Detecting Individual Sites Subject to Episodic Diversifying Selection

Ben Murrell[1,2], Joel O. Wertheim[3], Sasha Moola[2], Thomas Weighill[2], Konrad Scheffler[2,4], Sergei L. Kosakovsky Pond[4]*

[1] Biomedical Informatics Research Division, eHealth Research and Innovation Platform, Medical Research Council, Tygerberg, South Africa, [2] Computer Science Division, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa, [3] Department of Pathology, University of California San Diego, La Jolla, California, United States of America, [4] Department of Medicine, University of California San Diego, La Jolla, California, United States of America

Abstract

The imprint of natural selection on protein coding genes is often difficult to identify because selection is frequently transient or episodic, i.e. it affects only a subset of lineages. Existing computational techniques, which are designed to identify sites subject to pervasive selection, may fail to recognize sites where selection is episodic: a large proportion of positively selected sites. We present a mixed effects model of evolution (MEME) that is capable of identifying instances of both episodic and pervasive positive selection at the level of an individual site. Using empirical and simulated data, we demonstrate the superior performance of MEME over older models under a broad range of scenarios. We find that episodic selection is widespread and conclude that the number of sites experiencing positive selection may have been vastly underestimated.

Citation: Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, et al. (2012) Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet 8(7): e1002764. doi:10.1371/journal.pgen.1002764

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received January 17, 2012; **Accepted** April 26, 2012; **Published** July 12, 2012

Funding: This research was supported in part by the National Institutes of Health (AI47745 and AI57167, AI74621, GM093939); a Joint DMS/NIGMS Mathematical Biology Initiative through Grant NSF-0714991; a University of California San Diego Center for AIDS Research (UCSD CFAR)/NIAID Developmental Award to SLKP (AI36214); a National Institutes of Health Training Fellowship to JOW (AI43638); the Bioinformatics, Statistical Analysis, and Evolutionary Core of the UCSD CFAR; the National Research Foundation of South Africa; and Europeaid Grant number SANTE/2007/147-790 from the European Commission. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: spond@ucsd.edu

Introduction

Following the introduction of computationally tractable codon-substitution models [1,2] nearly two decades ago, there has been sustained interest in using these models to study the past action of natural selection on protein coding genes. Positive selection can be inferred whenever the estimated ratio (ω) of non-synonymous (β) to synonymous (α) substitution rates significantly exceeds one (reviewed in [3] and [4]). In the original models, the ω ratio was shared by all sites in an alignment, providing little power to detect the signature of positive selection. Indeed, even among classical examples of positively selected genes [5,6,7], most substitutions are expected to be neutral or deleterious [8]. Consequently, relatively few genes in which mean ω estimates are significantly greater than one are expected to exist, e.g. only 35/8079 were found in a human - chimpanzee genome-wide comparison [9].

Random effects codon-substitution models [10] permitted ω to vary from site to site, which made it possible to identify instances when positive selection had acted only upon a small proportion of sites. Such site-level models can detect which positions in a sequence alignment may have been influenced by diversifying positive selection, e.g. [11,12]. However, these models posit that diversifying selective pressure at each site remains constant throughout time, i.e. affects most lineages in the phylogenetic tree, (Figure 1A), and there are very few cases where this assumption is biologically justified (see [13,14,15,16] for examples of models that allow selection to vary throughout the tree). When a site evolves under purifying selection on most lineages, site methods which assume ω is constant over time may be unable to identify any episodic positive selection, since they will likely infer $\omega < 1$ [17]. It has been noted that positive selection is more readily identified in smaller alignments: counterintuitively, including additional sequences may cause sites to no longer be detected [18,19]. This phenomenon could be readily explained by purifying selection on some lineages masking the signal of positive selection on others.

We present a mixed effects model of evolution (MEME), based on the broad class of branch-site random effects phylogenetic methods recently developed by our group [20]. MEME allows the distribution of ω to vary from site to site (the fixed effect) and also from branch to branch at a site (the random effect, Figure 1B). Our approach provides a qualitative methodological advance over existing approaches which integrate site-to-site and lineage-to-lineage rate variation, e.g. the branch-site methods [17] or codon-based covarion models [13]. MEME can reliably capture the molecular footprints of both episodic and pervasive positive selection, a task for which current models are not well suited. Using empirical sequence data sets spanning diverse taxonomic categories and gene functions, along with comprehensive simulations, we demonstrate that MEME matches the performance of traditional site methods when natural selection is pervasive, and that MEME reliably identifies episodes of diversifying evolution affecting a small subset of branches at individual sites, where site methods often report purifying selection at the same site. For most

