been acting on a protein since the divergence of the species being compared. An alternative approach, based on a Poisson Random Field (PRF) model of polymorphism frequencies, uses the site frequency spectrum at a locus sampled from a population to deduce the average selective pressure for or against amino acid changes in a gene (Sawyer and Hartl 1992). PRF models can also be used to construct likelihood ratio tests of departure from neutrality (Bustamante et al. 2001). Like most comparative methods, however, both of these models typically assume that all the sites within a gene experience the same selective pressure against amino acid substitutions (but see the site-by-site likelihood tests of Yang et al. [2000]). Using the PRF method, for example, authors have estimated a very small "average" selection pressure against amino acid changes in *E. coli* genes: $\sigma \sim 10^{-8}$ (Hartl and Sawyer 1994). This value does not represent the arithmetic average of the true σ values across sites, but rather the best-fit constant value of σ that would make the PRF model consistent with observed sequence variation at polymorphic sites.

When evolutionary rates are estimated at *individual* residues (Yang 2000; Yang et al. 2000), however, we find great variation across sites. Moreover, direct experimental measurements of the fitness consequences of amino acid substitutions in micro-organisms reveals huge variation in selection pressures across the residues of an individual protein: a substantial proportion of substitutions is lethal and a substantial proportion has undetectable effect (Wertman et al. 1992; Wlocha et al. 2001; Zyle and De Visser 2001; Sanjuan et al. 2004). Therefore, it is not entirely clear how best to interpret the value of $\sigma \sim 10^{-8}$ estimated for *E. coli* genes using the PRF model, which assumes constant pressure at each residue.

Compared to dN/dS or σ estimated by the PRF model, codon volatility quantifies selection pressures in a very different, coarser manner. As discussed above, volatility essentially measures the number of sites in a gene that experience negative ($\sigma \gg u$) versus neutral ($\sigma \ll u$) versus positive selection. Given that amino acid changes at many sites in a protein sequence are lethal while changes at other sites have no effect, it is reasonable and meaningful to estimate the number of sites in the selected versus neutral regimes. But volatility is not sensitive to variation in selective pressures within either of these regimes. Hence, the volatility measure is in some ways a coarser description of selective pressure than PRF or dN/dS. One should not necessarily expect that volatility will correlate very strongly with dN/dS or PRF estimates, because the latter measures represent some sort of average σ over the entire gene, and are thus presumably sensitive to the full range of variation in σ. A measure of selective constraint based on codon volatility is therefore different from and complementary to measures based on dN/dS or the PRF model.

In our analysis of synonymous signatures of selection on proteins, we have treated sites as independent and freely recombining. The same independence assumption is made under the PRF model, and under most incarnations of dN/dS. In reality, sites within a gene are often tightly linked. Linkage increases the variance in estimates of selection pressures inferred by the PRF model (Akashi and Schaeffer 1997), and it is likely to have similar effects on inferences made from synonymous codon usage. The precise effects of linkage on estimates of selection remains a topic for future research. We also note that temporal variations in the selection pressure at a site may affect synonymous codon usage (Myers et al. 2005).

It is important to note that the most common model used to estimate dN/dS from orthologous nucleotide sequences does not itself reflect the true relationship between selection and volatility. dN/dS is often estimated by fitting maximum-likelihood parameters to a simplified Markov-chain model of sequence evolution that ignores population variability (Goldman and Yang 1994). Models that ignore population variability are perfectly reasonable approximations when comparing the sequences of divergent lineages; but such models fail to detect the effect of amino acid selection on synonymous codon usage. Such models consider only a single sequence that is assumed to represent the dominant genotype in the population at any time. Mutation and selection are modeled simultaneously by adjusting the transition rates between codon states in the sequence (Goldman and Yang 1994). As a result, in equilibrium, the number of transitions into a state per unit time must equal the number of transitions out of that state; and so equilibrium synonymous codon usage does not depend the strength of selection in these simplified models (Goldman and Yang 1994). In fact, such models typically require as parameters the specification of the equilibrium codon usage (Goldman and Yang 1994), and so they clearly cannot be used to predict equilibrium codon usage. Therefore, simulations of sequence evolution based on these simplified models (such as the non-frequency-dependent simulations of Zhang [2004] or those of Nielsen and Hubisz [2005] or Dagan and Graur [2004]) will fail to detect the relationship between selection and codon usage. By contrast, more realistic simulations that account for population variability (such as the frequency-dependent simulations of Zhang [2004] and the Fisher-Wright simulations in this work) will properly reflect the relationship between selection and codon usage in a replicating population.

