Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models

Asif U. Tamuri,* Mario dos Reis,*,† and Richard A. Goldstein*,[1]
*Medical Research Council National Institute for Medical Research, London, NW7 1AA, United Kingdom, and †Department of Genetics, Evolution, and Environment, University College London, London, WC1E 6BT, United Kingdom

ABSTRACT Estimation of the distribution of selection coefficients of mutations is a long-standing issue in molecular evolution. In addition to population-based methods, the distribution can be estimated from DNA sequence data by phylogenetic-based models. Previous models have generally found unimodal distributions where the probability mass is concentrated between mildly deleterious and nearly neutral mutations. Here we use a sitewise mutation–selection phylogenetic model to estimate the distribution of selection coefficients among novel and fixed mutations (substitutions) in a data set of 244 mammalian mitochondrial genomes and a set of 401 PB2 proteins from influenza. We find a bimodal distribution of selection coefficients for novel mutations in both the mitochondrial data set and for the influenza protein evolving in its natural reservoir, birds. Most of the mutations are strongly deleterious with the rest of the probability mass concentrated around mildly deleterious to neutral mutations. The distribution of the coefficients among substitutions is unimodal and symmetrical around nearly neutral substitutions for both data sets at adaptive equilibrium. About 0.5% of the nonsynonymous mutations and 14% of the nonsynonymous substitutions in the mitochondrial proteins are advantageous, with 0.5% and 24% observed for the influenza protein. Following a host shift of influenza from birds to humans, however, we find among novel mutations in PB2 a trimodal distribution with a small mode of advantageous mutations.

WHEN a novel mutation appears in the genome of an organism, it may have three different effects on the fitness ($w = 1 + s$) of its carrier: The mutation may be deleterious ($s < 0$), reducing fitness through reduced fertility or survival rate. It may be neutral ($s \approx 0$), that is, having such a small effect on fitness that the fate of the mutant is mostly determined by random drift. Or the mutation may be advantageous ($s > 0$), increasing the fitness of its carrier by increasing its fertility or survival in its environment. The frequency distribution of the different types of mutants and their associated selection coefficients (s, also known as fitness effects) is a key issue in population genetics (Bustamante 2005; Eyre-Walker and Keightley 2007). The ultimate fate of a mutation, whether it will become fixed or lost in a population, depends on the strength of selection and on the effect of random drift due to finite population size. In fact, the fitness effect s and the population number N are so closely linked that normally the distribution is expressed in terms of the population scaled coefficient $S = 2Ns$.

Kimura (1968, 1983), in his neutral theory of molecular evolution, proposed that the dominant fraction (p_-) of all novel mutations would be highly deleterious, with a minority fraction ($p_0 = 1 - p_-$) being neutral. When organisms colonize a new habitat or are subject to environmental change, the opportunity for adaptive evolution would arise, and a fraction ($p_+ = 1 - p_0 - p_-$) of novel mutations would be advantageous. The magnitudes of these fractions for a protein-coding gene would depend on the protein in question; functionally important or structurally constrained proteins (such as the histones) would be characterized by a very large fraction of deleterious mutations ($p_- \gg p_0$), while structurally less constrained proteins (such as the fibrinopeptides) would have a larger fraction of neutral mutations ($p_0 > p_-$). Extensions to Kimura's theory have been made, including considering the contribution of nearly neutral mutations to the evolutionary process (Ohta 1973, 1992; Kimura 1983). Under this latter extension, there is a spectrum of nearly neutral mutations ranging

doi: 10.1534/genetics.111.136432
Manuscript received November 3, 2011; accepted for publication December 18, 2011
Available freely online through the author-supported open access option.
[1]Corresponding author: Mathematical Biology, National Institute for Medical Research, Mill Hill, London NW7 1AA, United Kingdom. E-mail: richard.goldstein@nimr.mrc.ac.uk

from slightly deleterious to slightly advantageous, with the neutrality of a given change dependent on the population size; evolutionary trajectories consist of a balance between slightly deleterious and slightly advantageous substitutions. Others have argued that, even under more typical conditions, adaptive substitutions would be frequent, the greater probability of fixation compensating for their relative rarity among mutations (Gillespie 1994).

Akashi (1999) considered that under a neutral model the distribution of S among novel mutations could be bimodal, with the modes centered around highly deleterious and neutral mutations. During adaptive episodes, the distribution would have three modes, with a small additional mode centered around advantageous mutations. Because deleterious mutations have a vanishingly small probability of becoming fixed in a population, most substitutions (*i.e.*, fixed mutations) would be neutral. In this case, the distribution of S among substitutions would be unimodal and centered around neutral mutations. During an adaptive episode, natural selection would drive many positively selected mutations quickly to fixation. In this case, the distribution of substitutions would be bimodal, with modes centered around nearly neutral and advantageous substitutions.

While the effect of mutations can be studied experimentally, these studies are difficult to perform on higher organisms and too insensitive to observe any but the largest fitness effects (Eyre-Walker and Keightley 2007). Due to these limitations, alternative approaches have been developed that estimate the distribution of fitness effects from biological sequence data. Much of the work on estimation of the distribution of S from DNA sequence data has been based at the population level (*e.g.*, Sawyer and Hartl 1992; Bustamante *et al.* 2002). These methods usually work with allele data from different individuals within a population, and the level of polymorphism within the population and the number of fixed differences with an outgroup species are used to estimate the distribution. These methods look at the evolutionary process over relatively short periods of time and thus normally use approximate mutation models such as the infinite alleles model (Kimura 1969, 1983, p. 43). More recently, phylogenetic methods that look at the evolutionary process over longer periods of time have been used to estimate the distribution of selection coefficients (Nielsen and Yang 2003; Yang and Nielsen 2008; Rodrigue *et al.* 2010). Although these use more realistic mutation models than the population-based methods, they ignore polymorphism and assume that all the observed differences among species are fixed. These two approaches sometimes result in different conclusions; population-based methods can yield an extremely large fraction of adaptive changes (Fay *et al.* 2001), especially in *Drosophila* (Sawyer *et al.* 2003, 2007), while phylogenetic methods often result in more modest estimates of p_+ (Nielsen and Yang 2003; Rodrigue *et al.* 2010). Similarly, population methods find the distribution of slightly deleterious mutations falling off leptokurtically, that is, more rapidly than exponentially (such as in a gamma distribution with $\alpha < 1$) (Eyre-Walker *et al.* 2006), while evolutionary models often yield a more rounded distribution ($\alpha > 1$) (Nielsen and Yang 2003; Rodrigue *et al.* 2010). It is not clear whether these differences represent the different methodologies and the approximations that they make or the details of the particular organisms under study. Worryingly, the evolutionary models fail to yield a substantial amount of lethal mutations (Nielsen and Yang 2003; Rodrigue *et al.* 2010) that would be expected on the basis of mutation experiments (Wloch *et al.* 2001; Sanjuan *et al.* 2004; Hietpas *et al.* 2011) and have been obtained by population-based studies (Piganeau and Eyre-Walker 2003; Yampolsky *et al.* 2005; Eyre-Walker *et al.* 2006).

One of the difficulties in estimating the distribution of selection coefficients is the complex nature of the selective constraints, even within a single protein, representing a range of functional, structural, and physiological requirements. Certain locations, such as those involved in protein functionality, may be invariant, while other locations may have a wide latitude in the amino acids compatible with that position. It is not only the magnitude of the selective constraints that vary from one location to another; one position may be constrained to hydrophobic residues, another constrained to residues that can take part in hydrogen bonding interactions, and a third requiring a certain degree of flexibility. The types of substitutions that can occur can be substantially different, even among locations that are changing at similar rates. Different approaches have addressed this issue to various degrees. For instance, Nielsen and Yang (2003) considered that the overall rate of substitutions could vary from one location to another, but considered that this rate variation would affect all possible substitutions equally; that is, slowly varying locations were as unrestricted in the amino acids as rapidly varying locations. Thorne *et al.* (2007) relaxed the standard assumption of independent sites, considering the selective constraints imposed by the need to maintain a stable well-defined structure; this was estimated using protein structure prediction algorithms, despite their construction being motivated by a quite different problem. Rodrigue *et al.* (2010) adapted a mixture-model approach that grouped locations under similar selective constraints and developed more specific models for characterizing these different types of locations; each individual location was then represented by a mixture of these models (Koshi and Goldstein 1998). The available data determined the number of components in the mixture that could be justified.

The most specific characterization of the substitution process was developed by Halpern and Bruno (1998), who proposed a sitewise phylogenetic model where evolution at each amino acid residue in a protein is characterized by a location-specific set of fitnesses and by the nucleotide-level mutation pattern. Although Halpern and Bruno demonstrated its utility for the estimation of evolutionary distances, use of the model has been limited, as the number of adjustable parameters required more data and computational

resources than have previously been available. Here we explore the use of this model in the estimation of the distribution of S. We are interested in assessing how the assumption of site-specific fitnesses may affect estimates of the shape of the distribution of S among novel mutations and substitutions. We apply a modified version of their model to a data set of 12 mitochondrial proteins in 244 mammalian species. We also apply this model to a data set of a polymerase protein from 401 influenza viruses isolated from avian and human hosts. As the human viruses are the product of a host shift event from an avian host (Taubenberger *et al.* 2005), this allows us to investigate the distribution of selection coefficients during a well-defined adaptive episode.