Author Summary

Identifying regions of protein coding genes that have undergone adaptive evolution is important to answering many questions in evolutionary biology and genetics. In order to tease out genetic evidence for natural selection, genes from a diverse array of taxa must be analyzed, only a subset of which may have undergone adaptive evolution; the same gene region may be under stabilizing or relaxed selection in lineages leading to other taxa. Most current computational methods designed to detect the imprint of natural selection at a site in a protein coding gene assume the strength and direction of natural selection is constant across all lineages. Here, we present a method to detect adaptive evolution, even when the selective forces are not constant across taxa. Using a variety of well-characterized genes, we find evidence suggesting that natural selection is generally episodic and that modeling it as such reveals that many more sites are subject to episodic positive selection than previously appreciated.

empirical data sets analyzed here, episodic selection appears to be the dominant form of adaptive evolution. The biological implications of this type of selection are discussed for each specific data set. We conclude by providing practical guidelines for applying MEME to biological data, and argue that while it is possible to reliably identify sites or branches subject to episodic diversifying selection, statistical power to detect individual branch-site pairs evolving adaptively is inherently limited by a small sample size available for such inference.

Methods

At its core, our approach uses phylogenetic models to describe the evolution of codon characters along a branch in a phylogeny by a continuous-time stationary Markov process. Given a phylogenetic tree τ, with B branches and a vector of relative branch length parameters $T=(t_i,i=1\dots B)$, the probability of changing from codon i to j at a site along branch b in time t_b, is recorded in the (i,j) element of the transition matrix $M_b(t_b)=e^{Qt_b}$, where Q is the rate matrix. The elements $Q=\{q_{ij}\}$ parameterize the instantaneous rate of substitution of codon i with codon j:

