Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change

Adam Eyre-Walker and Peter D. Keightley†*

*Centre for the Study of Evolution and School of Life Sciences, University of Sussex, Brighton, United Kingdom; and †Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

The prevalence of adaptive evolution relative to genetic drift is a central problem in molecular evolution. Methods to estimate the fraction of adaptive nucleotide substitutions (α) have been developed, based on the McDonald–Kreitman test, that contrast polymorphism and divergence between selectively and neutrally evolving sites. However, these methods are expected to give downwardly biased estimates of α if there are slightly deleterious mutations, because these inflate polymorphism relative to divergence. Here, we estimate α by simultaneously estimating the distribution of fitness effects of new mutations at selected sites from the site frequency spectrum and the number of adaptive substitutions. We test the method using simulations. If data meet the assumptions of the analysis model, estimates of α show little bias, even when there is little or no recombination. However, population size differences between the divergence and polymorphism phases may cause α to be over or underestimated by a predictable factor that depends on the magnitude of the population size change and the shape of the distribution of effects of deleterious mutations. We analyze several data sets of protein-coding genes and noncoding regions from hominids and Drosophila. In Drosophila genes, we estimate that approximately 50% of amino acid substitutions and approximately 20% of substitutions in introns are adaptive. In protein-coding and noncoding data sets of humans, comparison to macaque sequences reveals little evidence for adaptive substitutions. However, the true frequency of adaptive substitutions in human-coding DNA could be as high as 40%, because estimates based on current polymorphism may be strongly downwardly biased by a decrease in the effective population size along the human lineage.

Introduction

The role of advantageous mutations has been one of the most hotly debated questions in molecular evolution for nearly 40 years (Kimura 1983; Gillespie 1991), but only recently has the question become tractable with the publication of large amounts of within-species polymorphism data (Nielsen 2005; Eyre-Walker 2006). These data allow tests for the presence of selective sweeps based on a local reduction in polymorphism and/or an increase in linkage disequilibrium (Nielsen 2005), and many such examples have been published (e.g., Glinka et al. 2003; Akey et al. 2004; Wright et al. 2005; Voight et al. 2006; Sabeti et al. 2007; Williamson et al. 2007). Methods have also been developed that combine information on polymorphism and between-species divergence (Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Boyko et al. 2008) or that compare polymorphism levels in species with different effective sizes (Loewe et al. 2006) to estimate the proportion of adaptively driven substitutions (α). Estimates of α have varied considerably between taxa. For example, estimates for adaptive amino acid substitutions in protein-coding sequences are 10% or less for hominids (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005; Gojobori et al. 2007; Boyko et al. 2008) but are greater than 50% for some microorganisms and Drosophila (Smith and Eyre-Walker 2002; Nielsen and Yang 2003; Bierne and Eyre-Walker 2004; Andolfatto 2005; Charlesworth and Eyre-Walker 2006; Welch 2006; Shapiro et al. 2007; Maside and Charlesworth 2007) (reviewed in Eyre-Walker 2006).

Current methods to estimate the proportion of adaptive substitutions are based on the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991), which was itself developed from the Hudson–Kreitman–Aguadé test (Hudson et al. 1987). In the MK test, levels of within-population polymorphism at two categories of sites are compared with the corresponding levels of between-species divergence. For example, numbers of synonymous and nonsynonymous polymorphisms are compared with numbers of synonymous and nonsynonymous substitutions. Under the assumption that synonymous mutations are neutral and nonsynonymous mutations are either strongly deleterious, neutral or advantageous, α can be estimated from simple expressions contrasting these four quantities (Charlesworth 1994; Fay et al. 2001). Estimates will tend to be biased either upward or downward, however, if the true distributions of fitness effects of new mutations at either the synonymous or nonsynonymous sites differ from that assumed. In particular, slightly deleterious mutations that are subject to weak negative selection, such that $N_e s \approx 1$ (where N_e and s are the effective population size and selection coefficient, respectively), contribute proportionately more to polymorphism than to divergence. As a consequence, if some nonsynonymous mutations are slightly deleterious, α will tend to be underestimated (McDonald and Kreitman 1991; Eyre-Walker 2002). There is evidence that slightly deleterious mutations contribute to variation in many populations (see Charlesworth and Eyre-Walker 2007 and references therein), so it is desirable to account for slightly deleterious mutations within methods to infer α.

In order to control for the effects of slightly deleterious mutations, Fay et al. (2001) have suggested removing low-frequency variants from the data. Because slightly deleterious mutations segregate at lower average frequencies than neutral mutations, this method preferentially removes slightly deleterious mutations and is expected to reduce the underestimation of α. However, the procedure will fail to remove all slightly deleterious mutations and will also remove some effectively neutral mutations. Moreover, the frequency cut-off below which variants

Key words: adaptation, humans, drosophila, site frequency spectrum, nucleotide polymorphism.

E-mail: keightley.mbe2009@gmail.com.

Mol. Biol. Evol. 26(9):2097–2108. 2009
doi:10.1093/molbev/msp119
Advance Access publication June 17, 2009

should be removed is arbitrary. Although removing low-frequency variants tends to increase estimates of α in many data sets (Fay et al. 2001; Bierne and Eyre-Walker 2004; Charlesworth and Eyre-Walker 2006), theory suggests that this procedure is expected to lead to biased estimates, unless the distribution of fitness effects of new mutations is strongly L-shaped (i.e., leptokurtic) or the level of adaptive evolution is very high (Charlesworth and Eyre-Walker 2008). As a consequence, Charlesworth and Eyre-Walker (2008) have suggested that many current estimates of α may be underestimates. For example, in the case of the human–chimpanzee comparison, true values of α greater than 8% would be expected to generate estimates of less than 0 even if polymorphisms with minor allele frequencies below 15% are excluded (Charlesworth and Eyre-Walker 2008).

Here, we investigate an approach to account for the contribution of slightly deleterious mutations to polymorphism and divergence. We estimate the distribution of fitness effects of new deleterious mutations (DFE) from the polymorphism data, then use the inferred DFE to predict the numbers of substitutions originating from neutral and slightly deleterious mutations between two species. If the observed number of substitutions is greater than this expectation, we can attribute the difference between observed and expected to advantageous substitutions, yielding an estimate of α.

Three methods have been proposed to estimate the DFE based on polymorphism data alone (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Boyko et al. 2008). Eyre-Walker et al. (2006) and Boyko et al. (2008) use a diffusion approximation to fit a distribution of fitness effects to the site frequency spectrum (SFS, i.e., the allele frequency distribution) observed in a population sample, and Keightley and Eyre-Walker (2007) use transition matrix methods to generate expected distributions of allele frequencies and from these compute the likelihood of the SFS. They differ in their treatment of the effects of demographic changes, such as recent population expansion or contraction, which affect the SFS in ways that may resemble selection. Eyre-Walker et al. (2006) account for the effects of demography by a simple approximation, whereas Keightley and Eyre-Walker (2007) and Boyko et al. (2008) estimate the DFE within the context of demographic models in which past changes in population size are allowed. The methods broadly agree in their estimates of the parameters of the DFE (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Boyko et al. 2008).