Other Sources of Codon Bias

Our analysis of volatility, much like the theoretical underpinning of the PRF method and dN/dS, is based on the Fisher-Wright model of population genetics, and it specifically ignores sources of direct selection on synonymous codons. In this section, however, we will discuss other processes that induce codon bias and that may conflate estimates of selection on proteins.

Although it came as a surprise to early theorists (King and Jukes 1969), it is now clear that several processes induce unequal usage of synonymous codons. In micro-organisms, some sources of bias, such as biased nucleotide content, apply roughly equally to all genes in a genome. To the extent that other sources of codon bias apply equally across a genome, there is a straightforward randomization procedure to control for these biases when comparing the volatilities of genes in a genome (Plotkin et al. 2004). If, however, other sources of codon bias differ from gene to gene within a genome, they may (if not properly controlled for) introduce errors into estimates of relative selection pressures on proteins inferred from codon volatility (Plotkin et al. 2004). By the same token, selection on synonymous codons—particularly selection that varies from gene to gene—will also conflate estimates of selection obtained by classical techniques such as dN/dS (Sharp and Li 1987; Hirsh et al. 2005; Chamary et al. 2006).

One process that may vary across the genome is the transition/transversion mutation bias. Results on *S. cerevisiae*, whose genome exhibits variation in the tr/tv bias, suggest that this source of variable codon bias will not distort estimates of selection based on volatility: whether or not one accounts for the variation in the tr/tv bias across the genome of *S. cerevisiae* one obtains virtually the same rankings of volatility p-values ($r > 0.99$). The same result holds for the *E. coli* genome as well.

Aside from mutational biases, there are other sources of codon bias that vary from gene to gene in some organisms. In the yeast *S. cerevisiae*, researchers have observed that synonymous codon usage, measured by the Codon Adaptation Index (CAI) (Sharp and Li 1987), is correlated with a gene's expression level in laboratory conditions (Coghlan and Wolfe 2000). This correlation is thought to be caused by selection for translational efficiency and/or accuracy: a codon corresponding to a more abundant tRNA is expected to be translated more quickly (due to the higher probability per unit time that the appropriate tRNA will "find" the codon) and more accurately (since the correct tRNA will likely have the greatest chance of pairing if it is the most abundant). Considering this alternative source of biased codon usage, two questions should be asked: Do other sources of codon bias distort estimates of selection? And how can we control for these confounding factors? We do not have a truly satisfactory answer for either of these questions, but the discussion below may shed some light on the issues involved.

The degree to which other sources of codon bias distort estimates of selection on proteins will depend on the organism being studied. In some species, such as humans, codon frequencies exhibit a much weaker correspondence with tRNA abundances than in other species. In a species with a strong correspondence between codon usage and tRNA abundances, the extent to which variation in this source of codon bias affects volatility will depend on whether volatile codons are (un)preferred: if there is no correlation between volatility and tRNA abundances, then the other sources of codon bias will introduce only random error into volatility estimates, making the conclusions based on volatility less powerful but still reliable. If instead the preferred codons tend to have either high or low volatility, then this effect will introduce systematic errors into volatility estimates of selection (Stoletzki et al. 2005). In the latter case, in order to quantify how much codon usage bias is caused by volatility as opposed to other factors, one would require a method to quantify for individual genes the amount of codon bias due to these other factors. Unfortunately we do not have near the necessary level of predictive power for other sources of codon bias in any organism. Although expression levels are somewhat predictive of codon bias, expression levels do not explain most of the variation in codon bias in any genome studied thus far (Akashi 2001; Coghlan and Wolfe 2000). Until the various sources of biased codon usage can be reliably disentangled, we cannot reliably quantify the effects of these biases on estimates of selection obtained via volatility. Nor can we quantify the effects of these biases on dN/dS, which is also conflated by translational selection on synonymous sites.

The second question, how to control for other sources of codon bias, is also difficult to answer at present. As discussed above, a truly satisfactory method to control for other sources of bias would require us to quantify other sources of codon bias in a predictive manner for each gene. While this degree of precision is not currently possible, one approach is to assume that the codon bias measured by CAI is entirely independent of volatility, and to then control for CAI using partial correlations. For several reasons, we expect this approach to be conservative, as we illustrate using the yeast *S. cerevisiae* (we use this species as an example because it shows a strong preference for codons that match abundant tRNAs, and because we have reliable dN/dS values for almost two-thirds of its genes, calculated from multiple alignments of closely related species [Hirsh et al. 2005]). First, we note that dN/dS is itself strongly correlated with both CAI and gene expression levels in yeast (Pal et al. 2001), and it is likely therefore that any measure of selective constraint consistent with dN/dS will itself be strongly correlated with CAI and expression. Second, it is possible that the codon bias measured by CAI is in part *caused* by volatility (i.e., highly expressed genes tend to experience stronger purifying selection and therefore exhibit codon usage biased towards lower volatility), and so controlling for CAI would be inappropriate. Despite several biological hypotheses (Drummond et al 2005), there is no accepted mechanistic explanation for the correlation between CAI and dN/dS in yeast (Pal et al. 2001; Akashi 2001), and so it is unclear whether controlling for CAI is entirely appropriate.