Methods

In the following discussion we assume a Wright–Fisher model of random genetic drift (*e.g.*, Wright 1931). We work with idealized populations where the effective and the real population numbers are the same. Locations in a gene are assumed to evolve independently, and they do not interfere with each other. We assume the selection coefficients (s) involved in the model are small, so that simplifying approximations about relative fixation probabilities can be made. It is also assumed that mutation rates are sufficiently small in relation to the population size so that polymorphism is negligible and locations remain fixed most of the time (Crow and Kimura 1970, pp. 442–445). The evolutionary process is viewed over long periods so the time from appearance to fixation of a novel mutant is nearly instantaneous. These assumptions are necessary to simplify the mathematical treatment of the model as discussed below.

Basic model

We model the substitution rate of a codon location in a functional protein under the action of selection, mutation, and random drift as a time-continuous Markov process. We modify the model of Halpern and Bruno (1998), and we use the notation of Yang and Nielsen (2008). Let us write $I = i_1 i_2 i_3$ and $J = j_1 j_2 j_3$ for any two codons ($I \neq J$), where i_k is the nucleotide at the kth position of I. The Malthusian fitness of codon I at location K of the gene is $f_{I,K}$, so the selection coefficient for a mutant that transforms I into J is $s_{IJ,K} = f_{J,K} - f_{I,K}$. We write $S_{IJ,K} = F_{J,K} - F_{I,K} = 2N(f_{J,K} - f_{I,K})$ for the scaled selection coefficient, where N is the effective chromosomal number and $F_{I,K}$ is the scaled fitness. The substitution rate from I to J ($I \neq J$) at the location is

$$q_{IJ,K} = \begin{cases} \mu_{IJ} \dfrac{S_{IJ,K}}{1 - e^{-S_{IJ,K}}} & \text{if } S_{IJ,K} \neq 0 \\ \mu_{IJ} & \text{else,} \end{cases} \tag{1}$$

where μ_{IJ} is the neutral mutation rate, and $S/(1 - e^{-s})$ is the relative fixation probability of a selected mutation compared with a neutral one (Kimura 1983, Equation 3.14). If the mutation is advantageous ($S_{IJ} > 0$), then $q_{IJ} > \mu_{IJ}$, and if the mutation is deleterious ($S_{IJ} < 0$), then $q_{IJ} < \mu_{IJ}$. Thus the effect of natural selection is to accelerate or reduce the rate of substitution compared to the neutral mutation rate. The $q_{IJ,K}$ form the off-diagonal elements of a 64×64 rate matrix ($\mathbf{Q}$) whose diagonal elements are $q_{II,K} = -\sum_{J \neq I} q_{IJ,K}$.

The selection coefficients ($S_{IJ,K}$) describe the effect of selection on the amino acid at a given location K, due to the protein's structure and function of the protein. This contrasts with the model of Yang and Nielsen (2008), where the nonsynonymous to synonymous substitution rate ratio, ω, is used to account for the effect of selection at the protein level. When modeling site-specific selection, the inclusion of ω is unnecessary.

The location-specific fitnesses ($F_{I,K}$) can be modeled at the amino acid or codon levels. We can write $F_{J,K} = F_J^{\text{co}} + F_{J,K}^{\text{aa}}$, where F_J^{co} is the fitness of J due to the effect of selection on codon bias (*e.g.*, Bulmer 1991) and $F_{J,K}^{\text{aa}}$ is the fitness of the particular amino acid at the location. In this study, we assume that the selective constraints are dominated by selection on the amino acid and ignore the effect of selection on codon bias. Under this assumption $F_{J,K} = F_{J,K}^{\text{aa}}$.

Mutation at the nucleotide level: Consider a cycle of DNA replication occurring in a tiny time interval τ. The probability of observing a particular nucleotide i mutating into j ($i \neq j$) during interval τ is $p_{ij}(\tau) \simeq g_{ij}\tau$, where $g_{ij} (\geq 0)$ is the rate of change $i \rightarrow j$ per time unit. The probability that i will remain unchanged is $p_{\text{ii}}(\tau) \simeq 1 + g_{\text{ii}}\tau$, where $g_{ii} = -\sum_i g_{ij}$. Note that we are modeling the mutation of DNA *before* natural selection takes place. The probability that a triplet I of nucleotides will change into triplet $J (I \neq J)$ is $p_{IJ}(\tau) = \prod_k p_{i_k j_k}(\tau) \simeq \mu_{IJ}\tau$. Because the time interval τ is very small, $p_{\text{ii}}(\tau) \approx 1$, so we can ignore these probabilities in the product term and then solve for the mutation rate μ_{IJ} to get

$$\mu_{IJ} \approx \frac{\prod_{k, i_k \neq j_k} p_{i_k j_k}(\tau)}{\tau} = \frac{\prod_{k, i_k \neq j_k} \tau g_{i_k j_k}}{\tau} = \tau^{n-1} \times \prod_{k, i_k \neq j_k} g_{i_k j_k}, \tag{2}$$

where n is the number of changing nucleotides.

The rate constants g_{ij} can be defined under any nucleotide substitution model (*e.g.*, Yang 1994). Here we use the HKY85 model (Hasegawa *et al.* 1985), where $g_{ij} = \nu\kappa\pi_j^*$ for transitions and $g_{ij} = \nu\pi_j^*$ for transversions, $\pi_j^* (\geq 0, \ \sum_j \pi_j^* = 1)$ is the equilibrium frequency of nucleotide j (achieved under no selection), κ is the transition–transversion rate parameter, and ν is a scaling constant. The mutation rate $I \rightarrow J$ is thus

$$\mu_{IJ} = \tau^{n-1} \nu^n \kappa^{n_t} \prod_{k, i_k \neq j_k} \pi_{j_k}^*, \tag{3}$$

where n_t is the number of nucleotide transitions necessary to go from I to J. For example, if codons I and J differ by a single

transversion, then $\mu_{IJ} = \nu\pi^*_{j_k}$, while if they differ by two transitions at positions k and l, then $\mu_{IJ} = \nu\tau\kappa^2\pi^*_{j_k}\pi^*_{j_l}$. We can now combine Equations 3 and 1 to get

$$q_{IJ,K} = \left(\tau^{n-1}\nu^n\kappa^{n_t}\prod_{k,i_k\neq j_k}\pi^*_{j_k}\right) \times \frac{S_{IJ,K}}{1-e^{-S_{IJ,K}}}. \quad (4)$$

Parameter τ controls the rate at which multiple simultaneous nucleotide substitutions are allowed to occur in $I \to J$. For example, if $\tau = 10^{-1}$, then triple substitutions occur at a rate in the order of 10^{-2} compared to single substitutions. If $\tau = 0$, simultaneous substitutions are not allowed and Equation 4 reduces to a sitewise version of Equation 2 in Yang and Nielsen (2008). This multiple-substitutions model contrasts with that of Halpern and Bruno (1998), which is based on the probability of observing a random mutation in a nucleotide sequence at equilibrium.

We scale the substitution rates on the basis of the expected number of *neutral* mutations per site (Halpern and Bruno 1998). When there is no selection acting on the sequence, the neutral substitution rate is simply $q^0_{IJ} = \mu_{IJ}(I \neq J)$, and the expected equilibrium frequency of J is $\pi^0_J = \pi^*_{j_1}\pi^*_{j_2}\pi^*_{j_3}$. We thus set $\nu = 1/\sum_{I=1}^{64}\sum_{J=1}^{64}\pi^0_I\mu_{IJ}(I \neq J)$ so that the expected number of neutral substitutions per codon location is one (*i.e.*, $-\sum_I \pi^0_I q^0_{II} = 1$).

Equation 4 describes a reversible process at the codon level. The proof of reversibility can be obtained by the same argument of Yang and Nielsen (2008) and it is not shown here. We note that the equilibrium frequency of codon J at location K is

$$\pi_{J,K} = \frac{\pi^*_{j_1}\pi^*_{j_2}\pi^*_{j_3}e^{F_{J,K}}}{z}, \quad (5)$$

where $z = \sum_{J=1}^{64}\pi^*_{j_1}\pi^*_{j_2}\pi^*_{j_3}e^{F_{J,K}}$.

Maximum-likelihood estimation: Equation 4 can be used to construct the transition probability matrix $\mathbf{P}_t = P_{IJ}(t) = e^{t\mathbf{Q}}$ that gives the probability of change $I \to J$ after time t. This matrix can then be used to calculate the likelihood of a sequence alignment under a fixed tree topology, using established procedures (Felsenstein 1981; Yang 2006). The value of the model parameters that maximize the log-likelihood (ℓ) can then be found by numerical optimization.

Estimation of branch lengths: Estimation of branch lengths by maximum likelihood is computationally expensive. We estimate individual branch lengths using faster codon-based methods (*e.g.*, Yang and Nielsen 2008), and the estimated tree with fixed topology is then used in the likelihood calculation of the model. During the calculation, the branch lengths are multiplied by a constant c and the value of this constant is chosen to maximize the likelihood. Therefore, the final tree has branch lengths as the expected number of *neutral* substitutions per site, that is, the number of substitutions that would have accumulated if the sequence was a pseudogene. We can convert the branch lengths to the expected number of substitutions per codon in the following manner: For location K, the expected substitution rate at equilibrium is $\lambda_K = \sum_I \pi_{I,K} q_{II,K}$. The average substitution rate for the whole sequence is $\bar{\lambda} = \sum_K\sum_I \pi_{I,K} q_{II,K}/L_c$. For a pseudogene $\bar{\lambda} = 1$, while a gene under purifying selection would have $\bar{\lambda}<1$. For a branch of length b neutral substitutions per site, $\bar{\lambda}\, b$ represents the usual substitutions per codon.

