Purifying Selection, Drift, and Reversible Mutation with Arbitrarily High Mutation Rates

Brian Charlesworth*,[1] and Kavita Jain†

*Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom, and
†Theoretical Sciences Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore 560064, India

ABSTRACT Some species exhibit very high levels of DNA sequence variability; there is also evidence for the existence of heritable epigenetic variants that experience state changes at a much higher rate than sequence variants. In both cases, the resulting high diversity levels within a population (hyperdiversity) mean that standard population genetics methods are not trustworthy. We analyze a population genetics model that incorporates purifying selection, reversible mutations, and genetic drift, assuming a stationary population size. We derive analytical results for both population parameters and sample statistics and discuss their implications for studies of natural genetic and epigenetic variation. In particular, we find that (1) many more intermediate-frequency variants are expected than under standard models, even with moderately strong purifying selection, and (2) rates of evolution under purifying selection may be close to, or even exceed, neutral rates. These findings are related to empirical studies of sequence and epigenetic variation.

THE infinite sites model, originally proposed by Fisher (1922, 1930) and developed in detail by Kimura (1971), has been the workhorse of molecular population genetics for four decades. Its core assumption is that any nucleotide site segregates for at most two variants and that the mutation rate scaled by effective population size (N_e) is so low that new mutations arise only at sites that are fixed within the population (see also Charlesworth and Charlesworth 2010, p. 207). This assumption facilitates calculations of the theoretical values of some key observable quantities, such as the expected level of pairwise nucleotide site diversity or the expected number of segregating sites in a sample (Kimura 1971; Watterson 1975; Ewens 2004). In the framework of coalescent theory, this implies a linear relation between the genealogical distance between two sequences and the neutral sequence divergence between them, greatly simplifying methods of inference and statistical testing (Hudson 1990; Wakeley 2008).

There has recently been some discussion of how to go beyond the infinite sites assumption of a low scaled mutation rate, which breaks down for species with very large effective population sizes, including some species of virus and bacteria, and even eukaryotes such as the sea squirt and outbreeding nematode worms, resulting in "hyperdiversity" of DNA sequence variability within a population (Cutter *et al.* 2013). It is important to note, however, that this problem can arise even when the scaled mutation rate is relatively low, since then the proportion of neutral nucleotide sites that are currently segregating in a population (which depends on the scaled mutation rate) can be substantial when the population size is sufficiently large. For example, with a neutral mutation rate of u per site in a population of N breeding adults, the expected fraction of sites that are segregating in a randomly mating population is $f_s = \theta\ [\ln(2N) + 0.6775]$, where $\theta = 4N_e u$ (Ewens 2004, p. 298). Thus, with $\theta = 0.01$, a reasonable value for many species (Leffler *et al.* 2012), we have $f_s = 0.15$ even when N has the implausibly low value of 1 million. This implies that ~15% of new mutations are expected to arise at sites that are already segregating, suggesting a significant departure from the assumptions of the infinite sites model. (An alternative way of looking at this is to determine the expected number of new mutations that occur at a site while a preexisting mutation is segregating, which is of a similar magnitude to f_s—see *Appendix*, Equation A1.)

doi: 10.1534/genetics.114.167973
Manuscript received July 3, 2014; accepted for publication September 11, 2014; published Early Online September 16, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167973/-/DC1.
[1] Corresponding author: Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, King's Bldgs., W. Mains Rd., Edinburgh EH9 3FL, United Kingdom. E-mail: Brian.Charlesworth@ed.ac.uk

In addition, it has been known for nearly 20 years that sufficiently high scaled mutation rates at some or all sites in a sequence can lead to substantial departures from the infinite sites expectations for statistics such as Tajima's *D*, which are commonly used to detect deviations from neutral equilibrium caused by population size changes or selection (Bertorelle and Slatkin 1995; Aris-Brosou and Excoffier 1996; Tajima 1996; Yang 1996; Mizawa and Tajima 1997). This is because the occurrence of mutations at sites that are already segregating increases the pairwise diversity among sequences, but does not increase the number of segregating sites (Bertorelle and Slatkin 1995). The analysis of data on DNA sequence variation in hyperdiverse species thus requires methods that deal with this problem, and a number of population genetics models that contribute to this have already been developed (Desai and Plotkin 2008; Jenkins and Song 2011; Cutter *et al.* 2012; Jenkins *et al.* 2014; Sargsyan 2014).

Finally, analyses of the inheritance of epigenetic markers, such as methylated cytosines, have suggested that these can sometimes be transmitted across several sexual generations, but with rates of origination or reversion that are several orders of magnitude higher than the mutation rates of DNA sequences (Johannes *et al.* 2009; Becker *et al.* 2011; Schmitz *et al.* 2011; Lauria *et al.* 2014). In view of the current interest in the possible functional and evolutionary significance of epigenetic variation (Richards 2006; Schmitz and Ecker 2012; Grossniklaus *et al.* 2013; Klironomos *et al.* 2013), it seems important to develop models that can shed light on their population genetics, to understand the evolutionary forces acting on them.

The purpose of this article is to develop a relatively simple analytical framework for examining the consequences of high scaled mutation rates, in the framework of the classical random mating, finite population size model with forward and backward mutations in the presence of selection and genetic drift (Wright 1931, 1937). The approach is similar in spirit to the biallelic model used by Desai and Plotkin (2008), but with a focus on sample statistics that summarize properties of the site frequency spectrum, as well as on the expected rate of substitutions along a lineage. As has been found in previous coalescent-based treatments with neutrality (Bertorelle and Slatkin 1995; Aris-Brosou and Excoffier 1996; Cutter *et al.* 2012), the results derived below show that very large departures from the infinite sites model occur when the scaled mutation rate is sufficiently high, even when fairly strong purifying selection is acting, resulting in features of the data such as a large excess of intermediate-frequency variants. In addition, the signal of purifying selection on substitutions along a lineage can be obscured or even converted into a signal of positive selection. The findings have significant implications for the interpretation of the results of studies of both epigenetic variability and DNA sequence variability in species with large effective population sizes. Readers who are interested primarily in the main biological conclusions may wish to skip over the details of the derivations.

Analysis of the Model of Purifying Selection, Drift, and Mutation

Basic assumptions

We assume a randomly mating, diploid, discrete generation population with *N* breeding adults, and effective population size N_e. Over a long sequence of *m* nucleotide sites, each site has two alternative types, A_1 and A_2, with mutation rates *u* and *v* from A_2 to A_1 and vice versa. (With diploidy, this means that only three genotypes are present at a site: A_1A_1, A_1A_2 and A_2A_2.) A_1 and A_2 might correspond to AT versus GC base pairs, unpreferred versus preferred synonymous codons, or selectively favored versus disfavored nonsynonymous variants. If epigenetic variation is being considered, then A_1 and A_2 could be regarded as the methylated or unmethylated states of a nucleotide site or a differentially methylated region (or vice-versa). This approach, while undoubtedly oversimplified, avoids the problem of modeling mutation among all four basepairs, which is difficult to deal with except by making the unrealistic assumption of equal mutation rates in all directions (Ewens 2004).

If selection is acting, we assume semidominance, with A_2 having a selective advantage *s* over A_1 when homozygous and with the fitness of A_1A_2 being exactly intermediate, although our general conclusions are probably not strongly dependent on this assumption. There is complete independence among sites (*i.e.*, recombination is sufficiently frequent that linkage disequilibrium is negligible), and all evolutionary forces are weak, so that the standard results of diffusion approximations can be employed.

When the population is at statistical equilibrium, the probability distribution of variant frequencies over sites remains stationary and the mean numbers of sites in each possible state are constant over time, despite continual changes at individual sites (Charlesworth and Charlesworth 2010, pp. 270–272). At any given time, some sites are fixed for the A_1 type, some are fixed for A_2, and others segregate for both. Let the equilibrium proportion of sites that are fixed for A_1 and A_2 be f_{1f} and f_{2f}, respectively. The proportion of sites that are segregating is $f_s = 1 - f_{1f} - f_{2f}$.

Results for some important population parameters

These assumptions allow the use of Wright's stationary distribution formula (Wright 1931, 1937) to describe the probability density of the frequency *q* of A_2 at a site,

$$\phi(q) = C\,\exp(\gamma q)p^{\alpha-1}q^{\beta-1}, \qquad (1)$$

where $p = 1 - q$, $\alpha = 4N_eu$, $\beta = 4N_ev$, $\gamma = 2N_es$, and the constant *C* is such that the integral of $\phi(q)$ between $q = 0$ and $q = 1$ is equal to 1. It is convenient to write *u* in terms of the mutational bias parameter, κ; *i.e.*, $u = \kappa v$, so that $\alpha = \kappa\beta$.

Properties of the distribution

An explicit expression for *C* can be obtained by noting that the integral of the other terms on the right-hand side with respect to q is equal to the product of $\Gamma(\alpha)\,\Gamma(\beta)/\,\Gamma(\alpha+\beta)$ and the confluent hypergeometric function ${}_1F_1(a,\ b,\ z)$ (Abramowitz and Stegun 1965, p. 503), with parameters $a=\beta$, $b=\alpha+\beta$, $z=\gamma$, where

$$ {}_1F_1(a,b,z)=\sum_{i=0}^{\infty}\frac{z^i}{i!}\frac{(a)_i}{(b)_i} \tag{2}$$

and $(x)_0=1,\ (x)_i=x(1+x)(2+x)\dots\ (i-1+x)$ for $i\geq 1$ (Pochhammer's symbol).

This can be seen by expanding the exponential term in Equation 1 in powers of γq and integrating over the range 0 to 1 (Kimura *et al.* 1963).

Integrating Equation 1, we have

$$C=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{1}{{}_1F_1(\beta,\alpha+\beta,\gamma)}. \tag{3}$$

Furthermore, the jth moment of q around zero, obtained from the integral of $q^j\phi(q)$ between 0 and 1, is given by

$$M_j(q)=\frac{{}_1F_1(\beta+j,\alpha+\beta+j,\gamma)(\beta)_j}{{}_1F_1(\beta,\alpha+\beta,\gamma)\ \ (\alpha+\beta)_j}. \tag{4}$$

In particular, the mean frequency of A_2 is

$$\overline{q}=\frac{1}{(1+\kappa)}\ \frac{{}_1F_1(\beta+1,\ \alpha+\beta+1,\gamma)}{{}_1F_1(\beta,\alpha+\beta,\gamma)} \tag{5a}$$

and the mean frequency of A_1 is

$$\overline{p}=\frac{\kappa}{(1+\kappa)}\ \frac{{}_1F_1(\beta,\alpha+\beta+1,\gamma)}{{}_1F_1(\beta,\alpha+\beta,\gamma)}. \tag{5b}$$

Approximations for small β: Approximations to these expressions for the case when $\beta << 1$ and κ is of order 1 are derived in the *Appendix*. Equations A3a and A3b imply that

$$\overline{q}=\frac{1}{[1+\kappa\exp(-\gamma)]}+O(\beta). \tag{6}$$

The left-hand side of Equation 6 is equivalent to the fraction of sites that carry A_2 in a random sequence sampled from the population; if A_1 and A_2 correspond to unpreferred and preferred codons, respectively, this measures the frequency of preferred codons, *Fop* (McVean and Charlesworth 1999). With epigenetic variation, if A_1 and A_2 correspond to methylated and unmethylated states, respectively, $\overline{q}$ measures the fraction of unmethylated sites or regions in a random genome.

