Evolutionary Models Accounting for Layers of Selection in Protein-Coding Genes and their Impact on the Inference of Positive Selection

Nimrod D. Rubinstein,†,[1,2] Adi Doron-Faigenboim,†,[1] Itay Mayrose,†,[3] and Tal Pupko*,[1,2]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel

[2]National Evolutionary Synthesis Center, Durham, North Carolina

[3]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

†These authors contributed equally to this work.

***Corresponding author:** E-mail: talp@post.tau.ac.il.

Associate editor: Manolo Gouy

Research article

Abstract

The selective forces acting on a protein-coding gene are commonly inferred using evolutionary codon models by contrasting the rate of nonsynonymous substitutions to the rate of synonymous substitutions. These models usually assume that the synonymous substitution rate, Ks, is homogenous across all sites, which is justified if synonymous sites are free from selection. However, a growing body of evidence indicates that the DNA and RNA levels of protein-coding genes are subject to varying degrees of selective constraints due to various biological functions encoded at these levels. In this paper, we develop evolutionary models that account for these layers of selection by allowing for both among-site variability of substitution rates at the DNA/RNA level (which leads to Ks variability among protein-coding sites) and among-site variability of substitution rates at the protein level (Ka variability). These models are constructed so that positive selection is either allowed or not. This enables statistical testing of positive selection when variability at the DNA/RNA substitution rate is accounted for. Using this methodology, we show that variability of the baseline DNA/RNA substitution rate is a widespread phenomenon in coding sequence data of mammalian genomes, most likely reflecting varying degrees of selection at the DNA and RNA levels. Additionally, we use simulations to examine the impact that accounting for the variability of the baseline DNA/RNA substitution rate has on the inference of positive selection. Our results show that ignoring this variability results in a high rate of erroneous positive-selection inference. Our newly developed model, which accounts for this variability, does not suffer from this problem and hence provides a likelihood framework for the inference of positive selection on a background of variability in the baseline DNA/RNA substitution rate.

Key words: evolutionary models, positive selection, Ka/Ks, codon models, maximum likelihood.

Introduction

One of the main applications of molecular evolutionary models is to detect the intensity of selective pressures acting on genes and on specific sites within them. A common application is to use codon-based models for examining the evolution of protein-coding genes, in which the selection intensity is inferred by contrasting the rate of nonsynonymous (amino acid altering; Ka) to the rate of synonymous (silent; Ks) nucleotide substitutions. The underlying assumption in these analyses is that selective forces acting on protein-coding genes operate at the protein level only, whereas synonymous substitutions are free from selection and reflect the neutral rate of evolution. In such a case, Ks is homogenous across codon positions, Ka is heterogeneous, and the inference of site-specific Ka/Ks ratios is based solely on Ka variation. Accordingly, protein-coding sites showing a Ka/Ks ratio significantly lower than one are regarded as undergoing purifying selection, suggesting they are functionally or structurally important. Protein-coding sites with Ka/Ks significantly greater than one are indicative of positive Darwinian selection, suggesting adaptive evolution (Hurst 2002; Yang 2005). Finally, sites evolving with a Ka/Ks ratio not significantly different than one are regarded as free from selection at the protein level. Clearly, such inferences are obtained at the phylogenetic context and as such are only able to detect relatively old events of adaptation or diversification characterized by repeated events of positive selection. Other scenarios of adaptation require different methodologies for their detection. In this respect, a complementary approach for detecting recent events of adaptation is to analyze sequences at the population level, and indeed many such methods have been and are still being developed (e.g., Tajima 1989; McDonald and Kreitman 1991; Sabeti et al. 2007; Yi et al. 2010).

In standard model-based inference approaches for Ka/Ks, variation of Ks among sites is not explicitly accounted for. However, variation of Ks is expected if in addition to the protein level, selection also operates at the DNA and/or RNA level, with varying intensities among sites. A large body of evidence demonstrating selection at the DNA and RNA levels of protein-coding genes has been assembled over the years. Translational accuracy and efficiency

is a major purifying selective force operating on codon usage, mainly in unicellular organisms that maximize growth (Ikemura 1985; Sharp and Li 1987), and has been indicated to vary along the sequence (Akashi 1994; Zhou et al. 2009). Splicing regulatory elements within coding sequences are also known to exhibit significant degrees of conservation (Baek and Green 2005; Pagani et al. 2005; Xing and Lee 2005; Goren et al. 2006). Synonymous mutations may affect the mRNA structure stability and thus impact phenotype and fitness (e.g., Nackley et al. 2006; Kudla et al. 2009). Overlapping genes also exhibit reduced rates of synonymous substitutions (Miyata and Yasunaga 1978; Rogozin et al. 2002; Chamary et al. 2006), indicating selective constraints. In addition, various types of *cis*-elements embedded within coding regions, such as antisense RNAs (Katayama et al. 2005; He et al. 2008; Mercer et al. 2008), microRNAs (annotated in miRBase, Griffiths-Jones et al. 2008), and nucleosome-binding motifs (Segal et al. 2006; Warnecke et al. 2008), are expected to be under purifying selection at the DNA and RNA levels. In a genomewide survey, Hellmann et al. (2003) estimated that ~39% of synonymous sites in primates are subject to substantial purifying selection. In addition to purifying synonymous selection, indications of positive synonymous selection have been reported (e.g., Resch et al. 2007). These observations suggest that realistic codon models should explicitly account for among-site synonymous substitution rate variation.

Evolutionary models that consider selection at synonymous sites have been developed for population genetics data (Bulmer 1991; McVean and Charlesworth 1999). More recently at the phylogenetic level, several codon models that account for variability in Ks have also been developed. Some aim to estimate preferred or unpreferred synonymous codon substitutions (Nielsen et al. 2007; Zhou et al. 2010) and thus rely on an a priori tabulation of the codons to preferred and unpreferred categories. Accurate tabulation of codon preference, however, depends on availability of gene expression data as codon preference is most biased in highly expressed genes (Sharp and Li 1987; Zhou et al. 2010). A major limitation to such an approach is that, at least for multicellular organisms in which gene expression varies from tissue to tissue and perhaps even between cells, preferred and unpreferred codons are unknown and cannot be simply derived. Yang and Nielsen (2008) developed a parameter-rich codon evolutionary model that aims to distinguish mutation, drift, and selection when analyzing the evolutionary forces acting on silent mutations. Both Pond and Muse (2005) and Mayrose et al. (2007) developed models that account for site-to-site variation of both the synonymous and the nonsynonymous substitution rates. These models, however, do not follow the established paradigm for the inference of positive selection, which compares the fit of two nested models to the data (in terms of maximum likelihood): one that explicitly allows for positive selection and one that does not (e.g., the M8 vs. M8a models implemented in PAML, Yang 2007). Rather, both Ka and Ks at each site vary according to independent gamma distributions, and consequently, positive selection is implicitly allowed in each site since the multivariate Ka, Ks distribution will have a fraction with Ka/Ks > 1. Thus, a measure of statistical support for the presence of positive selection is not provided by these models. In addition, these models rely on the traditional separation between Ka and Ks, whereas selective forces acting at the nucleotide level most likely operate simultaneously on both synonymous and nonsynonymous mutations (e.g., important elements at the DNA or RNA levels are expected to be maintained by selection not only against synonymous mutations but also against nonsynonymous ones). Finally, Scheffler et al. (2006) developed a model that introduces a synonymous substitution rate parameter that also affects nonsynonymous substitutions. The authors used this model to show that it reduces the rate of false inference of positive selection when recombination is present in the data. It is not clear, however, why synonymous substitution rates, which supposedly reflect DNA-/RNA-level constraints, should be constrained to vary only among codons rather than among codon sites.