Nevertheless, we have examined the relationship between volatility and dN/dS while controlling for CAI. We find that even when controlling for CAI there remains a highly significant partial rank correlation between volatility p-values and dN/dS in yeast ($p < 10^{-22}$) (Plotkin et al. 2005). We interpret this result as evidence that volatility is measuring selective constraints above and beyond any signal inherent in CAI Partial correlations are unreliable, however, when controlling for a noisy variable (Drummond et al. 2006). Although CAI is measured without noise, CAI is a noisy proxy for translation rates. An alternative approach proposed by Drummond et al. (2006), principal component regression, cannot be applied in the case of only two predictor variables, since the PCA will choose axes weighted equally by each predictor variable. Therefore, we performed a principal component rank regression on three predictor variables, and we find that volatility p-values explain roughly the same amount of independent variation in dN/dS as is explained by CAI or mRNA expression levels (8.2%, 11.1%, and 9.3%, respectively). This analysis further supports our conclusion that volatility is correlated with selective constraints on yeast proteins, above and beyond any signal determined by CAI or expression levels.

Finally, we note that there may be direct selection on synonymous codons in order to evade mis-translation (Konopka 1985). Since mis-translation is most likely to occur between a codon and an anticodon that differ by a single nucleotide, the definition of volatility (Eq. 1) is appropriate for measuring the selective pressure for or against mis-translation. The strength of this type of selection on synonymous codons would not depend on θ and σ but, rather, on the mis-incorporation rate of tRNA (which is far higher than the mutation rate) and the detriment of mis-translation (which is likely far lower than that of most mis-sense mutations). It is difficult at present to measure the molecular parameters of tRNA mis-incorporation and its fitness effects; so it is unclear how much of a volatility signal arises from mis-translation avoidance versus selection on mis-sense mutations. Whatever the strength of the mis-translation signal, however, the volatility of a gene would still reflect the degree to which there is selection to conserve, or not to conserve, the (translated) protein sequence.

Practical Limitations of Volatility

In this paper we have presented a theoretical analysis of codon usage and selection under the Fisher-Wright model. The Fisher-Wright model accounts for only the most basic of processes that influence molecular evolution. Nevertheless, there is great deal of empirical evidence indicating that the volatility of a gene, controlling for its ammo acid sequence, is indeed correlated with the selective constraint it experiences. Aside from highly significant correlations between volatility p-values and dN/dS in bacterial species and yeast (Plotkin et al. 2004), volatility also reflects a range of other features known to correlate with selection on proteins. In *S. cerevisiae*, for example, volatility p-values are strongly correlated with the essentiality of genes, the number of their protein-protein interactions, and the degree to which they are preserved throughout the eukaryotic kingdom (Plotkin et al. 2006). Furthermore, volatility is significantly elevated among the known antigens and surface proteins (which experience positive selection) in the pathogens *Mycobacterium tuberculosis*, *Plasmodium falciparum*, and influenza A virus (Plotkin and Dushoff 2003; Plotkin et al. 2004). In fact, elevated volatility has helped researchers to identify new families of antigenic proteins expressed on the surface of *P. falciparum* and infected erythrocytes (Winter et al. 2005). In addition, volatility is significantly depressed in the genes essential for growth of *M. tuberculosis*, as well as in the genes conserved between related *Mycobacterium* species (Plotkin et al. 2004). Therefore, despite potential confounding sources of codon bias that cannot at present be controlled for with appropriate accuracy, in practice volatility-based methods produce estimates of selection pressures that are consistent with our understanding of protein evolution over a diverse range of micro-organisms.

Several authors have suggested that the empirical results above are artifacts caused by the length or amino acid composition of rapidly evolving proteins (Nielsen and Hubisz 2005; Dagan and Graur 2004; Friedman and Hughes 2005; Sharp 2005; Stoletzki et al. 2005). These suggestions are patently incorrect. The empirical results summarized above are all based on "volatility p-values," which control exactly for each gene's amino acid sequence (Plotkin et al. 2004). A gene's amino acid content or length cannot possibly bias its volatility p-value toward zero or one. Related claims that results using volatility are artifacts of horizontal gene transfers (Sharp 2005), although potentially valid, are inconsistent with empirical evidence (see Supplementary Information).

