J Mol Evol (2008) 67:418–426
DOI 10.1007/s00239-008-9146-9

Divergence and Polymorphism Under the Nearly Neutral Theory of Molecular Evolution

John J. Welch · Adam Eyre-Walker · David Waxman

Received: 11 January 2008 / Accepted: 14 July 2008 / Published online: 26 September 2008

Abstract The nearly neutral theory attributes most nucleotide substitution and polymorphism to genetic drift acting on weakly selected mutants, and assumes that the selection coefficients for these mutants are drawn from a continuous distribution. This means that parameter estimation can require numerical integration, and this can be computationally costly and inaccurate. Furthermore, the leading parameter dependencies of important quantities can be unclear, making results difficult to understand. For some commonly used distributions of mutant effects, we show how these problems can be avoided by writing equations in terms of special functions. Series expansion then allows for their rapid calculation and, also, illuminates leading parameter dependencies. For example, we show that if mutants are gamma distributed, the neutrality index is largely independent of the effective population size. However, we also show that such results are not robust to misspecification of the functional form of distribution. Some implications of these findings are then discussed.

J. J. Welch (✉)
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK
e-mail: J.J.Welch@ed.ac.uk

J. J. Welch · A. Eyre-Walker · D. Waxman
Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Falmer, Brighton BN1 9QG, UK

A. Eyre-Walker
e-mail: A.C.Eyre-Walker@sussex.ac.uk;

D. Waxman
e-mail: D.Waxman@sussex.ac.uk

Keywords Genetic drift · Distribution of mutant effects · Neutrality index · Special functions

Introduction

The neutral and nearly neutral theories of molecular evolution placed interpopulation divergence and intrapopulation polymorphism within a single explanatory framework, attributing both to the action of genetic drift (Kimura 1983; Ohta and Gillespie 1996). While alternative theories, emphasizing positive selection, and/or linkage effects, continue to receive attention (e.g., Gillespie 1991, 2001), the drift-based theories remain central to the study of molecular evolution.

Neutral theory, in the strict sense, assumes that most mutants are either strongly deleterious or wholly neutral, with only the latter contributing to divergence or polymorphism. This assumption yields tractable equations, and easily interpretable results, but is almost certainly unrealistic. The nearly neutral theory, by contrast, assumes that selection coefficients are drawn from a continuous range, with a large class of mildly deleterious mutants—an assumption that has a great deal of empirical support (e.g., Eyre-Walker et al. 2002, 2006; Piganeau and Eyre-Walker 2003; Yampolsky et al. 2005; Loewe and Charlesworth 2006; Loewe et al. 2006; Eyre-Walker and Keightley 2007).

When selection coefficients are drawn from a continuous distribution, the equations for many quantities of interest involve an integral over this distribution, and this has some unfortunate consequences. First, on a purely practical level, when equations are implemented in a likelihood framework for parameter estimation, the estimation procedure must involve numerical integration, and

this can impose a significant computational burden, especially when multidimensional integrals are required (e.g., Nielsen and Yang 2003; Williamson et al. 2004). Second, the inclusion of a continuous distribution makes the equations more difficult to understand, and the predicted parameter dependencies of important quantities less transparent, than under strict neutrality. Finally, and more fundamentally, there is continuing debate about the functional form of the distribution of selection coefficients in nature (e.g., Nielsen and Yang 2003; Loewe and Charlesworth 2006; Eyre-Walker and Keightley 2007), and this calls into question the generality of conclusions reached by imposing an arbitrarily chosen functional form (Tachida 1996; Eyre-Walker 2002; Sawyer et al. 2003; Loewe and Charlesworth 2006; Woodhams 2006; Eyre-Walker and Keightley 2007).

Here we investigate the expected levels of divergence and polymorphism under the nearly neutral theory, when selection coefficients are drawn from a continuous distribution. The study has three main aims. First, it is shown that, for some commonly used distributions, the relevant likelihood equations can be written in terms of special functions; this means that the quantities of interest can be calculated rapidly without the need for numerical integration. Second, approximate forms of these expressions are shown to follow directly from the definitions of the special functions, and these approximations show clearly the leading dependencies on the parameters of biological interest. Finally, results from some different distributions of selection coefficients are compared, to test the robustness of the conclusions.

Expected Levels of Polymorphism and Divergence

Consider, first, expected levels of polymorphism and divergence at a collection of independent sites, when all mutants are subject to a common strength of selection. These results were derived in detail by Kimura (1962, 1969) and others (Ewens 1979; Sawyer and Hartl 1992; Hartl et al. 1994). Here we just give brief heuristic derivations.