In this work, we develop evolutionary codon models that distinguish between selection operating at the DNA and RNA levels and selection operating at the protein level. Our models allow testing for the presence of variability in the baseline DNA/RNA substitution rate among codon sites, and we show that such variability is common among vertebrate protein-coding genes. We construct two nested models that account for such variability, one that allows for the presence of positive selection and one that does not. By contrasting these models, it is possible to identify positive selection while accounting for variability in the baseline DNA/RNA substitution rate. Using simulations, we show that ignoring the spatial variation of DNA/RNA selective forces often leads to high rate of erroneous positive-selection inference. We note that since the synonymous substitution rate in our model varies among sites, it cannot be assumed to correspond to the neutral rate of evolution. Hence, equating Ka/Ks ratios higher than one with positive selection is not trivial. Nevertheless, we do stress that in the context of analyzing protein-coding genes, positive selection refers to the selection regime operating at the protein level, and we provide a discussion on how inferred Ka/Ks ratios higher than one, when spatial variation in the baseline DNA/RNA selective forces is present, should be interpreted.

Materials and Methods

A Multilayer Evolutionary Model

The codon model presented here is a continuous-time Markov process where the evolution of coding regions is represented as a multilayer process. Specifically, our model distinguishes between substitutions at the DNA/RNA level and substitutions at the protein level. These factors are captured within the Q matrix, which describes the instantaneous substitution rate from codon $I=i_1i_2i_3$ to codon $J=j_1j_2j_3$, where i_k is the nucleotide at the k'th codon site ($k = 1, 2,$ or 3) in codon I. For $I \neq J$, we define Q_{IJ} as follows:

$$Q_{IJ} = \begin{cases} \rho_k \alpha_{i_k j_k} \pi_J & I \text{ and } J \text{ differ by one synonymous substitution at codon site } k \\ \rho_k \alpha_{i_k j_k} \omega \pi_J & I \text{ and } J \text{ differ by one nonsynonymous substitution at codon site } k \\ 0 & I \text{ and } J \text{ differ by more than one nucleotide} \end{cases}. \quad (1)$$

The diagonal elements are determined by the constraint that each row in Q sums to zero. π_J is the frequency of codon J, which in our model is calculated using the product of the observed nucleotide frequencies at the three codon sites (the F3x4 model of Yang et al. 2000) but can also be calculated differently (e.g., the equilibrium frequency of the "target nucleotide," Muse and Gaut 1994). Selection pressure is modeled as a multiplicative effect on those instantaneous substitution rates. Specifically, selection at the DNA and RNA levels should operate on both synonymous and nonsynonymous mutations and hence affects all entries of Q and is represented by the ρ parameter. Selection at the protein level affects only nonsynonymous mutations and is represented by the ω parameter, which is the nonsynonymous-to-synonymous rate ratio. Both ρ and ω are treated as random variables sampled independently for each site from some prior distributions (see below for details). α_{ij} is the substitution factor to change from nucleotide i_K to nucleotide j_K. As in Yang and Nielsen (2008), the α_{ij} factors can be parameterized based on any time-reversible nucleotide substitution model, such as the general time reversible model (Yang 1994a) or the Hasegawa-Kishino-Yano (HKY) model (Hasegawa et al. 1985). Specifically, in any time-reversible nucleotide model, the instantaneous substitution rate from l to p equals $\alpha_{lp}\pi^*_p$, where π^*_p is the frequency of nucleotide p and the α_{lp} values describe the symmetrical part of the instantaneous rate matrix.

In the analyses presented in this study, the DNA- and RNA-level substitution process is based on the HKY model (Hasegawa et al. 1985). Thus, $\alpha_{lp}=\kappa$ if l and p differ by a transition and $\alpha_{lp}=1$ otherwise, where κ is the transition–transversion rate ratio.

Notably, in our model, we do not assume that all synonymous substitutions have the same rate. We also do not assume that all first codon sites evolve at a single rate, all second codon sites evolve at a single rate, and so forth. Rather, the evolution of a protein of length L codons is characterized by $3L$ values of ρ. Each such value is assumed to be sampled independently from some prespecified distribution, thus allowing these values to vary over different codon sites. This parameterization allows accounting for cases in which, for example, a certain site of a specific codon is important for mRNA structure stability, whereas another site in that codon site is free to vary.

Allowing for Site-to-Site Variation in the Substitution Process

Among-site rate variation is modeled by assuming that the baseline DNA/RNA substitution rate (ρ) and the nonsynonymous-to-synonymous rate ratio (ω) are random variables sampled independently for each site (codon and protein, respectively) from two independent distributions.

Similar to Yang et al. (2000), a beta + ω_s distribution is assumed over ω. Accordingly, ω is sampled from a discrete $B(p, q)$ distribution with C^ω rate categories with probability P_0, and assigned a value ω_s with probability $(1 - P_0)$. When $\omega_s=1$, only neutral and purifying selective forces are allowed. Alternatively, when ω_s can vary in the range $(1, \infty)$, positive Darwinian selection is allowed. These two options for ω parameterization are identical to the M8a and M8 models of Yang et al. (2000), respectively.

When no variation in the baseline DNA/RNA substitution rate is assumed, ρ equals 1 for all sites. Alternatively, when the baseline DNA/RNA substitution rate is allowed to vary, ρ is sampled from a gamma distribution $\Gamma(\alpha, \beta)$ approximated by C^ρ discrete rate categories (Yang 1994b). Due to the confounding effects between evolutionary rates and divergence times (Felsenstein 1981), the distribution of ρ is restricted to have mean one, which is facilitated by equating α and β resulting in a single shape parameter, α^ρ. This parameter determines the shape of the distribution of ρ, where the lower the α^ρ value the larger the variability in the baseline DNA/RNA substitution rates. All the above parameters are optimized using standard likelihood maximization procedures.