Nevertheless, there remain several factors that will likely limit the applicability of volatility in practice. Chief among these concerns is selection on synonymous sites for translational speed or accuracy, as discussed above. Despite the fact that volatility p-values are significantly correlated with dN/dS even when controlling for expression levels and CAI in yeast, translational selection certainly has the potential to conflate synonymous signatures of selection on proteins. This concern is all the more problematic considering that translational selection is strongest in organisms that experience large effective population sizes, which are required for volatility to detect negative selection.

In addition, our analysis has clarified and quantified a second, practical limitation of volatility: namely, tests for selection based on synonymous signals are not very powerful, requiring many sites under selection. As a result, except for rare circumstances, volatility cannot be expected to reliably detect positive selection on individual genes. Perhaps it is for this reason that most of the strong empirical results using volatility have been observed for groups of genes, such as families of antigens (Plotkin et al. 2004).

Our analysis has also clarified and quantified a third, practical limitation of volatility: the need for large effective population sizes or mutation rates. Volatility has virtually no power to detect negative selection when θ per site is much smaller than unity. Estimates of θ for microbes vary widely, and periods of hyper mutation will increase the relevant value of θ even when neutral site heterozygosity is small. But as we have shown, synonymous codon usage will not reflect selection on proteins in organisms, such as macroscopic species and probably some microbes as well, that consistently experience small population sizes and mutation rates.

Acknowledgments. We thank Daniel Fisher, Andrew Murray, and Michael Turelli for their input during the preparation of the manuscript. We also thank an anonymous referee for substantial conceptual input. J.B.P. acknowledges support from the Harvard Society of Fellows, the Milton Fund, and the Burroughs Wellcome Fund. M.M.D. acknowledges support from a Merck Award for Genome-Related Research.

References

Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11:660–666

Akashi H, Schaeffer SW (1997) Natural selection and the frequency distribution of "silent" DNA polymorphism in Drosophila. Genetics 146:295–307

Anisimova M, Bielawski JP, Yang Z (2001) The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. Mol Biol Evol 18:1585–1592

Berg O (1996) Selection intensity for codon bias and the effective population size of *Escherichia coli*. Genetics 142:1379–1382

Bjedov I, Tenaillon O, Gerard B, et al. (2003) Stress-induced mutagenesis in bacteria. Science 300:1404–1409

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

Bustamante CD, Wakely J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159:1779–1788

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Rev Genet 7:98–108

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303

Clark A, Glanowski S, Nielsen R, Thomas P, Kejariwal A, MA MT, Tanenbaum D, Civello D, Lu F, B BM, Ferriera S, Wang G, Zheng X, White T, Sninsky J, Adams M, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963

Coghlan A, Wolfe KH (2000) Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast 16:1131–1145

Crow JF, Kimura M (1970) An introduction to population genetics theory. Burgess, Minneapolis

Dagan T, Graur D (2004) The comparative method rules! codon volatility cannot detect positive Darwinian selection using a single genome sequence. Mol Biol Evol 22:496–500

Debry R, Marzluff WF (1994) Selection on silent sites in the rodent H3 historic gene family. Genetics 138:191–202

Denamur E, Lecointre G, Darlu P, OTenaillon CA, Sayada C, Sunjevaric I (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell 103:711–721

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102:14338–14343

Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23:327–337

Ewens W (2004) Mathematical populations genetics I. Springer-Verlag, New York

Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR, Venter JC, Fraser CM (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol 184:5479–5490

Friedman R, Hughes AL (2005) Codon volatility as an indicator of positive selection: Data from eukaryotic genome comparisons. Mol Biol Evol 22:542–543

Gibbons RJ, Kapsimalis B (1967) Estimates of the overall rate of growth of the intenstinal microflora for hamsters, Guinea pigs, and mice. J Bacteriol 93:510–512

Gillespie J (2001) Is the population size of a species relevant to its evolution? Evolution 55:2161–2169

Giraud A, Radman M, Matic I, Taddei F (2001) The rise and fall of mutator bacteria. Curr Opin Microbiol 4:582–585

Golding GB, Strobeck C (1982) Expected frequencies of codon use as a function of mutation rates and codon fitnesses. J Mol Evol 18:379–386

Goldman N, Yang Z (1994) Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Hahn MW, Mezey JG, Begun DJ, Gillespie JH, Kern AD, Langley CH, Moyle LC (2005) Codon bias and selection on single genomes. Nature 433:E1