The expected divergence, d, along a lineage of t generations, at a site where all mutants have the same selection coefficient, s, is the product of the number of mutants expected to appear, and their probability of reaching fixation, and so takes the form

$$d_0 = 2N\mu t \times \pi(s, N_e, N) \tag{1}$$

where N is the census population size, N_e the effective population size, μ the mutation rate per generation, and $\pi(s, N_e, N)$ the fixation probability. (Both here and elsewhere, we use a zero subscript to indicate a model where all mutants have the same selection coefficient). The expected level of polymorphism at the site is

$$p_0 = 4N_e\mu \int_0^1 \psi(x; s, N_e, N)k(x)dx \tag{2}$$

where $\psi(x; s, N_e, N)dx$ is the probability of mutant alleles segregating in the population between frequency x and frequency $x + dx$, and the function $k(x)$ describes the sampling of alleles from the population. The exact form of $k(x)$ will vary according to the measure of polymorphism that is required (e.g., mean heterozygosity, number of singletons, or total number of polymorphic sites). However, most quantities of interest can be represented by a sampling function comprising one or more terms of the form

$$k(x) \propto x^n (1-x)^m \tag{3}$$

where n and m are nonnegative integers. For example, to model heterozygosity, we would set $n = m = 1$ and specify $k(x) = 2x(1-x)$ (e.g., Kimura 1979). Sawyer and Hartl (1992) and Hartl et al. (1994) describe more complex sampling functions of the same form but with different values of n and m, and Nielsen et al. (2004) describe corrections that may be appropriate for real-world data. All results below use the generalized form of the sampling function, Eq. 3, and so apply to each of these particular cases.

Approximate forms of the remaining functions, $\pi(s, N_e, N)$ and $\psi(x; s, N_e, N)$, were obtained using diffusion analysis by Kimura (1962, 1969, Eqs. 13 and 37). Using these results allows us to write the expressions for d_0 and p_0 in terms of the scaled parameters:

$$\gamma \equiv 4N_e s \tag{4}$$

$$\theta \equiv 4N_e\mu \tag{5}$$

Namely

$$d_0(\gamma) = \mu t \frac{\gamma}{1-e^{-\gamma}} \tag{6}$$

$$p_0(\gamma) = \theta \int_0^1 \frac{1-e^{-\gamma(1-x)}}{1-e^{-\gamma}} x^{n-1}(1-x)^{m-1}dx \tag{7}$$

$$= \frac{\theta}{1-e^{-\gamma}} - \sum_{j=1}^{\infty} \frac{B(n, m+j)}{j!}(-\gamma)^j \tag{8}$$

In the last equation, the representation of $p_0(\gamma)$ as an infinite sum follows from expanding the term $e^{-\gamma(1-x)}$ in powers of $-\gamma(1-x)$ and using $\int_0^1 x^{n-1}(1-x)^{m-1}dx = B(n, m)$, which is the beta function (Abramowitz and Stegun 1965), which for integer arguments is $B(n, m) = (n-1)!(m-1)!/(n+m-1)!$.

To relax the assumption that all mutants are subject to the same strength of selection, we generalize Eqs. 6 and 7 by incorporating a distribution of scaled selection coefficients, denoted $F_i(\gamma)$:

$$d_i = \int_{-\infty}^{\infty} d_0(\gamma) F_i(\gamma) d\gamma \tag{9}$$

$$p_i = \int_{-\infty}^{\infty} p_0(\gamma) F_i(\gamma) d\gamma \tag{10}$$

The following sections evaluate these expressions for some important forms of $F_i(\gamma)$.

Strict Neutrality

To derive results for strict neutrality (Kimura 1983), assume that a proportion f of mutants is selectively neutral, with $\gamma = 0$, while the remaining proportion, $(1-f)$, is severely deleterious, and so contribute nothing to either divergence or polymorphism. With these assumptions, we reproduce standard results (Kimura 1983):

$$d_1 = \mu t f \tag{11}$$

$$p_1 = \theta f B(n, m+1) \tag{12}$$

with the subscript 1 denoting strict neutrality. The extent to which results from other distributions deviate from these neutral expectations can be quantified using the "neutrality index" (Rand and Kann 1996). The index, NI_i, is defined as the ratio of polymorphism to divergence when both quantities are standardized by their strictly neutral equivalents:

$$NI_i = \frac{p_i}{p_1}\frac{d_1}{d_i}, \tag{13}$$

and so is equal to unity under strict neutrality.

Single-Sided Gamma Distribution

The distribution most commonly used to describe deleterious mutations is the single-sided gamma distribution:

$$F_2(\gamma; |\overline{\gamma}|, \beta) = \frac{|\gamma|^{\beta-1} e^{-|\gamma|\beta/|\overline{\gamma}|} (\beta/|\overline{\gamma}|)^{\beta}}{\Gamma(\beta)}, \quad \gamma \le 0. \tag{14}$$

We have parameterized the distribution in terms of the absolute value of its mean, $|\overline{\gamma}|$, and a dimensionless shape parameter, β. The parameter β is related to the coefficient of variation of the distribution via CV $(\gamma) = \beta^{-1/2}$, and to the excess kurtosis via $\kappa(\gamma) = 6/\beta$. This distribution was used in the earliest work on the nearly neutral theory by Ohta (1977), who used the single-sided exponential distribution (equivalent to setting $\beta = 1$ in Eq. 14), and by Kimura (1979), who introduced the arbitrary shape parameter, β.