The four options presented above, homogenous or variable DNA/RNA substitution rates (ρH and ρV, respectively) with or without allowing for positive selection (M8 and M8a, respectively) define four models: M8a-ρH, M8-ρH, M8a-ρV, and M8-ρV. The M8a-ρH model can be used as a null model versus the M8-ρH model when testing for positive selection with no ρ variability (which is identical to the M8 vs. M8a model comparison implemented in the PAML package, Yang 2007). Similarly, the M8a-ρV model can be used as a null model versus the M8-ρV model when testing for positive selection with ρ variability. Finally, ρ variability can be tested by comparing either M8a-ρV versus M8a-ρH or M8-ρV versus M8-ρH. Each such model comparison can be performed using the likelihood ratio test (LRT) with one degree of freedom.

Estimating Site-Specific Selective Forces

Once the tree and model parameters are estimated, selective pressure at an individual site can be inferred. If DNA- or RNA-level selective forces are analyzed and the ρV model shows significantly better fit to the data than the ρH model, codon site–specific substitution rates can be used for inference of the degree of DNA- or RNA-level selection. In case positive selection is sought and the M8 model shows a significantly better fit to the data than the M8a model, protein-coding site–specific Ka/Ks ratios and their posterior probability to evolve under positive-selection

pressure can be used (Nielsen and Yang 1998). Specifically, the posterior probability that protein-coding site h evolved with a certain Ka/Ks ratio is as follows:

$$P(\omega^{(h)} = \omega_k | D_h, T, \theta) = \frac{P(D_h|T, \theta, \omega_k)P(\omega_k)}{\sum_{k'} P(D_h|T, \theta, \omega_{k'})P(\omega_{k'})}, \quad (2)$$

where D_h denotes the data at site h, T denotes the tree topology and the estimated branch lengths, θ denotes the set of model parameters, and $P(\omega_k)$ denotes the prior probability of ω_k. The summation in the denominator is over the $C^\omega + 1$ discrete rate categories of ω. In the ρV model, a distribution over ρ is assumed and the likelihood of each ω category accounts for all possible combinations of the ρ rate categories at each codon site:

$$P(D_h|T, \theta, \omega_k) = \sum_{(\rho_1,\rho_2,\rho_3)\in C^\rho} P(D_h|T, \theta, \omega_k, \rho_1, \rho_2, \rho_3)P(\rho_1)P(\rho_2)P(\rho_3). \quad (3)$$

The posterior expectation of the Ka/Ks value at each protein-coding site h is as follows:

$$E(\omega^{(h)}|D_h, T, \theta) = \sum_{k'} \omega_{k'} P(\omega^{(h)} = \omega_{k'}|D_h, T, \theta). \quad (4)$$

Data set Construction

To assess the extent of ρ variability in real data, we analyzed a large sample of protein-coding DNA sequences. Multiple sequence alignments (MSAs) were constructed as follows. All RefSeq (Pruitt et al. 2005) genes residing on human chromosome 1 were retrieved from the University of California–San Cruz (UCSC) hg18 database (Hsu et al. 2006). Whenever several isoforms were found for a certain locus, the largest one was used to represent the gene. We additionally filtered the data to include genes that were used by the Rhesus Macaque Genome Sequencing Consortium only (Gibbs et al. 2007). For the resulting list of genes, pregenerated MultiZ (Blanchette et al. 2004) alignments of 5-way mammals (Human-Chimpanzee-Rhesus-Mouse-Dog) and 17-way vertebrates (Human-Chimpanzee-Rhesus-Mouse-Rat-Rabbit-Dog-Cow-Armadillo-Elephant-Tenrec-Opossum-Chicken-Xenopus-Zebrafish-Tetraodon-Fugu) were downloaded from the UCSC genome browser using the Galaxy tool (Giardine et al. 2005). Each such alignment was constructed by concatenating the exon-specific alignments. The GBlocks program (Castresana 2000) was then used to remove badly aligned sites. Finally, alignments shorter than 50 codons were removed. This process resulted in 297 and 296 alignments for the 5-way and 17-way data sets, respectively. Not all the 17-way data sets included all 17 taxa, rather the average number of taxa in this data set was 11.32.

Tree topologies for the 5-way and 17-way data sets were derived from the National Center for Biotechnology Information taxonomy database (Wheeler et al. 000): (Rhesus, (Dog, Mouse), (Human, Chimpanzee)) for the 5-way data sets and (((Tetraodon,Fugu),Zebrafish),Xenopus,(Chicken,(Opossum,Armadillo,((Tenrec, Elephant), ((Cow,Dog),(((Rat,Mouse),Rabbit),((Human,Chimpanzee),Rhesus)))))))) for the 17-way data sets. Whenever a 17-way MSA did not include all 17 taxa, the corresponding missing taxa were pruned from the tree without altering the topology of the remaining taxa.

Simulation Studies

Simulation studies were conducted to assess both the ability of the ρV model to correctly detect cases of variability in codon site substitution rates and the impact of accounting for ρ variability on the inference of positive selection. For these tasks, sequence data were simulated with parameters $\kappa=2$ and $\omega\sim B(0.5, 2)$, without and with positive selection using $\omega_s=1$ and 2, respectively, with probability 0.2, and with different degrees of variability in codon site substitution rates using a range of α^ρ values (0.1, 0.2, 0.5, 0.8, 1.1, 1.4, 1.7, and 3.2), representing biologically realistic values based on real data (see Results). This amounted to a total of 16 model combinations. In addition, sequences with no variability in codon site substitution rates and without positive selection were simulated. In order not to limit our simulations to a specific phylogeny, 50 random trees of 20 taxa were generated according to a birth–death process using the Mesquite program (http://mesquiteproject.org) with default parameters (speciation rate 0.3 and extinction rate 0.1) and were scaled so that the total tree length equals 2. These trees were used as phylogenies for each set of simulated sequences. For each combination of model parameters, sequences were simulated along the 50 phylogenies to produce 50 data sets of indel-less 900-bp codon alignments. We used the 900-bp length as it is approximately the mode of MSA length distribution in the 5-way data set.

Source Code

The C++ code implementing the models described in this manuscript along with usage explanations are freely available without any restriction at http://www.tau.ac.il/~talp/multilayer.

