FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection

Ben Murrell,[1,2,3] Sasha Moola,[1,3] Amandla Mabona,[1,4] Thomas Weighill,[1] Daniel Sheward,[5] Sergei L. Kosakovsky Pond,[6] and Konrad Scheffler*[,1,6]

[1]Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa

[2]Biomedical Informatics Research Division, eHealth Research and Innovation Platform, Medical Research Council, Tygerberg, South Africa

[3]Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[4]Department of Mathematics and Applied Mathematics, University of Cape Town, Cape Town, South Africa

[5]Division of Medical Virology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[6]Department of Medicine, University of California, San Diego

***Corresponding author:** E-mail: kscheffler@ucsd.edu.

Associate editor: Beth Shapiro

Abstract

Model-based analyses of natural selection often categorize sites into a relatively small number of site classes. Forcing each site to belong to one of these classes places unrealistic constraints on the distribution of selection parameters, which can result in misleading inference due to model misspecification. We present an approximate hierarchical Bayesian method using a Markov chain Monte Carlo (MCMC) routine that ensures robustness against model misspecification by averaging over a large number of predefined site classes. This leaves the distribution of selection parameters essentially unconstrained, and also allows sites experiencing positive and purifying selection to be identified orders of magnitude faster than by existing methods. We demonstrate that popular random effects likelihood methods can produce misleading results when sites assigned to the same site class experience different levels of positive or purifying selection—an unavoidable scenario when using a small number of site classes. Our Fast Unconstrained Bayesian AppRoximation (FUBAR) is unaffected by this problem, while achieving higher power than existing unconstrained (fixed effects likelihood) methods. The speed advantage of FUBAR allows us to analyze larger data sets than other methods: We illustrate this on a large influenza hemagglutinin data set (3,142 sequences). FUBAR is available as a batch file within the latest HyPhy distribution (http://www.hyphy.org), as well as on the Datamonkey web server (http://www.datamonkey.org/).

Key words: **evolutionary model, coding sequence evolution, approximate Bayesian inference, parallel algorithms.**

Article

Introduction

Codon-based models of evolution have proved extremely useful for identifying sites evolving under selection in protein-coding genes (Anisimova and Kosiol 2009; Delport et al. 2009). These models use a probabilistic approach to infer whether the nonsynonymous substitution rate (β) at a specific site is faster or slower than the neutral rate, which is typically set to the synonymous rate (α) at the same site (or to the mean synonymous rate for the entire alignment). However, existing software tools are simply too slow to allow analysis of many large data sets that are currently available.

The codon-modeling literature has largely focused on two ways of inferring the selection parameters (α, β), either jointly or as the ratio $\omega = \beta/\alpha$. First, in fixed effects likelihood (FEL) models (Kosakovsky Pond and Frost 2005; Massingham and Goldman 2005) the parameters are inferred independently for each site. This approach avoids assumptions about the distribution of selection parameters over sites, yielding greater flexibility to describe such distributions. However, the absence of parametric assumptions means that evidence from one site cannot inform our expectations regarding another: The inference at an individual site is based only on the limited amount of data from that site. The effect of this is that point estimates of site-specific parameter values can be unreliable, although robust inference is still possible by taking the uncertainty about these point estimates into account (Kosakovsky Pond and Frost 2005). Furthermore, methods where the number of parameters increases with the number of observations can be asymptotically inconsistent (Felsenstein 2001).

By contrast, random effects likelihood (REL) models (Nielsen and Yang 1998; Kosakovsky Pond and Muse 2005) are designed to share information across sites by inferring a gene-specific distribution for the selection parameters, with the assumption that the rates at each site are an independent draw from this distribution. Site-specific distributions for the selection parameters can then be obtained by application of Bayes' rule. Many parametric forms have been investigated for the gene-specific distribution of selection parameters

(Yang 2000) but to make parameter estimation tractable and to obtain reliable point estimates of parameter values, all of them are restricted to a small number of parameters. In addition, the distribution is either discrete or is discretized to allow numerical computation of the likelihood. The number of discrete components must be small, because the computational complexity of the likelihood calculation increases linearly with the number of discrete components. It is worth emphasizing that the synonymous and nonsynonymous substitution rates are inherently continuous-valued quantities and that their discretization is an approximation for computational convenience; as we show in later sections, overly coarse discretizations can mislead inference.

Huelsenbeck et al. (2006) proposed a nonparametric Bayesian approach, which addresses both the choice of distribution over selection parameters and the discretization, but at a prohibitive computational cost. Data augmentation techniques have improved the speed of inference under complex models (Lartillot 2006; Rodrigue et al. 2008; de Koning et al. 2012), but they remain intractable for large alignments.

In this article, we introduce FUBAR (a Fast Unconstrained Bayesian AppRoximation), which exploits several computational shortcuts to speed up the detection of positive or purifying selection, and to relax the above REL restrictions, leading to improved robustness against model misspecification and permitting the analysis of large data sets for which selection analysis was previously intractable. The key idea is to precompute a number of conditional likelihoods, arranged on an a priori-selected grid of values for α and β (in contrast to existing REL methods that shift the locations of the $[\alpha, \beta]$ categories during optimization, depending on the data). Inference of selection parameters then proceeds without requiring further phylogenetic likelihood computation, instead repeatedly reusing the precomputed values. Our default recommendation for the number of grid points is 400: This is large and therefore finely discretized compared with the number of (α, β) categories in typical random effects approaches, for example, that of Kosakovsky Pond and Muse (2005), which uses nine categories. However, as evidenced by the speedups we obtain, it is vastly smaller than the number of likelihood calculations performed during either optimization-based or sampling-based inference in existing methods, regardless of whether they use fixed or random effects models.

Although we have also used this approach to obtain large speedups in fixed effects models and in random effects models employing rate distributions with a small number of parameters, one of its key features is that it allows the implementation of far more parametrically complex models without extra computational cost. For this reason, we see the greatest utility in Bayesian approaches that allow large numbers of parameters to be used without being subject to overparameterization. Here, we present such an approach: following conventional random effects models, the selection parameters at each site are drawn from a gene-specific distribution for (α, β), but instead of using a low-dimensional parametric form for this distribution we adopt the general bivariate discrete distribution, parameterized by a weight at each point of the grid (fig. 1), and imposing no further constraints on the individual weights. Using a hierarchical Bayesian framework, we assume a Dirichlet prior for the gene-specific distribution of rate class weights, and use a Markov Chain Monte Carlo (MCMC) approach to integrate over the uncertainty in the posterior gene-specific and site-specific distributions. We show that this approach is less vulnerable to model misspecification than existing approaches, while also running orders of magnitude faster.