Boyko et al. (2008) recently outlined a method for incorporating between-species divergence data to estimate the proportion of adaptive substitutions. In a manner similar to Boyo et al., we build on our previously described methods to infer parameters of the DFE (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007) by incorporating between-species divergence data. We estimate the proportion of adaptive substitutions by comparing the observed numbers of substitutions between a species pair with numbers of substitutions expected from the fixation of deleterious mutations with effects drawn from the estimated DFE. We investigate the performance of this method by simulations. We start by analyzing simulated data that conform to the method of analysis. We then investigate two departures from the model of analysis that are likely to induce biases. In the first, we model a change in the effective population size, leading to a change in the effectiveness of selection, between the divergence and polymorphism phases. In the second, we investigate the effects of tight linkage within loci. We apply the method to obtain estimates of α for protein-coding and noncoding DNA sequences in hominids and Drosophila and interpret these results in the light of our simulations.

Materials and Methods

The method simultaneously estimates the distribution of fitness effects of deleterious mutations and the proportion of adaptive substitutions between a species pair. We assume that the data include two classes of nucleotide sites, one evolving neutrally and a second subject to negative and positive selection. Mutation rates at the neutral and selectively evolving sites are assumed to be equal. We assume that polymorphisms at selectively evolving sites originate from new deleterious mutations that are semidominant with effects s sampled from some distribution, where s is defined as the difference in fitness between the homozygotes. These may include effectively neutral mutations. We disregard the contribution of strongly advantageous mutations to polymorphism, because these are expected to be both rare and to become rapidly fixed. We also assume that there is no contribution to polymorphism from sites under balancing selection. Sites are assumed to segregate independently and to have no more than two alleles segregating per locus.

Our method estimates parameters of the DFE using the polymorphism SFS, then uses this to estimate the proportion of substitutions driven by positive selection between species. We used the folded SFS and the folded expected allele frequency distribution as described by Keightley and Eyre-Walker (2007). We model the effects of recent demographic changes by assuming that the population starts at equilibrium, then experiences a step change in size, t generations in the past. We infer the distribution of effects of deleterious mutations by maximum likelihood following the method of Keightley and Eyre-Walker (2007). We assume that the distribution of effects of deleterious mutations, $f(s|a, b)$, is a gamma distribution with scale parameter a and shape parameter b (note that we use Roman rather than the more usual Greek characters for scale and shape because we use α to refer to the estimate of the fraction of adaptive substitutions, for consistency with several recent studies). We generate the expected allele frequency distribution by transition matrix methods for a given parameter combination, including s, N_1, and N_2 (population sizes before and after the size step change, respectively), t, and f_0, a parameter representing the frequency of unmutated sites, under the assumption that no more than two alleles can segregate at a site. We use this expected allele frequency distribution to compute the likelihood of the observed SFS data, integrating numerically over $f(s)$ for the selectively evolving site data, and assuming $s = 0$ for the neutrally evolving sites. We assume that numbers of derived

segregating alleles are binomially distributed conditional on the population frequency. We maximize the product of likelihoods for the neutral and selected site data as a function of the model parameters (a, b, N_2, t, f_0). In practice we assume a fixed N_1, so that the relevant demographic parameters estimated are N_2/N_1 and t/N_2. We obtain estimates of the selective effects of new mutations as the product Ns, where N is a measure of the effective population size experienced by the polymorphisms. This is obtained by averaging N_1 and N_2, weighted by the expected contributions of neutral mutations to allele frequency variation t generations after the population size change, that is,

$$N = \frac{N_1 w_1 + N_2 w_2}{w_1 + w_2}, \quad (1)$$

where $w_1 = N_1(1 - 1/(2N_2))^t$ and $w_2 = N_2(1 - e^{-t/(2N_2)})$. We use a weighting scheme rather than current or ancestral population size, because in some cases, there is effectively no information to estimate N_1 or N_2. For example, if there has been a population-size change a very long time ago, there is almost no information to estimate N_1. Similarly, if there has been a very recent population-size change, there is little information to estimate N_2.

We estimate the proportion of adaptive substitutions as follows. The divergence at the neutral sites estimates a quantity proportional to the mutation rate per site. The predicted divergence at the selected sites, in the absence of advantageous mutations, is proportional to the product of the mutation rate and the average fixation probability of a selected mutation, which we infer based on parameters estimated from the polymorphism data analysis. The difference between the observed and predicted divergences therefore estimates the divergence due to adaptive substitutions.

The fixation probability of a new mutation of selective effect $N_e s$ appearing as a single copy in a diploid population of census size N is

$$u(N_e, s) = \frac{1 - \exp(-2N_e s/(2N))}{1 - \exp(-2N_e s)}, \quad (2)$$

(Kimura 1957, 1983). Note that we define s as the difference in fitness between homozygotes, whereas Kimura defined this as $2s$. The census and effective population sizes are not distinguished in the transition matrix method, so equation (2) can be rewritten as

$$u(N, s) = \frac{1 - \exp(-s)}{1 - \exp(-2Ns)}, \quad (3)$$

and the rate of fixation from recurrent mutations is proportional to the product of the mutation rate per site and $2Nu(N, s)$. Let d_N and d_S be the numbers of selected (e.g., nonsynonymous) and neutral (e.g., synonymous) substitutions per site, respectively. Our estimate of the rate of adaptive substitutions at the selected sites then depends on the distribution of s, parameters of which are assumed to have been estimated. The difference between the observed and expected rates of selected substitutions can be written as $d_N - d_S \int_0^\infty 2Nu(N, s) f(s|a, b) \mathrm{d}s$, and the estimate of the proportion of adaptive substitutions is then,

$$\alpha = \frac{d_N - d_S \int_0^\infty 2Nu(N, s) f(s|a, b) \mathrm{d}s}{d_N}. \quad (4)$$

We have also developed a second approach to estimate α, based on the method to estimate the DFE of Eyre-Walker et al. (2006). Details of the method and results of its application are presented in the supplementary information, Supplementary Material online.