Adjustable parameters: One of the mutational bias parameters (π^*_j) is redundant as the nucleotide equilibrium frequencies obey the constraint $\sum_{j=1}^4 \pi^*_j = 1$. Similarly, only relative values of the fitness parameters ($F_{I,K}$) matter, so for each location, one of the fitness parameters can be set equal to zero. For a coding sequence with L_c codon locations, the model has six mutation parameters (τ, κ, c, and 3 π^*_j) and $(20 - 1)L_c$ values of $F_{I,K}$. Information from all codon locations is used by the likelihood method to estimate the value of the six mutation parameters. The variance of these parameters decreases with increased sequence length. The amino acid fitnesses are location specific and can be reliably estimated only for alignments of many sequences under reasonable levels of divergence (see below).

Unobserved amino acids: Only a few amino acid types are usually seen within a given alignment location. For a codon J coding for an unobserved amino acid, the maximum-likelihood estimate (MLE) of $F_{J,K}$ tends to $-\infty$. (Exceptions may exist when an unobserved amino acid may help facilitate substitutions between observed amino acids, such as pairs that cannot be connected by a single-base change.) Because $S_{IJ,K}/(1-e^{-S_{IJ,K}}) \to 0$ if $S_{IJ,K} \to -\infty$, then the corresponding columns of the rate matrix are zero. Rather than estimating $F_{J,K}$ for these unobserved amino acids, it is possible to fix these values to $-\infty$ and collapse the rate matrix accordingly. For example, for a location where only two amino acids encoded by two codons each are observed, the corresponding rate matrix would be of size 4×4 and only $2 - 1$ amino acid fitness parameters would be found by numerical optimization (Holder *et al.* 2008). This approximation greatly reduces computing time.

Distribution of selection coefficients: We calculate the distribution of selection coefficients among novel mutations and among substitutions. At equilibrium, the proportion of expected mutations with a given value of S among all mutants at all locations is

$$m^0(S) = \frac{\sum_K\sum_{I\neq J}\pi_{I,K}\mu_{IJ}\delta(S - S_{IJ,K})}{\sum_K\sum_{I\neq J}\pi_{I,K}\mu_{IJ}}, \quad (6)$$

where $\delta(S - S_{IJ,K}) = 1$ if $S - S_{IJ,K} = 0$ and $= 0$ otherwise. The proportion among substitutions is

$$m(S) = \frac{\sum_K\sum_{I\neq J}\pi_{I,K}q_{IJ,K}\delta(S - S_{IJ,K})}{\sum_K\sum_{I\neq J}\pi_{I,K}q_{IJ,K}}. \quad (7)$$

Note that the scaled Malthusian fitness, $F = 2Nf$, is related to the Darwinian fitness, w, by $w = e^{F/2N}$ (Crow and Kimura 1970, p. 8). For the wild type $w = 1$ and $F = 0$, and for a lethal mutant, $w = 0$ and $F = -\infty$; this means that the distribution of selection coefficients ranges from $-\infty$ to ∞. In experimental studies Darwinian fitnesses are normally used, and the distribution of fitness effects ranges from -1 to ∞. For $S \ll N$, selection coefficients and fitness effects are nearly identical.

Following Li (1978) we define an $I \rightarrow J$ mutation (or substitution) as deleterious if $S_{IJ,K} < -2$, as nearly neutral if $-2 < S_{IJ,K} < 2$, and as advantageous if $2 < S_{IJ,K}$. The proportions of the three types of mutations are p_-, p_0, and p_+, respectively. For example, the proportion of advantageous mutations among all substitutions is

$$p_+ = \int_2^\infty m(S)dS. \tag{8}$$

The uncertainty in the estimation of the distribution of S can be assessed by classical and parametric bootstrapping. In the classical bootstrap, we sample locations at random (with replacement) from the alignment and then we recalculate the distribution using Equations 6 and 7 to generate confidence intervals. In the parametric bootstrap, synthetic data are generated using the ML estimates from the real data set, and then all parameters are reestimated for the synthetic data using exactly the same procedure as for the real data (including estimation of the tree topology, branch lengths, global mutation parameters, and fitnesses). When the parametric model offers an adequate description of the real data, both the classical and the parametric bootstrap lead to similar results (Felsenstein 2003, Chap. 20).

Software implementation

The software implementation of the model is available from http://mathbio.nimr.mrc.ac.uk/. The program is able to utilize multicore and distributed architectures, making the estimation of global and site-specific parameters computationally tractable. Estimation of the parameters is done in three steps. As our program does not perform branch length estimation, we first optimize branch lengths under one of the codon substitution models available in other software. We use the FMutSel0 model in the program CODEML [PAML package (Yang 2007; Yang and Nielsen 2008)], using the branch-by-branch optimization option and empirical codon frequencies (method = 1 and estFreq = 0 in the CODEML control file). Second, we use our program to estimate the global parameters ($\boldsymbol{\pi}^*$, κ, τ, and c), using the approximate method described above, where the substitution rate matrix is collapsed by neglecting all unobserved residues. The MLEs of $\boldsymbol{\pi}^*$ and κ estimated by CODEML in the first step are used as starting values for the sitewise model in the second step. In the third step, the global parameters are fixed, but all fitnesses are reestimated, this time relaxing the assumption that $F_I = -\infty$ for unobserved amino acids. Likelihood calculation is thus performed using the full 64×64 substitution matrix. The fitness parameter of the most common amino acid at each location is fixed to $F_{I,L} = 0$, while the other fitness parameters are limited to $-20 < F < 20$. The $\mathbf{F}$ values are estimated three times for each site, once with all $F = 0$, once with F of unobserved residues set to -20, and once using random starting values (between -3 and 3). We used the MLE with the highest likelihood.

Results and Discussion

Statistical properties of the model

Consistency and normality of fitness estimators: In the general case, the likelihood function L, the probability of observing data $\mathbf{x} = (x_i)$ given parameters $\theta = (\theta_i)$, is given by the joint density $L(\mathbf{x} \mid \theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$. When the data (x_i) are independent (and under other regularity conditions), the maximum-likelihood estimator of θ, $\hat{\theta}$, is shown to be consistent and asymptotically normally distributed (*e.g.*, Stuart *et al.* 1999, Chap. 18). When the data are not independent, consistency and asymptotic normality may not be guaranteed. Estimation of the fitnesses for a particular location K, when the global parameters (τ, κ, c, and $\boldsymbol{\pi}^*$) are known, proceeds by maximizing a joint-likelihood function $L(\mathbf{x} \mid \theta) = f(x_1, x_2, \ldots, x_n \mid \theta)$, where x_i represents the observed codon in species i for the given location. These data are not independent (they are correlated according to the underlying tree structure) and the asymptotic properties of the fitness estimators are unclear.

We can employ Monte Carlo simulations to investigate the asymptotic properties of fitness estimators when the number of species sampled is increased. We follow two simple simulation strategies. For the fixed-height tree (FHT) approach, we started with a rooted 64-taxa symmetric tree with branch lengths {7.5, 3.75, 1.875, 0.9375, 0.46875, 0.46875} moving from the root of the tree to the leaves, for a tree height of 15 and total branch length of 105. The next tree in the series with 128 taxa is constructed by inserting a bifurcating node at the midpoint of the terminal branches, resulting in branch lengths of {7.5, 3.75, 1.875, 0.9375, 0.46875, 0.234375, 0.234375} and the tree height unchanged, but the total branch length is increased to 120. The 256-, 512-, and 1024-taxa trees are constructed using the same procedure. For the variable-height tree (VHT) approach, we start with a rooted 64-taxa symmetric tree where all branch lengths are equal to 0.25, for a tree height of 1.5 and a total branch length of 31.5. The 128-taxa tree is constructed by replacing each leaf with a bifurcating node with branch lengths of 0.25 leading to two new leaves. This results in an increase in the tree height to 1.75 and an increase in the total branch length to 63.5. This procedure is repeated to yield 256-, 512-, and 1024-taxa trees. We consider a location with two possible amino acids with equilibrium frequencies $\{\pi_1, \pi_2\} = \{0.015, 0.985\}$, $\{0.333, 0.667\}$, and $\{0.5, 0.5\}$; the frequencies of all other amino

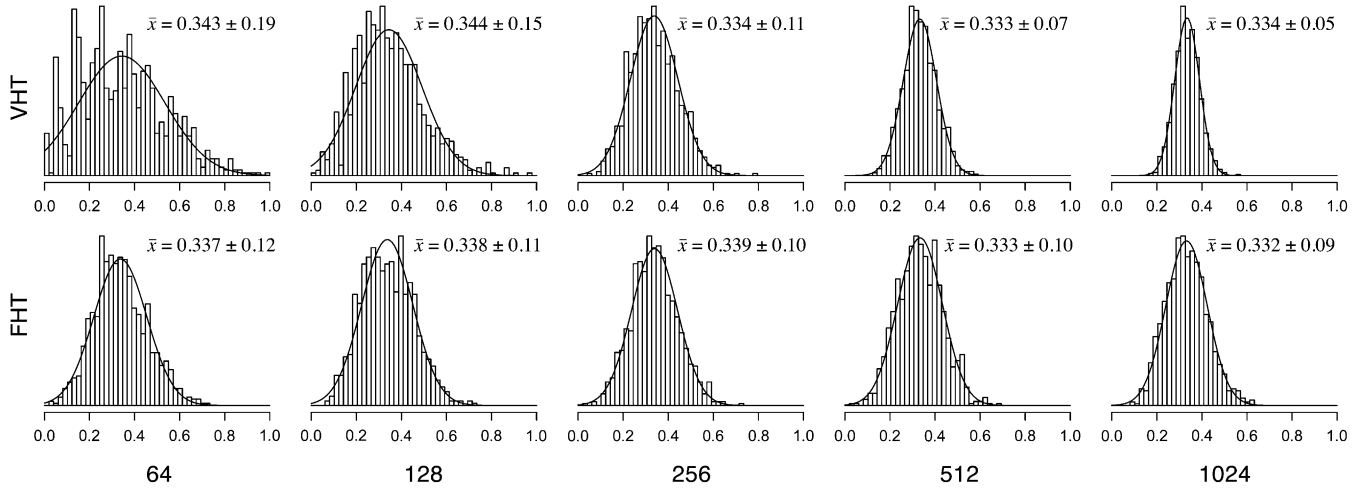

Figure 1 Consistency and normality of fitness estimators. The histogram is the estimated sampling distribution of $\hat{\pi}$ from 1000 simulations. The solid line represents the best-fitting normal distribution to the simulated data.

acids are set to zero. The global parameters are set to $\kappa = 2$, $\boldsymbol{\pi}^* = (0.25)$, and $\tau = 0$. For each setup 1000 sites are simulated, and π_1 is then estimated by ML with global parameters fixed to their true values.