New Approaches

Following Muse and Gaut (1994), we model the evolution of a particular site along a particular branch of the phylogenetic tree as a continuous-time Markov process, governed by the

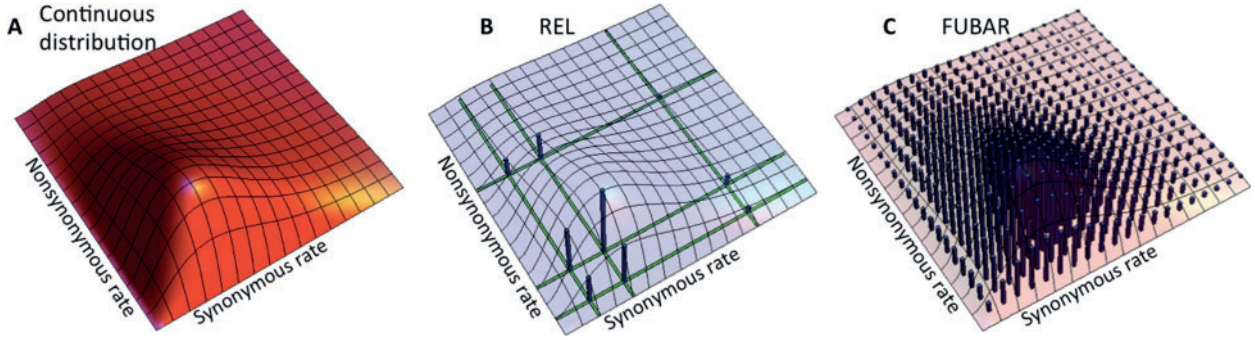

FIG. 1. The synonymous and nonsynonymous rates (α, β) are continuous model parameters that vary from one site to another, illustrated by a hypothetical distribution in (*A*). Typical random effects models, as exemplified by Dual REL (Kosakovsky Pond and Muse 2005) in (*B*) use a small number of discrete categories to approximate this continuous distribution, allowing the location (represented by the green bars) and the probability mass of the discrete points to vary; a change in the location of a point necessitates a re-evaluation of the phylogenetic likelihood function. FUBAR, in (*C*), uses a much denser grid of values chosen a priori, relying on the grid density to circumvent the need to move the parameter locations, and on MCMC to sample the weights assigned to each point. Without the need for movable grid lines, FUBAR needs to compute the conditional likelihood associated with each point only once, eliminating the bottleneck hindering traditional random effects models. Note that the uniform grid spacing depicted here is stylized. As the uncertainty in the selection parameters grows with their magnitude, FUBAR uses larger spacing for larger values (see text for details).

instantaneous rate matrix $Q = \{q_{ij}\}$, with elements that describe the rate of substitution of codon i with codon j:

$$q_{ij}(\alpha, \beta, \Pi, \mathcal{N}) = \begin{cases} \alpha\pi_j n_{ij}, & \delta(i,j) = 1,\ \mathrm{AA}(i) = \mathrm{AA}(j), \\ \beta\pi_j n_{ij}, & \delta(i,j) = 1,\ \mathrm{AA}(i) \neq \mathrm{AA}(j), \\ 0, & \delta(i,j) > 1, \\ -\sum_{k \neq i} q_{ik}, & i = j. \end{cases} \quad (1)$$

$\delta(i,j)$ counts the number of nucleotide differences between codons i and j, and AA(i) is the amino acid encoded by i. α and β are the rates of synonymous and nonsynonymous substitutions respectively. n_{ij} (comprising $\mathcal{N}$) are the nucleotide mutational biases, which we model using the 5-parameter general time reversible (GTR) nucleotide model (Tavaré 1986). π_j (comprising Π) denote the equilibrium frequency parameters.

We denote a phylogenetic tree $\mathcal{T}$, specifying both the tree topology and a branch length parameter, t_b, for every branch b. The probability of changing from codon i to j at a site along branch b in time t_b is recorded in the corresponding element of the transition matrix e^{Qt_b}. The likelihood of observing the site given the model parameters is calculated using Felsenstein's pruning algorithm (Felsenstein 1981). The goal of a selection analysis is to infer values for α and β for each site, and to provide a measure of evidence for the hypotheses that $\alpha > \beta$ or $\alpha < \beta$.

Modeling Variation in α and β

Calculating the Likelihood

The model used by FUBAR requires that the synonymous and nonsynonymous rates vary across sites. To achieve this, we follow Kosakovsky Pond and Muse (2005) and treat α and β as random effects, specifying a distribution from which they are drawn, and we integrate over that distribution to calculate the marginal likelihoods. To ensure identifiably, we require that $E[\alpha] = 1$. For computational tractability, these distributions are discrete. Furthermore, the sites are assumed to evolve independently, with the overall likelihood being the product of the site likelihoods. Thus, if x_k denotes the kth site of the alignment X, the overall likelihood can be calculated as:

$$p(X \mid \mathcal{T}, \Pi, \mathcal{N}, \theta) = \prod_k \sum_{\alpha,\beta} p(x_k \mid \alpha, \beta, \mathcal{T}, \Pi, \mathcal{N})\, p(\alpha, \beta \mid \theta), \quad (2)$$

where $p(\alpha, \beta \mid \theta)$ specifies the probability of each (α, β) combination, θ is a set of parameters governing this distribution, and $p(x_k \mid \cdot)$ is computed by the standard phylogenetic pruning algorithm (Felsenstein 1981).

Recycling Conditional Likelihoods

To prevent having to recalculate the conditional likelihoods $p(x_k \mid \alpha, \beta, \mathcal{T}, \Pi, \mathcal{N})$ in equation (2), we estimate the parameters that would affect them in advance and use the estimated values throughout. The equilibrium frequency parameters, $\hat{\Pi}$, are derived directly from nucleotide frequency counts using the CF3 × 4 estimator (Kosakovsky Pond et al. 2010). The nucleotide substitution rates, $\hat{\mathcal{N}}$, and the tree topology and branch length parameters, $\hat{\mathcal{T}}$, are fixed at the maximum likelihood estimates (MLEs) under a nucleotide model.

To construct a distribution over the selection parameters, a set of allowable values of α and β and their associated probabilities, $p(\alpha, \beta \mid \theta)$, must be specified. Random effects models typically specify parametric distributions for α and β as a function of θ. These distributions are either discrete or are subsequently discretized, in such a way that the allowable values of α and β depend on θ and therefore change at every step of the optimization or sampling procedure used to infer θ. As a result, existing random effects methods are forced to recompute the conditional likelihood many times during inference. We avoid this by fixing the locations of α and β to a prespecified grid (fig. 1).