Exact Results

From Eqs. 9 and 14, we have

$$d_2 = \mu t \int_0^{\infty} \frac{\gamma}{e^{\gamma} - 1} \frac{\gamma^{\beta-1} e^{-\gamma\beta/|\overline{\gamma}|} (\beta/|\overline{\gamma}|)^{\beta}}{\Gamma(\beta)} d\gamma \tag{15}$$

which can be expressed in terms of a special function. To obtain this we use

$$(e^{\gamma} - 1)^{-1} = e^{-\gamma} \sum_{j=0}^{\infty} e^{-j\gamma} \tag{16}$$

and interchange the order of summation and integration to obtain

$$\begin{aligned} d_2 &= \mu t \frac{(\beta/|\overline{\gamma}|)^{\beta}}{\Gamma(\beta)} \sum_{j=0}^{\infty} \int_0^{\infty} e^{-\gamma[1+j+\beta/|\overline{\gamma}|]} \gamma^{\beta} d\gamma \\ &= \mu t \frac{\Gamma(\beta+1)}{\Gamma(\beta)} (\beta/|\overline{\gamma}|)^{\beta} \sum_{j=0}^{\infty} [1 + j + \beta/|\overline{\gamma}|]^{-(1+\beta)} \\ &= \mu t \beta (\beta/|\overline{\gamma}|)^{\beta} \zeta(1+\beta, 1+\beta/|\overline{\gamma}|). \end{aligned} \tag{17}$$

Here

$$\zeta(\Upsilon, \alpha) = \sum_{j=0}^{\infty} (\alpha + j)^{-\Upsilon} \tag{18}$$

is the Hurwitz zeta function (Abramowitz and Stegun 1965). Equation 17, first obtained by Kimura (1979, Eq. 8), can be calculated rapidly using various well-established series approximations. The relevant numerical methods are implemented in commercially available symbolic mathematics software and in publicly available software libraries, such as the GNU Scientific Library (Galassi et al. 2006), which is also implemented in the R environment (R Development Core Team 2006).

A result similar to Eq. 17 can be obtained for polymorphism in terms of the infinite series:

$$p_2 = \theta (\beta/|\overline{\gamma}|)^{\beta} \sum_{j=1}^{\infty} \frac{B(n, m+j)}{j B(\beta, j)} \zeta(\beta + j, 1 + \beta/|\overline{\gamma}|). \tag{19}$$

The terms of this sum decrease in magnitude with j, and so for numerical calculation, approximations of arbitrary accuracy can be obtained by truncating the series at a suitable point.

An alternative is to replace the double integral of Eq. 10 with a one-dimensional integral. This can be done by defining a function $H_i(x)$:

$$H_i(x) \equiv \int_{-\infty}^{\infty} \frac{1 - e^{-\gamma(1-x)}}{1 - e^{-\gamma}} F_i(\gamma) d\gamma \tag{20}$$

such that $p_i = \theta \int_0^1 H(x) x^{n-1} (x)^{m-1} dx$. For the single-sided gamma distribution, this function can be calculated in closed form:

$$H_2(x) = (\beta/|\overline{\gamma}|)^{\beta}[\zeta(\beta, \beta/|\overline{\gamma}| + x) - \zeta(\beta, \beta/|\overline{\gamma}| + 1)]. \quad (21)$$

This exact result is defined for all positive β, but $\beta < 1$ requires analytic continuation (Fine 1951), and for $\beta = 1$ (Ohta 1977), we use $\lim_{\beta\to 1}[\zeta(\beta, a) - 1/(\beta - 1)] = -\Psi_0(a)$, where $\Psi_0(\bullet)$ is the digamma function (Abramowitz and Stegun 1965) yielding the exact result,

$$\lim_{\beta\to 1} H_2(x) = |\overline{\gamma}|^{-1}[\Psi_0(1/|\overline{\gamma}| + 1) - \Psi_0(1/|\overline{\gamma}| + x)]. \quad (22)$$

Approximations

To derive approximate expressions for the divergence and polymorphism, we require information about the typical magnitudes of β and $|\overline{\gamma}|$. Studies fitting gamma distributions to data from various taxa and loci have almost all agreed that $\beta < 1$ provides the best fit (Keightley 1994; Piganeau and Eyre-Walker 2003; Eyre-Walker et al. 2006; Loewe et al. 2006; Loewe and Charlesworth 2006; but see Nielsen and Yang 2003). This finding accords with the high kurtosis and a high concentration of mutants of negligible effect inferred from more direct approaches to estimating the distribution (Davies et al. 1999; Lynch et al. 1999; Eyre-Walker and Keightley 2007). Estimates of $|\overline{\gamma}|$ from bioinformatic studies have tended to be large, often of order 100, which again accords with results from more direct approaches (Keightley and Eyre-Walker 1999; Lynch et al. 1999; Loewe et al. 2006) and with broader surveys of selective constraint (Eyre-Walker et al. 2002; Subramanian and Kumar 2006). Seemingly contradictory estimates of order 1 have appeared in the literature (e.g., Bustamante et al. 2002; Sawyer et al. 2003), but this disagreement is only apparent, because these authors estimated a different quantity: the mean value of $N_e\,|s|$ for mutants eligible to become polymorphic or fixed, i.e., excluding severely deleterious mutants.

Together, these studies suggest that $\beta/|\overline{\gamma}| \ll 1$ will tend to hold in nature. When this is so, the Hurwitz zeta function in Eq. 17 is well approximated by the Riemann zeta function, $\zeta(1+\beta) = \sum_{j=1}^{\infty} j^{-(1+\beta)} \approx 1 + (2/3)^{\beta}/\beta$, and this yields

$$d_2 \approx \mu t \beta^{\beta+1}\zeta(1+\beta)|\overline{\gamma}|^{-\beta} \quad (23)$$

(Kimura 1979; see also Gillespie 1991).