Simulations

To investigate the performance of the inference procedure, we analyzed data generated by Monte Carlo simulation. For simulated data involving freely recombining sites, neutral and selected site polymorphism data for models that allowed recent changes in population size were simulated as described by Keightley and Eyre-Walker (2007). The proportion of neutral sites showing a difference, q_S, was computed from $q_S = (3/4)(1 - \exp(-4d_S/3))$, and the number of neutral differences was a binomial deviate with parameters q_S and the number of neutral sites. For the selected sites, which are subject to negative and positive selection, we computed the proportion of sites showing an adaptive fixation or a fixation due to the fixation of a deleterious mutation from

$$q_N = \frac{q'_N}{(1 - \alpha)}, \quad (5)$$

where q'_N is the proportion of sites with a fixed deleterious mutation, computed from

$$q'_N = \int_0^\infty \frac{3}{4}(1 - \exp(-4d_S u(\lambda N, s))) f(s|a, b) \mathrm{d}s. \quad (6)$$

In equation (6), $u()$ is the fixation probability of a new mutation with a selection coefficient s in an effective population of size λN, and N is the population size used to generate the polymorphism data. The parameter λ specifies whether the divergence and polymorphism phases have the same effective sizes ($\lambda = 1$), or the data are modeled under long-term population contraction ($\lambda < 1$) or expansion ($\lambda > 1$). The number of selected sites showing a nucleotide difference was a binomial variate with parameters q_N and the number of selected sites.

To investigate the performance of the method with little or no recombination under various demographic scenarios we performed simulations using the simulation package sfs_code (Hernandez 2008). This simulation package performs forward population genetic simulations under a variety of demographic models and distributions of mutational fitness effects. We assumed free recombination between loci and no recombination within loci, so varying rates of recombination were simulated by varying the number of loci, keeping the total number of sites simulated constant at 50,000. The simulations were run with population sizes of 500, with a burn-in of 10,000 generations to approach approximate mutation–selection–drift equilibrium. Parameters

Table 1
Details of the Human Data Sets

Data Set	No. Loci	DNA Type	No. Sites	Proportion of Sites Segregating		Divergence to Macaque
				Africa	Europe	
PGA	255	Intron	1,841,515	0.00350	0.00218	0.0550
		Zero-fold	124,876	0.00146	0.00107	0.0196
		Upstream	59,762	0.00350	0.00219	0.0486
		Downstream	67,831	0.00348	0.00189	0.0474
EGP	193	Intron	1,753,015	0.00319	0.00182	0.0553
		Zero-fold	109,974	0.00110	0.000769	0.0145
		Upstream	41,006	0.00381	0.00204	0.0527
		Downstream	48,829	0.00318	0.00195	0.0449

of the DFE were $b = 0.30$, mean $N_e s = 10{,}000$, and $\alpha_{true} = 0$. The sample size was 40 chromosomes, and the expected diversity and divergence at silent sites were 0.01 and 0.2, respectively, similar to the levels found in the Drosophila data set analyzed here. In the simulations with population size expansion or contraction, the population was increased or decreased by 2-fold N_1 generations in the past. For the simulations involving population subdivision, populations were duplicated xN generations in the past and then allowed to evolve independently before being both sampled together.

Data

Human-gene sequences and associated polymorphism frequency data were downloaded from the Environmental Genome Project (EGP) web site (University of Washington, Seattle, WA; URL: http://egp.gs.washington.edu August 2007; Livingston et al. 2004) and from the Program for Genomic Applications (PGA) web site (National Heart, Lung and Blood Institute SeattleSNPs, Seattle, WA; URL: http://pga.gs.washington.edu August 2007). These data sets are among the largest available for which there is minimal single nucleotide polymorphism (SNP) ascertainment bias. We analyzed polymorphism frequency data for African and European samples separately. As described by Keightley and Eyre-Walker (2007), we set a minimum number of alleles for each population and excluded sites that reported fewer than these minima (which resulted in the exclusion of about 5% of the polymorphic sites); for sites with greater than this minimum, alleles were randomly sampled without replacement. We analyzed 38 alleles for both the African and European PGA data sets and 42 and 34 for the African and European EGP data sets, respectively. Macaque sequences (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) orthologous to the EGP–PGA human sequences (which refer to version hg17 of the human genome) were obtained from the UCSC hg18/RheMac2 alignments (downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsRheMac2/) with the aid of the liftOver utility. We analyzed the human–macaque divergence rather than the human–chimpanzee in order to increase the power of the analysis. The macaque sequences were realigned to the hg17 sequences using MAVID (Bray and Pachter 2004). We extracted all intronic sequences, all exons, and the 500 bp upstream and downstream from the start and stop codon, respectively. For the neutral sites, we used intronic sites, because most intronic sites evolve only slightly more slowly than pseudogenes in mammals (Gaffney and Keightley 2006). To lessen the influence of selection on our neutral standard, we removed some intronic sites under purifying selection by excluding the first 6 and last 16 bp of each intron, which are involved in splicing. We further excluded putatively functional sites by excluding sites for which the probability of being in its most conserved state is >0.1 in the "PhastCons" hidden Markov model from the analysis of 17 vertebrate genomes (Siepel et al. 2005; data downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons17way/). We thereby excluded intronic nucleotides conserved deep into the vertebrate phylogeny, and observed a slight increase in the mean intronic evolutionary divergence and diversity. This method is likely to remove blocks of strongly negatively selected sites, but may fail to remove selected sites that have a more dispersed distribution (Asthana et al. 2007), and if some mutations at these sites behave as nearly neutral this would lead to overestimation of α. For selected sites, we used protein-coding sites that are zero-fold degenerate in both species. For all classes of sequence, we only included sites that are unlikely to be part of a hypermutable CpG dinucleotide, that is, not preceded by a C or followed by a G in either species (Kondrashov et al. 2006). Bases for which the macaque quality score was <20 were excluded (as suggested by Rhesus Macaque Genome Sequencing and Analysis Consortium 2007), as were bases opposite a gap, bases within stretches of ≥ 100 bp in which the divergence was >0.3, and bases that were not included in the human polymorphism survey. Statistics concerning numbers of loci, numbers of sites analyzed, and polymorphism and divergence levels are shown in table 1.

We also analyzed the *Drosophila melanogaster* polymorphism data set and the orthologous *Drosophila simulans* sequences for 397 loci described by Shapiro et al. (2007), which was kindly provided by Dr Joshua Shapiro. The *D. melanogaster* gene sequences refer to version 3 of the reference genome (Shapiro et al. 2007). We realigned the set of *melanogaster* sequences and the *simulans* sequences using MAVID (Bray and Pachter 2004), under a model of a *melanogaster* star phylogeny with *simulans* as an outgroup. We considered the African *melanogaster* alleles, which originated from Zimbabwe (10 alleles) and Botswana and Zambia (2 alleles each). We excluded sites for which fewer than 12 alleles were sequenced; if more than 12 alleles were sequenced at a site, we randomly sampled 12 alleles without replacement. We generated SFSs for zero-fold and 4-fold sites of exons and for introns for the *melanogaster* allele

Table 2
Details of the Drosophila Data Set

		No. Sites		Proportion of Sites Segregating		Divergence to *Simulans*	
Selected DNA Type	No. Loci	Selected	4-Fold	Selected	4-Fold	Selected	4-Fold
Zero-fold	397	185,023	43,788	0.00425	0.0478	0.0105	0.104
Intron	291	57,894	30,317	0.0406	0.0530	0.0892	0.103

sequences and corresponding divergences using the alignment of a randomly chosen *melanogaster* sequence to the *simulans* sequence. We used 4-fold sites as our neutrally evolving standard. It might be desirable to use short introns as the neutral standard (Halligan and Keightley 2006), but there were insufficient short intron data for this purpose. The mean length of the extracted introns was 128 bases. Details of the Drosophila data set are shown in table 2. We used the Jukes–Cantor method to correct divergences for multiple hits for all data sets.