Results

Assessing the Ability to Detect Variability in the Baseline DNA/RNA Substitution Rates

Simulations were conducted in order to assess the ability of our model to correctly detect variability in the baseline DNA/RNA substitution rate (ρ). To this end, we simulated sequences with considerable variability in ρ (obtained by setting α^ρ to 0.1) but without allowing positive selection. We then measured the percentage of data sets for which the M8a-ρV model was found to fit significantly better than the M8a-ρH model, according to the LRT at the 0.05 significance level. The M8a-ρV model significantly better fitted 100% of the data sets compared with the M8a-ρH model. This detection rate remained at 100% for increasing α^ρ values of 0.2, 0.5, 0.8, 1.1, 1.4, and 1.7 and decreased to 96% for an α^ρ value of 3.2. For sequences simulated assuming no ρ variability, the M8a-ρV fitted 6% of the data sets significantly better than the M8a-ρH model, which is

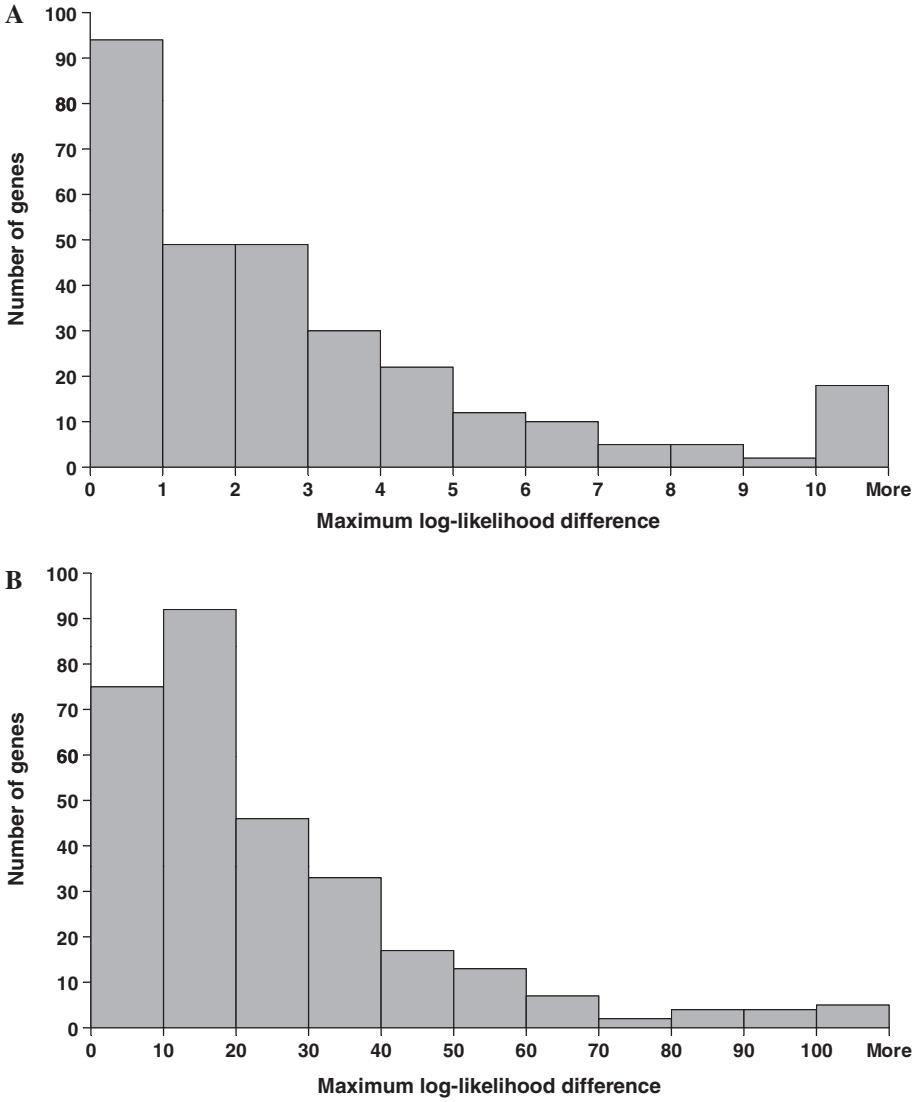

FIG. 1. Histograms of maximum log-likelihood difference obtained by comparing models M8a-ρV versus M8a-ρH on the (*A*) 5-way data and (*B*) the 17-way data. A value of 1.92 is considered significant at *P* value = 0.05.

expected as random error (*P* value = 0.46; binomial test assuming a 0.05 random error rate). We thus conclude that our test for ρ variability is accurately calibrated to detect variation in DNA/RNA substitution rates typical of real sequence data (discussed below).

Assessing the Extent of the Baseline DNA/RNA Substitution Rate Variability in Real Data

We next sought to estimate the extent of significant variability in the baseline DNA/RNA substitution rates in real sequence data. We constructed a data set of 297 five-way alignments of mammalian protein-coding genes (Human, Chimpanzee, Rhesus, Mouse, and Dog) mapped to human chromosome 1. LRT between the M8a-ρV and M8a-ρH models at the 0.05 significance level reveled that in 53% of the genes the null hypothesis of homogenous ρ can be rejected. After a 0.05-level false discovery rate (FDR) adjustment (Benjamini and Hochberg 1995), this value was reduced to 45% (fig. 1*A*). A similar result was obtained when the M8-ρV and M8-ρH models that allow for positive selection were compared (50% reduced to 44% after a 0.05-level FDR adjustment). Accordingly, the distribution of the α^{ρ} parameter that quantifies ρ variability among the genes explained significantly better by the ρV model was dominated by $\alpha^{\rho} < 2$ values (fig. 2*A*). When the alignments were augmented with more sequences by using the 17-way vertebrate data, this trend became even more pronounced. The ρV model showed a significantly better fit to 95% of the genes (remaining 95% after a 0.05-level FDR adjustment) when positive selection was not allowed (M8a models; fig. 1*B*). Similar to the 5-way data, the majority of 17-way genes explained better by the ρV model were found to be characterized by α^{ρ} values lower than 2 (fig. 2*B*). These results clearly suggest that variability in DNA-/RNA-level substitution rates is prevalent and can be detected even with relatively low taxonomic sampling. In addition, the ability to detect this signal is only marginally affected whether or not positive selection is allowed.