Equation (23) shows that the expected divergence will be approximately loglinear in $|\overline{\gamma}|$, with the slope determined by the shape parameter, i.e., that $\ln d_2 \approx -\beta \ln|\overline{\gamma}| + const.$ Furthermore, applying the same approximation to Eq. 19 shows that the same applies to polymorphism. Together, this means that we can approximate the neutrality index, Eq. 13, as follows:

$$\begin{aligned} NI_2 &\approx 1 + \sum_{j=2}^{\infty} \frac{\zeta(\beta + j)}{j\beta B(\beta, j)\zeta(\beta + 1)} \frac{B(m + j, n)}{B(m + 1, n)} \\ &\approx 1 + \beta \sum_{j=2}^{\infty} \frac{\zeta(j)}{j} \frac{B(m + j, n)}{B(m + 1, n)} + o(\beta^2) \end{aligned} \quad (24)$$

Equation 24 confirms the general finding that under the assumptions of the nearly neutral theory, $NI > 1$ will always hold (Rand and Kann 1996). This reflects the presence of weakly deleterious mutants, which are able to contribute to transient polymorphism but unlikely to reach fixation.

Figure 1 plots the approximate and exact forms of the divergence, polymorphism, and neutrality index, arbitrarily setting the polymorphism sampling parameters (Eq. 3) to $m = 1$ and $n = 8$, such as might be used to model the frequency of singletons in a sample of eight alleles (Hartl et al. 1994). The figure shows that the approximations are good when $|\overline{\gamma}|$ is not too small and β is not too large, and this is generally consistent with the empirical results discussed above.

From our approximations, Eqs. 23 and 24, some leading parameter dependencies follow directly. First, the neutrality index, NI_2 is shown to increase linearly with the shape parameter, β, but to be largely independent of the mean strength of selection $|\overline{\gamma}|$. Second, the effective population size, N_e, appears solely in the parameters $|\overline{\gamma}|$ and θ (Eqs. 4 and 5), and so we have

$$d_2 \propto N_e^{-\beta} \quad (25)$$

$$p_2 \propto N_e^{1-\beta} \quad (26)$$

$$NI_2 \approx \text{independent of } N_e \quad (27)$$

Equations (24) and (25)–(26) also show that the differences between neutrality and near neutrality are least marked when the shape parameter, β, is very small; this is because when $\beta \ll 1$, divergence and polymorphism will be relatively insensitive to changes in N_e, and the neutrality index will remain close to its neutral value of unity. This behavior can be understood by recalling that the kurtosis of the gamma distribution is given by $\kappa(\gamma) = 6/\beta$, and that a highly leptokurtic distribution approximates the situation under strict neutrality, in that mutations are concentrated in a peak around $\gamma = 0$ and in a tail of large negative values.

Partially Reflected Gamma Distribution

Single-sided distributions, such as Eq. 14, are unrealistic in that they contain no beneficial mutations and, so, embody the implicit assumption that populations will degenerate indefinitely (Gillespie 1995; Tachida 1996). A related

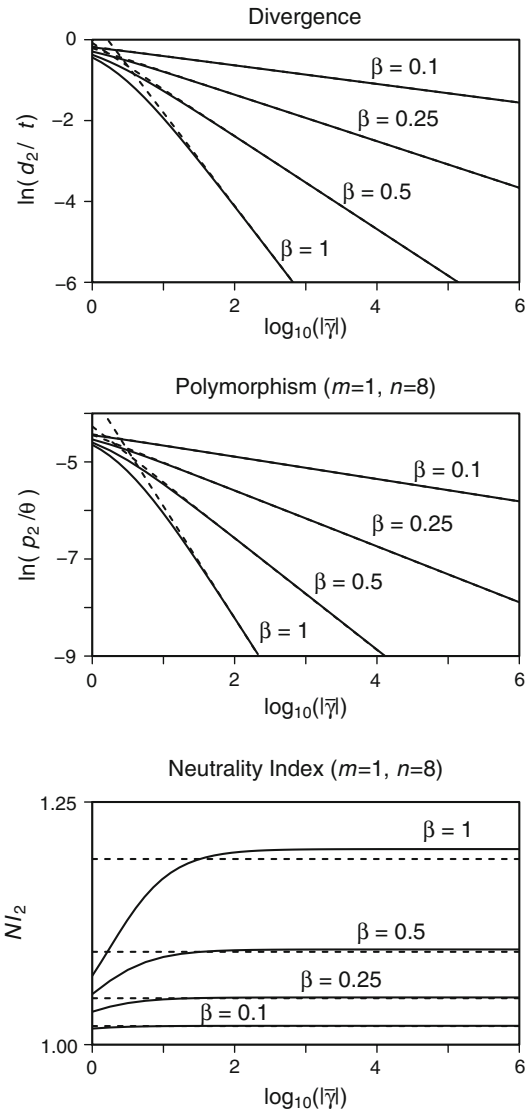

Fig. 1 Expected values of divergence, polymorphism, and the neutrality index under the assumptions of the nearly neutral theory, with deleterious selection coefficients drawn from a single-sided gamma distribution, Eq. 14. Solid lines show exact results obtained from Eqs. 11–13 and 17 and by numerical integration from Eqs. 10 and 20–22. Dashed lines show analytical approximations obtained from Eq. 23, and from Eqs. 8 and 24, truncating the series expansions after three terms in both cases. Results are shown for four different values of the shape parameter, β, and are plotted as a function of the mean absolute scaled selection coefficient, Eqs. 4 and 14