The analyses in this manuscript use a square $N \times N$ ($N = 20$) grid, with points used to represent negative selection ($\alpha > \beta$), neutral evolution ($\alpha = \beta$), and positive selection ($\alpha < \beta$). Along a given axis, 70% of the points are used to describe rates <1 (for $N = 20$, there are 14 points at 0, 1/14, 1/7, … 13/14), a point is placed at 1, and the remainder of the points are spaced out over [1, 50] using cubic steps (for $N = 20$, there are 5 such points at $1 + (pn)^3$ for $n = 1, 2, 3, 4, 5$ and $p = \sqrt[3]{49}/5$). The nonlinear spacing of values above 1 can be justified by the empirical observation that the variance of rate estimates generally mirror the magnitude of the rate, that is, faster rates are more difficult to estimate precisely. The cap of rate values at 50 can be similarly justified by noting that, for most empirical data sets, any values above 50 are essentially infinite. Our preliminary experiments with different grids (results not shown) indicated that the inference of which sites are under selection was relatively robust to the choice of the grid. The software implementation permits the user to choose N, and it is straightforward to modify the grid definition if desired.

Finally, we parameterize $p(\alpha, \beta \mid \theta)$ as the general bivariate distribution, such that θ is a vector containing a probability weight for each point on the grid.

Markov Chain Monte Carlo

We model the probability weight vector θ (and hence the gene-wide distribution of $[\alpha, \beta]$) as a draw from a symmetric Dirichlet hyperprior:

$$p(\theta \mid c) \sim \prod_{n=1}^{N} \theta_n^{(c-1)}, \quad (3)$$

where N is the number of points in the grid and θ_n are the elements of the probability weight vector. The concentration parameter $0 < c \leq 1$, which controls the "clumpiness" of the distribution, is set to 0.5 for all analyses, but can be tuned by the user.

To perform the MCMC sampling, we implemented the Metropolis algorithm in the HyPhy software package (Kosakovsky Pond et al. 2005), seeking to obtain a set of samples from the posterior distribution of θ, given the alignment. We begin in an initial state based on relative

cumulative weights assigned to each grid point derived from the precomputed conditional likelihoods at each site k: $w(\alpha,\beta) \sim \sum_k p(x_k \mid \alpha,\beta,\mathcal{T},\Pi,\mathcal{N})/C_k$, where $C_k = \sum_{(\alpha,\beta)} p(x_k \mid \alpha,\beta,\mathcal{T},\Pi,\mathcal{N})$. We multiplicatively perturb each weight by a value sampled uniformly from [0.8, 1.2]. The resulting vector, $\theta^{[0]}$, is normalized so that the elements sum up to 1.

To propose a change to $\theta^{[t]}$, we first randomly pick two elements of the vector (grid points). A perturbation η is sampled from a uniform distribution between 0 and $\max(\min[0.001, 1/S], W)$, where S is the number of sites in the alignment and W is the median weight in the initial state. We chose the upper bound as an empirically derived value to optimize the rate of chain mixing. We add η to the first element and subtract it from the second to obtain a proposed new state θ'. $\theta^{[t+1]}$ is then set to θ' with probability

$$a = \min\left[1, \frac{p(X \mid \theta')p(\theta')}{p(X \mid \theta^{[t]})p(\theta^{[t]})}\right], \quad (4)$$

and to $\theta^{[t]}$ otherwise. Here, $p(\theta^{[t]})$ is obtained from equation (3) and $p(X \mid \theta^{[t]})$ is the likelihood (eq. 2), calculated from our precomputed conditional likelihoods using matrix multiplication. The proposal distribution implied by this procedure is symmetric; hence, we have no need for a proposal ratio in equation (4). The resulting MCMC chain can be computed extremely efficiently, drawing $10^4 - 10^5$ samples/second, which is sufficient to produce almost identical site posteriors on separate runs after a few minutes of run time.

We assess MCMC convergence using potential scale reduction factors (PSRFs) and an effective sample size (Gelman et al. 2003) computed for the posterior probabilities of positive selection for each site. For all data sets tested, an MCMC chain length of 2×10^6 with the first half discarded as burn-in yielded good convergence (assessed by running 10 MCMC chains in parallel from random starting positions). Each chain is thinned to yield T samples from the posterior distribution (the default implementation sets $T = 100$). On the influenza analysis (discussed in later section), for example, all PSRFs were less than 1.03 and all effective sample sizes were more than 150. Our implementation in the HyPhy software package allows the user to specify the chain lengths, but does not use automated stopping because the MCMC step is not a computational bottleneck: instead, computation times are dominated by fitting the nucleotide model (parallelized using OpenMP) and precomputing the conditional likelihoods (parallelized using MPI).

Site-Specific Inference

The MCMC procedure yields a set of T samples $\{\theta^{[t]}\}$. For each sample, we calculate the site-specific posterior distribution of (α,β) using Bayes' theorem:

$$p(\alpha,\beta \mid x_i,\hat{\mathcal{T}},\hat{\Pi},\hat{\mathcal{N}},\theta^{[t]}) = \frac{p(x_i \mid \alpha,\beta,\hat{\mathcal{T}},\hat{\Pi},\hat{\mathcal{N}})p(\alpha,\beta \mid \theta^{[t]})}{\sum_{\alpha,\beta} p(x_i \mid \alpha,\beta,\hat{\mathcal{T}},\hat{\Pi},\hat{\mathcal{N}})p(\alpha,\beta \mid \theta^{[t]})}. \quad (5)$$

The posterior probability that positive selection occurred at a site is the total probability that $\beta > \alpha$, averaging over the samples:

$$p(\beta > \alpha \mid x_i,\hat{\mathcal{T}},\hat{\Pi},\hat{\mathcal{N}}) = \frac{1}{T}\sum_{\forall t}\sum_{\forall \beta>\alpha} p(\alpha,\beta \mid x_i,\hat{\mathcal{T}},\hat{\Pi},\hat{\mathcal{N}},\theta^{[t]}). \quad (6)$$

Bayes factors can be calculated straightforwardly:

$$\mathrm{BF}(i) = \frac{p(\beta > \alpha \mid x_i,\hat{\mathcal{T}},\hat{\Pi},\hat{\mathcal{N}})}{1 - p(\beta > \alpha \mid x_i,\hat{\mathcal{T}},\hat{\Pi},\hat{\mathcal{N}})} \Big/ \frac{p(\beta > \alpha)}{1 - p(\beta > \alpha)}, \quad (7)$$

where the prior probability $p(\beta > \alpha)$ in the denominator is calculated by summing over the $\beta > \alpha$ portion of all MCMC samples $\theta^{[t]}$.

Results

Power and False-Positive Rates

To assess the statistical properties of FUBAR, we compared power and false-positive rates between FUBAR and FEL, using a collection of simulated alignments where the values for α and β varied from one site to another. These data were simulated over phylogenies estimated from three empirical data sets of varying size: 23 encephalitis virus *env* sequences, 38 vertebrate rhodopsin sequences, and 212 camelid VHH sequences (see Murrell, Wertheim, et al. 2012 for details).