The leading term on the right-hand side of Equation 6 is identical to the Li–Bulmer equation commonly used in analyses of selection on codon usage (Li 1987; Bulmer 1991). This result is, however, often derived by assuming that nearly all sites are fixed and calculating the rate of flux between sites fixed for A_1 and A_2; $\overline{q}$ is then taken to be the frequency of sites that are fixed for A_2, with $1-\overline{q}$ representing the frequency of sites fixed for A_2 (Bulmer 1991). This raises the question of how good an approximation we obtain by neglecting the term of order β, when the infinite sites assumption is violated, so that a significant fraction of sites are in fact segregating for A_1 and A_2.

First, we note that it is immediately obvious from Equation 1 and Equations 5a and 5b that $\overline{q}$ with $\gamma=0$ is equal to $1/(1+\kappa)$, so that Equation 6 for this case is exact, as has long been known (Wright 1931). We can also obtain a first-order approximation to Equations A3a and A3b when $\gamma\neq 0$ by expanding in powers of β, which will be accurate when β is sufficiently small. Neglecting second-order and higher terms in β, as is also done in Equations 8a–8c, we obtain

$$\overline{q}\approx\frac{1-\beta\kappa g\exp(-\gamma)}{[1+\kappa\exp(-\gamma)]}, \tag{7a}$$

where

$$g=\frac{\left[\gamma+\kappa\exp(-\gamma)\sum_{i=1}^{\infty}\left(\gamma^i/i!\right)a_{i+1}+\sum_{i=2}^{\infty}\left(\gamma^i/i!\right)\left(a_{i+1}-a_i\right)\right]}{[1+\kappa\exp(-\gamma)]} \tag{7b}$$

and a_i is the harmonic series $1+1/2+1/3+\ldots+1/(i-1)$, with $i\geq 2$.

As shown after Equation A3b of the *Appendix*, the leading term in Equation 6 should provide a good approximation when $\beta\kappa\leq 0.1$.

Approximations for large γ: For examining what happens when γ becomes very large, it is useful to note that the Taylor's series expansion of Equation 5b for small β yields the expression

$$\overline{p}\approx\frac{\kappa\exp(-\gamma)}{1+\kappa\exp(-\gamma)}\left\{1+\beta\left[\sum_{i=1}^{\infty}\frac{\gamma^i}{i!}+\frac{\kappa\exp(-\gamma)}{1+\kappa\exp(-\gamma)}\sum_{i=1}^{\infty}\frac{\gamma^{i+1}\ln(i)}{(i+1)!}\right]\right\}. \tag{8a}$$

For large γ, this gives

$$\overline{p}\approx\ \frac{\kappa\exp(-\gamma)}{1+\kappa\exp(-\gamma)}\left[1+\frac{\beta\exp(\gamma)}{\gamma}\right]; \tag{8b}$$

i.e.,

$$\overline{p}\approx\frac{\beta\kappa}{\gamma}\left[1+O(\gamma^{-1})\right]. \tag{8c}$$

The first term on the right-hand site of Equation 8c is equivalent to the asymptotic expression for $\overline{p}$ with large γ given by Kimura *et al.* (1963). This implies that, for sufficiently large γ compared with β, the mean frequency of the disfavored variant is equal to its equilibrium frequency under mutation–selection balance with $s >> u$ in an infinite population, where $p = 2u/s = 2v\kappa/s$ (Haldane 1927), as expected intuitively. Numerical studies show that Equation 8b performs well for $\gamma > 1$ if $\beta << 1$, when it can give a good approximation when neither the leading term in Equation 6 nor the Kimura *et al.* (1963) large γ approximation is accurate (results not shown). Equation 8b implies that the leading term in Equation 6 is accurate when $\gamma << -\ln(\beta)$.

Frequencies of fixed sites: Second, the approximate frequencies of sites that are fixed for A_1 and A_2, $f_{1\mathrm{f}}$ and $f_{2\mathrm{f}}$, can be found from the integrals of $\phi(q)$ between 0 and $1/(2N)$ and 1– $1/(2N)$ and 1, respectively (Ewens 2004, p. 178). For large N, such that $\gamma N^{-1} << 1$, when q is close to zero we have $\phi(q) = C[q^{\beta-1} + O(\gamma N^{-1}) + O(\beta N^{-1})]$, and so

$$\begin{aligned} f_{1\mathrm{f}} &\approx \int_0^{1/(2N)} \phi(q)\mathrm{d}q \\ &= C[\beta^{-1}(2N)^{-\beta} + O(\gamma N^{-1}) + O(\beta N^{-1})] \end{aligned} \tag{9a}$$

$$\begin{aligned} f_{2\mathrm{f}} &\approx \textstyle\int_{1-1/(2N)}^{1} \phi(q)\mathrm{d}q \\ &= C\exp(\gamma)[(\beta\kappa)^{-1}(2N)^{-\beta\kappa} + O(\gamma N^{-1}) + O(\beta N^{-1})], \end{aligned} \tag{9b}$$

where the terms in $O(\gamma N^{-1})$ and $O(\beta N^{-1})$ can be neglected when N is sufficiently large (*cf.* Kimura 1981). Approximations for these expressions for small β can readily be obtained (see *Appendix*).

Nucleotide site diversity: Third, the expected pairwise nucleotide site diversity, π, can be obtained from the expectation of $2pq$, given by $2E\{q - q^2\}$. From Equation 4, we have

$$E\left\{q^2\right\} = \frac{\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)} \frac{{}_1F_1(\beta+2,\ \alpha+\beta+2,\gamma)}{{}_1F_1(\beta,\alpha+\beta,\gamma)}. \tag{10a}$$

The expectation of $E\{q - q^2\}$ is given by subtracting Equation 10a from Equation 5b. Using Equation 2 and simplifying, we obtain

$$E\{pq\} \quad = \frac{\alpha/(\alpha+\beta+1) + \sum_{i=1}^{\infty}\left(\gamma^i/i!\right)\left[(\beta+1)_i/(\alpha+\beta+1)_i - (\beta+1)_{i+1}/(\alpha+\beta+1)_{i+1}\right]}{\left[1+\kappa+\gamma+\sum_{i=2}^{\infty}\left(\gamma^i(\beta+1)_{i-1}/i!\ \ (\alpha+\beta+1)_{i-1}\right)\right]}. \tag{10b}$$

This is equal to one-half of the expected pairwise diversity per site, π. Using the same approach as for Equations 7 and 8, keeping only terms of order β we obtain

$$\pi \approx \frac{2\beta\kappa}{\gamma} \frac{[1 - \exp(-\gamma)]}{[1 + \kappa\exp(-\gamma)]}. \tag{11}$$

As expected, this is identical to equation 15 of McVean and Charlesworth (1999) for the infinite sites model at statistical equilibrium, where new mutations arise only at sites that are fixed either for A_1 or for A_2. When $\gamma >> \beta$, this term converges on the deterministic value under mutation–selection balance, $2\beta\kappa/\gamma$, which corresponds to the diversity expected at deterministic mutation–selection balance with $p = 2u/s$ (see above).

In the case of neutrality, Equation 10b reduces to the following expression (Charlesworth and Charlesworth 2010, p. 237):

$$\pi = \frac{2\beta\kappa}{(1+\kappa)[\beta(1+\kappa)+1]}. \tag{12}$$

As expected intuitively, the neutral diversity is always less than for the infinite sites model with a given value of β and κ, where Equation 11 with $\gamma = 0$ gives $\pi = 2\beta\kappa/(1 + \kappa)$, because some new mutations arise at sites that are already polymorphic; π approaches $2\kappa/(1 + \kappa)^2$ for large β, which is the value for an infinite population at equilibrium under reversible mutation between A_1 and A_2.

Rate of substitution along a lineage

Analytic results: The rate of substitution of new mutations along a lineage can be modeled as follows. Conditioning on a frequency q of the A_2 variant at a site in a given generation, there is an expected number of $2Nv\kappa q$ mutations per site from A_2 to A_1 and $2Nvp$ from A_1 to A_2. The corresponding probability that A_1 becomes fixed, conditional on p, is $Q_1(p) = [\exp(\gamma p) - 1]/[\exp(\gamma) - 1]$ (Kimura 1962). Conditioning on this fixation event, the probability that it is a new A_1 mutation that has been fixed is $1/(2Np)$. The expected number of new A_1 mutations that become fixed is thus equal to $v\kappa p^{-1}qQ_1(p)$. Similarly, the conditional probability that A_2 eventually becomes fixed is $Q_2(q) = [1 - \exp(-\gamma q)]/[1 - \exp(-\gamma)]$; the net expected number of new A_2 mutations that become fixed is $vpq^{-1}Q_2(q)$. (At first sight, it would seem that this procedure cannot be applied to mutations arising in the fixed classes and that these should be treated separately, but the argument given in the *Appendix* shows that it provides an accurate approximation for this situation as well.)

Integrating over all values of q, the net rate at which new mutations enter the population and become fixed is thus

$$\lambda = \nu \int_0^1 [\kappa p^{-1} q Q_1(p) + p q^{-1} Q_2(q)] \phi(q) \mathrm{d}q. \tag{13}$$

The terms involving functions of p and q in the integrand are

$$p^{-1} q Q_1(p) \phi(q) = \frac{C[\exp(\gamma p) - 1]}{[\exp(\gamma) - 1]} \exp(\gamma q) p^{\alpha - 2} q^{\beta} \tag{14a}$$

$$p q^{-1} Q_2(q) \phi(q) = \frac{C[1 - \exp(-\gamma q)]}{[1 - \exp(-\gamma)]} \exp(\gamma q) p^{\alpha} q^{\beta - 2}. \tag{14b}$$

The corresponding integrals are

$$I_1 = \left\{ \frac{\exp(\gamma) - {}_1F_1(\beta + 1, \alpha + \beta, \gamma)}{{}_1F_1(\beta, \alpha + \beta, \gamma)} \right\} \frac{\beta}{(\alpha - 1)[\exp(\gamma) - 1]} \tag{14c}$$

and

$$I_2 = \left\{ \frac{{}_1F_1(\beta - 1, \alpha + \beta, \gamma) - 1}{{}_1F_1(\beta, \alpha + \beta, \gamma)} \right\} \frac{\alpha}{(\beta - 1)[1 - \exp(-\gamma)]}. \tag{14d}$$

Note that ${}_1F_1(\beta - 1, \alpha + \beta, \gamma) - 1$ has a factor of $\beta - 1$, so that the term in $\beta - 1$ in the denominator of Equation 14d cancels. At first sight, Equation 14c appears to have a singularity at $\alpha = 1$. However, by using the relation ${}_1F_1(a, b, z) = {}_1F_1(b - a, b, z) \exp(z)$, we find that ${}_1F_1(\beta + 1, \alpha + \beta, \gamma) = {}_1F_1(\alpha - 1, \alpha + \beta, \gamma)\exp(\gamma)$, so that the numerator of Equation 14c contains a factor of $\alpha - 1$, which cancels the term in the denominator.