Hartl DL, Sawyer SA (1994) Selection intensity for codon bias. Genetics 138:227–234

Higgs P (1994) Error thresholds and stationary mutant distributions in multi-locus diploid genetics models. Genet Res Cambr 63:63–78

Hirsh AE, Fraser HB, Wall DP (2005) Adjusting for selection on synonymous sites in estimates of evolutionary distance. Mol Biol Evol 22:174–177

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer-RNAs and the occurrence of the respective codons in its protein. J Mol Biol 146:1–21

Kellis M, Patterson N, Endrizzi M, Birren B, Lander E (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788

Konopka AJ (1985) Theory of degenerate coding and informational parameters of protein coding genes. Biochimie 67:455–468

Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genomics. Hum Genet 1:539–559

LeClerc J, Li B, Payne WL, Cebula TA (1996) High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. Science 274:1208–1211

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–4404

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. Genet Res Cambr 23:23–25

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the ADH locus in *Drosophila*. Nature 351:652–654

Miyata T, Miyazawa S, Yashunaga T (1979) Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219–236

Myers LA, Ancel FD, Lachmann M (2005) Evolution of genetic potential. PloS Comput Biol 1:236–243

Nagylaki T (1992) Introduction to theoretical population genetics. Springer, Berlin

Nielsen R, Hubisz M (2005) Detecting selection needs comparative data. Nature 433:E6

Notley-McRobb L, Seeto S, Ferenci T (2001) Enrichment and elimination of *mutY* mutators in *Escherichia coli* populations. Genetics 162:1955–1062

Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci USA 96:12638–12643

Oliver A, Canton R, Campo P, Baquero F, Blazquez J (2000) High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. Science 288:1251–1254

Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931

Plotkin JB, Dushoff J (2003) Codon bias and frequency-dependent selection on the hemagglutinin epitopes of Influenza A virus. Proc Natl Acad Sci USA 100:7152–7157

Plotkin JB, Dushoff J, Fraser HB (2004) Detecting selection using a single genome sequence of *M. tuberculosis* and *P falciparum*. Nature 248:942–946

Plotkin JB, Dushoff J, Fraser HB (2005) Codon bias and selection on single genomes: reply. Nature 433:E7–E8

Plotkin JB, Fraser HB, Dushoff J (2006) Natural selection on the genome of *Saccharomyces cerevisiae* (in preparation)

Sanjuan R, Moya A, Elena S (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci USA 101:8396–8401

Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132:1161–1176

Sharp PM (2005) Gene "volatility" is most unlikely to reveal adaptation. Mol Biol Evol 22:807–809

Sharp PM, Li WH (1987) The codon adaptation index: A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Simonsen KL, Churchill GA, Aquadro CF (1995) Poperties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429

Sorensen M, Kurland C, Pedersen S (1989) Codon usage determines translation rate in *Escherichia coli*. J Mol Biol 207:365–377

Stoletzki N, Welch J, Hermisson J, Eyre-Walker A (2005) A dissection of volatility in yeast. Mol Biol Evol 22:2022–2026

Tajima F (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics 143:1457–1465

Tang H, Wyckoff GJ, Lu J, Wu C (2004) Universal evolutionary index for amino acid changes. Mol Biol Evol 21:1548–1556

Tenaillon O, Denamur E, Matic I (2004) Evolutionary significance of stress induced mutagenesis in bacteria. Trends Microbiol 12:264–270

Thompson CJ, McBride JL (1973) On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. Math Biosci 21:127–142

van Nimwegen E, Crutchfield J, Huynen M (1999) Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 96:9716–9820

Wertman K, Drubin D, Botstein D (1992) Systematic mutational analysis of the yeast ACT1 gene. Genetics 132:337–350

Wilke C (2001) Adaptive evolution on neutral networks. Bull Math Biol 63:715–130

Winter G, Kawai S, Haeggstrom M, Kaneko O, vonEuler A, Kawazu S, Palm D, Fernandez V, Walgren M (2005) SURFIN is a polymorphic antigen expression on *Plasmodium falciparum* merozoites and infected erythrocytes. J Exp Med 20l:1853–1863

Wlocha DM, Szafranieca K, Bortsb RH, Korona R (2001) Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. Genetics 159:441–452

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

Yampolsky LY, Stoltzfus A (2004) The exchangability of amino acids in proteins. Genetics 170:1459–1472

Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol 51:423–432

Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zhang J (2004) On the evolution of codon volatility. Genetics 16S:495–501

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366

Zyle C, DeVisser J (2001) Estimates of the rate and distribution of fitness effects of spontaneous mutation in *Saccharomyces cerevisiae*. Genetics 157:53–61