Table 1 demonstrates the superiority of FUBAR over FEL. At a posterior threshold of 0.9, FUBAR achieves very low false-positive rates on data that were simulated under neutrality, and has better power in 21 of 27 configurations (and equal power in a further 2). To achieve a fair comparison between tests with different measures of evidence—*P* values vs. posterior probabilities—the thresholds were adjusted so that both FEL and FUBAR achieve false-positive rates of 0.05 on neutral data. This makes the superiority of FUBAR even clearer. FUBAR has greater power in every case, and the difference is sometimes substantial, especially for lower values of $\omega > 1$.

Speed Comparisons

We performed speed comparisons of FUBAR against the FEL and REL analyses implemented in HyPhy. REL methods are typically computationally intensive and nontrivial to parallelize, precluding their use on very large alignments with many sequences. Fixed effects methods are faster and typically parallelized, so FEL was used as our primary point of reference. A very large HIV-1 *env* alignment was obtained from LANL, stripped of gaps and subsampled to create alignments of varying size. To investigate how computation time increases with the number of sites, we sampled 100 taxa randomly from the *env* alignment and created 5 alignments with 50, 100, 200, 400, and 800 randomly sampled codon sites, respectively. To investigate how computation time increases with taxa, we fixed the number of sites to 200 and sampled alignments with 25, 50, 100, 200, 400, and 800 taxa. All phylogenies

Table 1. Comparative Performance of FEL and FUBAR on Simulated Data.

Simulation	FP : Power		Power at FP = 0.05	
	FEL	FUBAR	FEL	FUBAR
Encephalitis virus env				
$\omega^+ = 1.25$	0.01:0.03	0.00:0.01	0.04	0.10
$\omega^+ = 1.5$	0.00:0.03	0.00:0.02	0.09	0.14
$\omega^+ = 1.75$	0.00:0.03	0.00:0.04	0.08	0.17
$\omega^+ = 2$	0.00:0.05	0.00:0.07	0.13	0.24
$\omega^+ = 3$	0.00:0.09	0.00:0.20	0.19	0.38
$\omega^+ = 5$	0.00:0.19	0.00:0.44	0.34	0.60
$\omega^+ = 8$	0.00:0.28	0.00:0.60	0.50	0.74
$\omega^+ = 12$	0.00:0.34	0.00:0.67	0.54	0.82
$\omega^+ = 16$	0.00:0.38	0.00:0.77	0.63	0.85
Vertebrate Rhodopsin				
$\omega^+ = 1.25$	0.01:0.07	0.00:0.04	0.07	0.12
$\omega^+ = 1.5$	0.01:0.08	0.00:0.08	0.08	0.18
$\omega^+ = 1.75$	0.01:0.13	0.01:0.15	0.14	0.26
$\omega^+ = 2$	0.01:0.19	0.01:0.27	0.13	0.37
$\omega^+ = 3$	0.01:0.32	0.01:0.57	0.34	0.59
$\omega^+ = 5$	0.01:0.48	0.01:0.80	0.51	0.88
$\omega^+ = 8$	0.01:0.67	0.01:0.96	0.74	0.98
$\omega^+ = 12$	0.00:0.71	0.00:0.99	0.80	1.00
$\omega^+ = 16$	0.00:0.76	0.00:0.99	0.88	1.00
Camelid VHH				
$\omega^+ = 1.25$	0.01:0.11	0.01:0.09	0.06	0.09
$\omega^+ = 1.5$	0.02:0.19	0.01:0.20	0.14	0.21
$\omega^+ = 1.75$	0.01:0.34	0.01:0.42	0.26	0.53
$\omega^+ = 2$	0.01:0.51	0.01:0.60	0.48	0.62
$\omega^+ = 3$	0.01:0.74	0.01:0.74	0.64	0.78
$\omega^+ = 5$	0.01:0.93	0.01:0.95	0.93	0.97
$\omega^+ = 8$	0.01:0.98	0.01:0.99	0.98	0.99
$\omega^+ = 12$	0.01:0.97	0.01:1.00	0.97	1.00
$\omega^+ = 16$	0.02:0.99	0.03:1.00	0.99	1.00

NOTE.—The rate of false positives (FP) and power are reported for a fixed nominal test *P* value of 0.05 for FEL, and a posterior threshold of 0.9 for FUBAR. To achieve a fair comparison between tests with different measures of evidence, power is also shown for the *P* value or posterior threshold that achieves FP of 0.05, estimated empirically from the distribution of *P* values or posteriors on the subset of sites evolving neutrally.

were estimated with FastTree 2 (Price et al. 2010) using the GTR nucleotide model. FEL and FUBAR were compared on a computing cluster, with the analyses running in parallel on 10 nodes each. FUBAR was consistently faster than FEL across all tested alignments. As can be seen in figure 2, FEL took from 3.3 times longer (214 s for FEL vs. 65 s for FUBAR) for the smallest alignment, to 19.5 times longer (1 h 2 min for FEL vs. 3 min for FUBAR) for the largest alignment, with the relative disparity increasing uniformly with alignment size. We also ran a discrete REL model, using three categories each for α and β and without parallelization, on the smallest and largest alignments. The running times were 22 min 25 s (20.7 times longer than FUBAR) and 35 h 29 min (709.7 times longer than FUBAR), respectively.

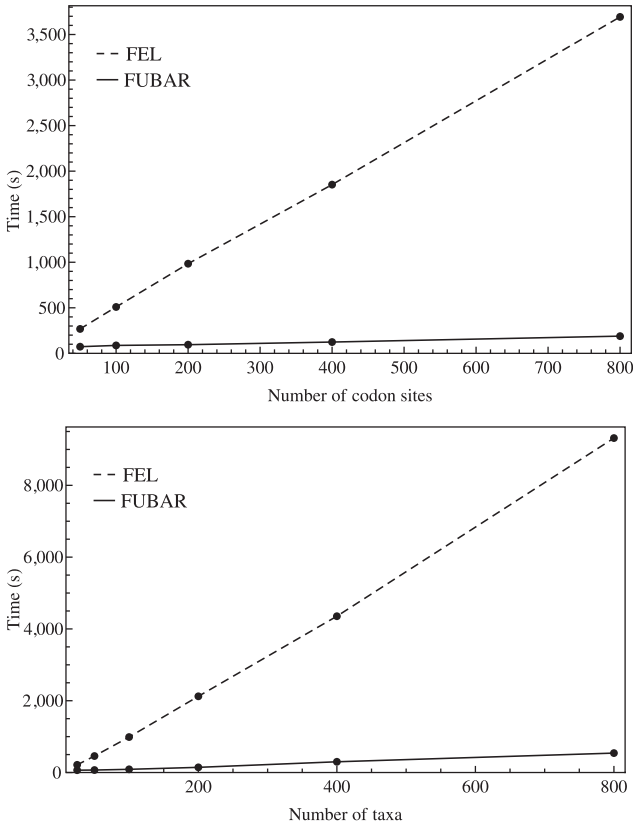

FIG. 2. Execution times for FEL and FUBAR as a function of the number of codon sites (top) and number of taxa (bottom).