The net rate of substitution is given by

$$\lambda = \nu(\kappa I_1 + I_2). \tag{15}$$

As γ approaches zero, Equations 14 and 15 imply that λ tends to $2\nu\kappa/(1 + \kappa)$; this is independent of the population size and is identical to the infinite sites expression with neutrality at statistical equilibrium under reverse mutation (Charlesworth and Charlesworth 2010, p. 274), as expected from the fact that the equilibrium neutral substitution rate is equal to the net mutation rate for any class of mutational model (Kimura 1968).

When α and β are sufficiently small, the main contributions to λ come from the two fixed classes, so that the initial frequencies of the new mutations can be equated to $1/(2N)$, when $O(\beta^2)$ terms in I_1 and I_2 are neglected. Using the above result that the frequencies of the fixed classes are equal to the infinite sites values multiplied by a factor $1 - O(\beta)$, the infinite sites expression for the case of selection is recovered, neglecting higher-order terms in β (equation 6.11 of Charlesworth and Charlesworth 2010, p. 275). Again, this implies that, as expected, the infinite sites model provides a good approximation for the rate of substitution with sufficiently small β.

Scaling relative to the neutral rate: There are two different ways in which we can determine the ratio of the value of λ with $\gamma > 0$ to that for a neutral standard, thereby removing the dependence on the mutation rate term in Equation 15. First, λ with selection can be compared to its value at statistical equilibrium with the same value of α and β. This would be appropriate for comparing rates of evolution at putatively neutral sites in a given genomic region with those at sites that are potentially under purifying selection, without making any corrections for differences in base composition; this is often done when comparing nonsynonymous and synonymous rates of substitution across different genes by statistics such as K_A/K_S. Second, λ with selection can be compared with the neutral rate conditioned on the same mean frequencies of A_1 and A_2 along the sequence as for the selected sites; this corresponds to methods that compare probabilities of substitution between the same pairs of nucleotides in contexts when these are putatively selected *vs.* putatively neutral (Halligan *et al.* 2004; Eory *et al.* 2010).

Numerical results for the population parameters

Numerical results generated from the above formulas are presented in Figure 1 and Figure 2. Figure 1 illustrates the dependence of the following variables on the scaled mutation rate (β) and the scaled intensity of selection (γ), assuming a mutational bias (κ) of 2 toward the deleterious variant at a site: the mean frequency per site of the deleterious variant A_1 ($\overline{p}$), the expected diversity (π), the expected proportion of sites that segregate for variants (f_s), and the above two measures of the rate of substitution relative to neutral expectation. Figure 2 illustrates the dependence of $\overline{p}$ and π on β at a finer scale, for different values of κ and γ. For clarity, the infinite sites values for $\overline{p}$ and the relative rates of substitution are not shown; with selection, the infinite sites values for these parameters are close to their values when $\beta = 0.002$.

With neutrality, the exact value of $\overline{p}$ is always equal to the infinite sites value and is independent of β for a given value of κ. With selection and low β (0.002 or 0.02), it can be seen that agreement with the infinite sites predictions is fairly good for both these values despite the fact that the proportion of sites that are segregating can be quite substantial with $\beta = 0.02$; the second-order approximation of Equations 7a and 7b gives very close agreement even for $\beta = 0.2$ with weak selection, but diverges for $\beta > 0.2$ when $\gamma > 0.5$ (results not shown), whereas the value of $\overline{p}$ departs quite seriously from the infinite sites values at $\beta = 0.2$ when $\gamma > 5$. A similar pattern of departure from the infinite sites value holds for π, except when selection is strong ($\gamma = 5$ or 50), when agreement is still good at $\beta = 2$; this is because the exact diversity and the infinite sites value both approach the deterministic value under mutation–selection balance when $1 << \gamma$ and $\beta << \gamma$ (see Equation 11). Somewhat surprisingly,

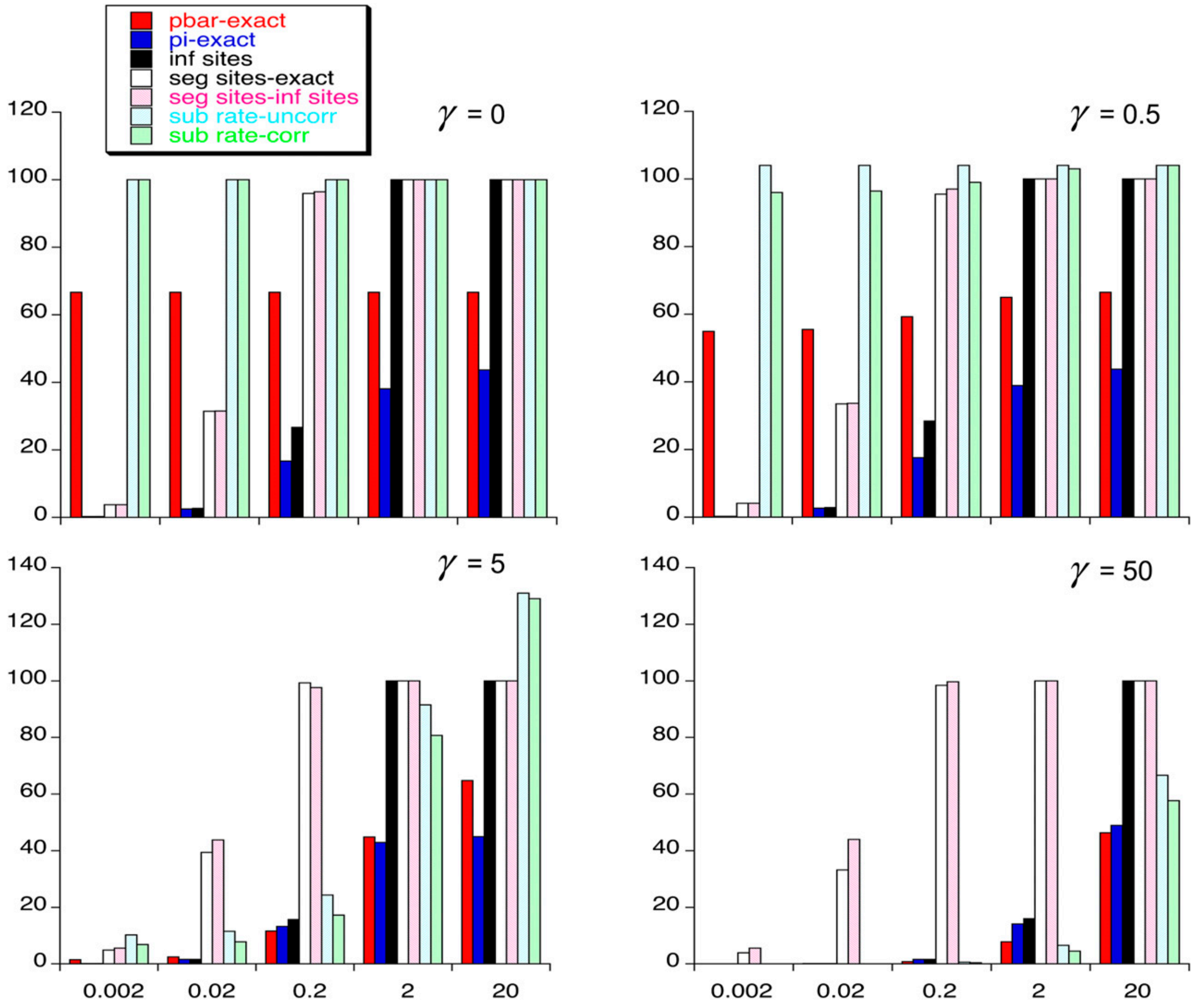

Figure 1 The vertical bars are the values (in percentages) of the mean frequency of A_1, $\overline{p}$ (red), π from Equation 10b (blue), π as given by the infinite sites model (black), the proportion of segregating sites from Equation 9 (white), the proportion of segregating sites under the infinite sites model (pink), the uncorrected rate of substitution relative to neutrality (light blue), and the corrected rate of substitution relative to neutrality (green).

the infinite sites and exact values of the proportion of sites that are segregating always agree well.

Perhaps the most interesting result to emerge is that, with $\kappa > 1$, the rate of substitution relative to neutral expectation can exceed one when there is moderate selection and mutational bias toward the deleterious variants. This has long been known to apply to the infinite sites model when the "uncorrected" relative rate is used and when there is mutational bias (Eyre-Walker 1992; McVean and Charlesworth 1999), which can cause serious problems for phylogenetic inferences concerning selective constraints (Lawrie *et al.* 2011). As shown in the *Appendix*, the "corrected" relative rate is always expected to be less than one under the infinite sites assumption (see Equation A8). But with sufficiently high β, the corrected relative rate can exceed one, even for $\gamma = 5$, and can be only just below one for lesser values of β. The reason for this seemingly paradoxical result is presumably the fact that nearly all sites are segregating if β is high; when $\overline{p}$ is sufficiently high because mutation and drift are overcoming selection, there is a substantial chance that a new mutation to the favorable variant A_2 can arise at a segregating site, which has a higher chance of fixation than a neutral variant and hence contributes to an elevated substitution rate. With sufficiently strong mutational bias, $\overline{p}$ can be $>>1/2$, so that the contribution from the enhanced fixation probability of favorable mutations outweighs the lower contribution from the fixation of deleterious mutations.