model that avoids this assumption is the "partially reflected" gamma distribution introduced by Piganeau and Eyre-Walker ((2003); Bulmer 1991). This distribution was derived from a mechanical model of evolution, in which deleterious mutations are generated from a gamma distribution, but in which back mutations from the deleterious state to the wild type are also permitted. These assumptions lead to the following equilibrium distribution of scaled selection coefficients:

$$F_3(\gamma;|\bar{\gamma}|,\beta)=\frac{|\gamma|^{\beta-1}e^{-|\gamma|\beta/|\bar{\gamma}|}(\beta/|\bar{\gamma}|)^{\beta}}{\Gamma(\beta)(1+e^{\gamma})} \tag{28}$$

which applies to both negative and positive γ (see plots in Piganeau and Eyre-Walker 2003). While Eq. 28 has been parameterized to match the single-sided Eq. 14, $|\bar{\gamma}|$ now represents the mean selection coefficient only when all loci are fixed for the beneficial allele, but because only weakly deleterious mutants can reach fixation, it will differ little from the true mean of Eq. 28.

Results

The divergence for the partially reflected gamma distribution is

$$\begin{aligned} d_2 &= 2\mu t\frac{(\beta/|\bar{\gamma}|)^{\beta}}{\Gamma(\beta)}\int_0^{\infty}\frac{\gamma^{\beta}e^{-\gamma\beta/|\bar{\gamma}|}}{e^{\gamma}-e^{-\gamma}}d\gamma \\ &= \mu t\beta\left(\frac{\beta}{2|\bar{\gamma}|}\right)^{\beta}\zeta(\beta+1,(1+\beta/|\bar{\gamma}|)/2). \end{aligned} \tag{29}$$

where we used the expansion $(e^{\gamma}-e^{-\gamma})^{-1}=e^{-\gamma}\sum_{j=0}^{\infty}e^{-2j\gamma}$. By similar means, expected polymorphism and the function H_3 (Eq. 20) are found to be

$$p_3=\theta\left(\frac{\beta}{2|\bar{\gamma}|}\right)^{\beta}\sum_{j=0}^{\infty}\frac{2^{1-j}B(n,m+j)}{jB(\beta,j)}\zeta\left(\beta+j,(1+\beta/|\bar{\gamma}|)/2\right) \tag{30}$$

$$H_3(x)=\left(\frac{\beta}{2|\bar{\gamma}|}\right)^{\beta}[\zeta(\beta,(\beta/|\bar{\gamma}|+x)/2)-\zeta(\beta,1+(\beta/|\bar{\gamma}|-x)/2)] \tag{31}$$

Approximations for these expressions can be derived using $\zeta(x,1/2+\varepsilon)\approx\zeta(x,1/2)=(2^x-1)\zeta(x)$ (Truesdell 1950). Then, following the same procedures used for the single-sided case, we find

$$d_3\approx(2-2^{-\beta})d_2 \tag{32}$$

$$NI_3\approx 1+\beta\sum_{j=2}^{\infty}\frac{2(1-2^{-j})\zeta(j)}{j}\frac{B(m+j,n)}{B(m+1,n)} \tag{33}$$

As such, the leading parameter dependencies are identical to the single-sided cases. This shows that the conclusions above are robust to the inclusion in the model of weakly beneficial substitutions.

Single-Sided Lognormal Distribution

The gamma distributions investigated above are flexible and widely used. They also have limited theoretical justification, arising from simple models of selection on quantitative traits (Martin and Lenormand 2006; Gu 2007a, b). But a theoretical case can also be made for other

distributions. For example, it follows from the central limit theorem that normal or lognormal distributions might apply if each mutant has many pleiotropic effects on independent components of fitness (Sawyer et al. 2003; Loewe and Charlesworth 2006).

There is also evidence from bioinformatic studies that the gamma distribution might not be the most appropriate choice. For example, Nielsen and Yang ((2003); Yang et al. 2000) fitted various distributions to divergence data from animal mitochondrial genes and found that the best-fitting distribution was normal, with a class of invariable sites (see also Sawyer et al. 2003). However, the normal distribution was not a significant improvement over a gamma with a shape parameter of $\beta \approx 3$, and the gamma distribution generally approximates a normal when $\beta \gg 1$ (although in this case our approximate results would not apply).

A different conclusion was drawn by Loewe and Charlesworth (2006), who fitted various distributions to polymorphism data and estimates of the lethal mutation rate from *Drosophila*. They found that a normal distribution could not adequately fit the polymorphism data and that a gamma distribution regularly underpredicted the lethal mutation rate. A lognormal distribution, by contrast, gave a good fit to both kinds of data (although, again, its improvement over the gamma was not formally significant).

Because of the success of the lognormal distribution in the study by Loewe and Charlesworth (2006), and to investigate the robustness of the conclusions reached above, we now examine expected levels of divergence and polymorphism when scaled selection coefficients are log-normally distributed.