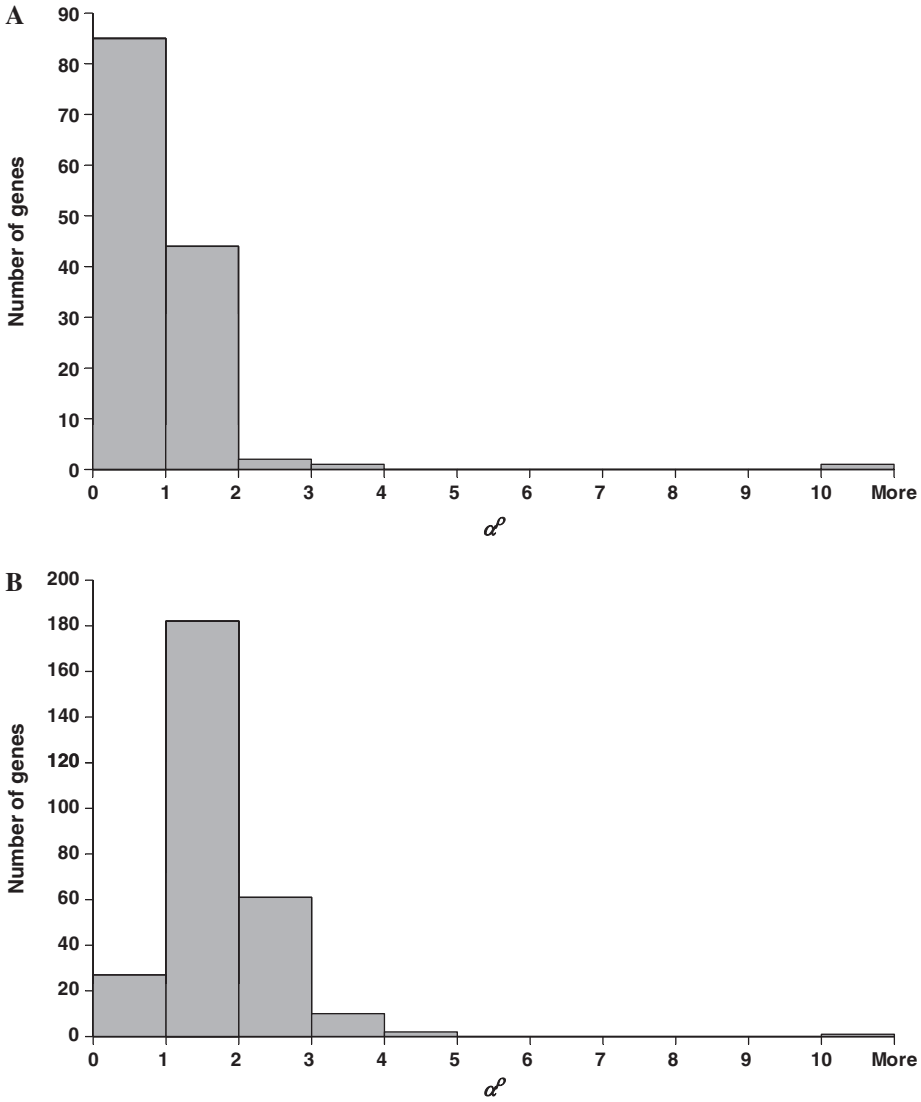

FIG. 2. Distribution of the α^{ρ} parameter that quantifies the intensity of the among-site variability of the baseline DNA/RNA substitution rate. The values were obtained by the M8a-ρV model for the (*A*) 5-way and (*B*) 17-way genes that were found to be significantly better explained by the M8a-ρV model compared with the M8a-ρH model.

In the results above, a gamma distribution was assumed to model the variability of the baseline DNA/RNA substitution rate. We tested whether this assumption is justified by comparing three variants of the ρV model, in which the variation of ρ is assumed to follow a gamma, a beta, or a uniform distribution. We analyzed all 5-way genes in which the ρV model was found to fit the data significantly better than the ρH model for all of the assumed ρ distributions. The Akaike Information Criterion (Akaike 1974) values showed a better fit for a gamma distribution compared with either a beta distribution or a uniform distribution (~96% and 100% of the compared data sets; P values $< 10^{-12}$ and 10^{-7}, respectively; Wilcoxon signed-rank test). We note that this only means that a gamma distribution most suitably captures the variation at the DNA-/RNA-level selective forces and not that the selective forces operating at the DNA/RNA level actually give rise to a gamma distribution (see Felsenstein 2001 for a related discussion).

The Effect of Accounting for Variability in the Baseline DNA/RNA Substitution Rate on the Inference of Positive Selection

Establishing that variability in the baseline DNA/RNA substitution rate is prevalent and our model has the power to detect it, we next assessed how it affects inference of positive selection. We compared the performances of the ρV and the ρH models in detecting positive selection when proteins evolve under variable degrees of DNA/RNA substitution rates. We treated this issue as a classification problem and hence simulated data with ($\omega_s = 2$) and without ($\omega_s = 1$) positive selection, with varying intensities of variability in the DNA/RNA substitution rates. We then used these data to perform a receiver operating characteristic (ROC) analysis. Subsequently, we used the area under the receiver operating characteristic curve (AUC) to assess the performance of the different models across the whole range of P values that are used as the threshold to