As was previously shown by McVean and Charlesworth (1999) for the infinite sites model, the equilibrium diversity with selection can also considerably exceed the neutral equilibrium value with the same mutational parameters, when there is a mutational bias toward deleterious alleles (see also Kondrashov *et al.* 2006). For example, in Figure 1, with $\gamma = 5$ and $\beta = 2$, $\pi = 0.43$ but is 0.38 for the neutral case; with $\beta = 20$ and $\gamma = 50$ the values are 0.49 and 0.44,

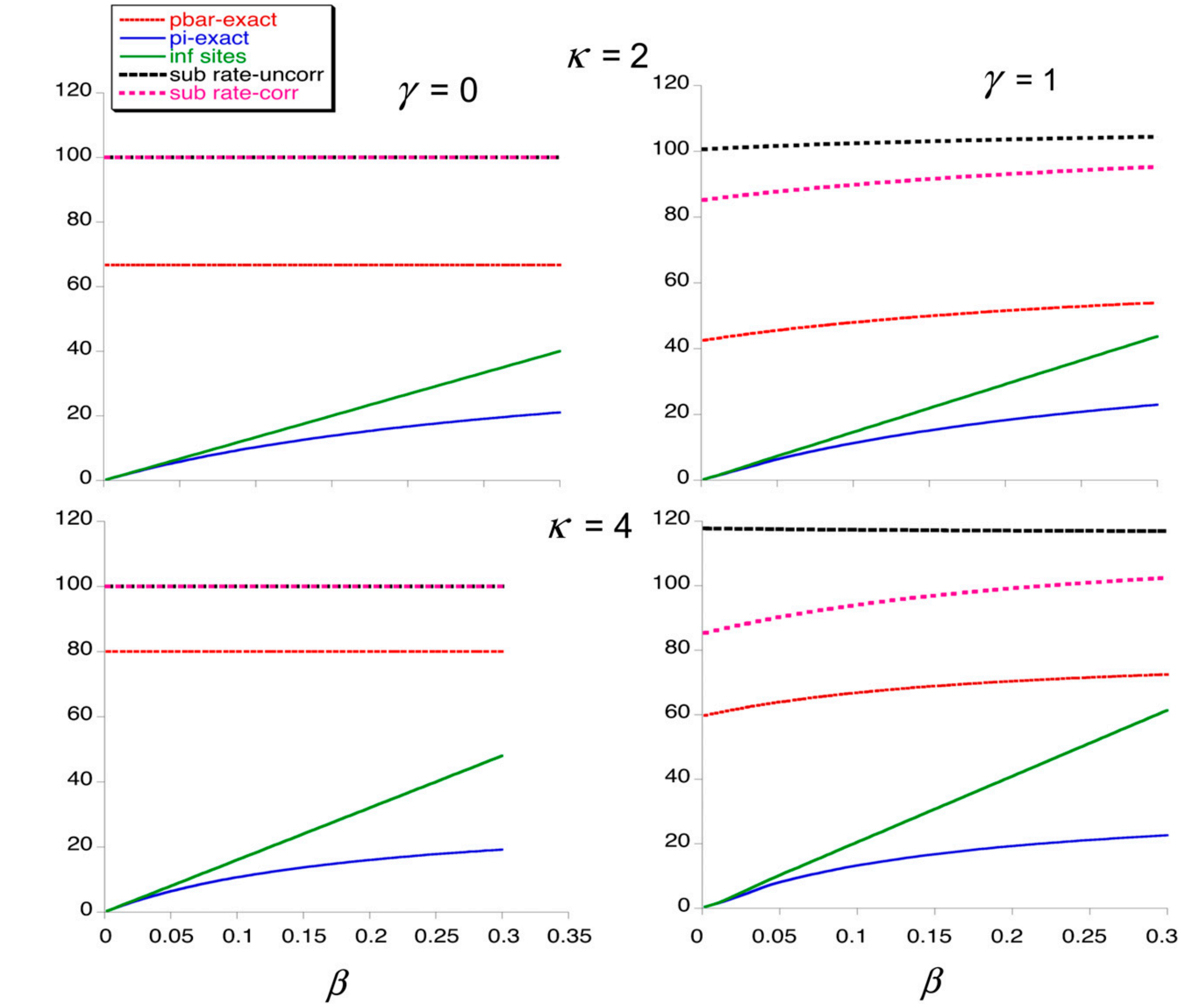

Figure 2 The curves are the values (in percentages) as functions of β for the mean frequency of A_1, $\overline{p}$ (red, dashed curve), π from Equation 10b (blue, solid curve), π as given by the infinite sites model (green, solid curve), the uncorrected rate of substitution relative to neutrality (black, dashed curve), and the corrected rate of substitution relative to neutrality (pink, dashed curve).

respectively. In this case, there is no meaningful way of correcting for differences in base composition between the neutral and selected sites when there are substantial departures from the infinite sites assumption, since the diversity in the neutral case is not related to the mean allele frequency in a simple way.

Properties of a Sample from a Population

Analytic results

This raises the question of the extent to which the properties of a sample of alleles from a population are affected by deviations from the infinite sites assumption. With the above model, the probability that a sample of n alleles segregates for k A_2 variants at a site and $n - k$ copies of A_1 can be obtained from the corresponding binomial distribution with parameter q, integrated over $\phi(q)$, takes the form

$$p(k) = \binom{n}{k} C \int_0^1 \exp(\gamma q)\ \ (1-q)^{\alpha+n-k-1} q^{\beta+k-1} \mathrm{d}q, \tag{16a}$$

where C is given by Equation 3 (McVean and Charlesworth 1999; Desai and Plotkin 2008).

Using the properties of the confluent hypergeometric function, this yields

$$p(k) = \binom{n}{k} \frac{{}_1F_1(\beta+k, \alpha+\beta+n, \gamma)(\beta)_k(\alpha)_{n-k}}{{}_1F_1(\beta, \alpha+\beta, \gamma)(\alpha+\beta)_n}, \quad (0<k<n) \tag{16b}$$

$$p(0) = \frac{{}_1F_1(\beta, \alpha+\beta+n, \gamma)(\alpha)_n}{{}_1F_1(\beta, \alpha+\beta, \gamma)(\alpha+\beta)_n} \quad (16c)$$

$$p(n) = \frac{{}_1F_1(\beta+n, \alpha+\beta+n, \gamma)(\beta)_n}{{}_1F_1(\beta, \alpha+\beta, \gamma)(\alpha+\beta)_n}. \quad (16d)$$

The proportion of sites that are observed to be segregating is

$$p_{\text{seg}} = 1 - p(0) - p(n). \quad (16e)$$

The conditional site frequency spectrum (SFS) for segregating sites can be obtained by dividing Equation 16b by Equation 16e. The folded SFS for segregating sites [which describes the numbers of variants of either type at frequencies 1–0.5n + 1 (n odd) or 0.5n (n even)] can also easily be obtained.

Equations 16a–16e can readily be used to obtain the theoretical values of standard sample statistics, such as the diversity per site (π) (Tajima 1983), Watterson's $\theta_{\text{w}} = p_{\text{seg}}/a_n$ (Watterson 1975), and Tajima's D (Tajima 1989b), using the standard formulas for these quantities. A well-known problem with Tajima's D is the fact that its magnitude is strongly dependent on both the level of variability in the population and the length of sequence used to estimate it (Tajima 1989b). Langley *et al.* (2014) proposed the use of the summary statistic $\Delta_\pi = (\pi - \theta_{\text{w}})/\theta_{\text{w}}$ for measuring the extent of departure of the SFS from the infinite sites neutral equilibrium expectation, which should not suffer from these problems. Another summary statistic for this purpose is provided by the proportion of singleton variants among segregating sites, given by

$$p_{\text{sn}} = \frac{[p(1) + p(n-1)]}{p_{\text{seg}}}. \quad (16f)$$

(This is closely related to the widely used D statistic of Fu and Li 1993.)

Numerical results

Use of the series expression for the confluent hypergeometric function allows rapid computation of all relevant statistics; to avoid overflow when γ is large, however, it is necessary to use logarithms of the individual terms and partial sums of the series (which can be done, since the selection model is defined such that $\gamma > 0$). A FORTRAN program is available on request to B. Charlesworth.

Table 1 displays some examples of such computations, for the case of a mutational bias of 2 toward deleterious mutations, for a subset of the parameter values used in Figure 1. The expected π values are not shown, since these are the same as the population diversities given in Figure 1. Figure 3 show the folded SFSs for some chosen examples, using a sample size of 20 alleles. It can be seen that a high β value (20) means that the proportion of sites that are found to be segregating (p_{seg}) is effectively 100%, even for γ as high as 50 and a sample size (n) of 20. A moderate β value (0.2) behaves similarly in the neutral case with a sample size of 200, but otherwise is associated with a p_{seg} of <80% (p_{seg} is as low as 13% for n = 20 and γ = 50). With neutrality or weak selection ($\gamma \leq 5$), moderate or high values of β cause a distortion of the SFS toward a much lower proportion of singletons (p_{sn}) and higher Tajima's D and Δ_π than is expected with the infinite sites model. Even for γ = 50, a very low p_{sn} and a positive D are found when β = 20. This reflects the tendency of high β values to push the distribution of q toward intermediate frequencies, which has long been known (Wright 1931). Some analytical approximations for p_{seg} are derived in supporting information, File S1.

Discussion

The results described above have some important implications for the interpretation of data on DNA sequence variation and evolution when there is hyperdiversity; *i.e.*, the scaled mutational parameter (β in the notation used here) is sufficiently large that the infinite sites model does not accurately describe patterns of variation within populations. Recent surveys of DNA sequence polymorphisms show that that such hyperdiversity is more common than previously thought, even in multicellular organisms (Cutter *et al.* 2013). In addition, given the evidence from studies of organisms like *Arabidopsis thaliana* and maize that epigenetic variants such as methylated cytosines can be transmitted fairly stably through meiosis, but have origination and disappearance rates that are several orders of magnitude higher than those of nucleotide variants (Johannes *et al.* 2009; Becker *et al.* 2011; Schmitz *et al.* 2011; Lauria *et al.* 2014), the patterns described above are relevant to population-level studies of some classes of epigenetic variants.

Distortion of the SFS with hyperdiversity

As was pointed out ~20 years ago in the context of human mitochondrial DNA sequence variability (Bertorelle and Slatkin 1995; Aris-Brosou and Excoffier 1996; Tajima 1996; Yang 1996), a major effect of a high scaled mutation rate (β in the notation used here) is that more intermediate-frequency variants will be present at polymorphic sites in a sample from a population than under the equilibrium infinite sites model. In particular, for a stationary population at equilibrium between drift and the input of neutral or nearly neutral mutations, the expected values of Tajima's D statistic (D_{T}) and the Δ_π statistic proposed by Langley *et al.* (2014) are positive rather than slightly negative or zero, respectively, as expected under the infinite sites model (Tajima 1989b)—see Figure 1 and Table 1. This reflects the fact that the expected value of the pairwise diversity per site (π) is greater than the expected value of the measure of diversity based on the number of segregating sites at a locus (θ_{w}). As can be seen from Table 1, this effect is quite noticeable even for β as low as 0.02 when selection is absent or weak, and small positive values of D_{T} and Δ_π are found with neutrality even when β = 0.002 (of the order of 1% with n = 200).

Table 1 Sample statistics for the reversible mutation model (κ = 2)

		n = 20				n = 200			
	β	p_{seg}	p_{sn}	D_{T}	Δ_{π}	p_{seg}	p_{sn}	D_{T}	Δ_{π}
γ = 0	0.02	0.088	0.287	0.038	0.016	0.142	0.159	0.089	0.040
	0.2	0.533	0.219	0.322	0.109	0.728	0.086	0.745	0.326
	2.0	0.966	0.062	1.181	0.399	1.000	0.001	1.252	1.226
	20	0.999	0.007	1.637	0.553	1.000	0.000	3.570	1.567
γ = 0.5	0.02	0.094	0.289	0.029	0.009	0.152	0.159	0.076	0.034
	0.2	0.559	0.213	0.348	0.117	0.753	0.080	0.848	0.373
	2.0	0.970	0.055	1.248	0.421	1.000	0.001	2.924	1.284
	20	0.999	0.007	1.649	0.556	1.000	0.000	3.586	1.574
γ = 5.0	0.02	0.071	0.441	−0.654	−0.226	0.146	0.228	−0.843	−0.380
	0.2	0.511	0.319	−0.241	−0.081	0.787	0.104	−0.032	−0.014
	2.0	0.990	0.025	0.950	0.532	1.000	0.000	3.443	1.511
	20	0.999	0.004	1.755	0.592	1.000	0.000	3.722	1.634
γ = 50	0.02	0.014	0.845	−1.539	−0.586	0.063	0.478	−1.820	−0.851
	0.2	0.129	0.795	−1.650	−0.564	0.478	0.351	−1.826	−0.806
	2.0	0.744	0.408	−0.967	−0.327	0.998	0.005	−0.171	−0.169
	20	1.000	0.000	1.329	0.736	1.000	0.000	4.270	1.875

p_{seg} is the proportion of sites that are segregating, p_{sn} is the proportion of singletons among segregating sites in a sample of size n, D_{T} is the mean of Tajima's D for a sequence of 450 bp, and $\Delta_{\pi} = (\pi - \theta_{\mathrm{w}})/\theta_{\mathrm{w}}$, where $\theta_{\mathrm{w}} = p_{\mathrm{seg}}/a_n$ and $a_n = 1 + 1/2 + \ldots + 1/(n - 1)$. All statistics were calculated from Equations 16a–16f.