Mean Estimates and Bootstrapping

For humans, we took weighted averages of the estimates for the two data sets (PGA and EGP) within populations (Africa and Europe). To combine the results for the African and European populations, we also took averages, this time weighted by the numbers of alleles sampled in each population. To obtain estimates of standard errors (SEs) and confidence limits (CLs), we bootstrapped the data sets by locus 200 times. Estimates for the two human populations are likely to be correlated, because the same loci were surveyed. To obtain joint bootstrap estimates for Africa and Europe, we therefore generated paired bootstrap data sets consisting of the same lists of loci, resampled with replacement, for each population.

Results

Simulations

In order to verify the method to estimate α, we analyzed simulated nucleotide polymorphism and divergence data for cases in which the simulation model and analysis model agreed. In this case the effective population size under which the nucleotide divergence data were simulated (N_0) equaled the most recent effective population size (N) used to generate the polymorphism data. In the simulations, we allowed recent population expansions or contractions from population sizes N_1 to N_2. The results of these simulations (table 3) suggest that under the scenarios simulated the method recovers simulated values of α with minimal bias. Although reassuring, it is perhaps not surprising that parameter estimates show little bias when the data are simulated under the model assumed by the analysis method. We previously showed that our method to estimate DFE parameters (Keightley and Eyre-Walker 2007) is robust to substantial departures from the demographic model assumed, including three-epoch scenarios. This implies that α will also be recovered with little bias. However, there are two key assumptions that might be broken in real data.

First, the method assumes that the effective population size experienced by the polymorphism data is the same as the mean effective population size during the divergence phase. If the effective population size in the divergence phase differs from that in the polymorphism phase, then bias is expected because nearly neutral deleterious mutations can make larger or smaller relative contributions to polymorphism and divergence (McDonald and Kreitman 1991; Eyre-Walker 2002). The extent of this bias can be inferred as follows: Let the average (harmonic mean) of the effective population size during the divergence phase be N_D and the effective population size during the polymorphism phase (i.e., the average effective population size since the age of the oldest coalescence) be N_P. Under the assumption that the DFE can be described by a gamma distribution with a shape parameter b and a mean absolute selection coefficient of $E(s)$, neither of which change if the effective population size changes, it can be shown that the average fixation probability of a new mutation is proportional to $(1/N_D)^b$. This relationship was first inferred by Chao and Carr (1993), based on two special cases derived by Ohta (1977) and Kimura (1979), and has recently been generalized by Welch et al. (2008). The expected (true) number of selected-site substitutions (per site) during the divergence phase is then

$$d_{N_{true}} = \frac{2\mu tk}{(1 - \alpha_{true})N_D^b}, \quad (7)$$

where μ is the mutation rate, t is the time of divergence, k is a constant, and α_{true} is the true proportion of adaptive substitutions. Given our estimate of the shape parameter b and effective population size N_P from the analysis of polymorphism data, the estimated number of selected-site substitutions originating from effectively neutral deleterious mutations during the divergence phase is

$$d_{N_{est}} = \frac{2\mu tk}{N_P^b}. \quad (8)$$

This will be an underestimate of the true number of effectively neutral substitutions if $N_P > N_D$, and an overestimate if $N_P < N_D$. From equations (7) and (8), the estimated value of α is therefore

$$\alpha_{est} = \frac{d_{N_{true}} - d_{N_{est}}}{d_{N_{true}}}. \quad (9)$$

Equation (9) simplifies to

$$\alpha_{est} = 1 - (1 - \alpha_{true})\lambda^{-b}, \quad (10)$$

where $\lambda = N_P/N_D$. We evaluated this formula for a range of values of α_{true}, λ and values of b compatible with estimates for real data (Eyre-Walker et al. 2006; Keightley and

Table 3
Simulation Results for Cases in Which the Analysis and Simulation Models Agree

Demographic Scenario	Simulated Values		Mean Estimate [SD]
	b	α	α
Constant population	0.2	0	−0.0037 [0.057]
	1		−0.0040 [0.085]
	0.2	0.25	0.24 [0.052]
	1		0.26 [0.074]
	0.2	0.5	0.52 [0.049]
	1		0.49 [0.048]
	0.2	0.75	0.78 [0.013]
	1		0.76 [0.021]
2-Fold expansion	0.2	0	−0.065 [0.11]
	1		−0.14 [0.17]
	0.2	0.25	0.19 [0.11]
	1		0.19 [0.15]
	0.2	0.5	0.47 [0.047]
	1		0.46 [0.079]
	0.2	0.75	0.76 [0.026]
	1		0.72 [0.042]
2-Fold contraction	0.2	0	−0.023 [0.057]
	1		−0.062 [0.089]
	0.2	0.25	0.23 [0.055]
	1		0.21 [0.051]
	0.2	0.5	0.50 [0.020]
	1		0.49 [0.051]
	0.2	0.75	0.75 [0.013]
	1		0.76 [0.038]

Parameters of the simulation: $N_1 = 50$, $t = 50$, $E(s) = 0.1$, $f_0 = 0.95$, $d_S = 0.1$. There were 10,000 neutral and selected sites and 20 replicates per parameter combination.

Table 4
Evaluation of equation (10) to Infer Bias in Estimating α in the Presence of a Change in Long-Term Population Size Compared with Simulation Results (*sim*)

α_{true}	b	λ	α_{est} (eq. 10)	α_{est} (sim) [SD]
0	0.1	1/8	−0.23	−0.18 (0.15)
		1/2	−0.072	−0.032 (0.12)
		2	0.067	0.12 (0.10)
		8	0.19	0.26 (0.076)
	0.35	1/8	−1.1	−1.2 (0.53)
		1/2	−0.27	−0.36 (0.32)
		2	0.22	0.19 (0.17)
		8	0.52	0.52 (0.058)
0.25	0.1	1/8	0.077	0.085 [0.12]
		1/2	0.20	0.28 [0.088]
		2	0.30	0.36 [0.075]
		8	0.39	0.42 [0.044]
	0.35	1/8	−0.55	−0.65 [0.37]
		1/2	0.044	0.031 [0.17]
		2	0.41	0.43 [0.080]
		8	0.64	0.61 [0.049]
0.5	0.1	1/8	0.38	0.42 [0.058]
		1/2	0.46	0.51 [0.050]
		2	0.53	0.57 [0.037]
		8	0.59	0.62 [0.037]
	0.35	1/8	−0.035	−0.014 [0.17]
		1/2	0.36	0.39 [0.082]
		2	0.61	0.62 [0.053]
		8	0.76	0.75 [0.029]

The parameters were: $E(s)$ = -0.1, nonsynonymous divergence = 0.1, and there were 20 replicates per parameter combination.