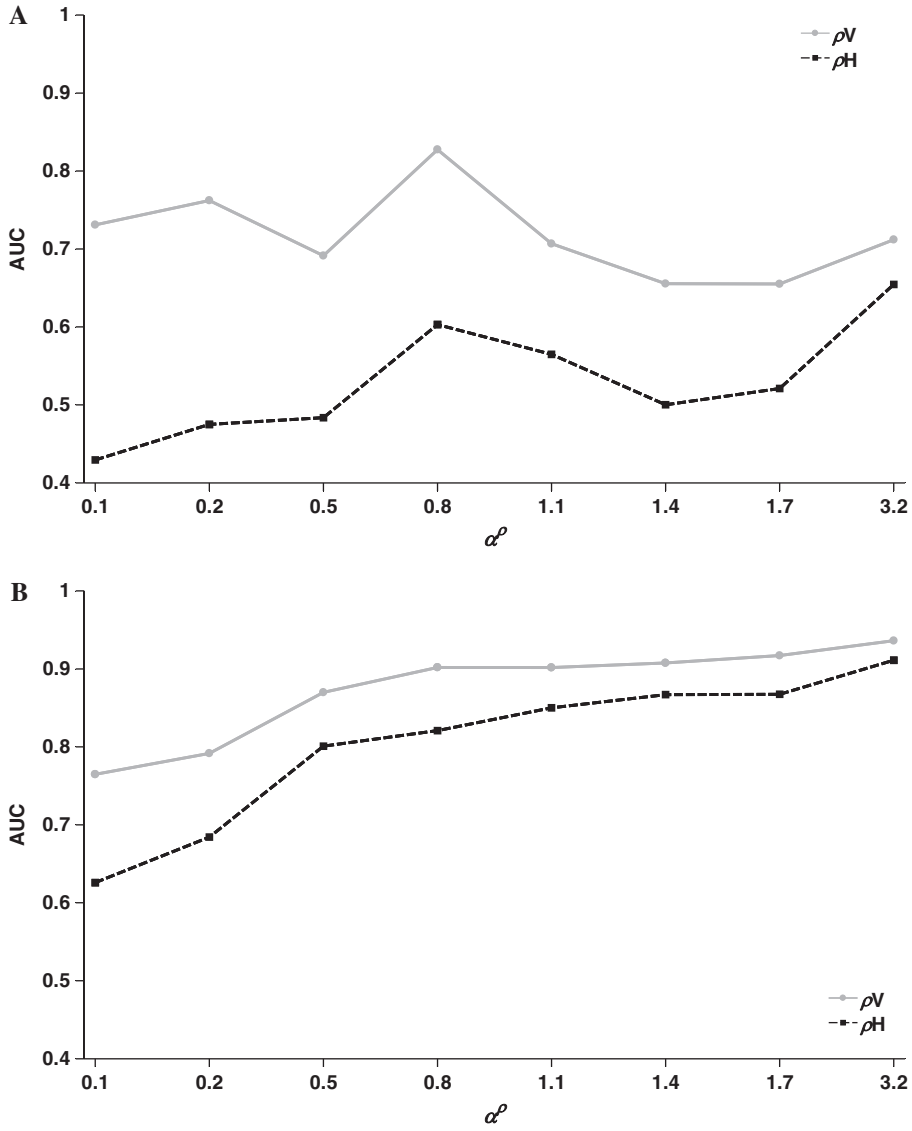

FIG. 3. The impact of accounting for baseline DNA/RNA substitution rate variability on the inference of positive selection. AUC values of the ROC analysis comparing the inference of positive selection according to the: (*A*) analysis of the whole gene using the LRT and (*B*) site-specific inference using posterior probabilities of Ka/Ks $>$ 1, under the ρH and ρV models, for varying degrees of DNA/RNA substitution rates, determined by the α^{ρ} parameter.

determine statistical significance. An AUC of 1 indicates optimal performance (i.e., the model can perfectly distinguish cases of positive selection from those of no positive selection), whereas an AUC of 0.5 indicates poor performance, comparable with that of a random predictor.

When sequences were simulated with strong variability in the baseline DNA/RNA substitution rates ($\alpha^{\rho}=0.1$), the ρH model obtained a poor AUC of 0.43. In contrast, the ρV model obtained an AUC of 0.73 (fig. 3A). The difference in AUCs obtained by the two models remained considerable through the entire range of ρ variability intensities ($\alpha^{\rho}=0.2$, 1.1, 0.5, 0.8, 1.1, 1.4, 1.7, and 3.2; fig. 3A). We thus conclude that in cases of strong-to-moderate variability in the baseline DNA/RNA substitution rates, the detection of positive selection is more accurate when ρ variability is accounted for, regardless of the *P* value used as threshold for invoking significant positive selection.

In practical analyses, positive selection is inferred when the *P* value of the LRT comparing the M8 versus M8a models is lower than 0.05. We thus additionally measured the percentage of false positive-selection inference of the ρH and ρV models at the 0.05 significance level for the above simulated data with the different intensities of codon site substitution rates and without positive selection ($\omega_s=1$). In the case of strong ρ variability ($\alpha^{\rho}=0.1$), positive selection is erroneously inferred in 96% of the data sets using the ρH model. In contrast, the ρV model erroneously infers positive selection in only 4% of the simulated data sets (fig. 4). The rate of false detection of the ρH model gradually declines through milder variabilities in codon site substitution rates (fig. 4), finally dropping to 22% for ρ variability simulated using $\alpha^{\rho}=3.2$. In contrast, the rate of false detection of the ρV model was nearly constant across the range of α^{ρ} values and did not exceed 8% (expected as random error,

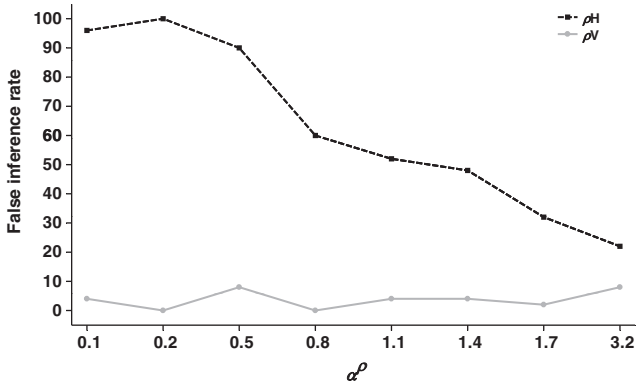

FIG. 4. The rate of falsely inferred positive selection at the 0.05 significance level of the M8 versus M8a LRT *P* value, obtained for the ρH and ρV models, for varying degrees of DNA/RNA baseline substitution rates, determined by the α^{ρ} parameter.

P value = 0.24; binomial test assuming a 0.05 random error rate). These results suggest that, often, apparent statistically significant positive selection may be an artifact obtained due to considerable variation in ρ when it is ignored.

Next, we tested the ability of the models to detect positive selection in specific sequence sites. We first performed an ROC analysis by varying the site-specific posterior probabilities of Ka/Ks $>$ 1 obtained by the M8 models (ρH and ρV; fig. 3*B*). In this case, the true site-specific Ka/Ks ratios are known from the simulation process, and hence protein-coding sites that were simulated with Ka/Ks $>$ 1 were regarded as under positive selection. In the case of strong ρ variability ($\alpha^{\rho}=0.1$), the ρH model obtained an AUC of 0.63 whereas the ρV model obtained an AUC of 0.76. The superiority of the ρV model over the ρH model in terms of AUC remained consistent yet gradually diminished through weaker degrees of variability in codon site substitution rates, culminating in 0.94 versus 0.91 for the ρV and the ρH models, respectively, for data simulated with moderate ρ variability ($\alpha^{\rho}=3.2$; fig. 3*B*). These results demonstrate that the ρV model has greater power in detecting protein-coding site–specific positive-selection forces than the ρH model.

We next sought to better understand what causes erroneous inference of site-specific positive selection by the ρH model. To this end, we analyzed data sets that were simulated under the ρV model without positive selection. We compared the average baseline DNA/RNA substitution rate in protein-coding sites that were erroneously inferred to be under positive selection (according to the ρH model) with that of the remaining protein-coding sites. Sites in which positive selection is erroneously inferred were found to have a significantly higher average baseline substitution rate compared with the remaining sites (*P* value $< 10^{-4}$ for all codon site rate variability intensities; paired *t*-test). This comparison suggests that sites that evolve under high baseline DNA/RNA substitution rates experience many nonsynonymous substitutions not driven by positive selection but rather by weaker selective constraints at the DNA/RNA level. The ρH model averages the Ks across the entire protein, which leads to an underestimation of Ks for such sites. Thus, the Ka/Ks ratio is artificially inflated at these sites, which leads to false positive-selection inference (see Discussion).

The ROC analyses performed above illustrate how powerful the compared models in detecting positive selection are, either according to the LRT or according to the site-specific posterior probabilities of $\omega>1$. Notwithstanding, it is also informative to study how accurately the models infer ω ratios in the range (0,1). To this end, we measured the correlation between the simulated site-specific ω values and their inferred posterior expectations, according to the ρV and ρH models. Similarly, we also measured the correlation between the simulated codon site–specific substitution rates (i.e., ρ values) and their inferred posterior expectations according to the ρV model. For data simulated without positive selection and with strong variability in the baseline DNA/RNA substitution rates ($\alpha^{\rho}=0.1$), the ρV and the ρH models showed similar correlations for the ω ratios (Pearson correlation coefficients 0.38 and 0.36, for the ρV and ρH models, respectively; *P* value $< 10^{-15}$ for both correlations). This indicates that accounting for DNA-/RNA-level selective forces does not come at the expense of reduced accuracy in the inference of protein-level selective forces.