With very high values of β, a positive Tajima's D can occur even with quite strong purifying selection (a scaled selection parameter γ of 50), as shown in Table 1. A SFS with an excess of intermediate-frequency variants at loci across the genome is usually interpreted as indicating a recent population bottleneck or a subdivided population (*e.g.*, Staedler *et al.* 2009). False positive results for tests for bottlenecks and/or subdivision may thus be obtained if infinite sites rather than finite sites models are applied to hyperdiverse populations or epigenetic variation, even when moderately strong purifying selection is acting. Given that very small positive mean values across sites of statistics such Δ_{π} can be statistically significant with genomic-scale data and large sample sizes, caution should be used when applying infinite sites predictions to such data sets. The suggested criterion for hyperdiversity of π or θ_{w} of 5% for using finite sites models rather than the infinite sites model (Cutter *et al.* 2013) may be too high for such data.

Relation to the data

This raises the question of whether there is indeed evidence for the expected pattern of a skew of the SFS spectrum toward intermediate-frequency variants. In the study of *Caenorhabditis* sp. 5 by Cutter *et al.* (2012), where the within-population diversity at synonymous sites is ~0.08, Tajima's D values for "scattered" samples [where one allele per locus was sampled from each of 13 locations, to minimize departure from the standard coalescent process (Wakeley 2000)] were nearly all positive, with a mean of 0.28. This is consistent with the coalescent simulations of Cutter *et al.* (2012), who used the SIMCOAL2 program of Laval and Excoffier (2004) with a finite sites model with equal mutation rates among all four possible nucleotides (A. Cutter and L. Excoffier, personal communication). The model used here gives an expected value of Tajima's D of ~0.10 with γ = 0 or 0.5 and a mutational bias of 2, assuming a sample size of 13 and 150 bp per locus (corresponding approximately to the numbers of synonymous sites in the study). At least qualitatively, this species thus fits the expectation under hyperdiversity for DNA sequence variability.

In contrast, the synonymous SFS in the much more hyperdiverse species *Caenorhabditis brenneri* is biased toward low-frequency variants, with a mean Tajima's D of –0.56 over 23 loci with an average of ~150 bp per locus (Dey *et al.* 2103, table S3), again using scattered sampling. Similarly, in the only detailed survey of epigenomic variation published to date, that of ~200 northern European accessions of *A. thaliana* (Schmitz *et al.* 2013, supplementary table 9), the SFS for single methylated *vs.* nonmethylated cystosines is also highly skewed toward low-frequency variants. The lack of linkage disequilibrium between this class of variants and SNPs suggests that these epigenetic variants are not caused by nucleotide site variants associated with methylation status, but represent true heritable epiallelic variation (Schmitz *et al.* 2013).

There are several possible reasons for this sharp disagreement between the theoretical predictions and these observations. One is that demographic effects, such as a recent population expansion, mean that predictions based on the assumption of a stationary population are overwhelmed by the well-known excess of rare variants associated with expansion (Tajima 1989a; Slatkin and Hudson 1991). This is ruled out for the case of epigenetic variation in *A. thaliana*, because the SFS for SNPs is far less biased toward rare variants (Schmitz *et al.* 2013), but remains possible for *C. brenneri*. The second possibility is that purifying selection is sufficiently strong to skew the SFS toward rare variants. This seems unlikely in the case of *C. brenneri*, where the estimates of the overall γ for synonymous sites suggest a value close to 0.5 (Dey *et al.* 2103), which is

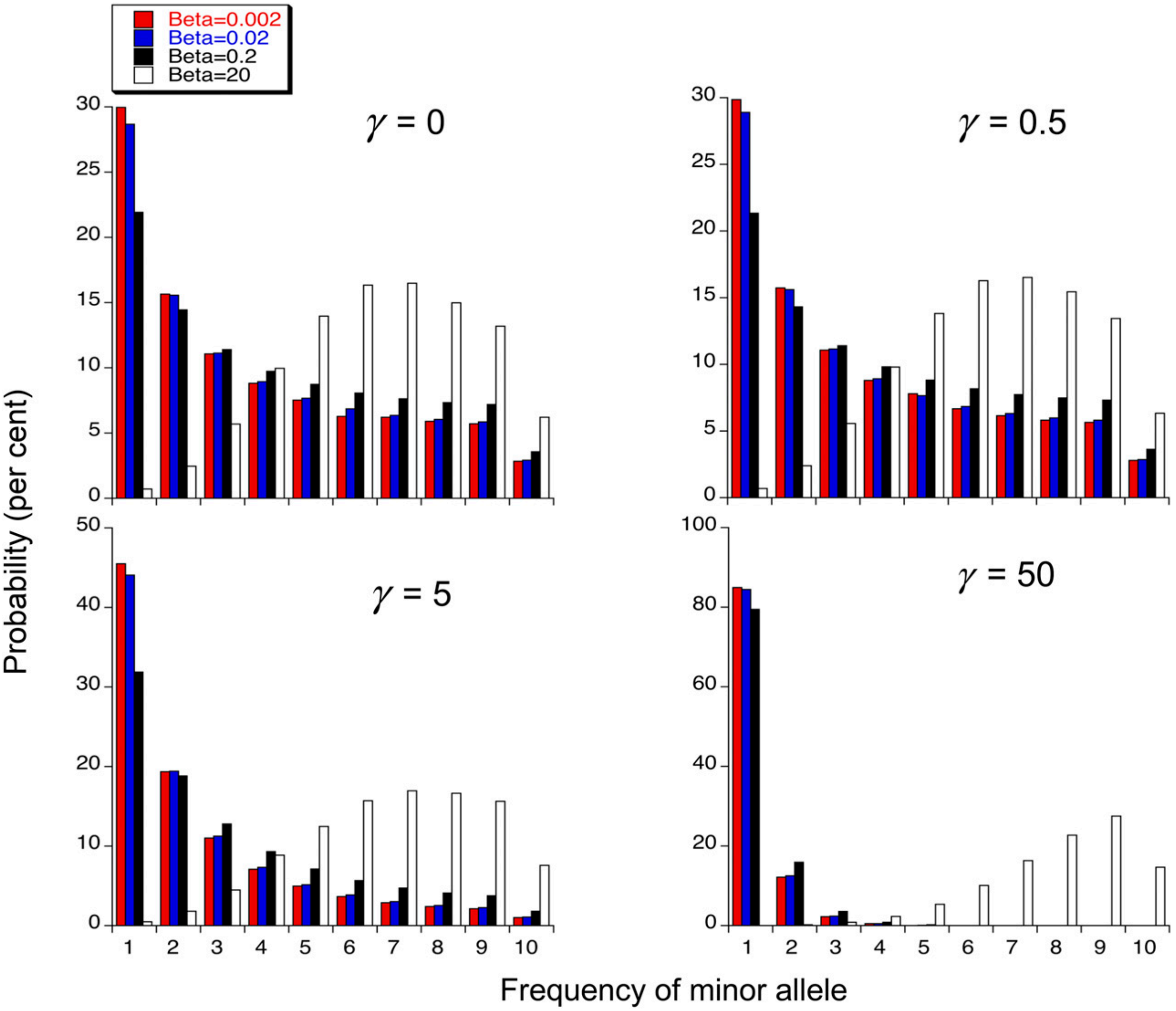

Figure 3 The vertical bars are the values (in percentages) of the probabilities of finding the minor allele in a sample of 20 at the frequencies indicated on the *x*-axis, for different values of β and γ.

insufficient to cause a skew toward rare variants (see Table 1). This explanation is more plausible for the *A. thaliana* example, since high levels of methylation of cytosines are nonrandomly distributed across the genome and are especially prevalent in transposable element sequences where methylation is important for their silencing (Schmitz *et al.* 2013). It is therefore very likely that the methylated states in such sequences are favored by selection. Another possibility is that methylation is selectively neutral, and the differences between genomic regions simply reflect different levels of mutational bias, either toward or against methylation. Calculations using the biallelic model show that extreme mutational bias at neutral or nearly neutral sites can overcome the skew of the SFS toward intermediate-frequency variants (results not shown). The published results of mutation accumulation experiments in *A. thaliana* (Becker *et al.* 2011; Schmitz *et al.* 2011) do not shed much light on the question of the extent of the direction and magnitude of mutational bias, since the experimental design ascertains sites for which at least one of the mutation accumulation lines contains a methylated cytosine at the site in question. It is thus strongly biased toward detecting variants at which the original state was methylation, making it hard to determine the rate of mutation toward methylation. Distinguishing between these possible interpretations is a challenging task and will require the use of numerical models that incorporate past population size changes and population structure.

Limitations of the biallelic model

It is important to note that the biallelic model used here, which is similar to that used by Bertorelle and Slatkin (1995) and Desai and Plotkin (2008), is likely to underestimate the effect of hyperdiversity on the SFS, since the presence of more than two variants at a segregating site will result in higher π but not θ_w. On the other hand, the infinite alleles assumption, apparently used by Aris-Brosou and Excoffier (1996), means that the upper limit to π is 1, whereas in reality there is a maximum of four segregating variants per site, leading to an upper limit to π of 3/4 (when all four variants are present at equal frequencies), as opposed to 1/2 for the biallelic model used here. Given the almost universal existence of mutational biases toward transitions *vs.* transversions and for GC to AT *vs.* AT to GC mutations, the upper limit is in practice likely to be considerably <1, so that the biallelic model with modest mutational bias probably provides a reasonably good guide to the values of measures of skew in the SFS. There has been much discussion of the question of why silent nucleotide site diversity does not a span a much wider range than is commonly seen (see Leffler *et al.* 2012 for a recent account). The saturation of diversity at a level considerably $<3/4$ when there is mutational bias may well be a contributing factor.

An intermediate situation is provided by assuming a K-allele model (Ewens 2004, pp. 192–200) with $K = 4$, corresponding to equal mutation rates among all four nucleotide states at a site (Tajima 1996; Yang 1996; Desai and Plotkin 2008). Under neutrality the exchangeability of the different nucleotides under this model means that the probability density $\phi(q_i)$ for the frequency q_i of a variant of type i (i = 1–4) is proportional to $(1-q_i)^{\theta-1}q_i^{(\theta/3)-1}$, where θ is the net mutation rate per site; *i.e.*, $\phi(q_i)$ follows a β-distribution with parameters θ and $\theta/3$ (Tajima 1996). With semidominant selection with type i having a selective advantage s over all other variants, which are assumed to be selectively equivalent to each other, this expression is simply multiplied by $\exp(\gamma q_i)$.