Figure 1 shows the results of the simulations for $\pi_1 = 0.333$ for both tree strategies and for 64–1024 taxa. In both cases, as the number of taxa is increased, the standard error of $\hat{\pi}_1$ decreases, and the sampling distribution increasingly resembles a normal distribution. The standard error decreases much faster for the VHT than for the FHT strategy. The VHT strategy resembles the case of a biologist who samples additional, more divergent outgroup taxa that root more deeply in the tree. The FHT strategy resembles the case of a biologist who samples additional, similar species from the same genera, thus adding a modest amount of extra information. Using $\pi_1 = 0.015$ and $\pi_1 = 0.5$ and increasing the number of amino acids observed at the location (four, seven, or eight) yield the same trends (not shown). We note that under the invariance principle of ML, $\hat{\pi} = f(\hat{F})$ (Equation 5); therefore, estimating $\hat{\pi}$ or $\hat{F}$ leads to the same inference.

Distribution of selection coefficients for simulated data: As seen above, a large number of species of reasonable divergence are necessary to estimate the equilibrium frequencies (and fitnesses) for the codons within each location in a protein. A more important question is whether the distribution of selection coefficients can be estimated adequately for moderate data sets. We tested the robustness of estimates of p_-, p_0, and p_+ by generating synthetic data sets that explore the breadth of the high-dimensional parameter space of the model. Specifically, we studied how the estimates were affected by different distributions of site-specific fitnesses, varying the number of taxa and varying mutation rates.

To ensure that the generated data sets were reasonably realistic, the set of observed residues at each site was determined by randomly choosing a location from a mitochondrial genome alignment (described below). One thousand sites were sampled and, for each site, those residues not observed in the sampled location had their fitnesses fixed to $-\infty$. We then sampled the site-specific fitness for each residue from an underlying distribution. These are the "known" fitnesses. To explore the effect of different distributions of site-specific fitnesses (F), we considered three different distributions: (i) a gamma distribution with $\alpha = 2$ and $\beta = 1$, (ii) a normal distribution with $\mu = 0$ and $\sigma = 2$, and (iii) a normal distribution with $\mu = 0$ and $\sigma = 5$. Each distribution of F leads to a distinct distribution of S. Using these fitness values, we synthesized three data sets on the variable-height tree with 256 taxa.

To investigate the effect of varying sample sizes, we created data sets with 64, 128, and 192 taxa by sampling from the 256 sequences generated under the normal distribution ($\sigma = 5$) in the previous step. For each sample, the tree topology and branch lengths were estimated. We also simulated data with the same fitnesses on a 4096-taxa tree to examine the benefit of having many more taxa.

To test the effect of increased or reduced mutation rate, two data sets were synthesized using the original fitnesses drawn from the normal ($\sigma = 5$) distribution. One set was generated with twice the mutation rate of the original 256-taxa tree, while the other had half the mutation rate of the original tree.

The site-specific fitnesses for each of the nine generated sets of sequences were reestimated by ML using our model, fixing the global parameters to their true values. As each synthesized data set was created with known global parameters [$\kappa = 2$, $\boldsymbol{\pi}^* = (0.25)$, and $\tau = 0$] and site-specific fitnesses, the true proportion of deleterious, neutral, and advantageous mutations is also known. Table 1 shows the proportions p_-, p_0, and p_+ of mutations calculated using the known fitnesses and compares them to the proportions obtained by estimating the fitnesses by ML. We found that in all cases the proportions of different types of mutations

can be readily estimated, as well as the general shape of the distribution of S (not shown).

These tests demonstrated the difficulties of estimating the fitnesses for very deleterious mutations. For example, an amino acid with fitness $F = -10$ at a location has an equilibrium frequency of $\pi = 4.5 \times 10^{-5}$ (see Equation 5). That is, we would expect to sample sequences from ~22,000 species to see this amino acid once at the location. Therefore, it is not possible to distinguish between $F = -20$ and $F = -10$, and we report the distribution of S from -10 to 10. However, our tests showed that with more taxa and more evolutionary time, we can recover more closely the shape of the curve for very deleterious mutations.

Analysis of real data

We use two real data sets to estimate the distribution of fitness effects. The first data set is an alignment of the 12 protein genes on the heavy strand of the mitochondrial genome of 244 placental mammal species. The alignment is constructed with PRANK (Loytynoja and Goldman 2008) and edited manually to removed small gappy regions at the end tails of some of the mitochondrial protein genes. The alignment is 3598 codons long. The tree topology is estimated by ML with RAxML, using the GTR+Γ model (Yang and Kumar 1996; Stamatakis *et al.* 2005).

The second data set is an alignment of the PB2 gene of 401 influenza viruses isolated from 80 human and 321 avian hosts. The alignment is 759 codons long. The data set and the tree topology are described in Tamuri *et al.* (2009). The human viruses are monophyletic and they are thought to be the product of a shift from an avian (the natural reservoir) to a mammalian host sometime around 1882–1913, before the 1918 influenza pandemic (dos Reis *et al.* 2009). The PB2 gene codes for a subunit of the virus polymerase complex. The polymerase genes seem to be involved in host adaptation, and there is evidence of several amino acid substitutions after the host shift (Taubenberger *et al.* 2005; dos Reis *et al.* 2011). Tamuri *et al.* (2009) identified 25 locations in PB2 where amino acid equilibrium frequencies are different between the viruses of the two hosts. To accommodate this observation, we first perform estimation of the fitnesses and global parameters for all residues in the protein. In a second step, a nonhomogeneous model that assumes different fitnesses for avian and human viruses is fitted to the 25 adaptive locations. For example, consider a location L that is one of the 25 adaptive locations. The substitution rate between codons I and J along the branches linking the viruses found in the human host (H) is given by

$$q_{IJ,L}^{(\mathrm{H})} = \begin{cases} \mu_{IJ} \dfrac{S_{IJ,L}^{(\mathrm{H})}}{1 - e^{-S_{IJ,L}^{(\mathrm{H})}}} & \text{for } S_{IJ,L}^{(\mathrm{H})} \neq 0 \\ \mu_{IJ} & \text{else,} \end{cases} \tag{9}$$

where $S_{IJ,L}^{(\mathrm{H})} = F_{J,L}^{(\mathrm{H})} - F_{I,L}^{(\mathrm{H})}$ are the location- and host-specific selection coefficients. Similarly, the substitution rate at the adaptive locations and along the branches linking the avian viruses is $q_{IJ,L}^{(\mathrm{A})}$ and the avian-specific fitnesses are $F_{J,L}^{(\mathrm{A})}$. Therefore, for each adaptive location, $2 \times 19 = 38$ fitnesses parameters are estimated, 19 for each host. For a nonadaptive location, $q_{IJ,K}^{(\mathrm{H})} = q_{IJ,K}^{(\mathrm{A})}$ and $F_{IJ,K}^{(\mathrm{H})} = F_{IJ,K}^{(\mathrm{A})}$. The distribution of selection coefficients during evolution in the avian host is calculated using Equations 6 and 7, with $\pi_{I,K}^{(\mathrm{A})}$, $q_{IJ,K}^{(\mathrm{A})}$, and $S_{IJ,K}^{(\mathrm{A})}$. For the distribution of selection coefficients following the host shift, we consider that, immediately after the host shift, the equilibrium frequencies ($\pi_{I,K}^{(\mathrm{A})}$) will reflect the frequencies characteristic of avian viruses. At this point, however, the substitution rates ($q_{IJ,K}^{(\mathrm{H})}$) and the resulting fitnesses ($F_{IJ,K}^{(\mathrm{H})}$) will reflect the situation in the human host. Therefore, at this host shift instant we have

$$m_{\mathrm{HS}}^{0}(S) = \frac{\sum_K \sum_{I \neq J} \pi_{I,K}^{(\mathrm{A})} \mu_{IJ} \delta\left(S - S_{IJ,K}^{(\mathrm{H})}\right)}{\sum_K \sum_{I \neq J} \pi_{I,K}^{\mathrm{A}} \mu_{IJ}} \tag{10}$$

and

$$m_{\mathrm{HS}}(S) = \frac{\sum_K \sum_{I \neq J} \pi_{I,K}^{(\mathrm{A})} q_{IJ,K}^{(\mathrm{H})} \delta\left(S - S_{IJ,K}^{\mathrm{H}}\right)}{\sum_K \sum_{I \neq J} \pi_{I,K}^{\mathrm{A}} q_{IJ,K}^{\mathrm{H}}}. \tag{11}$$