The correlations for the codon site–specific substitution rates were higher, with Pearson correlation coefficients of 0.63, 0.61, and 0.76, for the first, second, and third codon sites, respectively; (*P* value $< 10^{-15}$ for all correlations). The higher correlation obtained for the third codon site may be explained by the fact that it is the least constrained among the three codon sites, and it thus experiences more substitutions and therefore has a stronger phylogenetic signal. In the same vein, the slightly lower correlations for the second codon site (compared with that of the first) is somewhat expected as it is the most constrained among the three codon sites and hence the least informative.

Case Study: The TRIM5α Gene

To exemplify the impact of accounting for possible ρ variability on the inference of positive selection, we chose the alpha isoform of the TRIM5 protein (TRIM5α) as a test case. TRIM5 is a member of the tripartite motif family in primate genomes. It includes a Really Interesting New Gene finger domain, a B-box domain, and a coiled-coil domain. In addition, an SPRY domain was found in the alpha isoform (Reymond et al. 2001). Rhesus cells were found to be resistant to HIV-1 infection, and it was found that TRIM5α is responsible for this restriction (Hatziioannou et al. 2004). This restriction factor was previously found to evolve under positive selection (Sawyer et al. 2005).

We first analyzed the TRIM5α alignment assuming homogenous ρ among codon sites. Positive selection was significantly supported (maximum log-likelihood difference of 91.25; *P* value $< 10^{-40}$; comparing the M8-ρH and M8a-ρH models), in accordance with Sawyer et al. (2005). Of its 493 protein-coding sites, 31 positively selected sites (Ka/Ks $>$ 1) were inferred at a posterior probability higher than 0.95 and

FIG. 5. Projection of the selection intensity values onto the primary sequence of the TRIM5α protein inferred by (*A*) the ρH model and (*B*) the ρV model. Inferred site-specific selective forces are colored according to the color scale: Purifying selection is colored in shades of burgundy, neutral selection is colored in white, and positive selection is colored in shades of yellow. Bins 1 and 2 represent sites with posterior probability above 0.99 and 0.95, respectively, to evolve with Ka/Ks $>$ 1. The two species-specific restriction determinants are indicated by red boxes.

20 sites at a posterior probability higher than 0.99 (fig. 5*A*). We then tested whether this gene evolved under variable ρ among codon sites comparing the M8a-ρV and M8a-ρH models. Indeed, the null hypothesis of among–codon site ρ homogeneity was rejected (75.15 maximum log-likelihood difference, P value $< 10^{-33}$). We thus tested again for positive selection, this time comparing the M8-ρV and M8a-ρV models. Positive selection was again detected, albeit with a lower log-likelihood difference (37.22 maximum log-likelihood difference, P value $< 10^{-17}$). However, 39 positively selected sites were inferred at a posterior probability higher than 0.95 and 23 sites at a posterior probability higher than 0.99 (fig. 5*B*). Among these 23 sites are those that comprise the SPRY domain, which was experimentally shown to account for the restrictive function of TRIM5α to HIV-1 in rhesus cells (Hatziioannou et al. 2004). This test case thus exemplifies that accounting for the baseline DNA/RNA substitution rate variability does not eliminate the signal of positive selection, if such is present in the data, but most probably results with more reliable inference.

Discussion

Codon Models Accounting for Layers of Selection in Protein-Coding Genes

Codon models have widespread use for the detection of selective forces. It is becoming evident that protein-coding genes may encode functions at the DNA and RNA levels, and hence any type of mutation, synonymous and nonsynonymous, may result with profound phenotypic consequences (e.g., Duan et al. 2003; Pagani et al. 2005; Nackley et al. 2006; Kimchi-Sarfaty et al. 2007; Kudla et al. 2009). This suggests that DNA-/RNA-level selection intensity should vary from site to site. Accordingly, substitution rates are expected to vary both among and within codon sites. This observation motivated us to explicitly incorporate such selection layers into codon evolutionary models. This extension to existing codon models is expected to have major implications, including the ability to detect and quantify codon site–specific selection (including positive selection), ancestral sequence reconstruction, the inference of phylogenetic trees, and accurate estimation of dating events.

In the multilayer evolutionary models developed here, selection at the DNA/RNA level equally affects both synonymous and nonsynonymous mutations. This satisfies the underlying biological assumption according to which constraints at the DNA and RNA level also affect the protein level. Such parameterization is more realistic compared with previous models, which used independent Ka and Ks distributions (Pond and Muse 2005; Mayrose et al. 2007). Our parameterization further allows explicit testing for the presence of positive selection using model comparisons. Nevertheless, extending the model to account for rate dependencies among sites (e.g., Mayrose et al. 2007), heterotachy (e.g., Lopez et al. 2002; Penn et al. 2008), and lineage-specific changes in selection intensities (Yang and Nielsen 2002) is warranted.

The baseline DNA/RNA substitution rate was found to be significantly variable in approximately half of the mammalian genes analyzed, emphasizing the justification for using our developed methodology. We note that our models do not explicitly distinguish between selection and mutation aside from estimating transition–transversion rate ratio. However, strong variability of mutation forces within a single gene should be rare and is thus considerably less likely to contribute to variability in the DNA/RNA substitution rates than varying selective forces. Notwithstanding, context-dependent mutation effects may be responsible for part of the baseline DNA/RNA variation in substitution rates. Clearly, more

studies are needed to understand the relative importance of the various factors that contribute to such a phenomenon (e.g., codon bias, mRNA structure stability, splicing signals, and context-dependent mutation effects) and how they change in various taxonomical groups.

The Inference of Positive Selection when the Baseline DNA/RNA Substitution Rate Varies among Sites

Classically, positive selection is invoked for protein-coding sites that experience more fixations of nonsynonymous mutations than would be expected by chance, that is, random fixations due to drift of neutral mutations. Ks is commonly used as a proxy for the expectation of the rate of such random fixations. However, if synonymous mutations are under purifying selection, their fixation rate no longer reflects the fixation rate of neutral mutations. This begs the question: How can positive selection be inferred under such settings?