Following Tajima (1996), these assumptions allow simple analytical formulas for the sample statistics used above to be obtained for the case of neutrality: $\pi = \theta/[1 + (4\theta/3)]$, $p_{\mathrm{seg}} = 1 - [S_{n-1}(\theta/3)/S_{n-1}(4\theta/3)]$, and $p_{\mathrm{sn}} = n\theta S_{n-2}(\theta)/p_{\mathrm{seg}}S_{n-1}(4\theta/3)$, where $S_k(x) = (1+x)(2+x)\ \ldots\ (k+x)$. These can be compared with the statistics obtained from the biallelic model in Table 1, setting θ to the equilibrium infinite sites neutral diversity with reverse mutation $2\beta\kappa/(1 + \kappa) = 4\beta/3$ (with $\kappa = 2$) to obtain comparable net scaled mutation rates per site. As expected, for very low θ, the two models yield similar results, but even with $\beta = 0.02$ the four-allele model gives noticeably higher expected values of Tajima's D and Δ_π; *e.g.*, with a sample size of 20 and $\beta = 0.02$, the values of Tajima's D and Δ_π are 0.069 and 0.022, respectively, *vs.* 0.038 and 0.016 for the biallelic model. With a sample size of 20 and $\beta = 0.2$, the values of Tajima's D and Δ_π for the four-allele model are 0.61 and 0.20, respectively, compared with 0.32 and 0.11 for the biallelic model; values of D and Δ_π much greater than twice the biallelic values can be generated by the four-allele model when β is large, reaching 4.7 and 1.6, respectively, with $\beta = 20$. The proportion of singletons behaves rather differently under the four-allele model; it can even increase with β up to some upper limit, after which it declines and is always higher than for the biallelic model (*e.g.*, 0.36 *vs.* 0.22, respectively, for $\beta = 0.2$ and $n = 20$; and 0.17 *vs.* 0.01 for $\beta = 20$). This behavior presumably reflects the fact that there are four possible variants at each site that can behave as singletons in the case of the four-allele model, and the above formula simply sums over the probabilities that each one of these is a singleton, regardless of the status of the other three possible variants at the same site. A statistic such as Δ_π is thus probably a better summary of the skew of the SFS than the proportion of singletons when a substantial fraction of polymorphic sites segregate for more than two variants, unless variants are collapsed into biallelic alternatives such as GC *vs.* AT base pairs (for example, Evans *et al.* 2014).

For studying situations with multiple alleles per nucleotide and nonequilibrium demography, numerical methods such as that of Zeng (2010) will be needed.

Some other implications

One difficulty with interpreting the results of population surveys of epiallelic variation is that it is impossible to know whether sites that lack epigenetic marks in all individuals sampled are potentially capable of acquiring them. This means that the denominator in per-site statistics such as π and θ_{w} is unknown, making it hard to apply standard population genetics methods to these kinds of data. Fortunately, however, with high β values (>0.2), nearly all sites capable of mutation will be found to be segregating in a large sample, even with a scaled selection strength as high as $\gamma = 5$; thus, the majority of sites capable of epimutations can be identified from population surveys, unless strong purifying selection is acting. Population surveys could, therefore, be a valuable tool for the characterization of the epigenome.

Another finding that is relevant for both hyperdiverse DNA sequence variation and hypermutable epigenetic variation is the fact that substitution rates for sites under purifying selection may be close to or even greater than rates at neutral sites with high β values. As described above, this may occur even after correcting for the effects of differences in base composition between neutral and selected sites (Figure 1 and Figure 2). This lack of sensitivity of substitution rates to the strength of purifying selection is consistent with the patterns described by Cutter *et al.* (2013), where there is only a weak relation between codon usage bias and a measure of synonymous site divergence in the hyperdiverse species *Ciona savigni*. Similarly, diversity at sites subject to weak purifying selection is expected to show a nonlinear pattern of relationship with γ, such that π increases with γ when sites are close to neutral and then declines again as γ approaches or exceeds 1; the range of γ values over which there is an increase is broader for large β (Figure 2). Synonymous diversity of genes in *C. brenneri* does indeed show a quadratic relation with the frequency of optimal codons, such that genes with ~50% optimal codons have the highest diversity values (A. Cutter, personal communication).

With γ values typical of those reported from studies of selection on codon usage ($\gamma \leq 1$), the standard Li–Bulmer equation (Li 1987; Bulmer 1991) tends to overestimate the expected level of codon bias, as measured by the mean frequency of the favored allelic type ($\overline{q}$), when $\beta > 0.02$. For example, with $\gamma = 1$ and $\beta = 0.2$, the exact value of $\overline{q}$ from Equation 5a is 0.49 compared with the Li–Bulmer infinite sites prediction of 0.58, while the second-order approximation from Equations 7a and 7b gives 0.44. Analyses of codon usage in hyperdiverse species that use codon usage data to estimate γ (see Sharp *et al.* 2010) should probably use the exact expression. It is interesting in this context to note that there is only a small difference in the mean level of codon usage bias between *C. brenneri* and *C. remanei*, despite an approximately threefold difference in synonymous site diversity (A. Cutter, personal communication). This raises the question of whether the purifying selection model used here is appropriate for codon usage or whether a model of stabilizing selection (Kimura 1981) is more realistic, since the latter means that γ is insensitive to N_e over a wide range of parameter values, provided that there is mutational bias (Charlesworth 2013). The behavior of this model with hyperdiversity would, therefore, be worth studying.

Acknowledgments

We thank Deborah Charlesworth, Asher Cutter, Kai Zeng, and two anonymous reviewers for their comments on the manuscript and Asher Cutter and Laurent Excoffier for valuable discussions.

Literature Cited

Abramowitz, M., and I. A. Stegun, 1965 *Handbook of Mathematical Functions*. Dover, New York.

Aris-Brosou, S., and L. Excoffier, 1996 The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. 13: 494–504.

Becker, C., J. Hagmann, J. Mueller, D. Koenig, O. Stegle *et al.*, 2011 Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature 480: 245–249.

Bertorelle, G., and M. Slatkin, 1995 The number of segregating sites in expanding human populations with implications for estimates of demographic parameters. Mol. Biol. Evol. 12: 887–892.

Bulmer, M. G., 1991 The selection-mutation-drift theory of synonomous codon usage. Genetics 129: 897–907.

Charlesworth, B., 2013 Stabilizing selection, purifying selection and mutational bias in finite populations. Genetics 194: 955–971.

Charlesworth, B., and D. Charlesworth, 2010 *Elements of Evolutionary Genetics*. Roberts & Co., Greenwood Village, CO.

Cutter, A. D., G.-X. Wang, H. Ai, and Y. Peng, 2012 Influence of finite-sites mutations, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity. Mol. Ecol. 21: 1345–1359.

Cutter, A. D., R. Jovelin, and D. Dey, 2013 Molecular hyperdiversity and evolution in very large populations. Mol. Ecol. 22: 2074–2095.

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

Dey, A., C. K. W. Chan, C. G. Thomas, and A. D. Cutter, 2013 Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. Proc. Natl. Acad. Sci. USA 110: 11056–11060.

Eory, L., D. L. Halligan, and P. D. Keightley, 2010 Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol. Biol. Evol. 27: 177–192.

Evans, B. J., K. Zeng, J. A. Esselstyn, B. Charlesworth, and D. J. Melnick, 2014 Reduced representation genome sequencing suggests low diversity on the sex chromosomes of Tonkean macaque monkeys. Mol. Biol. Evol. 31: 2425–2440.

Ewens, W. J., 2004 *Mathematical Population Genetics. 1. Theoretical Introduction*. Springer-Verlag, New York.

Eyre-Walker, A., 1992 The effect of constraint on the rate of evolution in neutral models with biased mutation. Genetics 131: 233–234.

Fisher, R. A., 1922 On the dominance ratio. Proc. R. Soc. Edinb. 42: 321–341.

Fisher, R. A., 1930 The distribution of gene ratios for rare mutations. Proc. R. Soc. Edinb. 50: 205–220.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Grossniklaus, U., B. Kelly, A. Ferguson-Smith, M. Pembry, and S. Lindquist, 2013 Transgenerational epigenetic inheritance: How important is it? Nat. Rev. Genet. 14: 228–235.

Haldane, J. B. S., 1927 A mathematical theory of natural and artificial selection. Part V. Selection and mutation. Proc. Camb. Philos. Soc. 23: 838–844.

Halligan, D. L., A. Eyre-Walker, P. Andolfatto, and P. D. Keightley, 2004 Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. Genome Res. 14: 273–279.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7: 1–45.

Jenkins, P. A., and Y. S. Song, 2011 The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. Theor. Popul. Biol. 80: 158–173.

Jenkins, P. A., J. W. Mueller, and Y. S. Song, 2014 General triallelic frequency spectrum under demographic models with variable population size. Genetics 196: 295–311.

Johannes, F., E. Porcher, F. Texeira, V. Saliba-Colombani, M. Simon *et al.*, 2009 Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet. 6: e1000530.

Kimura, M., 1962 On the probability of fixation of a mutant gene in a population. Genetics 47: 713–719.

Kimura, M., 1968 Evolutionary rate at the molecular level. Nature 217: 624–626.

Kimura, M., 1971 Theoretical foundations of population genetics at the molecular level. Theor. Popul. Biol. 2: 174–208.

Kimura, M., 1981 Possibility of extensive neutral evolution under stabilizing selection with special reference to non-random usage of synonymous codons. Proc. Natl. Acad. Sci. USA 78: 454–458.

Kimura, M., T. Maruyama, and J. F. Crow, 1963 The mutation load in small populations. Evolution 48: 1303–1312.

Klironomos, F., J. Berg, and S. Collins, 2013 How epigenetic mutations can affect genetic evolution: model and mechanism. BioEssays 35: 571–578.

Kondrashov, F. A., A. Y. Ogurtsov, and A. S. Kondrashov, 2006 Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. J. Theor. Biol. 240: 616–626.

Langley, S. A., G. H. Karpen, and C. H. Langley, 2014 Nucleosomes shape DNA polymorphism and divergence. PLoS Genet. 10: e1004457.

Lauria, M., S. Piccinni, R. Pirona, G. Lund, A. Viotti *et al.*, 2014 Epigenetic variation, inheritance, and parent-of-origin effects of cytosine methylation in maize (*Zea mays*). Genetics 196: 653–666.

Laval, G., and L. Excoffier, 2004 SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20: 2485–2487.

Lawrie, D. S., D. A. Petrov, and P. W. Messer, 2011 Faster than neutral evolution of constrained sequences: the complex interplay of mutational biases and weak selection. Genome Biol. Evol. 3: 383–395.

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel *et al.*, 2012 Revisiting an old riddle: What determines genetic diversity levels within a species? PLoS Biol. 10: e1001388.