Eyre-Walker 2007; Boyko et al. 2008), and compared the results with simulations (table 4). For cases with $\lambda > 1$ ($\lambda < 1$) (i.e., there has been a population expansion [contraction]), there is an excess (deficit) of nearly neutral deleterious mutant substitutions compared with that expected from the polymorphism data, so α is over(under)estimated. The over(under)estimation is most serious for high values of the shape parameter of the gamma distribution (b), because a higher proportion of mutations fall into the nearly neutral class. The results suggest that the extent of over(-under)estimation can be quite serious if there has been a major population expansion (contraction).

The second key assumption that might be violated in real data is that of free recombination. To investigate the consequences of this we ran a series of sfs_code simulations (Hernandez 2008) with varying levels of recombination. In each simulation, there was free recombination between genes but no recombination within genes. We simulated five recombination scenarios: 100 genes of 500 bp, 50 genes of 1,000 bp, 10 genes of 5,000 bp, 2 genes of 25,000 bp, and 1 gene of 50,000 bp. We simulated a stationary population size with no adaptive evolution and a population size that had undergone a 2-fold expansion or a 2-fold contraction N_1 generations in the past.

For all parameter combinations, the method appears to estimate the shape parameter of the distribution of fitness effects with little bias, although estimates of the mean effect are very noisy (data not shown), as has been noted previously (Keightley and Eyre-Walker 2007) (table 5). With free and restricted recombination, α is estimated with almost no bias when population sizes are stationary. As expected, α is overestimated if there has been population-size expansion and underestimated if there has been contraction. However, the degree of bias does not seem to depend on the level of recombination as long as there is not very restricted linkage. With complete linkage or very restricted linkage, α is overestimated relative to the expectations with free recombination. This is probably due to the fact that the SFS becomes discontinuous, with some frequency classes overrepresented, making it difficult to fit any model. It therefore seems that the method is fairly robust to the effects of linkage as long as there is not complete linkage.

Estimates of the Distribution of Effects of Deleterious Mutations in Coding and Noncoding DNA of Humans and Drosophila

The method simultaneously infers the distribution of fitness effects of new neutral and deleterious mutations and the proportion of adaptive substitutions. Estimates of the mean effect of a deleterious mutation are strongly influenced by large-effect alleles that are rare in any reasonably sized population sample, so they tend to have wide confidence intervals. We therefore focus on comparing the estimated proportions of mutations in different $N_e s$ ranges (table 6), which are less sensitive to strongly deleterious mutations.

In humans, our results suggest that 30–40% of amino acid-changing mutations behave as effectively neutral ($0 < N_e s < 1$), whereas in Drosophila, only about 6% behave as effectively neutral. Conversely, in Drosophila about 90% of amino acid–changing mutations are so strongly selected

Table 5
Simulations with Varying Levels of Recombination and Population Size Expansion/Contraction or Population Subdivision

No. of Loci	Locus Length (bp)	Parameter Estimates [SE]	
		B	α_{est}
Stationary population size			
100	500	0.31 [0.013]	−0.014 [0.037]
50	1,000	0.32 [0.013]	0.068 [0.031]
10	5,000	0.30 [0.014]	0.0082 [0.035]
2	25,000	0.31 [0.016]	0.14 [0.039]
1	50,000	0.39 [0.022]	0.33 [0.041]
Population size expansion			
100	500	0.31 [0.010]	0.16 [0.024]
50	1,000	0.32 [0.012]	0.18 [0.026]
10	5,000	0.30 [0.011]	0.15 [0.031]
2	25,000	0.31 [0.015]	0.25 [0.034]
1	50,000	0.32 [0.024]	0.30 [0.041]
Population size contraction			
100	500	0.33 [0.015]	-0.092 [0.042]
50	1,000	0.32 [0.016]	-0.15 [0.052]
10	5,000	0.31 [0.017]	-0.17 [0.048]
2	25,000	0.36 [0.023]	0.019 [0.055]
1	50,000	0.36 [0.030]	0.039 [0.065]
Population subdivision 4N generations in the past			
50	1,000	0.32 [0.0089]	0.060 [0.023]
Population subdivision 2N generations in the past			
50	1,000	0.30 [0.0080]	0.018 [0.026]
Population subdivision N generations in the past			
50	1,000	0.32 [0.0078]	0.061 [0.025]

The simulated values of α and b were 0 and 0.3, respectively.

that they almost never become fixed ($N_e s > 10$), whereas the corresponding figure for human populations is about 50%. These results have been reported previously for similar polymorphism data sets (Keightley and Eyre-Walker 2007). For both humans and Drosophila, there are large differences in log L between models with and without selection at zero-fold sites, and the proportion of mutations with $N_e s > 10$ is highly significantly different from zero (table 6).

Widespread purifying selection at zero-fold sites, albeit less effective in humans than Drosophila, contrasts with the estimates for the strength of selection on mutations in the 500-bp regions upstream and downstream of the start and stop codon in humans (tables 6). For most data sets, differences in log L between models with and without selection are nonsignificant, and the proportions of mutations estimated to have $N_e s > 1$ are generally very small. This implies that our analysis, which is based solely on the SFSs, gives little evidence for deleterious mutations in these regions, a somewhat surprising result, because sequences involved in the control of gene expression are thought to be concentrated in these regions (Veyrieras et al. 2008).

In Drosophila, the results strongly support purifying selection in introns (differences in log L from no selection model = 26; $P < 0.001$). Furthermore, the intronic nucleotide divergences between *D. melanogaster* and *D. simulans* are markedly lower than at 4-fold sites (table 2). However, the selective effects of mutations in introns tend to be markedly weaker than at zero-fold sites, because relatively few intronic mutations have selective effects $N_e s > 100$.

As expected, in both African populations of humans and Drosophila, the demographic parameter estimates point to recent population expansion, and parameter estimates are similar to those reported previously from analysis of similar data sets (Keightley and Eyre-Walker 2007, table 4).