Additionally (table 2), we used 16 alignments from a previous paper by our group (Murrell, Wertheim, et al. 2012), ranging in size and divergence level to provide a sense of a real-world speedup that could be realized by FUBAR. We compared FUBAR, FEL, REL, and the M2 (3 rate classes) and M8 (9 rate classes) models implemented in PAML v4.16. FUBAR and FEL were run on 10 processors (a number readily available even to researchers on a desktop). REL was run using 3×3 rate classes using built-in OpenMP parallelization in HyPhy (potentially using up to 9 processors). Finally, PAML was run on a single processor—to our knowledge no parallel version of the package exists—using the faster (by branch) optimization procedure (Yang 2000). All analyses were performed on systems equipped with 16-core 64-bit AMD Opteron 6272 processors running CentOS 6, and relied on gcc 4.4.6 to compile the source code.

Similar to the results in figure 2, FUBAR is the fastest of all methods except on the smallest alignments (e.g., the Primate Lysozyme alignment), and the benefit to using FUBAR becomes increasingly apparent with larger data sets, where, for example, PAML can run two orders of magnitude slower.

Robustness to Model Misspecification

Prior to FUBAR, random effects models typically used a small number of site categories to capture rate variation from one site to another. We wanted to investigate how empirical Bayesian inference behaves when the model is misspecified, and, in particular, when the model is too simple to

Table 2. Run Time Comparisons between Different Selection Detection Methods on 16 Empirical Data Sets, Sorted on the Duration of the FUBAR Run.

Data Set	Taxa	Codons	Mean Divergence Subs/Site	FUBAR Run Times (s)	Run Times (Times Slower than FUBAR)			
					FEL	REL	PAML M2a	PAML M8
Echinoderm H3	37	111	0.33	40	5.1	*12.0*	7.1	*46.1*
Flavivirus NS5	18	342	0.48	45	8.6	4.5	9.3	*25.5*
Drosophila adh	23	254	0.26	53	3.4	4.0	2.7	4.3
West Nile virus NS3	19	619	0.13	58	6.1	5.9	*37.2*	*105.5*
Hepatitis D virus Ag	33	196	0.29	59	4.0	3.3	*10.1*	*22.4*
Primate lysozyme	19	130	0.08	62	0.5	3.0	0.7	1.8
Vertebrate rhodopsin	38	330	0.34	62	*12.0*	4.9	8.4	*18.2*
Japanese encephalitis virus env	23	500	0.13	68	4.8	8.8	1.6	4.0
Mamallian β-globin	17	144	0.38	74	1.5	8.4	2.3	5.6
Abalone sperm lysin	25	134	0.43	78	1.9	3.9	3.7	9.3
HIV-1 vif	29	192	0.08	84	2.6	3.8	2.3	4.5
Salmonella recA	42	353	0.04	102	2.1	2.9	2.6	*12.3*
Camelid VHH	212	96	0.27	120	6.3	*17.2*	*141.0*	*311.1*
Diatom SIT	97	300	0.54	136	*10.2*	5.1	*21.5*	*19.3*
Influenza A virus H3N2 HA	349	329	0.04	210	*15.0*	*14.4*	*221.1*	*616.4*
HIV-1 rt	476	335	0.08	278	*15.2*	*14.4*	Ø[a]	Ø[a]

NOTE.—Run times that are at least 10 times greater than those of FUBAR are italicized, and those at least 100 times greater are underlined.
[a]PAML reported an error regarding too many ambiguities in the data set.

accommodate the data, as this is almost universally true of most models for real data sets. An example of this is the M2a model implemented in PAML (Wong et al. 2004), which postulates three categories for ω $(= \beta/\alpha)$. We simulated 10 replicate alignments of 1,000 sites each, using a constant $\alpha = 1$ (i.e., no synonymous rate variation, as is assumed in PAML), but with β taking values of 0.2 (50% of sites), 1 (30%), 3 (10%), and 11 (10%). This represents a situation where most sites are under purifying selection or evolving neutrally, whereas a smaller proportion of sites are under either weak or strong positive selection. The use of four site categories is seemingly a small violation of the M2a model, whose alternative model allows the following three categories: one purifying $(\omega^- < 1)$, one neutral $(\omega^N = 1)$, and one positive selection $(\omega^+ > 1)$. The point of this setup is that, in biological reality, the strength of positive selection is not constant across all sites experiencing positive selection—if this causes problems for M2a, it is reasonable to assume that coarse discretization is also problematic for many other models and not just for sites under positive selection.

The positive selection site category used by M2a must attempt to accommodate both the $\omega = 3$ and the $\omega = 11$ sites, and the resulting MLE (averaged over 10 replicates) is $\bar{\omega}^+ = 7.6$ (SD 0.64). The evidence in favor of positive selection at a specific site is determined by the ratio $\frac{P(\omega^+ \mid X)}{1-P(\omega^+ \mid X)} \approx \frac{P(\omega^+ \mid X)}{P(\omega^N \mid X)}$ between the posterior probability of it belonging to the positive selection category and that of it belonging to a different category (LHS); in this example, the latter is dominated by the probability of the site belonging to the neutral category (RHS). For any given gene-specific distribution (acting as a prior for the site-specific distribution), this ratio is proportional to the likelihood ratio $\frac{P(x_i \mid \omega^+)}{P(x_i \mid \omega^N)}$, i.e., the ratio between the likelihoods evaluated at $\omega = 7.6$ and at $\omega = 1$: this represents the contribution from the data at the site in question. The true peak of the likelihoods for most sites of interest is between these values, declining to either side. For some sites, the likelihood at $\omega = 1$ is higher than at $\omega = 7.6$, and vice versa for other sites. See figure 3 (top) for a visual depiction.

The effect of this (fig. 3, bottom) is that, among sites simulated with $\omega = 3$, M2a reports strong evidence in favor of positive selection (posterior probability > 0.90) for 41% of sites, but strong evidence against selection (posterior probability < 0.10) for 43% of sites. Instead of resulting in increased uncertainty (which would yield moderate posteriors), the slight model misspecification causes M2a to report incorrect inferences with high confidence. Discussion of what we would hope for should go in Discussion. In contrast, the dense conditional likelihood grid of FUBAR allows it to infer the presence of both $\omega > 1$ categories in the data and to base its site-specific inference on likelihoods evaluated much closer to the peak near $\omega = 3$. Of sites simulated with $\omega = 3$, 82% were detected with posteriors > 0.90, 0.4%—with posteriors < 0.50, and none with posteriors < 0.10. The mean posterior probability of $\omega > 1$ across sites simulated with $\omega = 3$ was 0.94 for FUBAR versus 0.49 for M2a.