Li, W.-H., 1987 Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. J. Mol. Evol. 24: 337–345.

McVean, G. A. T., and B. Charlesworth, 1999 A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74: 145–158.

Mizawa, K., and F. Tajima, 1997 Estimation of the amount of genetic variation when the mutation rate varies among sites. Genetics 147: 1959–1964.

Richards, E. J., 2006 Inherited epigenetic variation: revisiting soft inheritance. Nat. Rev. Genet. 7: 395–401.

Sargsyan, O., 2014 A framework including recombination for analyzing the dynamics of within-host HIV genetic diversity. PLoS Genet. 9: e87655.

Schmitz, R. J., and J. R. Ecker, 2012 Epigenetic and epigenomic variation in *Arabidopsis thaliana*. Trends Plant Sci. 17: 149–154.

Schmitz, R. J., M. D. Schultz, M. G. Lewsey, R. C. O'Malley, M. A. Urich *et al.*, 2011 Transgenerational epigenetic instability is a source of novel methylation variants. Science 334: 369–373.

Schmitz, R. J., M. D. Schultz, M. A. Urich, J. P. Nery, M. Pelizola *et al.*, 2013 Patterns of population epigenomic diversity. Nature 495: 193–198.

Sharp, P. M., L. B. Emery, and K. Zeng, 2010 Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. B 365: 1203–1212.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 143: 579–587.

Staedler, T., B. Haubold, C. Merino, W. Stephan, and P. Pfaffelhuber, 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182: 205–216.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in a finite population. Genetics 105: 437–460.

Tajima, F., 1989a The effect of change in population size on DNA polymorphism. Genetics 123: 597–601.

Tajima, F., 1989b Statistical method for testing the neutral mutation hypothesis. Genetics 123: 585–595.

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.

Wakeley, J., 2000 The effects of subdivision on the genetic divergence of populations and species. Evolution 54: 1092–1101.

Wakeley, J., 2008 *Coalescent Theory. An Introduction*. Roberts & Co., Greenwood Village, CO.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

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

Wright, S., 1937 The distribution of gene frequencies in populations. Proc. Natl. Acad. Sci. USA 23: 307–320.

Yang, Z., 1996 Statistical properties of a DNA sample under the finite-sites model. Genetics 144: 1941–1959.

Zeng, K., 2010 A simple multiallele model and its application to identifying preferred–unpreferred codons using polymorphism data. Mol. Biol. Evol. 27: 1327–1337.

Communicating editor: M. A. Beaumont

Appendix

The Expected Number of New Mutations That Arise at a Segregating Site

Assume that we have a nucleotide site that is segregating for a neutral mutation that arose at an initial frequency of $1/(2N)$. Let the probability that this variant mutates to an alternative nucleotide be u per generation (this includes the possibility that it reverts to the ancestral state); let the probability that the ancestral variant mutates to another state be v (this includes the possibility that the mutation is identical in state to the variant that is already segregating). If the frequency of the mutation in the population in a given generation is x, the expected total number of mutational events is $2N[ux + v(1 - x)]$. The expected time that the original mutation spends in the frequency interval x to $x + \mathrm{d}x$ is given approximately by $4N_e/(1 - x)$ for $0 < x \leq 1/(2N)$ and $2N_e/(Nx)$ for $1/(2N) < x \leq 1$ (Ewens 2004, p. 160). The total expected number of new mutations that arise during the sojourn of the mutation in the population is thus

$$4N_{\mathrm{e}}\left\{\int_0^{1/(2N)} \frac{2[ux+v(1-x)]}{(1-x)}\,\mathrm{d}x + \int_{1/(2N)}^{1} \frac{[ux+v(1-x)]}{Nx}\,\mathrm{d}x\right\} \approx 4N_{\mathrm{e}}[u+v\ln(2N)]. \tag{A1}$$

Approximations to Equations 5a and 5b with Small α and β

Equation 5a is equivalent to

$$\bar{q} = \frac{\left[1+\sum_{i=1}^{\infty}\left(\gamma^i(\beta+1)_i/i!(\alpha+\beta+1)_i\right)\right]}{\left[1+\kappa+\gamma+\sum_{i=2}^{\infty}\left(\gamma^i(\beta+1)_{i+1}/i!(\alpha+\beta+1)_{i+1}\right)\right]}. \tag{A2}$$

We can write terms of the form $(\beta + i - j)/(\alpha + \beta + i - j)$ as $1 - [\beta\kappa/(i - j)] + O(\beta^2)$; keeping only $O(\beta)$ terms, we have

$$\bar{q} \approx \frac{\left\{1+\sum_{i=1}^{\infty}(\gamma^i/i!)\prod_{j=1}^{i}[1-\beta\kappa/(i+1-j)]\right\}}{\left\{1+\kappa+\gamma+\sum_{i=2}^{\infty}(\gamma^i/i!)\prod_{j=1}^{i-1}[1-\beta\kappa/(i-j)]\right\}} \tag{A3a}$$

or

$$\bar{q} \approx \frac{\{1+\sum_{i=1}^{\infty}(\gamma^i/i!)\exp(-\beta\kappa a_{i+1})\}}{\{1+\kappa+\gamma+\sum_{i=2}^{\infty}(\gamma^i/i!)\exp(-\beta\kappa a_i)\}}, \tag{A3b}$$

where $a_i = 1 + 1/2 + 1/3 + \ldots 1/(i - 1)$ $(i \geq 2)$. The exponential terms in the numerator and denominator of Equation A3b can thus be replaced by $1 + O(\beta)$, yielding Equation 6 of the main text.

In Equations 7a and 7b, $a_{i+1} < i$ for $i > 1$, the first summation in the numerator of g is less than the sum of $\gamma^{\,i}(i - 1)!$, so that the sum is $<\gamma \exp(\gamma)$. Similarly, $a_{i+1} - a_i = 1/i$, so that the second summation is $<\exp(\gamma) - (1 + \gamma)$. It follows that g is positive and $<\kappa\gamma + \exp(\gamma) - 1$. This is multiplied by $\exp(-\gamma)$ in the numerator of Equation 5a, to obtain the multiplicand of $\beta\kappa$, yielding $\kappa\gamma \exp(-\gamma) + 1 - \exp(-\gamma) < \kappa + 1 - \exp(-\gamma)$. The contribution of $-\beta\kappa g \exp(-\gamma)$ to the numerator of Equation 7a is thus negative and smaller in magnitude than $\beta\kappa$ $(1 + \kappa)$. The leading term in Equation 6 should provide a good approximation when $\beta\kappa$ is ≤ 0.1.

Approximations for the Frequencies of the Fixed Classes

Assuming that $\alpha << 1$ and $\beta << 1$, and employing the approximations used in Equation A3b, we find that

$$f_{1\mathrm{f}} \approx \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left\{\frac{1+\kappa}{[1+\kappa+\gamma+\sum_{i=2}^{\infty}(\gamma^i/i!)\exp(-\beta\kappa a_i)]}+O(\beta^2)\right\}\times\ \beta^{-1}(2N)^{-\beta}[1+O(\gamma N^{-1})+O(\beta N^{-1})]. \tag{A4a}$$

Similarly,

$$f_{2\mathrm{f}} \approx \frac{\Gamma(\alpha+\beta)\exp(\gamma)}{\Gamma(\alpha)\Gamma(\beta)}\left\{\frac{1+\kappa}{[1+\kappa+\gamma+\sum_{i=2}^{\infty}(\gamma^i/i!)\exp(-\beta\kappa a_i)]}+O(\beta^2)\right\} \\ \times\ (\beta\kappa)^{-1}(2N)^{-\beta\kappa}[1+O(\gamma N^{-1})+O(\beta N^{-1})]. \tag{A4b}$$

We can use the fact that $\Gamma(1 + x) = x\Gamma(x)$ to write $\Gamma(x) = \Gamma(1 + x)/x$. For small x, the representation of the gamma function as an infinite product (Abramowitz and Stegun 1965) implies that $\Gamma(1 + x) = (1 - cx) + O(x^2)$, where $c \approx 0.577$ is Euler's constant, and similarly for the other gamma integrals. We can thus approximate $\Gamma(x)$ for small x by $1/x$, and the term involving gamma functions in Equations 9a and 9b is then $\kappa\beta/(1 + \kappa)[1 + O(\beta)]$. Using a similar approximation to that used in Equations 7a and 7b, and neglecting higher-order terms in β, we obtain

$$f_{1f} \approx \frac{\kappa \exp(-\gamma)(2N)^{-\beta}}{[1+\kappa\exp(-\gamma)]}\left\{1+\frac{\beta\kappa h\exp(-\gamma)}{[1+\kappa\exp(-\gamma)]}\right\}[1+O(\gamma N^{-1})+O(\beta N^{-1})] \tag{A4c}$$

$$f_{2f} \approx \frac{(2N)^{-\beta\kappa}}{[1+\kappa\exp(-\gamma)]}\left\{1+\frac{\beta\kappa h\exp(-\gamma)}{[1+\kappa\exp(-\gamma)]}\right\}[1+O(\gamma N^{-1})+O(\beta N^{-1})], \tag{A4d}$$

where

$$h=\sum_{i=2}^{\infty}\frac{\gamma^i}{i!}a_i. \tag{A4e}$$

The higher-order terms in β vanish when $\gamma = 0$, suggesting that these expressions are good approximations when β and γ are both small. More rigorously, for finite i, a_i is less than some constant A, which is approximately equal to $\ln(i - 1)$. If terms in $i > k$ can be neglected in the sum that defines h, $h < \ln(k)\exp(\gamma)$, so that $h \exp(-\gamma) < \beta\kappa \ln(k)$, where $\ln(k)$ is a small multiple of one unless γ is very large.

In addition, for arbitrary γ, the terms involving $(2N)^{-\beta}$ in Equations A4c and A4d are equal to $1- \beta \ln(2N) + O(\beta^2)$ and $1- \beta\kappa \ln(2N) + O(\beta^2)$, respectively. Provided that $\ln(2N)$ is of order one, f_{f1} and f_{f2} are each equal to their respective infinite sites value, multiplied by a factor $1 - O(\beta)$, implying that the infinite sites values provide a good approximation unless $\beta >> 0$.

Fixations of Mutations

Consider first the case of A_2 to A_1 mutations that arise at a site that was initially fixed for A_2. We approximate the frequency of this fixed class, f_{2f}, by the integral in Equation 9b. The fixation probability, Q_1, of an A_1 mutation with initial frequency $1/(2N)$ when N is large is $\gamma (2N)^{-1}[\exp(\gamma) - 1]^{-1} + O[\gamma^2(2N)^{-2}]$, so that the net number of new A_2 mutations that arise in a given generation and are expected to become fixed is $2N\kappa\nu f_{2f}\{\gamma/(2N)[\exp(\gamma) - 1]^{-1} + O[\gamma^2(2N)^{-2}]\} = \kappa\nu f_{2f}\{\gamma [\exp(\gamma) - 1]^{-1} + 2N\, O[\gamma^2(2N)^{-2}]\}$. Using the same approximation for Q_1 and the fact that q is close to one in Equation 9b, the corresponding formula from Equations 13 and 14a is

$$\kappa\nu\int_{1-1/(2N)}^{1}Q_1(p)p^{-1}q\phi(q)\mathrm{d}q=\kappa\nu\left\{\gamma[\exp(\gamma)-1]^{-1}\int_{1-1/(2N)}^{1}\phi(q)\mathrm{d}q+O[\gamma^2(2N)^{-2}]\right\}. \tag{A5}$$

Provided that $2N$ is sufficiently large in relation to γ, so that the higher-order terms in $\gamma(2N)^{-1}$ can be ignored, the two results are equivalent.