Mammalian mitochondrial data: Figure 2 shows the distribution of fitness effects for novel mutations and substitutions for the mammalian mitochondria data set. The distribution of S among novel mutations clearly shows a multimodal distribution with one large peak around nearly neutral mutations ($-2 < S < -2$) and with another peak corresponding to highly deleterious mutations ($S < -10$). This second peak includes all mutations to amino acids that have not been observed at a given position and that therefore

Table 1 Monte Carlo simulation of the distribution of selection coefficients

	p_-	p_0	p_+
Gamma distribution ($\alpha = 2$, $\beta = 1$)			
Known	0.864	0.133	0.0034
Estimated	0.882	0.115	0.0029
Normal distribution ($\mu = 0$, $\sigma = 2$)			
Known	0.897	0.099	0.0040
Estimated	0.908	0.089	0.0030
Normal distribution ($\mu = 0$, $\sigma = 5$)			
Known	0.964	0.034	0.0019
Estimated	0.969	0.030	0.0015
64 taxa	0.968	0.032	0.0014
128 taxa	0.967	0.031	0.0013
192 taxa	0.966	0.034	0.0009
4096 taxa	0.966	0.033	0.0019
Half mutation rate	0.968	0.030	0.0014
Doubled mutation rate	0.965	0.033	0.0016

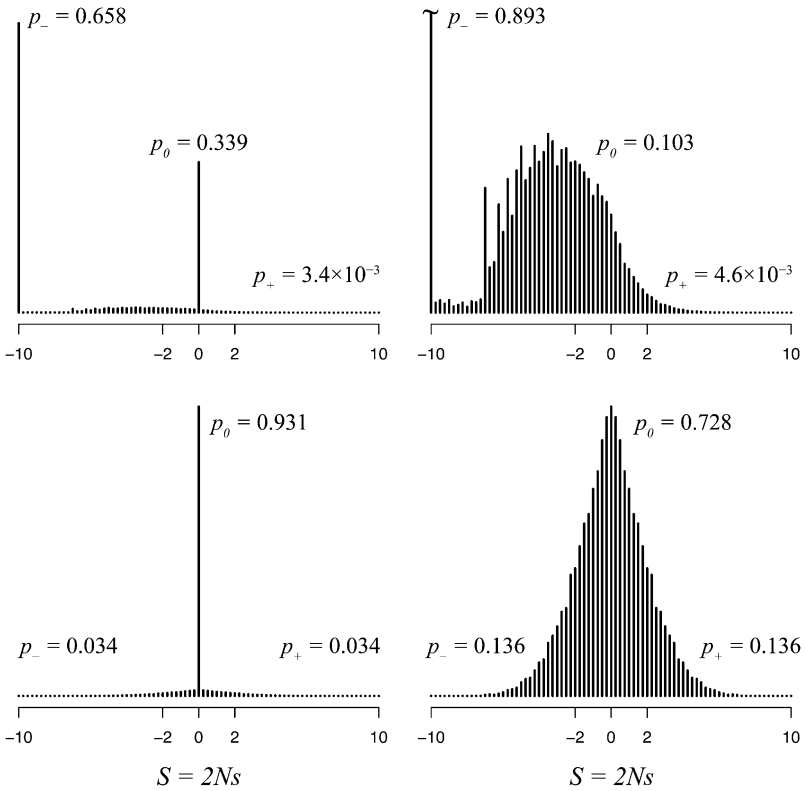

Figure 2 Distribution of selection coefficients in mammalian mitochondrial proteins estimated by ML. The heights of the histogram bars are calculated according to Equations 6 and 7. Distributions are shown for all mutations (top left), nonsynonymous mutations (top right), all substitutions (bottom left), and nonsynonymous substitutions (bottom right).

have the minimum allowed value of $F_{IJ,K} = -20$. Among substitutions, a main peak centered at neutral mutations dominates, and no substantial fraction of highly deleterious or highly advantageous ($10 < S$) substitutions is observed.

We observe that ~66% of mutations are deleterious ($S < -2$), similar to the fraction of deleterious mutations estimated in humans (Fay *et al.* 2001; Eyre-Walker *et al.* 2002). Approximately 52% of the mutations are strongly deleterious ($S < -10$), comparable with that estimated for humans (Fay *et al.* 2001) as well as with the fraction of mutations observed to be lethal in experimental studies of vesicular stomatitis virus (Sanjuan *et al.* 2004) and yeast (Wloch *et al.* 2001). We observe ~34% of mutations to be nearly neutral ($-2 < S < 2$), again similar to the fraction estimated by population-based methods in other data sets (*e.g.*, Eyre-Walker *et al.* 2002; Subramanian and Kumar 2006). Our estimates of the number of advantageous changes are modest, representing 0.5% of the nonsynonymous mutations and 14% of the nonsynonymous substitutions. This is in rough agreement with a number of population-based studies of human evolution (*e.g.*, Chimpanzee Sequencing and Analysis Consortium 2005), although some studies have estimated much larger fractions for humans (Fay *et al.* 2001) and *Drosophila* (Sawyer *et al.* 2003, 2007). In general, our numbers correspond to what would be expected in a nearly neutral evolutionary model (Akashi 1999).

The estimated values for the global mutation parameters for the sitewise mutation selection model (swMutSel0) fitted to the mammalian data are listed in Table 2. The equilibrium base frequencies ($\boldsymbol{\pi}^*$) are similar but not identical to those estimated with the FMutSel0 model by PAML, which neglects changes in base composition resulting from the selective constraints acting at the amino acid level. The value of τ, representing the tendency for simultaneous multiple-base substitutions, indicates that the proportions of single, double, and triple changes are 99.4%, 0.58%, and 0.002% respectively. The optimization procedure is likely to result in an overestimation of the frequency of multiple mutations. Mutations between two amino acids that are not convertible by a single-base change (*e.g.*, phenylalanine {TTT, TTC} to asparagine {AAT, AAC}) can result either through multiple-base changes or through a transient intermediate amino acid (such as tyrosine {TAT, TAC}). Our procedure, as described above, estimates τ while making the assumption that unobserved amino acids at any location, including possible intermediates, are incompatible with the selection constraints. This increases the requirement for multiple-base changes, increasing our estimate of τ. Even with this bias, our estimation of the multiple-substitution rate is more modest than proportions derived from simpler codon models applied to a more comprehensive protein data set (Kosiol *et al.* 2007) and may indicate either differences in the evolutionary process for mitochondrial DNA or biases that result when site-specific selective constraints are inadequately modeled.

Influenza PB2 data: Figure 3 shows the distribution of fitness effects for the influenza PB2 gene evolving in the avian host, and Figure 4 shows the distribution following

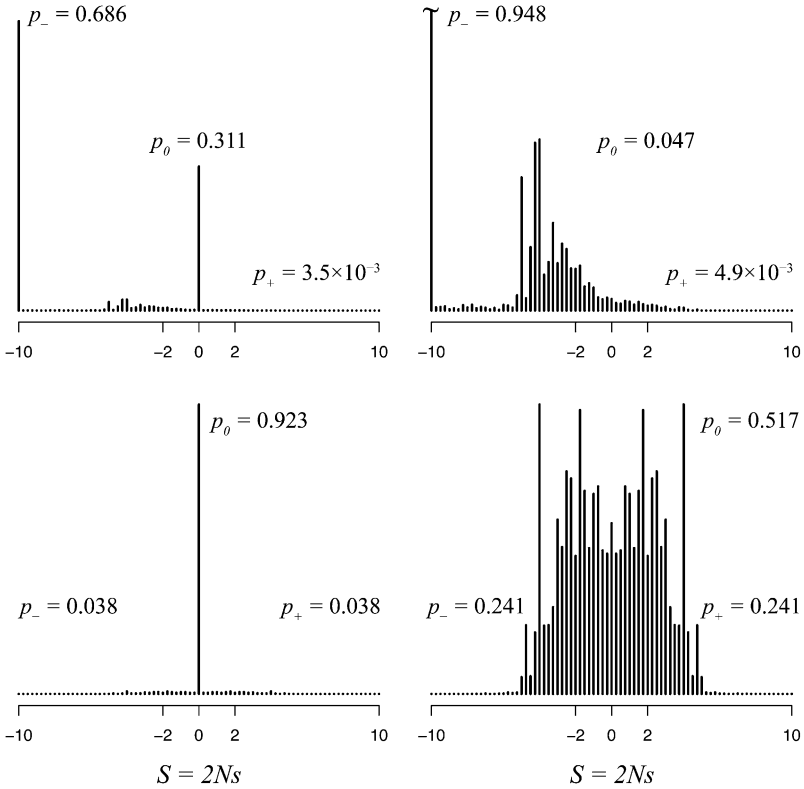

Figure 3 Distribution of selection coefficients in the PB2 gene of influenza for avian viruses at adaptive equilibrium. Distributions are shown for all mutations (top left), nonsynonymous mutations (top right), all substitutions (bottom left), and nonsynonymous substitutions (bottom right).

a well-defined adaptive event: the host shift to humans. As in the mitochondrial case, the distribution of S among mutations at adaptive equilibrium shows a multimodal distribution, with two main modes centered around nearly neutral ($-2 < S < 2$) and highly deleterious ($S < -10$) mutations. Among substitutions, the distribution is dominated by a main peak centered on neutral mutations. Interestingly, at the host shift event, we find two well-defined peaks among substitutions,