A Large Empirical Example—Influenza A Virus Hemagglutinin

To demonstrate the use of FUBAR, we analyzed a collection of global human influenza A virus (IAV) hemagglutinin subtype 3 (H3) sequences from the NCBI Influenza Virus Database (http://www.ncbi.nlm.nih.gov/genomes/FLU/, last accessed July 2012). The influenza hemagglutinin glycoprotein (HA)

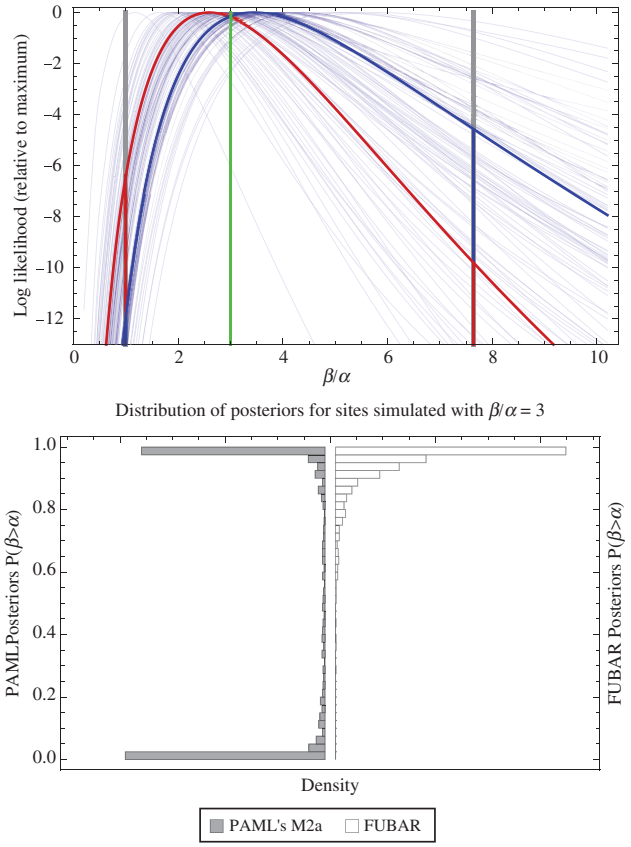

FIG. 3. Site-specific inference under misspecified models. (Top) 100 log-likelihood curves as functions of $\omega (= \beta/\alpha)$ for a set of simulated sites (see text for description). Vertical lines indicate the value $\omega = 3$ under which the sites were simulated, along with the values for the neutral and positive selection site categories ($\omega = 1$ and $\omega \approx 7.6$, respectively) used by the M2a model in PAML. The value of the positive selection site category does not match that under which the sites were simulated, due to the presence of other sites under stronger positive selection. The only evidence considered by M2a when classifying a site into the neutral or positive selection category is the value of the likelihood function at $\omega = 1$ and the value at $\omega = 7.6$. With the peaks of the likelihood functions between these options, the model becomes overconfident, assigning strong evidence either for positive selection (exemplified by the blue curve) or against it (exemplified by the red curve), even when this conclusion is incorrect. (Bottom) Histograms of posterior site-specific probabilities of positive selection calculated for sites simulated under a true $\omega = 3$. M2a (left) confidently identifies positive selection in nearly half of these sites, but also incorrectly declares strong evidence against positive selection in half. FUBAR (right) detects most of the sites, and does not claim strong evidence for incorrect conclusions.

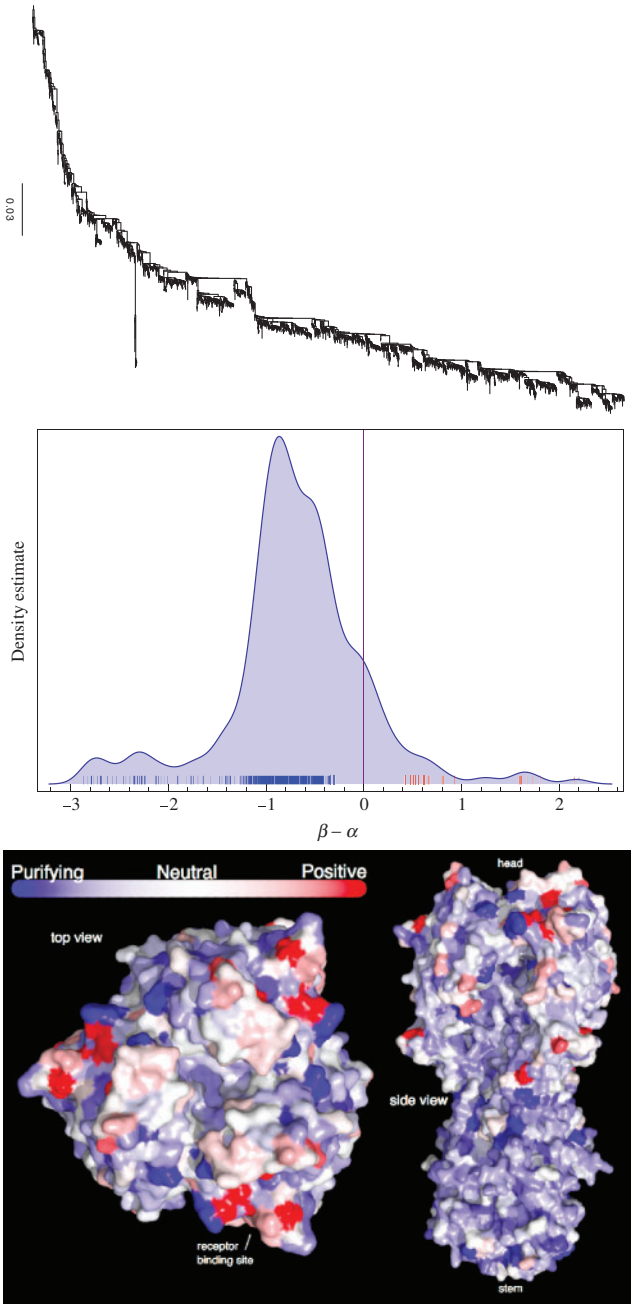

FIG. 4. Influenza hemagglutinin analysis. (Top) The H3 phylogeny with 3,142 coding sequences. (Middle) The smoothed histogram of $\beta - \alpha$ across H3, with the greatest density at mild purifying selection ($\beta < \alpha$), and fewer sites under positive selection ($\beta > \alpha$). The notches depict sites with posteriors greater than 0.9 for positive (red) or purifying (blue) selection. (Bottom) The inferred $\beta - \alpha$ values mapped to the HA protein (PDB 3ZTJ; Corti et al. 2011), displayed from two viewpoints. Red regions with stronger diversifying selection are likely involved in immune escape. These primarily occur on the "head" of the protein, with mostly purifying selection on the membrane proximal stem. See text for further detail.

mediates the entry of the virus into cells and is the target of neutralizing antibodies.