Estimates of Frequencies of Adaptive Substitutions in Humans and Drosophila

Taken at face value, our results suggest striking differences in the proportion of adaptive substitutions between humans and Drosophila (table 6), as others have noted previously (reviewed in Eyre-Walker 2006). In humans, the mean estimate for α for amino acid substitutions is −0.00, compared with 0.52 in Drosophila (tables 6 and 7). In humans, there are therefore apparently very few adaptive substitutions apparent in protein-coding genes, a finding that we shall discuss in relation to the simulation results. It should be noted that the CLs on our estimates for humans are quite wide, so moderately high levels of adaptive evolution cannot be ruled out. It should also be noted that our analysis uses current polymorphism data in humans to infer the DFE, and this DFE to infer adaptive evolution along the

Table 6
Estimates of Percentages of Mutations in Different $N_e s$ Ranges, the Differences in Log Likelihood (Δlog L) between Models with and without Selection, and Estimates of the Proportion of Adaptive Substitutions, α

Species	Population	DNA Type	Percentage of Mutations in $N_e s$ Range [SE]				ΔLog L	α [SE]
			0–1	1–10	10–100	>100		
Human	Africa PGA	Zero-fold	35 [6]	9 [5]	12 [9]	44 [10]	92	0.06 [0.18]
	Europe		38 [7]	18 [8]	25 [10]	19 [15]	44	0.01 [0.21]
	Africa EGP		30 [5]	6 [4]	6 [7]	58 [9]	97	−0.12 [0.23]
	Europe		29 [6]	22 [9]	33 [9]	17 [14]	42	0.02 [0.25]
	Africa PGA	Upstream	97 [8]	3 [6]	0 [2]	0 [2]	0.1	0.01 [0.18]
	Europe		88 [11]	6 [8]	5 [3]	1 [5]	0.6	0.03 [0.15]
	Africa EGP		100 [5]	0 [5]	0 [0]	0 [0]	0	−0.04 [0.12]
	Europe		100 [5]	0 [4]	0 [2]	0 [0]	0	−0.04 [0.11]
	Africa PGA	Downstream	83 [12]	6 [12]	6 [4]	5 [7]	2.0	0.05 [0.17]
	Europe		78 [16]	20 [15]	2 [4]	0 [7]	1.8	0.18 [0.21]
	Africa EGP		100 [6]	0 [1]	0 [0]	0 [4]	0	−0.24 [0.09]
	Europe		100 [5]	0 [4]	0 [1]	0 [0]	0	−0.24 [0.15]
Drosophila	Africa	Zero-fold	6 [1]	7 [1]	17 [5]	70 [5]	1,800	0.52 [0.06]
		Intron	70 [12]	23 [12]	7 [5]	0 [4]	26	0.23 [0.14]

Table 7
Summary of ML Estimates of α for Human-Macaque, as Weighted Averages across Populations and Data Sets, and Their 95% CL

DNA Type	α [95% CL]
Zero-fold	−0.00 [−0.30, 0.24]
Upstream	−0.01 [−0.11, 0.28]
Downstream	−0.03 [−0.17, 0.22]

human and macaque lineages. The implications of this are discussed below.

Estimates of the frequencies of adaptive substitution in the 500-bp noncoding DNA stretches upstream and downstream of the start codon in humans are also all nonsignificantly different from zero (tables 6 and 7). Unfortunately, we do not yet have upstream and downstream data for Drosophila. In Drosophila introns, the results suggest that adaptive substitutions are common, $\alpha = 20\%$, which is significantly different from zero ($P = 0.04$).

Table 8
Comparison between Estimates of α Obtained Using the Method of FWW and the Inference Method Described in This Paper

	Method of Inference		
Data Set	FWW—All	FWW ≥ 15%	This Paper
Drosophila zero-fold	−0.01	0.22	0.52
Drosophila introns	0.07	0.11	0.23
PGA African zero-fold	−0.20	−0.05	0.06
PGA Euro zero-fold	−0.41	−0.21	0.01
EGP African zero-fold	−0.35	−0.15	−0.12
EGP Euro zero-fold	−0.67	−0.30	0.02
PGA African upstream	−0.14	−0.08	0.01
PGA Euro upstream	−0.03	−0.08	0.03
EGP African upstream	−0.26	−0.02	−0.04
EGP Euro upstream	−0.23	−0.24	−0.04
PGA African downstream	−0.02	0.05	0.05
PGA Euro downstream	−0.01	0.13	0.18
EGP African downstream	−0.25	−0.47	−0.24
EGP Euro downstream	−0.40	−0.25	−0.24

FWW—all: using the complete polymorphism data set; FWW ≥ 15% excluding variants segregating at a frequency of <15%.

Comparison to Other Methods

Fay et al. (2001) (FWW) have suggested the simple correction of removing polymorphisms that segregate at frequencies below a certain value in order to partially control for the effects of slightly deleterious mutations in the estimation of α. We have compared our method with the FWW method either using all polymorphisms, or excluding those below 15%. Estimates of α appear to asymptote if cutoff frequencies are substantially above 15% (Charlesworth and Eyre-Walker 2008).

In almost all cases (table 8), FWW tends to give lower estimates of α than our method, and this is particularly evident in data sets in which we infer that a substantial fraction of segregating polymorphisms originate from slightly deleterious mutations. For example, in the Drosophila zero-fold data, we infer that the proportion of mutations with effects $N_e s$ between 1 and 10 exceeds the proportion with effects between 0 and 1, and consequently our method gives substantially higher values of α than the FWW method.

Besides being downwardly biased, there is another potential problem with excluding rare variants, a loss of data and hence an increase in the variance of the estimate of α. On the other hand, our method requires the estimation of extra parameters to correct for the fixation of slightly deleterious mutations. To investigate whether the new method differs from FWW in this respect, we estimated the SE for α using the FWW method by bootstrapping, including all polymorphisms and excluding polymorphisms below 15%, for the Drosophila zero-fold data. Surprisingly the SEs are very similar: FWW with all polymorphisms, SE = 0.07; excluding SNPs below 15%, SE = 0.07, whereas the SE from our method is 0.06. This suggests that including information from the DFE does not lead to an increase in the variance of the estimate of α, compared with the FWW method.

Discussion

The approach we describe estimates the proportion of adaptive substitutions by a two-stage process. Parameters of the distribution of effects of deleterious mutations are first estimated from polymorphism data. These parameter estimates are then used to predict the expected number of substitutions originating from deleterious mutations, and this is compared with the observed number of substitutions to estimate α. Its principal advantage is that it will give unbiased estimates of α if the assumptions of the model are met, whereas previous methods are generally expected to underestimate α, even if low-frequency polymorphisms are excluded (Charlesworth and Eyre-Walker 2008). The method also removes the need to set an arbitrary threshold below which low-frequency polymorphisms are excluded. However, the method requires us to assume that the DFE comes from some family of distributions (gamma in this case) and requires a specific demographic model.

We have also implemented a method to estimate α that uses estimates of the DFE obtained by the method of Eyre-Walker et al. (2006). Generally, the results from this approach agree closely with the results obtained using the method implemented here to estimate the DFE (Keightley and Eyre-Walker 2007), both in the proportion of mutations estimated within each range of $N_e s$, and α (supplementary tables 1 and 2, Supplementary Material online). This is perhaps not surprising, given that the two methods agree closely in their estimates of the DFE (see Keightley and Eyre-Walker 2007, table 6, supplementary table 1, Supplementary Material online). There are advantages and disadvantages to this alternative method. Rather than modeling changes in demography, demographic effects are corrected by an approximation. This simplification potentially allows the method to be applied to more involved data structures, which are not easily accommodated within the model given here, or more complex situations in which it is not possible to model demography explicitly; one such example is given below. The second method also appears to be slightly more robust to restricted recombination.