The single-sided lognormal distribution is nonzero for $\gamma < 0$ and in this range has the form

$$F_4(\gamma; |\overline{\gamma}|, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2|\gamma|}} \exp\left(\frac{[\ln|\gamma| - \ln|\overline{\gamma}|] + \sigma^2/2]^2}{2\sigma^2}\right). \tag{34}$$

We have parameterized this distribution with its mean, $|\overline{\gamma}|$, but it is commonly parameterized with the mean of the associated normal distribution, $E[\ln|\gamma|] = \ln|\overline{\gamma}| - \sigma^2/2$, or in other ways (Loewe and Charlesworth 2006). The second parameter, σ^2, which is the variance of the associated normal distribution, functions like the shape parameter, β, of the gamma distribution, with the coefficient of variation and excess kurtosis both increasing with σ^2. In detail, we have $\mathrm{CV}(\gamma) = (\mathrm{e}^{\sigma^2} - 1)^{1/2}$, and $\kappa(\gamma) = \left[|\overline{\gamma}|(\mathrm{e}^{\sigma^2} - 1)\right]^{-4}$ $(\mathrm{e}^{6\sigma^2} - 4\mathrm{e}^{3\sigma^2} + 6\mathrm{e}^{\sigma^2} - 3)$, which, unlike the equivalent expression for the gamma distribution, depends on the mean as well as the shape parameter.

To understand the expected divergence under the log-normal distribution, consider the following crude approximation:

$$\begin{aligned} \ln(d_4/\mu t) &\approx \ln\left(\int_0^1 F_4(\gamma; |\overline{\gamma}|, \sigma^2) d\gamma\right) \\ &\approx -\frac{\ln|\overline{\gamma}|}{\sigma}\sqrt{\frac{2}{\pi}}\left[1 + \frac{\ln|\overline{\gamma}| - \sigma^2}{\sqrt{2\pi}\sigma}\right] + const, \end{aligned} \tag{35}$$

where the constant is $\sigma/\sqrt{2\pi}\left(1 - \sigma/\sqrt{2\pi}\right) - \ln 2$ (Fig. 2). This approximation was obtained from Eq. 9 by treating mutants with $|\gamma| < 1$ as strictly neutral, and all others as severely deleterious, and then using a series expansion of the complementary error function (Abramowitz and Stegun 1965).

Equation 35 shows that when $\ln|\overline{\gamma}|$ and σ^2 are very close in value, then linear approximations relating $\ln(d_i/\mu t)$ to $\ln|\overline{\gamma}|$ are quite similar in the gamma and lognormal cases, with the shape parameters governing the slopes. But in the lognormal case, this linear approximation will be accurate over a very limited range of parameter space, and so the slope of the relationship will vary with N_e in general. This is confirmed in Fig. 2, where exact results are shown, with σ^2 values chosen to match the curves in Fig. 1. In addition to the curvature in the plots for the divergence and polymorphism, it is clear that the neutrality index does not approach a constant value when selection coefficients are lognormally distributed but, instead, continues to increase with $|\overline{\gamma}|$, and therefore N_e.

Discussion

We have derived expressions for the expected levels of divergence and polymorphism under the nearly neutral theory, with the assumption that the distribution of scaled selection coefficients is either gamma (containing only deleterious mutants) or partially reflected gamma (including back mutations of beneficial effect). These results have been given in exact forms (Eqs. 17–21 and 29–31) which can be calculated rapidly and accurately, without the need for numerical integration. Results have also been presented in approximate forms (Eqs. 23–27 and 32–33), which show clearly the leading dependencies on important parameters.

A particularly interesting result is the expected value of the neutrality index, Eq. 13, which we have shown to be of the form

$$NI \approx 1 + \beta K \tag{36}$$

where β is the shape parameter of the gamma distribution and K is a constant determined by the way in which polymorphism is measured (see Eqs. 3, 24, and 33). Under

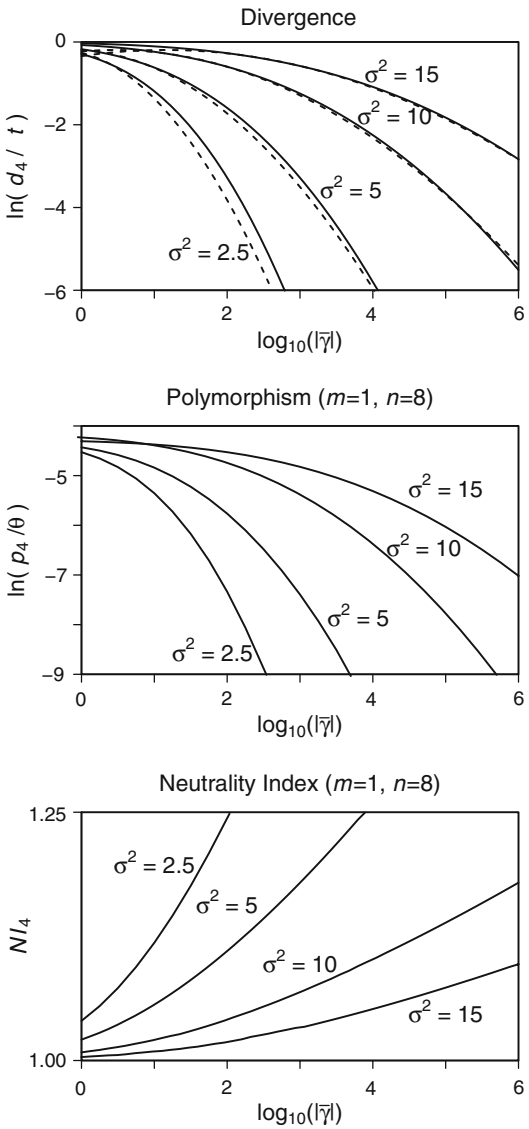

Fig. 2 Expected values of divergence, polymorphism, and the neutrality index when selection coefficients are drawn from a single-sided lognormal distribution, Eq. 34. Solid lines show exact results obtained by numerical integration, and the dashed line shows the crude approximation of Eq. 35. Other details match Fig. 1

these assumptions, therefore, the neutrality index is quite independent of the strength and efficacy of selection on deleterious mutants, and this has a number of interesting implications.