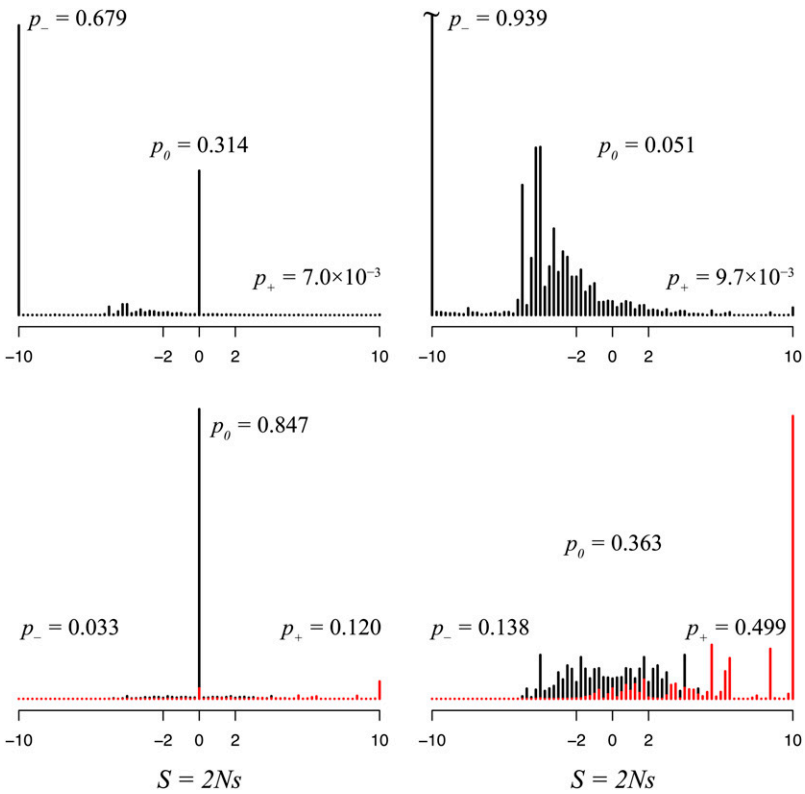

Figure 4 Distribution of selection coefficients in the PB2 gene of influenza for human viruses immediately after a host shift from bird. Distributions are shown for all mutations (top left), nonsynonymous mutations (top right), all substitutions (bottom left), and nonsynonymous substitutions (bottom right). The contributions from the 25 sites under different selective constraints in the two hosts are shown in red; the contributions from other sites are shown in black.

Table 2 Parameters in swMutSel0 and FMutSel0

		swMutSel0	FMutSel0
Mammal data	$\hat{\pi}^*$(T, C, A, G)	0.19, 0.27, 0.48, 0.06	0.17, 0.25, 0.52, 0.06
	$\hat{\kappa}$	7.06	6.97
	ω	—	0.05991
	$\hat{\tau}$	0.06257	—
Influenza data	$\hat{\pi}^*$(T, C, A, G)	0.24, 0.20, 0.37, 0.20	0.23, 0.19, 0.37, 0.21
	$\hat{\kappa}$	7.86	7.77
	ω	—	0.062
	$\hat{\tau}$	0.09199	—

one peak centered around neutral substitutions and another peak of highly advantageous substitutions ($10 < S$). We estimate that 12% of all substitutions and 50% of all nonsynonymous substitutions are advantageous at the host shift event. These results are in agreement with an adaptive model as pointed out by Akashi (1999).

There has been much discussion in the literature about the relative contributions of nearly neutral and advantageous substitutions to the evolutionary process (*i.e.*, Kimura 1983 *vs.* Gillespie 1994). We suggest that the distribution of S is not constant in time but changes as organisms undergo adaptation through novel environments, with the relative contributions of nearly neutral and advantageous mutations dependent on the particular evolutionary scenario. It seems sensible to think that organisms go through phases of mostly neutral and mostly adaptive episodes.

Estimates for the influenza global mutation parameters are listed in Table 2. As for the mitochondrial data, the equilibrium base frequencies ($\boldsymbol{\pi}^*$) are similar but not identical to those estimated with the FMulSel0 model. The value of τ is of the same order of magnitude as that in the mitochondrial case, indicating nearly the same proportions of single, double, and triple substitutions.

The parameteric form of the distribution of S: Extreme value theory has been used to show that, under a wide range of conditions, the distribution of selection coefficients for advantageous nonsynonymous mutations should be exponential (Gillespie 1984; Orr 2003). This prediction has been questioned on the basis of simulations of the evolution of RNA (Cowperthwaite *et al.* 2005), which yielded a distribution with an overabundance of slightly adaptive mutations. As shown in Figure 5, we observe that the distribution of S for advantageous mutations ($S > 0$) matches an exponential distribution for both the mammalian and the influenza data; a fit of the data between $0 < S < 5$ to $m_0(S) \sim \exp(-\beta S)$ yields an exponent of $\beta = 0.924$ (95% C.I.: 0.904–0.941) for mammals and $\beta = 0.688$ (95% C.I.: 0.630–0.733) for influenza, both in agreement with the results of extreme value theory.

Previous work analyzing intraspecies variation has suggested that the distribution of nonsynonymous deleterious mutations is leptokurtic, that is, having a faster initial fall-off followed by a longer tail, such as a gamma distribution with shape parameter $\alpha < 1$. For example, Eyre-Walker *et al.* (2006) analyzed human SNPs and fit the resulting nonsynonymous deleterious mutations to a gamma distribution with $\alpha = 0.23$. In contrast, Nielsen and Yang (2003) carried out an interspecies study of primate mitochondrial proteins, fitting a reflected gamma to the distribution of S. The reflected gamma distribution around zero is simply $\Gamma_R(S \mid \alpha, \beta) = \Gamma(-S \mid \alpha, \beta)$ for $S < 0$. They estimated $\alpha = 3.22$, far from leptokurtic. Their model does not seem biologically realistic, as it suggests that different selective constraints at different locations in the protein act to reduce the overall substitution rate without affecting the resulting equilibrium distribution of amino acids at that location. Our distribution of selective coefficients with $S < 0$ clearly does not fit a reflected gamma distribution. We can, however, fit a reflected gamma distribution to the more limited range of moderately deleterious mutations ($-7 < S < -2$) as shown in Figure 5. Over this range, our results more closely resemble the distribution obtained by Nielsen and Yang, with $\alpha = 3.601$ (95% C.I.: 2.921–4.298) and $\beta = 0.817$ (95% C.I.: 0.643–0.987). The distribution of S for the influenza data is highly multimodal between $-7 < S < -2$, so we do not attempt to fit these data to a reflected gamma as in the mammalian case.

Although our results on nearly neutral and advantageous mutations and substitutions roughly correspond to previous results obtained with evolution-based methods, we observe a large fraction of highly deleterious mutations ($S < -10$), better matching the number of experimentally observed lethal mutations. It is not surprising that previous analyses have had trouble estimating these highly deleterious mutations. Nielsen and Yang (2003) explicitly did not allow residues to be less or more favored at different locations, allowing changes only in the overall substitution rates; all substitutions are allowed at all but perfectly conserved locations. Rodrigue *et al.* (2010) consider models of selection at each location that are mixtures of various components; this averaging effect reduces the ability to identify highly unfavorable amino acids at specific positions.

Uncertainties in the estimation of the distribution of S: We estimate the uncertainties and biases in our approach by using the classical and parametric bootstrap approaches. The classical bootstrap is used to generate error bars for the MLEs obtained from the real data, and the parametric bootstrap is used to generate simulated replicates of the

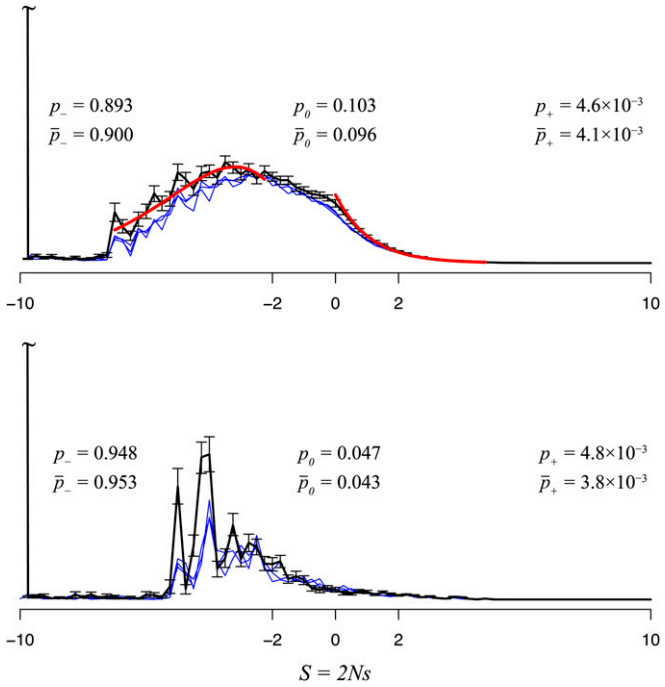

Figure 5 The parametric form of S and bootstrap analysis of the data. The mammalian data are shown at the top and the influenza data at the bottom. The black curves display the distribution of selection coefficients for nonsynonymous mutations including error bars obtained by classic bootstrapping. Red curves show best exponential fit for advantageous mutations ($0 < S < 5$) and best gamma distribution fit for moderately deleterious mutations ($-7 < S < -2$). Blue curves represent the distributions obtained from the parametric bootstrap analysis from three synthetic data sets. p and $\bar{p}$ are the proportions for the real data and the average for the parametric bootstraps, respectively.

distribution of S. The distributions of selective coefficients for nonsynonymous mutations for three parametric bootstrap data sets are compared with the results for the real data in Figure 5. As would be expected, the distributions are extremely similar for the neutral and advantageous substitutions. The general trends for the deleterious mutations are similar, although it appears that the calculations have a tendency to overestimate the magnitude of S for the deleterious mutations. This is not overly surprising, as this would result if the fitness of the extremely infrequent amino acids were underestimated and where their omission from the observed data reflects lack of evolutionary time rather than biological impossibility. This discrepancy may also be caused by our optimizing the tree branch lengths under the site-invariant FMutSel0 model rather than our sitewise model (Halpern and Bruno 1998). These differences have minimal effect on the fraction of mutations and substitutions that are deleterious, neutral, and advantageous.