In our analysis, we have summed data across genes, but this can potentially lead to biases if the effective population

size varies across the genome and the SFS is correlated with effective population size. For example, regions of the genome with the highest effective population size are expected to have the highest silent site diversity, but relatively low levels of nonsilent diversity. As a consequence, the SFS will be dominated by genes that appear to be subject to relatively high levels of selective constraints, and genomewide estimates of α will then be overestimates (Smith and Eyre-Walker 2002; Welch 2006; Shapiro et al. 2007). Due to computational limitations, the method principally described in this paper needs to be run on data summed across genes. However, we can use the alternative method described in the supplementary material to estimate α based on the method to estimate the DFE of Eyre-Walker et al. (2006). In this analysis, we allow each gene to have its own effective population size and mutation rate and assume that the demographic parameters and the distribution of s are shared across genes (Eyre-Walker et al. 2006). We have run this analysis on the Drosophila zero-fold data, for which there is a highly significant negative correlation between the apparent effectiveness of selection, as measured by the ratio of the number of nonsynonymous to synonymous polymorphisms, and the effective population size across the genome, as measured by silent site diversity, correcting for the obvious nonindependence using the method of Piganeau and Eyre-Walker (2009) (Adam Eyre-Walker, unpublished results). The resulting estimate of α is almost identical to the estimate that used summed data (data not shown). It therefore seems that summing data across genes will generally not lead to serious biases.

Our method to infer the DFE attempts to account for recent demographic changes, but our method to estimate α assumes, like all current MK test–based inference methods, that the strength and effectiveness of selection has remained constant during the evolutionary time period under analysis (McDonald and Kreitman 1991; Eyre-Walker 2002). We have derived a formula to predict the extent of over or underestimation of α if there have been long-term changes in effective population size. This depends on the magnitude of the population-size change and the shape of the distribution of effects of new mutations. If there has been a population-size contraction, the presence of slightly deleterious mutations is expected to lead to an underestimation of α, because some slightly deleterious mutations that currently segregate would not have segregated or become fixed in the past. Conversely, if there has been a population-size expansion, some weakly selected deleterious mutations would have become fixed in the past, while contributing proportionally less to polymorphism, potentially leading to an overestimation of α. However, population-size expansion is also likely to lead to a round of adaptive evolution via back- and compensatory mutation (Gillespie 1994; Charlesworth and Eyre-Walker 2007). If there are back and compensatory mutations, it is unclear whether α would be under or overestimated. It is therefore important to appreciate that population-size expansion or contraction is expected to lead to biased estimates of α, but the magnitude and direction will depend on the relative numbers of slightly deleterious and slightly advantageous mutations.

Huerta-Sanchez et al. (2008) have recently shown that fluctuating selection can mimic aspects of directional selection, even when the net selection coefficient of a new mutation is zero. Fluctuating selection increases the fixation probability of a new mutation but decreases the chance that a mutation will be found segregating in a sample of sequences. Fluctuating selection also tends to skew the SFS toward rare alleles. As a consequence, if fluctuating selection is common, our method will tend to overestimate α to an even greater extent than the FWW method. However, Huerta-Sanchez et al. (2008) only considered fluctuating selection models in which there are large changes in the strength of selection on a frequent basis. It will be of interest to investigate this model further.

We assume that advantageous mutations do not contribute to the level of polymorphism. This is a reasonable approximation if the strength of selection is strong, because strongly advantageous mutations must be rare, and even the most strongly selected advantageous mutations contribute only twice the level of heterozygosity as a neutral mutation (Kimura 1983). However, if advantageous mutations are weakly selected, we may underestimate the proportion of adaptive substitutions (Boyko et al. 2008). This is because weakly advantageous mutations contribute to polymorphism, implying that fewer effectively neutral deleterious mutations are required to explain a given level of diversity, which in turn implies that we expect fewer effectively neutral substitutions. There are recent estimates of the strength of selection acting upon new advantageous mutations, but these have differed substantially. For example, $N_e s$ for advantageous mutations has been estimated in Drosophila to be ~40 (Andolfatto 2007), ~350 to ~3,500 (Eyre-Walker 2006), ~2,000 (Li and Stephan 2006; Jensen et al. 2008), and ~10,000 (MacPherson et al. 2007) (the last three estimates were obtained by multiplying the estimate of s by an effective population size of 1,000,000). The extent by which weakly advantageous mutations may cause underestimation of the proportion of substitutions is therefore still unclear.

The results of the analysis of human and Drosophila polymorphism and divergence data need to be interpreted in the light of the fact that estimates of α can be strongly biased by changes in effective population size. Our estimates of α for Drosophila are largely consistent with previous estimates for *D. melanogaster* and other species. We estimate that ~50% of amino acid substitutions between *D. melanogaster* and *D. simulans* at zero-fold sites have been driven by positive selection. This agrees with previous estimates in *D. simulans* (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Welch 2006; Haddrill et al. 2008). Our estimate is similar to one estimate (Andolfatto 2007) that uses *D. melanogaster* polymorphism data, but rather higher than two other estimates (Bierne and Eyre-Walker 2004; Shapiro et al. 2007). Our estimate of α for Drosophila zero-fold sites is also similar to estimates for protein-coding genes in the *Drosophila americana* group (Maside and Charlesworth 2007), which used the method of Bierne and Eyre-Walker (2004) (which typically gives similar estimates to FWW), and is also similar to estimates in *Drosophila miranda* based on FWW and the method of Bierne and Eyre-Walker (2004) (Bachtrog 2008). An exception is estimates of α from *D. miranda* which are ~10% (Bachtrog and Andolfatto 2006). The similarity of the estimates among most Drosophila species allows us to

tentatively conclude that a high rate of adaptive protein evolution is general to Drosophila. The alternative explanation of a long-term population expansion in all species seems less likely. Our estimate of α for Drosophila introns, which indicates widespread adaptive evolution, is similar to those obtained by Andolfatto (2005) and Haddrill et al. (2008), who used the FWW method, excluding low-frequency variants. The consistency is explained by the relatively low frequencies of slightly deleterious mutations segregating in Drosophila introns.