We reconstructed the phylogeny (fig. 4) for 3,142 complete H3 nucleotide sequences isolated from Humans using FastTree 2 (Price et al. 2010). The FUBAR selection analysis (which we restricted to 10 CPUs, just as for the timing comparisons) took one and a half hours. Figure 4 shows the distribution of $\beta - \alpha$ across HA, with the mode at mild purifying selection ($\beta < \alpha$), and with a minority of sites under positive selection ($\beta > \alpha$). We use $\beta - \alpha$ rather than the posterior $P(\beta > \alpha)$ because, with so many sequences, the posteriors can confidently report positive selection even when it is very weak, and so we examine the estimated magnitude of positive selection instead. As a measure of the magnitude of selection, β/α is very skewed (due to unreliability in estimates of this

ratio when α is small), but $\beta - \alpha$, with neutrality at 0, is more amenable to visualization. All sites described below are codon sites, given in H3 numbering (Winter et al. 1981), as opposed to the antigenic regions of the protein commonly referred to as "sites" in the influenza literature (but which we will refer to as "regions" here).

Codon sites under positive selection are almost exclusively localized to the globular head. Using $\beta - \alpha > 1$ as a working definition of strong positive selection, 11 codons were identified. Of these, seven sites (138, 145, 157, 194, 225, 226, and 229) are clustered in and around the receptor-binding site and fall broadly within three of the classical, major antigenic regions (regions A, B, and D; Wiley et al. 1981; Caton et al. 1982). Interestingly, site 226 projects into the receptor-binding pocket, and amino acid substitutions at this position can alter the receptor specificity (α (2, 6) vs. α (2, 3)) and consequently tissue tropism (Rogers et al. 1983). Sites 50 and 53 fall within region C (with site 45 located in close proximity). The remaining site under strong positive selection (site 3), does not lie within a previously defined antigenic region, and is likely located near the base of the membrane-proximal stem. The location of positively selected sites predominantly within the receptor binding site and antigenic regions is consistent with previous observations (Bush et al. 1999; Shih et al. 2007), and likely reflects selection for receptor binding avidity (Hensley et al. 2009) and immune escape.

The majority of sites under strong purifying selection are located within the stem. Antibodies to the HA stem are less common, but have nevertheless been shown to be able to mediate neutralization by inhibiting viral fusion with the host cell (Okuno et al. 1993; Varecková et al. 2003). This is consistent with the identification of broadly crossreactive antibodies that target this region (Ekiert et al. 2009; Sui et al. 2009; Wang et al. 2010; Corti et al. 2011), and reinforces the hemagglutinin stem as an attractive target for influenza vaccines.

Interestingly, HA2 site 172 is under extremely strong purifying selection ($\beta - \alpha = -23.5022$), although its function is not clear.

Of the sites under strong purifying selection in the globular head, sites 165, 187, 218, and 222 are clustered together in the quaternary structure at the protomer interface of the globular head, potentially representing a more accessible target for cross-neutralizing antibodies. Although site 165 represents an N-linked glycosylation site, which could potentially shield this region from antibody binding, it is also conceivable that the glycan may contribute to epitope formation. Several potent and broadly crossneutralizing HIV antibodies (PG9/PG16-like and PGT128-like antibodies) are dependent on both a peptide and a glycan component for binding (McLellan et al. 2011; Pejchal et al. 2011), providing a precedent for this mode of recognition.

Discussion and Conclusion

It is strikingly evident from figure 2 how slowly the computation time required by FUBAR increases with data set size. This means that it is particularly useful for analyzing very large data sets, for which selection analysis is simply not feasible using traditional methods. This is illustrated by our IAV example, which is, to our knowledge, the largest alignment analyzed for evidence of positive selection using phylogenetic codon-substitution models. However, FUBAR even offers a speed advantage on small data sets, along with its superior statistical performance in cases of model misspecification. Successful applications of the FUBAR implementation on Datamonkey that have already been published include a study of positive selection in the sugarcane mosaic virus (Li et al. 2013), porcine parvoviruses (Cadar et al. 2013), and hepatitis E virus (Smith et al. 2012).

Phylogenetic models of evolution have long employed computational shortcuts to speed up likelihood optimization, some of which we have adopted here. One widespread example involves the equilibrium frequencies: an estimate of the equilibrium frequency parameters, $\hat{\Pi}$, is often counted directly off the sequence data, invoking a stationarity assumption to reduce the number of parameters that need to be optimized (Kosakovsky Pond et al. 2010). This works because inference under the model appears not to be very sensitive to the typical magnitude of the deviations of these estimators from those derived using maximum likelihood (Kosakovsky Pond et al. 2010). Another example is that estimates of the nucleotide substitution rates, $\hat{\mathcal{N}}$, may be calculated using a simpler model—such as a codon model that does not allow site-to-site variability in selection intensity—and then fixed for the optimization of the more complicated model (Kosakovsky Pond and Frost 2005). This works for the same reason as the shortcut estimate $\hat{\Pi}$: inference is not usually affected. However, the justification for these shortcuts is merely empirical and their admissibility should be investigated further—comparing inference under the shortcuts to the full Bayesian solution—and situations where they mislead inference (if any) should be characterized.

Other work has used a variety of computational and statistical shortcuts to improve the computational efficiency of inference under phylogenetic models of evolution. Quang et al. (2008) pre-estimate a number of amino acid profiles from a large database, and then an analysis on a new alignment of interest proceeds by inferring weights for each profile. However, the goal of this method is to estimate the phylogeny, which requires recomputing the conditional likelihoods whenever the branch lengths or tree are modified during maximization, so the approach cannot exploit the likelihood recycling to the same degree as FUBAR. Fast phylogeny inference methods (e.g., FastTree—Price et al. 2010) employ fixed discretizations to handle site to site rate variation. Rather than computing the conditional likelihoods only once, which is prevented because they optimize over the phylogeny, they hard-assign each site to 1 of 20 rate classes, and only compute the likelihoods for those sites at those rate classes. This hard-assignment reduces the likelihood calculation to an approximation, but one that does not appear to have a negative impact on phylogeny inference.

Another common shortcut used here is to estimate the relative branch lengths under a simple model and fix them, although the overall tree length is still allowed to vary. This is adopted in the fixed effects models of Kosakovsky Pond and

Frost (2005) and in the Bayes empirical Bayes (BEB) approach of Yang et al. (2005). The BEB approach acknowledges that uncertainty about parameter values exists, but distinguishes between parameters for which these uncertainties matter (and where it is integrated out using a Bayesian approach) and parameters—such as relative branch lengths, nucleotide substitution rates, and equilibrium frequencies—for which it is sufficient to use the MLEs, ignoring the uncertainties in these point estimates. This approximation has been shown not to affect BEB inference on typical data sets (Scheffler and Seoighe 2005).