As might be expected given the close correspondence of the distributions for advantageous nonsynonymous mutations, the fit of the distribution of positive ($0 < S < 5$) selective coefficients for the three bootstrap data sets to an exponential yields values ($\bar{\beta} = 0.953$) similar to that obtained with the real mitochondrial data ($\beta = 0.924$). Although there are differences in the reflected gamma distribution fit for deleterious mutations for the mitochondrial data ($\bar{\alpha} = 5.611$ and $\bar{\beta} = 1.551$ in contrast to $\alpha = 3.601$ and $\beta = 0.817$), the results are still far from leptokurtic.

Comparison of the derived distribution of S with the results of parametric bootstrap simulations indicates that our phylogenetic analysis is able to successfully characterize the distribution of positive and near-neutral changes, although it overestimates the effect of deleterious mutations. The latter limitation is not unexpected: If an amino acid is rare or not observed at all at a given location, it is difficult to estimate how frequently it would be found at equilibrium. This is not an issue for representing substitutions, as these mutations would be extremely unlikely to occur.

Validations, assumptions, and limitations of the model

The model presumes that we know the true alignment for the selected data sets. Both the mammalian mitochondrial genes, which are well conserved, and influenza PB2 data sets produce good quality alignments. We also assume that the true phylogenetic tree is known. It has been shown, however, that small variations in tree topologies have minor impact on the parameters of phylogenetic models (Yang *et al.* 1994), and we would not expect it to significantly affect our calculations of selection coefficients. Additionally, we tested the effect of tree topology uncertainty during our parametric bootstrap analysis by restimating the tree topology for each replicate. Although the trees estimated for the bootstrapped data sets were different, but similar, to that of the real data sets, they did not have a major impact on our estimated distribution of S.

The analysis assumes that the various global and location-specific parameters are constants throughout the evolutionary process, with the exception of the host shift event explicitly included in the model for influenza PB2. The assumption that $F_{IJ,K} = 2Nf_{IJ,K}$ is a constant is based on assumptions regarding both the population size N and the fitness parameters $f_{IJ,K}$. Our analysis of the mitochondrial data set and PB2 evolving in an avian host assumes that the amino acid distribution is at equilibrium with respect to fixed selective constraints, resulting in a distribution of selective effects for accepted mutations symmetric around zero. This assumption explicitly eliminates the role of changes in selection and population size in adaptive evolution. The fitness parameters could change because of a number of effects. First, the structure, function, or physiological context of the protein could change. We have restricted these effects by considering mitochondrial proteins and PB2 from influenza. In neither case is there gene duplication that could lead to neofunctionalization that might result in changes in function or physiological context. We assume that the gene sequences are related by a single tree and recombination is absent, which is the consensus for both mammalian mitochondria (Lynch 2007) and influenza genes (Boni *et al.* 2008). It is well recorded that structural change is extremely slow relative to sequence change

(Aronson and Royer 1994). This does not mean that local structures might not change; these changes are more likely to occur in the exposed loop sections of the proteins, where there is reduced selective constraint. These would, therefore, likely result in small shifts in the neutral and near-neutral parts of the distribution of selection coefficients. Our analysis of the host shift effects on influenza PB2 demonstrates that changes in selective constraints can be explicitly included in the modeling, especially if information about the shift can be obtained independently (as, for instance, when there is a change in the host of a pathogen at a specific branch of the phylogenetic tree). Models of selection that include changes in selective constraints in a more general manner (such as covarion models, see Galtier 2001; Penny *et al.* 2001) could be used, but would result in even greater computational complexity.

Second, the selection constraints at a location might change due to substitutions that occur in other regions of the protein, that is, through the invalidation of our assumption that different locations in the protein evolve independently. Significant effort has been made in looking for such correlations between the substitution processes at interacting sites. The difficulty of this problem, the rather few examples where such effects have been substantiated, and the overall success of the independent sites assumption compared with models where it is relaxed (Rodrigue *et al.* 2006; Kleinman *et al.* 2010; Lakner *et al.* 2011) suggest that this effect is not likely to be large. This effect is even less likely to occur in the population-based models, as the timescales relevant to these studies are too short for many substitutions at other sites to occur. These population-based studies, however, are complicated by the interaction between the population dynamics that occur at different locations, such as interference between fixations of different mutations (Hill and Robertson 1966; Stephan 1995; Kirby and Stephan 1996) and genetic hitchhiking (Maynard-Smith and Haigh 1974; Barton 2000). Simulations of these phenomena in computational models might allow further reconciliation of the results of these types of studies.

It must also be pointed out that codon locations in a protein are tightly linked, and this can have a sizeable effect on the estimation of selection coefficients (Bustamante 2005). Our condition of independence implicitly assumes free recombination among locations. This is certainly not true for either the influenza or the mammal data sets analyzed. In particular, selection coefficients involving highly advantageous mutations are expected to be underestimated (Bustamante 2005). It is not clear at present how phylogenetic models could incorporate the assumption of linkage, and most works that have attempted to estimate the distribution of S from phylogenetic data have worked with the assumption of independence (*e.g.*, Nielsen and Yang 2003; Yang and Nielsen 2008; Rodrigue *et al.* 2010). Cartwright *et al.* (2011) studied the problem through simulation, but using a much simpler substitution model.

The assumption that $F_{IJ,K}$ is a constant also assumes that the effective population number has remained constant across lineages in the influenza and mammalian phylogenies. Both humans and *Drosophila* have undergone recent increases in population and expansion into new evolutionary niches (Merriwether *et al.* 1991; Glinka *et al.* 2003), possibly explaining why some (*e.g.*, Fay *et al.* 2001; Sawyer *et al.* 2003, 2007) but not all (Chimpanzee Sequencing and Analysis Consortium 2005) population-based studies of these groups yield a higher degree of adaptive evolution than observed here. This assumption could be relieved at the expense of additional parameters in the model as suggested by Nielsen and Yang (2003). Influenza viruses evolving in humans present oscillating population numbers, with population bottlenecks of low genetic diversity at the beginning and end of epidemic seasons (Rambaut *et al.* 2008). Because our model currently does not incorporate these variations in effective population number in mammals and influenza, our estimated fitnesses should be interpreted as averages over evolutionary timescales. We are currently exploring ways to incorporate variations in the effective population number in our model, but this is expected to be computationally challenging.

We also assume that the global mutation parameters (τ, κ, and the $\boldsymbol{\pi}^*$) do not vary across locations and across the tree. This assumption is unlikely to be strictly true; observed base compositions are known to be significantly different in different lineages. Differences in the equilibrium base composition in influenza, in particular, have been documented by the authors (dos Reis *et al.* 2009). Changes in the equilibrium base frequencies of, for instance, 10%, could result in similar changes in the estimate of $q_{IJ,K}$. However, $q_{IJ,K}$ is a steeply varying function of $F_{IJ,K}$, meaning that the changes expected in the latter quantity would be small.

Although a likelihood-ratio test for the effect of selection on codon bias is significant in both data sets ($P \ll 0.01$) (for details of the test see Yang and Nielsen 2008), we estimate fitnesses only at the amino acid level and explicitly ignore selection at the synonymous codon level, as estimation of the 60 global codon-level fitnesses would be a computationally onerous task. In mammals, selection on codon usage is very weak (Yang and Nielsen 2008; dos Reis and Wernisch 2009); similarly, selection on codon bias is negligible in influenza viruses (Shackelton *et al.* 2006). For this reason, in these data sets, selection coefficients for codon bias are expected to be small and within the nearly neutral interval, with a negligible effect on the shape of the distribution of selection coefficients among novel mutations and substitutions.

Conclusion

The dominant method of generating distributions of fitness effects has relied on a combination of intra- and interspecies variation. More recently, these population-based approaches have been joined by phylogenetic analyses that attempt to make a connection between the evolutionary process and population dynamics. These latter analyses offer a few

specific advantages. Perhaps the biggest advantage is an ability to look at a different timescale, allowing us to explore the relationship between population variation and evolutionary change. Second, the range of different organisms that can be studied is greatly increased, compared with the relatively few species (*e.g.*, humans, *Drosophila*, and yeast) where sufficient data exist to model population variation. Third, although both approaches involve making particular assumptions, the assumptions are different. Comparisons between results obtained with the different methods can provide insight into the nature and validity of these assumptions. Fourth, the substitution model can be elaborated to include additional effects, such as changes in selective constraints, population size, mutation rates at different points in evolution, or a relaxation of certain assumptions such as independence of sites. These extensions will become increasingly feasible as our sequence data, computational resource, and biological understanding continue to increase. Fifth, it is not necessary to prespecify a functional form for the distribution of S. This means that it is possible to decompose the evolutionary process and ask specific questions, such as the distribution of fitness effects involving changes to proline in helical regions of the protein.