We argue that although Ks can no longer be used to approximate the neutral substitution rate, the Ka/Ks ratio is still indicative of positive selection at the protein level. Our justification for this claim is as follows. Assume that a certain protein-coding site is not constrained by selection at the protein level (i.e., amino acid replacements at this site have no effect on the protein structure or function). Further assume that strong selective constraints operate at the DNA/RNA level of this site (e.g., due to RNA structure constraints). Ks at such a site should be lower than the neutral substitution rate, and Ka should be equal to Ks since synonymous and nonsynonymous mutations are equally affected by the DNA-/RNA-level selection and no additional constraints are imposed on nonsynonymous mutations compared with synonymous ones. The ρV model will thus infer Ka/Ks $= 1$ for that site, correctly indicating neutral evolution at the protein level, even though that site is under purifying selection at the DNA/RNA level. The ρH model, on the other hand, would most likely infer this site to be under purifying selection because the Ka/Ks ratio will be lower than 1 (i.e., the site-specific Ks value will be overestimated due to the homogenous Ks assumption).

Assume now that there is a diversifying selective force promoting fixations of nonsynonymous mutations at a certain protein-coding site and similar to the previous case, the site is also under purifying selection at the DNA/RNA level. Such a site evolves pleiotropically under this scenario, where selection both at the DNA/RNA level as well as at the protein level affects the same mutations, perhaps even differently (i.e., opposing them at the nucleotide level yet promoting them at the protein level). Ka at that site may be either higher or lower than the cross-sequence average Ks but definitely higher than the site-specific Ks. The ρV model will thus infer Ka/Ks > 1, correctly indicating positive diversifying selection at the protein level, which means an advantage for fixation of nonsynonymous mutations over synonymous ones. This conclusion would be true even if Ka at that site is lower than the cross-sequence average Ks.

When the DNA/RNA selective force is not accounted for, a protein-coding site is only regarded to be subject to selection at the protein level, be it purifying, neutral, or positive. In this framework, positive selection is invoked even though some of all the possible nonsynonymous mutations are probably deleterious and purified. Similarly, in our model, a protein-coding site is considered positively selected even though a fraction of the nonsynonymous mutations are subject to purifying selection stemming from the DNA-/RNA-level constraints. It is thus justified to invoke positive selection according to the Ka/Ks ratio using the ρV model, emphasizing the term "positive selection at the protein level" in lieu of the shorter term "positive selection."

Notwithstanding, there may be certain scenarios in which it is very difficult to distinguish between positive selection at the protein level and spatial variability of selective forces at the DNA/RNA level. For example, the first and third sites of a certain codon may be constrained by purifying selection to maintain an RNA structure, whereas the second site may be free to mutate since both at the DNA/RNA and the protein level it is free from constraints. In such a case, even the ρV model is expected to have difficulties in distinguishing positive selection from DNA/RNA selective forces. Although this scenario is quite biologically farfetched, it may certainly be present in our simulations and thus, at least in part, explain why the ρV model does not have optimal positive-selection inference power (as indicated by the AUC range of 0.66–0.83; fig. 3A). This scenario is also consistent with the above observation that the correlation between simulated and inferred substitution rates at the second codon site is the lowest among the three codon sites.

In practical terms, when positive selection is sought, we first recommend testing for variability of selective forces at the DNA/RNA level by comparing the ρV and the ρH models (the M8a model should suffice for that purpose). If variability in the baseline DNA/RNA substitution rate is ascertained, we recommend testing for positive selection by comparing the M8-ρV versus M8a-ρV models. Only if ρ homogeneity cannot be rejected, then the M8-ρH versus M8a-ρH test should be used.

Acknowledgments

We thank Ziheng Yang for extensive discussions and critical reading of an early version of this manuscript. We also thank Marcy Uyenoyama, Rasmus Nielsen, the associate editor, and two anonymous reviewers for helpful comments and suggestions. A.D.-F. and N.D.R. were supported by the Safra Bioinformatics Foundation at Tel Aviv University. N.D.R. and T.P. are supported by the National Evolutionary Synthesis Center (NESCent; NSF #EF-0905606). This study was supported by an Israel Science Foundation grant 878/09 to T.P.

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE T Automat Contr.* AC-19:716–723.

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* 136:927–935.

Baek D, Green P. 2005. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci U S A*. 102:12813–12818.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J Roy Stat Soc B*. 57:289–300.

Blanchette M, Kent WJ, Riemer C, et al. (12 co-authors). 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. 14:708–715.

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

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540–552.

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

Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum Mol Genet*. 12:205–216.

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

Felsenstein J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J Mol Evol*. 53:447–455.

Giardine B, Riemer C, Hardison RC, et al. (13 co-authors). 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*. 15:1451–1455.

Gibbs RAJ, Rogers MG, Katze R, et al. (176 co-authors). 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences—The complex definition of enhancers and silencers. *Mol Cell*. 22:769–781.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. 36:D154–D158.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22:160–174.

Hatziioannou T, Perez-Caballero D, Yang A, Cowan S, Bieniasz PD. 2004. Retrovirus resistance factors Ref1 and Lv1 are species-specific variants of TRIM5alpha. *Proc Natl Acad Sci U S A*. 101:10774–10779.

He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense transcriptomes of human cells. *Science* 322:1855–1857.

Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res*. 13:831–837.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC Known Genes. *Bioinformatics* 22:1036–1046.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet*. 18:486.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2:13–34.

Katayama S, Tomaru Y, Kasukawa T, et al. (31 co-authors). 2005. Antisense transcription in the mammalian transcriptome. *Science* 309:1564–1566.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315:525–528.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324:255–258.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol*. 19:1–7.

Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T. 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics* 23:i319–i327.

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

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

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A*. 105:716–721.

Miyata T, Yasunaga T. 1978. Evolution of overlapping genes. *Nature* 272:532–535.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11:715–724.

Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 314:1930–1933.

Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. *Mol Biol Evol*. 24:228–235.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.

Pagani F, Raponi M, Baralle FE. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A*. 102:6368–6372.

Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput Biol*. 4:e1000214.

Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*. 22:2375–2385.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 33:D501–D504.

Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol*. 24:1821–1831.

Reymond A, Meroni G, Fantozzi A, et al. (14 co-authors). 2001. The tripartite motif family identifies cell compartments. *EMBO J*. 20:2140–2151.

Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV. 2002. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet*. 18:228–232.

Sabeti PCP, Varilly B, Fry J, et al. (264 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.

Sawyer SL, Wu LI, Emerman M, Malik HS. 2005. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci U S A*. 102:2832–2837.

Scheffler K, Martin DP, Seoighe C. 2006. Robust inference of positive selection from recombining coding sequences. *Bioinformatics* 22:2493–2499.

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* 442:772–778.

Sharp PM, Li WH. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 15:1281–1295.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet*. 4:e1000250.

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA. 2000. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 28:10–14.

Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci U S A*. 102:13526–13531.

Yang Z. 1994a. Estimating the pattern of nucleotide substitution. *J Mol Evol*. 39:105–111.

Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39:306–314.

Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci U S A*. 102:3179–3180.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19:908–917.

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

Yi X, Liang Y, Huerta-Sanchez E, et al. (64 co-authors). 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.

Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol Biol Evol*. 27:1912–1922.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol*. 26:1571–1580.