For example, Presgraves (2005) studied a set of 98 protein-coding loci from *Drosophila melanogaster* and showed that the neutrality index correlated negatively with the local recombination rate. The level of recombination is an important determinant of local N_e values, and so this correlation was interpreted as an effect of within-genome variation in the efficacy of selection (Presgraves 2005). Equation 36 allows us to make a further inference, because it shows that if deleterious mutations are gamma distributed, with or without back mutation, then a correlation between NI and N_e cannot be attributed to mildly deleterious mutants alone. However, the correlation could be explained if a nonnegligible fraction of substitutions was strongly adaptive. (This can be shown by adding a constant number of adaptive substitutions to the divergence, before calculating the neutrality index as in Eq. 24 or 33). The conclusion that adaptive substitutions would create a dependency of the neutrality index on N_e follows under quite general conditions. For example, it does not depend on the rate of adaptive substitution itself increasing with N_e (Gillespie 2001), nor does it rely on the direct effects of genetic hitchhiking – although both effects would increase the reported correlation (Gillespie 2001). If Presgraves' (2005) result does indeed imply high rates of adaptive substitution, then this would be consistent with other lines of evidence suggesting widespread adaptive substitution in *D. melanogaster* (Eyre-Walker 2006).

It is important to note that Eq. 36, like all the results above, assumed demographic stability, i.e., that N_e remained constant throughout the divergence, and that polymorphism is at equilibrium. However, because we have derived the leading dependencies of all quantities on N_e (Eqs. 25–27), our results can also show how certain tests of the neutral theory are misled by demographic change. Consider, for example, the behavior of the neutrality index under a simple model of population expansion. Let us assume that N_e took a single constant value for a proportion $0 \leq q \leq 1$ of the total divergence time, and then increased by a factor z, to zN_e, for the remaining period. This higher N_e value will then govern the levels of polymorphism observed. Under this scenario, our results show that the neutrality index is expected to be

$$NI \approx \frac{1+\beta K}{1+q(z^{\beta}-1)}. \tag{37}$$

It follows that when population expansion is substantial and/or recent (i.e., when z is large or q close to unity), then the neutrality index is expected to be <1. This has important implications, because $NI < 1$ is generally taken as a signature of widespread adaptive substitution (McDonald and Kreitman 1991; Rand and Kann 1996). Such demographic artifacts in tests of positive selection have previously been investigated numerically (e.g., Eyre-Walker 2002; Charlesworth and Eyre-Walker 2007), but Eq. 37 allows them to be studied analytically. In addition, because Eq. 37 contains just three parameters, only one of which (β) could be locus-specific, it could also be used to test the hypothesis of population expansion in a formal likelihood framework.

While expressions such as Eqs. 36 and 37 are pleasingly simple, another conclusion of the present work is that such relationships might not hold unless the distribution of

mutant effects is adequately described by a gamma distribution. In particular, we have shown that the behavior of polymorphism and divergence is qualitatively different when selection coefficients are lognormally distributed (Fig. 2, Eq. 35).

It is important to ask, therefore, how far empirical evidence allows us to choose between the various functional forms. As has been mentioned, the lognormal was preferred over the gamma distribution in the recent study by Loewe and Charlesworth (2006), and its greater success was attributed to its weightier tail, with a correspondingly higher concentration of lethal mutants (Loewe and Charlesworth 2006). This feature of the distribution has wide empirical support, both from bioinformatic approaches (Nielsen and Yang 2003; Sawyer et al. 2003; Loewe and Charlesworth 2006), and from mutation accumulation studies (Keightley 1994; Elena et al. 1998; Sanjuan et al. 2004; Eyre-Walker and Keightley 2007).

But high concentrations of severely deleterious mutants can be modeled using distributions other than the lognormal, by adding a parameter such as f that appears in Eqs. 11 and 12 (Sawyer et al. 2003; Nielsen and Yang 2003; Eyre-Walker et al. 2006). And while less elegant than fitting a continuous distribution to all mutants, this approach may be more realistic, as direct studies have sometimes found the distribution to be bimodal, containing peaks of both weakly and strongly deleterious effects (Elena et al. 1998; Sanjuan et al. 2004; Eyre-Walker and Keightley 2007). Furthermore, Nielsen and Yang (2003) and Eyre-Walker et al. (2006) found that the inclusion of such a parameter made little difference to estimates of the gamma shape parameter, β, obtained from divergence and polymorphism data.

However, the qualitative differences between Fig. 1 and Fig. 2 cannot be attributed to differences in the proportion of strongly deleterious mutants. Instead, these differences are probably best explained by another feature of the lognormal distribution: its suppression of probability density around the origin, in the region of strict neutrality. Unlike the concentration of effectively lethal mutants, empirical evidence of this aspect of the distribution is much more equivocal. Loewe and Charlesworth's (2006) own method could not reject models where a discrete class of strictly neutral mutants was added to the lognormal, and experimental approaches have limited power to resolve very small selection coefficients (Eyre-Walker and Keightley 2007).