In humans, there have been several previous comparisons of levels of polymorphism with between-species divergence of protein-coding sequences (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005; Gojobori et al. 2007; Boyko et al. 2008). In common with our analysis, these studies have inferred low levels of adaptive evolution in protein-coding sequences. It might be that hominids have very low rates of adaptive evolution, possibly as a consequence of their low effective population size. There is evidence that the effectiveness of natural selection is weaker in hominids than other mammals (Li et al. 1987; Eyre-Walker and Keightley 1999; Keightley, Lercher, Eyre-Walker 2005; Keightley, Kryukov, et al. 2005), particularly in the human lineage after the divergence from the common ancestor with chimpanzee (Bakewell et al. 2007). However, it is also possible that the apparent low level of adaptive evolution is an artifact, brought on by a difference between the current effective population size of humans and the ancestral population size in hominids. As we have shown above, if $N_D > N_P$, then α is expected to be underestimated, and this underestimation can be substantial. Based on an analysis of DNA-sequence data from multiple loci, Burgess and Yang (2008) have recently concluded that the effective sizes of ancestral hominid populations, including the human–chimp ancestor, were 5–10 times higher than current effective population sizes of humans, chimps, and gorillas. If we assume, for simplicity, that the effective population size during the evolution of humans and macaques was at this higher size until very recently, then we can predict the true value of α, given the observed value (about 0.05). Assuming a shape parameter (b) for the DFE of 0.2 (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Boyko et al. 2008), the true value of α would therefore be 0.31 if the $N_D = 5N_P$ and 0.40 if $N_D = 10N_P$. These values are therefore only slightly lower than the α value estimated for protein-coding genes in Drosophila. Estimates of b for human noncoding DNA are very imprecise, so we are uncertain about the extent of underestimation of α for these DNA sequence types. It also should be noted, finally, that adaptive substitutions may be frequent in noncoding DNA of hominids, but with the available data, adaptive substitutions are difficult to detect by methods based on the MK test against a background of many nearly neutral substitutions.

Supplementary Material

Supplementary tables 1 and 2 and supplementary method are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Joshua Shapiro for providing the Drosophila intron sequence data and Ryan Hernandez for providing sfs_code prior to publication. We are grateful to Andrea Betancourt, Penny Haddrill, Dan Halligan, John Welch, Brian Charlesworth and anonymous reviewers for helpful comments, and the Biotechnology and Biological Sciences Research Council for support.

Software availability: A web-server to infer the distribution of effects of new mutations and α based on the methods described above is available at P.D.K.'s website: http://homepages.ed.ac.uk/eang33/. Routines to estimate α based on Eyre-Walker et al. (2006) have been implemented within the DoFE program for Mac, Linux, and Windows operating systems, available at http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Welcome.html.

A.E.W. and P.D.K. formulated the basic method; P.D.K. implemented the method presented in this paper, and A.E.W. implemented the method presented in the supplementary material. P.D.K. assembled the data; P.D.K. and A.E.W. performed the analyses and wrote the paper.

Literature Cited

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2:e286.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature. 437:1149–1152.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial substitutions in the Drosophila genome. Genome Res. 17:1755–1762.

Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007. Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci USA. 104:12410–12415.

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

Bachtrog D, Andolfatto P. 2006. Selection, recombination and demographic history in *D. Miranda*. Genetics. 174:2045–2059.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA. 104:7489–7494.

Bierne N, Eyre-Walker A. 2004. Genomic rate of adaptive aminoacid substitution in Drosophila. Mol Biol Evol. 21:1350–1360.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. ssessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4:e1000083.

Bray N, Pachter L. 2004. MAVID: constrained ancestral alignment of multiple sequences. Genome Res. 14:693–699.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. 25:1979–1994.

Chao L, Carr DE. 1993. The molecular clock and the relationship between population size and generation time. Evolution. 47:688–690.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res. 63:213–227.
Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. Mol Biol Evol. 23:1348–1356.
Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proc Natl Acad Sci USA. 104:16992–16997.
Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol. 25:1007–1015.
Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.
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. 1999. High genomic deleterious mutation rates in hominids. Nature. 397:344–347.
Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness of new deleterious amino acid mutations in humans. Genetics. 173:891–900.
Fay J, Wycoff GJ, Wu C-I. 2001. Positive and negative selection on the human genome. Genetics. 158:1227–1234.
Gaffney D, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. PLoS Genet. 2:e204.
Gillespie JH. 1991. The causes of molecular evolution. Oxford (United Kingdom): Oxford University Press.
Gillespie JH. 1994. Substitution processes in molecular evolution. III. Deleterious alleles. Genetics. 138:943–952.
Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics. 165:1269–1278.
Gojobori J, Tang H, Akey JM, Wu C-I. 2007. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. Proc Natl Acad Sci USA. 104:3907–3912.
Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding DNA in *Drosophila simulans*. Mol Biol Evol. 25:1825–1834.
Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. Genome Res. 16:875–884.
Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics. 24:2786–2787.
Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics. 116:153–159.
Huerta-Sanchez E, Durrett R, Bustamante CD. 2008. Population genetics of polymorphism and divergence under fluctuating selection. Genetics. 178:325–337.
JensenJD Thornton KR, Andolfatto P. 2008. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. PLoS Genet. 4:1–11.
Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 177:2251–2261.
Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res. 15:1373–1378.
Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol. 3:282–288.
Kimura M. 1957. Some problems of stochastic processes in genetics. Ann Math Stat. 28:882–901.
Kimura M. 1979. Model of effectively neutral mutations in which selective constraint is incorporated. Proc Natl Aca Sci USA. 76:3440–3444.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Kondrashov FA, Ogurtsov AY, Kondrashov AS. 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.
Li HP, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2:1580–1589.
Li W-H, Tanimura M, Sharp PM. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol. 25:330–342.
Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. 2004. Pattern of sequence variation across 213 environmental response genes. Genome Res. 14:1821–1831.
Loewe L, Charlesworth B, Bartolomé C, Nöel V. 2006. Estimating selection on non-synonymous mutations. Genetics. 172:1079–1092.
Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genome-wide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics. 177:2083–2099.
Maside X, Charlesworth B. 2007. Patterns of molecular variation and evolution in *Drosophila americana* and its relatives. Genetics. 176:2293–2305.
McDonald JH, Kreitman M. 1991. Adaptive evolution at the *Adh* locus in Drosophila. Nature. 351:652–654.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Gene. 39:197–218.
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.
Ohta T. 1977. Extension of the neutral mutation drift hypothesis. In: Kimura M, editor. Molecular evolution and polymorphism. Mishima: National Institute of Genetics. p. 148–167.
Piganeau G, Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. PLoS ONE. 4:e4396.
Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science. 316:222–234.
Sabeti PC, Varilly P, Fry B, et al. (12 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. Nature. 449:913–918.
Shapiro JA, Huang W, Zhang C, et al. (12 co-authors). 2007. Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci USA. 104:2271–2276.
Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature. 415:1022–1024.
Veyrieras J, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 4:e1000214.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

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

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

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet. 3: e90.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005. The effects of artificial selection on the maize genome. Science. 308:1310–1314.

Zhang L, Li W-H. 2005. Human SNPs reveal no evidence of frequent positive selection. Mol Biol Evol. 22:2504–2507.

Barbara Holland, Associate Editor

Accepted June 11, 2009