This approach can be applied to any set of proteins with a sufficiently large and diverse set of homologs. The resulting distribution of fitness effects constitutes a signature of the selective constraints and could provide interesting perspectives on individual proteins and their physiological context. The relative proportion of deleterious, neutral, and advantageous mutations could depend on the protein structure and function, reflecting such distinctions such as whether the protein is globular, membrane, or unstructured; cytosolic or excreted; signaling, enzymatic, or immunological; or solitary or a member of a larger gene family.

Finally, as has been pointed out by Thorne *et al.* (2007), the connection with population dynamics has the potential to reform our modeling of sequence evolution. Substitution models have predominantly been phenomenological, representing the results of the evolutionary process (an accepted substitution) rather than the mechanics of how those results occurred. The opportunity to provide a firmer basis for these models by connecting it to population processes can result not only in better models, but also in ones that can be used to understand biological systems, populations, and evolutionary processes. The model presented here bridges the gap between population genetics and substitution models of sequence evolution. Since it was originally introduced by Halpern and Bruno (1998), this site-specific codon-based evolutionary model has seen limited use; the large number of adjustable parameters results in the need for a significant amount of sequence data as well as computational resources. These two limitations are becoming less onerous. With the modifications described here, and with the availability of powerful parallel computing systems, it is now possible to obtain realistic estimates of the distribution of selection coefficients from phylogenetic data.

Acknowledgments

We thank Ziheng Yang and Simon Whelan for helpful discussions. A.U.T. is supported by a studentship grant from the Wellcome Trust, United Kingdom. M.d.R. is supported by a research grant awarded to Ziheng Yang by the Biotechnology and Biological Sciences Research Council, United Kingdom. R.A.G. is supported by the Medical Research Council, United Kingdom.

Literature Cited

Akashi, H., 1999 Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene 238: 39–51.

Aronson, H., and W. Royer, Jr., 1994 Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. Protein Sci. 3: 1706–1711.

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B Biol. Sci. 355: 1553–1562.

Boni, M. F., Y. Zhou, J. K. Taubenberger, and E. C. Holmes, 2008 Homologous recombination is very rare or absent in human influenza A virus. J. Virol. 82: 4807–4811.

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

Bustamante, C. D., 2005 Population genetics of molecular evolution, pp. 63–99 in *Statistical Methods in Molecular Evolution*. Springer-Verlag, New York.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature 416: 531–534.

Cartwright, R. A., N. Lartillot, and J. L. Thorne, 2011 History can matter: non-Markovian behavior of ancestral lineages. Syst. Biol. 60: 276–290.

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

Cowperthwaite, M. C., J. J. Bull, and L. A. Meyers, 2005 Distributions of beneficial fitness effects in RNA. Genetics 170: 1449–1457.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.

dos Reis, M., and L. Wernisch, 2009 Estimating translational selection in eukaryotic genomes. Mol. Biol. Evol. 26: 451–461.

dos Reis, M., A. J. Hay, and R. A. Goldstein, 2009 Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 'Spanish' influenza pandemic virus. J. Mol. Evol. 69: 333–345.

dos Reis, M., A. Tamuri, A. J. Hay, and R. A. Goldstein, 2011 Charting the host adaptation of influenza viruses. Mol. Biol. Evol. 28: 1755–1767.

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

Eyre-Walker, A., P. D. Keightley, N. G. Smith, and D. Gaffney, 2002 Quantifying the slightly deleterious mutation model of molecular evolution. Mol. Biol. Evol. 19: 2142–2149.

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

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

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368–376.

Felsenstein, J., 2003 *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Galtier, N., 2001 Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18: 866–873.
Gillespie, J., 1984 Molecular evolution over the mutational landscape. Evolution 38: 1116–1129.
Gillespie, J., 1994 *The Causes of Molecular Evolution*. Oxford University Press, London/New York/Oxford.
Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics 165: 1269–1278.
Halpern, A. L., and W. J. Bruno, 1998 Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15: 910–917.
Hasegawa, M., H. Kishino, and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22: 160–174.
Hietpas, R. T., J. D. Jensen, and D. N. Bolon, 2011 Experimental illumination of a fitness landscape. Proc. Natl. Acad. Sci. USA 108: 7896–7901.
Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. 8: 269–294.
Holder, M. T., D. J. Zwickl, and C. Dessimoz, 2008 Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363: 4013–4021.
Kimura, M., 1968 Evolutionary rate at the molecular level. Nature 217: 624–626.
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., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK/London/New York.
Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of variation at the white gene of *Drosophila melanogaster*. Genetics 144: 635–645.
Kleinman, C. L., N. Rodrigue, N. Lartillot, and H. Philippe, 2010 Statistical potentials for improved structurally constrained evolutionary models. Mol. Biol. Evol. 27: 1546–1560.
Koshi, J. M., and R. A. Goldstein, 1998 Models of natural mutations including site heterogeneity. Proteins 32: 289–295.
Kosiol, C., I. Holmes, and N. Goldman, 2007 An empirical codon model for protein sequence evolution. Mol. Biol. Evol. 24: 1464–1479.
Lakner, C., M. T. Holder, N. Goldman, and G. J. Naylor, 2011 What's in a likelihood? Simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. Syst. Biol. 60: 161–174.
Li, W. H., 1978 Maintenance of genetic variability under the joint effect of mutation, selection and random drift. Genetics 90: 349–382.
Loytynoja, A., and N. Goldman, 2008 Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320: 1632–1635.
Lynch, M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.
Maynard-Smith, J., and G. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.
Merriwether, D. A., A. G. Clark, S. W. Ballinger, T. G. Schurr, H. Soodyall *et al.*, 1991 The structure of human mitochondrial DNA variation. J. Mol. Evol. 33: 543–555.
Nielsen, R., and Z. Yang, 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., 1973 Slightly deleterious mutant substitutions in evolution. Nature 246: 96–98.
Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23: 263–286.
Orr, H. A., 2003 The distribution of fitness effects among beneficial mutations. Genetics 163: 1519–1526.
Penny, D., B. J. McComish, M. A. Charleston, and M. D. Hendy, 2001 Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. 53: 711–723.
Piganeau, G., and A. Eyre-Walker, 2003 Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc. Natl. Acad. Sci. USA 100: 10335–10340.
Rambaut, A., O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger *et al.*, 2008 The genomic and epidemiological dynamics of human influenza A virus. Nature 453: 615–619.
Rodrigue, N., H. Philippe, and N. Lartillot, 2006 Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. 23: 1762–1775.
Rodrigue, N., H. Philippe, and N. Lartillot, 2010 Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl. Acad. Sci. USA 107: 4629–4634.
Sanjuan, R., A. Moya, and S. F. Elena, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc. Natl. Acad. Sci. USA 101: 8396–8401.
Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.
Sawyer, S. A., R. J. Kulathinal, C. D. Bustamante, and D. L. Hartl, 2003 Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. J. Mol. Evol. 57(Suppl. 1): S154–S164.
Sawyer, S. A., J. Parsch, Z. Zhang, and D. L. Hartl, 2007 Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc. Natl. Acad. Sci. USA 104: 6504–6510.
Shackelton, L. A., C. R. Parrish, and E. C. Holmes, 2006 Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62: 551–563.
Stamatakis, A., T. Ludwig, and H. Meier, 2005 Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.
Stephan, W., 1995 Perturbation analysis of a two-locus model with directional selection and recombination. J. Math. Biol. 34: 95–109.
Stuart, A., J. K. Ord, and S. Arnold, 1999 *Advanced Theory of Statistics: Classical Inference and the Linear Model*, *Vol. 2A*. Arnold, London.
Subramanian, S., and S. Kumar, 2006 Higher intensity of purifying selection on >90 mutation rates. Mol. Biol. Evol. 23: 2283–2287.
Tamuri, A. U., M. dos Reis, A. J. Hay, and R. A. Goldstein, 2009 Identifying changes in selective constraints: host shifts in influenza. PLoS Comput. Biol. 5: e1000564.
Taubenberger, J. K., A. H. Reid, R. M. Lourens, R. Wang, G. Jin *et al.*, 2005 Characterization of the 1918 influenza virus polymerase genes. Nature 437: 889–893.
Thorne, J. L., S. C. Choi, J. Yu, P. G. Higgs, and H. Kishino, 2007 Population genetics without intraspecific data. Mol. Biol. Evol. 24: 1667–1677.
Wloch, D. M., K. Szafraniec, R. H. Borts, and R. Korona, 2001 Direct estimate of the mutation rate and the distribution of fitness effects in the yeast *Saccharomyces cerevisiae*. Genetics 159: 441–452.
Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.
Yampolsky, L. Y., F. A. Kondrashov, and A. S. Kondrashov, 2005 Distribution of the strength of selection against amino acid replacements in human proteins. Hum. Mol. Genet. 14: 3191–3201.
Yang, Z., 1994 Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39: 105–111.

Yang, Z., 2006 *Computational Molecular Evolution*. Oxford University Press, Oxford.
Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591.
Yang, Z., and S. Kumar, 1996 Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13: 650–659.
Yang, Z., and R. Nielsen, 2008 Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. 25: 568–579.
Yang, Z., N. Goldman, and A. Friday, 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11: 316–324.

Communicating editor: J. J. Bull