Because of this uncertainty, at present we have reason to remain skeptical of any quantitative conclusion that relies on the assumption that the distribution of selection coefficients can be well described by any single functional form (Tachida 1996; Eyre-Walker 2002; Sawyer et al. 2003; Loewe and Charlesworth 2006; Woodhams 2006; Eyre-Walker and Keightley 2007). Clearly, we also require a more detailed knowledge of the distribution of mutant effects in nature. Numerical methods, such as those reported here, combined with model selection techniques, should aid progress toward this goal.

Acknowledgments It is a pleasure to thank Laurence Loewe, Andrea Betancourt, and Fraser Lewis for their help with this work.

References

Abramowitz M, Stegun I (1965) Handbook of mathematical functions. Dover, New York

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

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL (2002) The cost of inbreeding in Arabidopsis. Nature 416:531–534

Charlesworth J, Eyre-Walker A (2007) The other side of the neutral theory, evidence of slightly advantageous back-mutations. Proc Natl Acad Sci USA 104:16992–16997

Davies EK, Peters AD, Keightley PD (1999) High frequency of cryptic deleterious mutations in *Caenorhabditis elegans*. Science 285:1745–1747

Elena SF, Ekunwe L, Hajela N, Oden SA, Lenski RE (1998) Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. Genetica 102(103):349–358

Ewens WJ (1979) Mathematical population genetics. Springer, Berlin

Eyre-Walker A (2002) Changing effective population size and the McDonald Kreitman test. Genetics 162:2017–2024

Eyre-Walker A (2006) The genomic rate of adaptive evolution. Trends Ecol Evol 21:569–575

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–618

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D (2002) Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol 19:2142–2149

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

Fine NJ (1951) Note on the Hurwitz zeta-function. Proc Am Math Soc 2:361–364

Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, Rossi F (2006) GNU scientific library reference manual—revised second edition (v1.8). Network Theory, Bristol

Gillespie JH (1991) The causes of molecular evolution. Oxford University Press, Oxford

Gillespie JH (1995) On Ohta's hypothesis: most amino acid substitutions are deleterious. J Mol Evol 40:64–69

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

Gu X (2007a) Stabilizing selection of protein function and distribution of selection coefficients among sites. Genetica 130:93–97

Gu X (2007b) Evolutionary framework for protein sequence evolution and gene pleiotropy. Genetics 175:1813–1822

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

Keightley PD (1994) The distribution of mutation effects on viability in Drosophila melanogaster. Genetics 138:1315–1322

Keightley PD, Eyre-Walker A (1999) Terumi Mukai and the riddle of deleterious mutation rates. Genetics 153:515–523

Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713–719

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903

Kimura M (1979) Model of effectively neutral mutations in which selective constraint is incorporated. Proc Natl Acad Sci USA 76:3440–3444

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Loewe L, Charlesworth B (2006) Inferring the distribution of mutational effects in Drosophila. Biol Lett 2:426–430

Loewe L, Charlesworth B, Bartolomè C, Nöel V (2006) Estimating selection on non-synonymous mutations. Genetics 172:1079–1092

Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L, Willis J (1999) Perspective: spontaneous deleterious mutation. Evolution 53:645–663

Martin G, Lenormand T (2006) A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60:893–907

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

Nielsen R, Yang Z (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20:1231–1239

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168:2373–2382

Ohta T (1977) Extension to the neutral mutation random drift hypothesis. In: Kimura M (ed) Evolution and polymorphism. National Institute of Genetics, Mishima, pp 148–167

Ohta T, Gillespie JH (1996) Development of neutral and nearly neutral theories. Theor Popul Biol 49:128–142

Piganeau G, Eyre-Walker A (2003) Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc Natl Acad Sci USA 100:10335–10340

Presgraves DC (2005) Recombination enhances protein adaptation in *Drosophila melanogaster*. Curr Biol 15:1651–1656

R Development Core Team (2006) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org)

Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol Biol Evol 13:735–748

Sanjuan R, Moya A, Elena SF (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

Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003) Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. J Mol Evol 57:S154–S164

Subramanian S, Kumar S (2006) Higher intensity of purifying selection on >90% of the human genes revealed by the intrinsic replacement mutation rates. Mol Biol Evol 23:2283–2287

Tachida H (1996) Effects of the shape of distribution of mutant effect in nearly neutral mutation models. J Genet 75:33–48

Truesdell C (1950) On the addition and multiplication theorems for the special functions. Proc Natl Acad Sci USA 36:752–757

Welch JJ (2006) Estimating the genome-wide rate of adaptive protein evolution in Drosophila. Genetics 173:821–837

Williamson S, Fledel-Alon A, Bustamante CD (2004) Population genetics of polymorphism and divergence for diploid species with arbitrary dominance. Genetics 168:463–475

Woodhams M (2006) Can deleterious mutations explain the time dependency of molecular rate estimates? Mol Biol Evol 23:2271–2273

Yampolsky LY, Kondrashov FA, Kondrashov SA (2005) Distribution of the strength of selection against amino acid replacements in human proteins. Hum Mol Genet 14:3191–3201

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