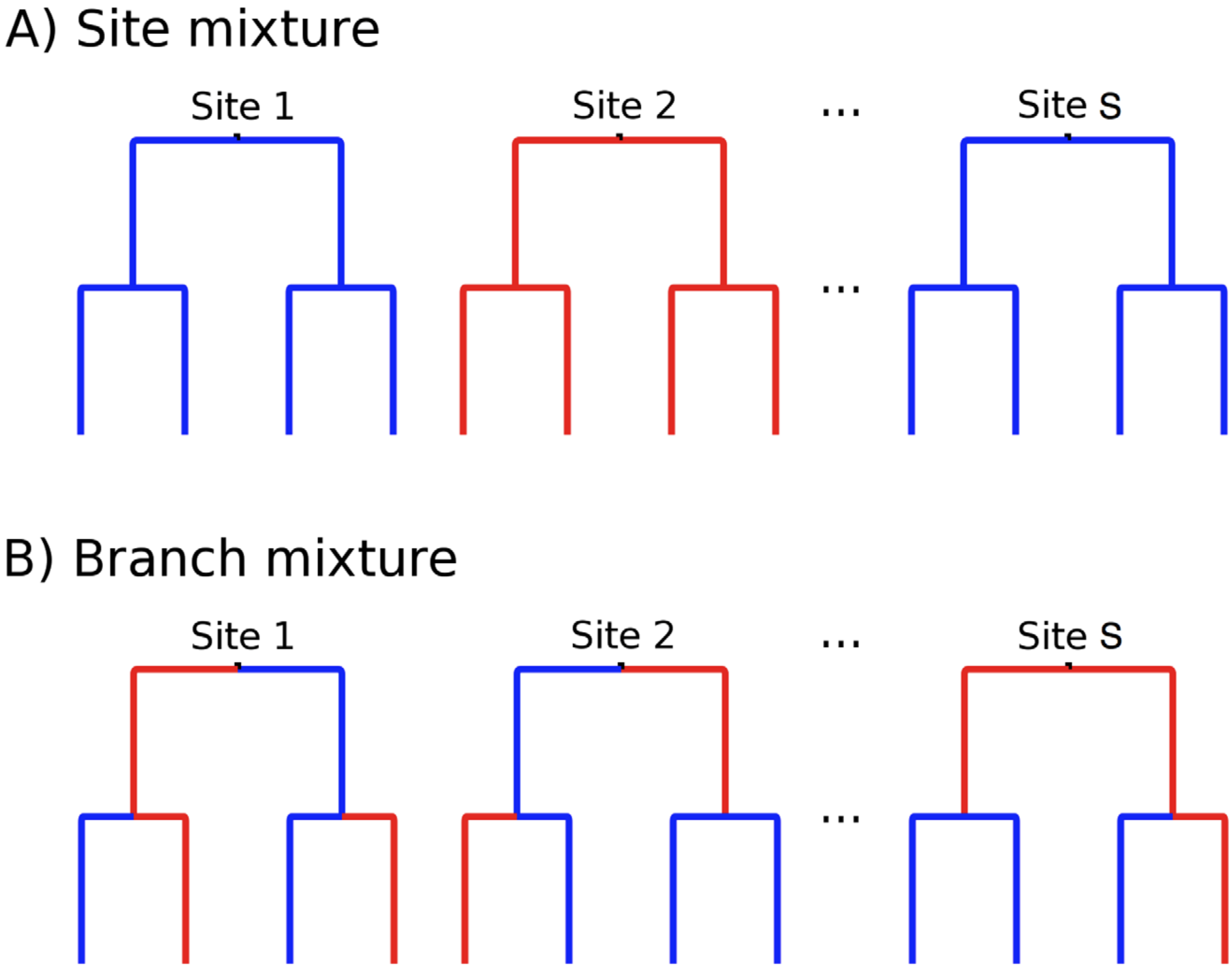

Figure 1. The standard random effects approach and samples. A) The standard random effects approach, in which the rates vary randomly over sites but are constant over branches. Different values of ω are showed in different colors. B) Samples from our new random effects approach [20], used by MEME, in which the rate on each branch is drawn independently of the rate on any other branch. All possible assignments of rates to sites are considered.
doi:10.1371/journal.pgen.1002764.g001

$$q_{ij}(\alpha,\beta,\Pi,\Theta)=\begin{cases}\alpha\theta_{ij}\pi_{ij}, & \delta(i,j)=1,\ \mathrm{AA}(i)=\mathrm{AA}(j),\\ \beta\theta_{ij}\pi_{ij}, & \delta(i,j)=1,\ \mathrm{AA}(i)\neq\mathrm{AA}(j),\\ 0, & \delta(i,j)>1,\\ -\sum\limits_{k\neq i}q_{ik}, & i=j.\end{cases}$$

$\delta(i,j)$ counts the number of nucleotide differences between codons i and j. α and β parameterize the rates of synonymous and non-synonymous substitutions, respectively. θ_{ij} (comprising Θ) are the nucleotide mutational biases, which we model using the 5-parameter general time reversible nucleotide model. π_{ij} (comprising Π) denote the equilibrium frequency parameters. Our estimate (denoted throughout as $\hat{\Pi}$) uses nine position-specific frequency parameters for the target nucleotides [1], corrected for the absence of stop codons using the CF3×4 estimator [21]. The likelihood of observing the site is calculated using the pruning algorithm [22] given the data, the tree (τ), the instantaneous rate matrix (Q), and the branch lengths (T).

To model the evolution of a site in an alignment in a manner that treats the non-synonymous rate (β) at each branch b as a random draw from one of K selective categories, we introduce a variable, c_b, which can take values from $1\ldots K$. An assignment of categories to all B branches, is described by the configuration vector $C=(c_1,\ldots,c_B)$ of branch categories. We assume that the category on each branch is independent of that on all other branches, and that each category has an associated probability, $p(c_b)$, for each branch. Next, we seek to marginalize the likelihood of each site D over all branch configuration vectors:

$$p(D)=\sum_C p(C)p(D|C)$$

Since this sum is over possible configurations, it has B^K terms, and would appear infeasible, unless B is small. However, if we assume that branch categories are independent, $p(C)=\Pi_{b=1}^{B}p(c_b)$, then the sum can be computed directly using the pruning algorithm by replacing the transition matrices with mixtures of transition matrices (see [20] for the derivation). If M_b is the transition matrix on branch b, and we denote Felsenstein's algorithm, which computes the probability of observing D given a transition probability matrix for every branch, as $\mathcal{F}(M_1,\ldots,M_B)$, then:

$$\begin{aligned}p(D)&=\sum_C\mathcal{F}(M_1^{c1},\ldots,M_B^{cB})p(C)\\&=\mathcal{F}\left(\sum_{c=1}^{K}p(c_1)M_1^{c1},\ldots,\sum_{c=1}^{K}p(c_B)M_B^{cB}\right),\end{aligned}\qquad(1)$$

where $M_b^{c_b}$ associates a transition matrix at each branch with a category. We have thus constructed a tractable model where the process at every branch is a random draw from a set of K categories.

In [20], we used this result to develop a model where each branch had a set of ω values and proportion parameters common to all sites. The goal was to identify lineages with a proportion of sites evolving with $\omega>1$. Here, we let each site have a set of free parameters governing the strength of selection for two discrete categories, and weights for each category, and these parameters are shared for all branches at that site. The goal is to detect sites where a proportion of lineages are evolving with $\omega>1$.