The following argument can be used for the other end of the frequency range. In this case, there is no contribution from the class fixed for A_1 mutations, whose frequency is f_{1f} as given by Equation 9a, to the fixation of new A_1 mutations. The corresponding formula from Equations 13 and 14a is

$$\kappa\nu\int_{0}^{1/(2N)}Q_1(p)p^{-1}q\ \phi(q)\ \mathrm{d}q=\kappa\nu\ \ (2N)^{-1}\left\{1+O[(1+\gamma)(2N)^{-1}]\right\}. \tag{A6}$$

Again, provided that $2N$ is sufficiently large in relation to γ, the two results are equivalent. Parallel arguments can be used for the fixation of new A_2 mutations.

The Relative Rate of Substitution Under the Infinite Sites Assumption

At equilibrium between mutation, drift, and selection, the frequencies of sites fixed for A_1 and A_2 under the infinite sites model are approximated by $\kappa \exp(-\gamma)/[1 + \kappa \exp(-\gamma)]$ and $1/[1 + \kappa \exp(-\gamma)]$, respectively (Li 1987; Bulmer 1991; McVean and Charlesworth 1999). Averaging over the contributions from mutations arising at each class of fixed sites, taking into account their respective fixation probabilities, the equilibrium rate of nucleotide substitution is then

$$\lambda(\gamma) = \frac{2\kappa v \gamma}{[1+\kappa \exp(-\gamma)][\exp(\gamma)-1]} \tag{A7a}$$

(Charlesworth and Charlesworth 2010, p. 275).

If we consider neutral mutations arising at fixed sites with the same frequencies of A_1 and A_2 variants as the selected sites (*i.e.*, with the same base composition), the substitution rate is

$$\lambda(0) = \frac{\kappa v[1+\exp(-\gamma)]}{[1+\kappa \exp(-\gamma)]}. \tag{A7b}$$

The ratio $R(\gamma) = \lambda(\gamma)/\lambda(0)$ gives the rate of substitution of selected mutations relative to neutral expectation, conditioning on the same base composition; we have

$$R(\gamma) = \frac{2\gamma}{[\exp(\gamma)-\exp(-\gamma)]}. \tag{A8}$$

It is easily seen that $R = 1$ at $\gamma = 0$ and decreases as γ increases.

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167973/-/DC1

Purifying Selection, Drift, and Reversible Mutation with Arbitrarily High Mutation Rates

Brian Charlesworth and Kavita Jain

DOI: 10.1534/genetics.114.167973

File S1

Purifying selection, drift and reversible mutation with arbitrarily high mutation rates

Supporting Information

Brian Charlesworth and Kavita Jain

Analytical approximations for the fraction of segregating sites

The proportion p_{seg} of segregating sites is given by

$$p_{seg} = 1 - p(n) - p(0) \tag{S1.1}$$

where

$$p(k) = \int_0^1 dq \, \binom{n}{k} q^k (1-q)^{n-k} \phi(q) \tag{S1.2}$$

and

$$\phi(q) = C e^{\gamma q} q^{\beta-1} (1-q)^{\alpha-1} \tag{S1.3}$$

Moments of the frequencies q and p: We first consider $p(n) = \overline{q^n}$ which is given by

$$\begin{aligned} p(n) &= \frac{(\beta)_n}{(\alpha+\beta)_n} \, \frac{{}_1F_1(n+\beta, n+\alpha+\beta, \gamma)}{{}_1F_1(\beta, \alpha+\beta, \gamma)} && \text{(S1.4)} \\ &= \frac{(\beta)_n}{(\alpha+\beta)_n} \, \frac{1 + \sum_{j=1}^{\infty} G_j^{(n)} \frac{\gamma^j}{j!}}{1 + \sum_{j=1}^{\infty} G_j^{(0)} \frac{\gamma^j}{j!}} && \text{(S1.5)} \end{aligned}$$

where

$$G_j^{(n)} = \frac{(n+\beta)_j}{(n+\alpha+\beta)_j} \tag{S1.6}$$

and $(a)_j$ is the Pochhammer's symbol. For $\alpha, \beta \to 0$ with $\kappa = \alpha/\beta$ finite, we have

$$G_j^{(n)} \approx \begin{cases} 1 - \alpha(H_{n+j-1} - H_{n-1}) \ , \ n > 0 \\ \frac{1}{1+\kappa} \ (1 - \alpha H_{j-1}) \ , \ n = 0 \end{cases} \tag{S1.7}$$

where $H_j = \sum_{k=1}^{j} (1/k)$ is the jth Harmonic number. Also, we can write

$$\frac{(\beta)_n}{(\alpha+\beta)_n} \approx \frac{1}{1+\kappa} \ (1 - \alpha H_{n-1}) \tag{S1.8}$$

Substituting the above approximations in the expression for $p(n)$ and keeping terms to order α, we finally obtain

$$p(n) \approx \frac{1}{1+\kappa e^{-\gamma}} \left(1 - \alpha H_{n-1} e^{-\gamma} - \frac{\alpha e^{-\gamma}}{1+\kappa e^{-\gamma}} (S_1(n) + \kappa S_2(n))\right) \tag{S1.9}$$

where

$$S_1(n) = \sum_{j=1}^{\infty} (H_{n+j-1} - H_{j-1}) \frac{\gamma^j}{j!} \overset{\gamma \gg 1}{\sim} \frac{e^{\gamma}}{\gamma} \left(n + c_1 \gamma^{-1}\right) \tag{S1.10}$$

$$S_2(k) = e^{-\gamma} \sum_{j=1}^{\infty} H_{n+j-1} \frac{\gamma^j}{j!} \overset{\gamma \gg 1}{\sim} \ln \gamma \tag{S1.11}$$

We note that the dependence on n appears at order α. Thus in the infinite sites model where these terms are neglected, all the moments of fraction q are equal. Setting $n = 1$ and 2 in the above equations reproduces the results for $\bar{q}$ in (7a) and (7b), and for $\overline{q - q^2}$ in (13) (after dividing by 2) given in the main text. In the neutral case, we have

$$p(n) \approx \frac{1 - \alpha H_{n-1}}{1+\kappa} \tag{S1.12}$$

while in the strong selection limit, using the asymptotic results for the sums $S_1(n)$ and $S_2(n)$, we get

$$1 - p(n) = \frac{1}{1+\kappa^{-1} e^{\gamma}} \left(1 + \frac{\alpha}{\kappa} (H_{n-1} + \frac{n e^{\gamma}}{\gamma})\right) \tag{S1.13}$$

For $\gamma \to \infty$, the above expression shows that $1 - p(n) \to \alpha n/\gamma$.

We next consider $p(0) = \overline{(1-q)^n}$ which is given by

$$p(0) = \frac{(\alpha)_n}{(\alpha+\beta)_n} \frac{{}_1F_1(\beta, n+\alpha+\beta, \gamma)}{{}_1F_1(\beta, \alpha+\beta, \gamma)} \tag{S1.14}$$

For $\alpha, \beta \to 0$ but arbitrary n and j, we can write

$$\frac{(\beta)_j}{(n+\alpha+\beta)_j} \approx \beta \frac{(j-1)!(n-1)!}{(n+j-1)!} \tag{S1.15}$$

Using the above approximation and as before, keeping terms to order α, we find that

$$p(0) \approx \frac{\kappa e^{-\gamma}}{1+\kappa e^{-\gamma}} \left(1 - \frac{\alpha}{\kappa} H_{n-1} + \frac{\alpha(n-1)!}{\kappa} S_3(n) + \frac{\alpha S_2(0)}{1+\kappa e^{-\gamma}}\right) \tag{S1.16}$$

where

$$S_3(n) = \sum_{j=1}^{\infty} \frac{\gamma^j}{j(n+j-1)!} \overset{\gamma \gg 1}{\sim} \frac{e^\gamma}{\gamma^n} \tag{S1.17}$$

In the case of neutrality, we have

$$p(0) = \frac{\kappa - \alpha H_{n-1}}{1+\kappa} \tag{S1.18}$$

and in the strong selection limit, we get

$$p(0) = \frac{1}{1+\kappa^{-1} e^\gamma} \left(1 - \frac{\alpha}{\kappa} (H_{n-1} - \frac{(n-1)! e^\gamma}{\gamma^n})\right) \tag{S1.19}$$

For $\gamma \to \infty$, the fraction $p(0) \to \alpha(n-1)!/\gamma^n$.

Segregating site fraction (p_{seg}): Using the above results, we can now look at the behavior of p_{seg}. For $\gamma = 0$, both $p(0)$ and $1 - p(n)$ contribute equally (in magnitude) to give

$$p_{seg} = \frac{2\alpha H_{n-1}}{1+\kappa} \tag{S1.20}$$

Since $H_n \sim \ln n + \gamma_{EM}$ for large n, the proportion of segregating sites increases logarithmically with the sample size in the neutral case. For $\beta = 0.02$, the above expression gives $p_{seg} = 0.094$ and 0.156 for $n = 20$ and 200 respectively which are close to the data in Table 1 of the main text. In the strong selection limit, for large γ, we have

$$p_{seg} \approx \frac{\alpha n}{\gamma} \tag{S1.21}$$

which increases linearly with the sample size.

One can also look at the $\beta \to \infty$ limit. For the neutral case, we have

$$p_{seg} = 1 - \frac{(\alpha)_n + (\beta)_n}{(\alpha+\beta)_n} \tag{S1.22}$$

For $n \ll \alpha, \beta$, we can write

$$\frac{(\alpha)_n}{(\alpha+\beta)_n} \approx \left(\frac{\kappa}{1+\kappa}\right)^n \tag{S1.23}$$

while for $n \gg \alpha, \beta$, using Stirling's approximation $s! \sim \sqrt{2\pi s}(s/e)^s$, we get

$$\frac{(\alpha)_n}{(\alpha+\beta)_n} \approx \frac{(\alpha+\beta-1)!}{(\alpha-1)!}\ n^{-\beta} \tag{S1.24}$$

Using these approximations, we find that

$$1 - p_{seg} = \begin{cases} \frac{1+\kappa^n}{(1+\kappa)^n}\ ,\ n \ll \alpha, \beta \\ (\alpha+\beta-1)!\ (\frac{1}{(\alpha-1)! n^\beta} + \frac{1}{(\beta-1)! n^\alpha})\ ,\ n \gg \alpha, \beta \end{cases} \tag{S1.25}$$

Thus in small samples (relative to scaled mutation rates), p_{seg} approaches unity exponentially fast while for larger samples, the approach is algebraic.