Inferring a gene-specific distribution of selection parameters allows information to be pooled across many sites, potentially resulting in improved power to detect selection at individual sites. Indeed, when we performed a FEL analysis (which does not do any information pooling) of the simulated data of the "Robustness to Model Misspecification" section, we detected only 31% of the sites simulated with $\beta/\alpha = 3$ at the $P < 0.05$ level, with a false-positive rate of 1.4% on the neutral and purifying sites, compared with 0.5% false-positives for FUBAR. The improved power of REL methods (including FUBAR) is not surprising in a simulation study where the data were generated using a gene-specific distribution of exactly the type assumed by these methods, but it seems reasonable to expect that information pooling should also be beneficial when analyzing biological data, provided the distributional assumptions used to do this do not cause problems due to model misspecification. Here, we have demonstrated a scenario in which traditional REL models suffer from exactly this problem: when the strength of selection is sufficiently heterogeneous across different sites in the same selective "category," inference can be severely misleading. This happens because the gene-specific distributions used by traditional REL models have highly restrictive parametric forms, using only a small number of parameters and discrete components that may not always match biological reality. FUBAR avoids this problem by using a highly flexible and therefore far less restrictive distributional form that is more robust against model misspecification.

Historically, models of evolution have been hampered by a large number of biologically unrealistic constraints, often necessitated by computational and/or statistical considerations. Examples include neglecting synonymous rate variation, confining sites to a small number of rate classes, assuming that different nucleotide and/or amino acid pairs have equal exchangeabilities, and assuming independence between different sites. Some of these constraints have already been lifted, while others are still in place. Bayesian approaches such as the one presented here offer a solution to the statistical problem of overparameterization in maximum likelihood methods; in conjunction with more efficient computational approaches this opens the door to using more biologically realistic models with larger numbers of parameters and hence fewer restrictions. In particular, our grid-based methodology has broad application potential: for instance, random effects approaches are used by DEPS (Kosakovsky Pond et al. 2008), EDEPS and MEDS (Murrell, de Oliveira, et al. 2012) to model directional selection—where the substitution rate toward a specific amino acid is elevated at a specific site—and by Branch-site REL (Kosakovsky Pond et al. 2011) and MEME (Murrell, Wertheim, et al. 2012) to model selection that varies across lineages. Grid-based variants of these methods could be constructed, allowing a large number of nonneutral categories, which should improve the statistical performance of these methods.

Acknowledgments

This work was supported in part by the National Institutes of Health grants AI47745, AI57167, AI74621, and GM093939, the UC Laboratory Fees Research Program grant 12-LR-236617, the National Research Foundation of South Africa, the University of Cape Town's University Research Council, and Europeaid grant SANTE/2007/147-790 from the European Commission.

References

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol.* 26:255–271.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol.* 16: 1457–1465.

Cadar D, Cságola A, Kiss T, Tuboly T. 2013. Capsid protein evolution and comparative phylogeny of novel porcine parvoviruses. *Mol Phylogenet Evol.* 66:243–253.

Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* 31:417–427.

Corti D, Voss J, Gamblin SJ, et al. (23 co-authors). 2011. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* 333:850–856.

de Koning APJ, Gu W, Castoe TA, Pollock DD. 2012. Phylogenetics, likelihood, evolution and complexity (PLEX). *Bioinformatics* 28: 2989–2990.

Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform.* 10:97–109.

Ekiert DC, Bhabha G, Elsliger MA, Friesen RHE, Jongeneelen M, Throsby M, Goudsmit J, Wilson IA. 2009. Antibody recognition of a highly conserved influenza virus epitope. *Science* 324:246–251.

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.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. Bayesian data analysis (Texts in Statistical Science). 2nd ed. Boca Raton: Chapman & Hall/CRC.

Hensley SE, Das SR, Bailey AL, et al. (11 co-authors). 2009. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science* 326:734–736.

Huelsenbeck JP, Jain S, Frost SW, Pond SKL. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A.* 103:6263–6268.

Kosakovsky Pond S, Delport W, Muse SV, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One* 30:e11230.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 22:1208–1222.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. Hyphy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol.* 28:3033–3043.

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

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza a virus. *Mol Biol Evol.* 25:1809–1824.

Lartillot N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. *J Comput Biol.* 13:1701–1722.

Li Y, Liu R, Zhou T, Fan Z. 2013. Genetic diversity and population structure of sugarcane mosaic virus. *Virus Res.* 17:242–246.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169:1753–1762.

McLellan JS, Pancera M, Carrico C, et al. (47 co-authors). 2011. Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* 480:336–343.

Murrell B, de Oliveira T, Seebregts C, Kosakovsky Pond SL, Scheffler K; on behalf of the Southern African Treatment and R. N. S. Consortium. 2012. Modeling HIV-1 drug resistance as episodic directional selection. *PLoS Comput Biol.* 8:e1002507.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8:e1002764.

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.

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.

Okuno Y, Isegawa Y, Sasao F, Ueda S. 1993. A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. *J Virol.* 67:2552–2558.

Pejchal R, Doores KJ, Walker LM, et al. (31 co-authors). 2011. A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science* 334:1097–1103.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Rodrigue N, Philippe H, Lartillot N. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. *Bioinformatics* 24:56–62.

Rogers GN, Paulson JC, Daniels RS, Skehel JJ, Wilson IA, Wiley DC. 1983. Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. *Nature* 304:76–78.

Scheffler K, Seoighe C. 2005. A Bayesian model comparison approach to inferring positive selection. *Mol Biol Evol.* 22:2531–2540.

Shih AC, Hsiao TC, Ho MS, Li WH. 2007. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc Natl Acad Sci.* 104:6283–6288.

Smith DB, Vanek J, Ramalingam S, Johannessen I, Templeton K, Simmonds P. 2012. Evolution of the hepatitis E virus hypervariable region. *J Gen Virol.* 93:2408–2418.

Sui J, Hwang WC, Perez S, et al. (14 co-authors). 2009. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nat Struct Mol Biol.* 16:265–273.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences. In: Miura RM, editor. Vol. 17. Providence (RI): American Mathematical Society. p. 57–86.

Varecková E, Mucha V, Wharton SA, Kostolanský F. 2003. Inhibition of fusion activity of influenza A haemagglutinin mediated by HA2-specific monoclonal antibodies. *Arch Virol.* 148:469–486.

Wang TT, Tan GS, Hai R, Pica N, Petersen E, Moran TM, Palese P. 2010. Broadly protective monoclonal antibodies against H3 influenza viruses following sequential immunization with different hemagglutinins. *PLoS Pathog.* 6:e1000796.

Wiley DC, Wilson IA, Skehel JJ. 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* 289:373–378.

Winter G, Fields S, Brownlee GG. 1981. Nucleotide sequence of the haemagglutinin gene of a human influenza virus H1 subtype. *Nature* 292:72–75.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol.* 51:423–432.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22: 1107–1118.