The MEME test for episodic diversifying selection

The fitting of MEME to an alignment of coding sequences proceeds in three stages:

First, the MG94×REV codon model with an alignment-wide $\omega=\beta/\alpha$ is fitted to the data using parameter estimates under a GTR nucleotide model as initial values. Although in some cases nucleotide branch lengths may be a good approximation to codon branch lengths [23,24], recent results indicate that in other instances, nucleotide models can significantly underestimate branch lengths and possibly bias downstream inference [25]. The resulting maximum likelihood estimates, $\hat{\Theta}$ and $\hat{t}_b$, for each branch $b\in 1\ldots B$, are used in the site-by-site analyses in the next two steps. Thus we are assuming that the relative branch length and mutational bias parameters are shared across sites and are well approximated by those estimated under a simpler codon model. However, the absolute branch lengths also depend on the site- and model-specific rate parameters below.

Second, at each site, we first fit the alternative random effects model of lineage-specific selective pressure with two categories of β: $\beta^-\leq\alpha$ and β^+ (unrestricted). The probability ($p(c_b)$ in equation 1) that branch $b\in 1\ldots B$ is evolving with $\beta^b=\beta^-$, is $0\leq q^-\leq 1$, and the complementary probability that it is evolving with $\beta_b=\beta^+$ is $q^+=1-q^-$. By equation 1, the phylogenetic likelihood at a site, marginalized over all 2^B possible joint assignments of β_b, is equivalent to computing the standard likelihood function with the following mixture transition matrix for each branch b:

$$\begin{aligned}&M_b(\alpha,\beta^-,\beta^+,q^-)=q^-e^{Q\left(\alpha,\beta^-;\hat{\Theta},\hat{\Pi}\right)\hat{t}_b}+\\&(1-q^-)e^{Q\left(\alpha,\beta^+;\hat{\Theta},\hat{\Pi}\right)\hat{t}_b}.\end{aligned}\qquad(2)$$

Consequently, the alternative substitution model includes four parameters for each site, inferred jointly from all branches of the tree: β^-,β^+,q^- and α. These form the fixed effects component of the model. Estimating α separately for each site accounts for the site-to-site variability in synonymous substitution rates [26].

Lastly, at every site, we fit the model from the previous step, but with $\beta^+\leq\alpha$: our null model. Using simulated data, we determined that an appropriate asymptotic test statistic for testing most worst-case null of of $\beta^+=\beta^-=\alpha$ is a 0.33 : 0.3 : 0.37 mixture of χ_0^2,χ_1^2 and χ_2^2 (see Text S1). Mixture statistics of this form often arise in hypothesis testing where model parameters take values on the boundaries of the parameter space, and closed-form expressions for mixing coefficients are difficult to obtain [27].

Throughout the manuscript, we compare MEME to the fixed effects likelihood approach, introduced in [24] (see Text S1 for motivation). The procedure used by FEL differs from MEME in that a single pair of α,β rates are fitted at each site (no variation over branches) in Step 2, and the test in Step 3 is to determine if $\alpha\neq\beta$. Positive selection is inferred by FEL when $\hat{\beta}>\hat{\alpha}$ and the p-value derived from the LRT is significant, based on the χ_1^2 asymptotic distribution.

Detecting individual branches subject to diversifying selection at a given site

If the LRT indicates that a particular site (s) is subject to episodic diversifying selection, it may be of interest to explore which branches at that site have undergone diversification. The empirical Bayes (EB) procedure originally used to identify individual sites subject to diversifying selection in random effects models [28], can be readily adapted here. To compute the empirical posterior probability at branch b that $\beta = \beta^{+} > \alpha$, we apply Bayes' theorem, using D_s to denote the data at site s and $\hat{\Theta}$ to denote all the maximum likelihood parameter estimates from the alternative MEME model fitted to site s:

$$P\left[\beta_b = \beta^{+} | D_s, \hat{\Theta}\right] = \frac{P(D_s | \beta_b = \beta^{+})(1 - q^{-})}{P(D_s | \beta_b = \beta^{+})(1 - q^{-}) + P(D_s | \beta_b = \beta^{-}) q^{-}}.$$

To compute the two likelihood terms $P(D_s | \beta_b = \beta^{+})$ and $P(D_s | \beta_b = \beta^{-})$, we use $q^{-} = 0$ and $q^{-} = 1$, respectively, for the model assigned to branch b in equation 2. The rest of the branches employ the matrices fitted under the alternative model of MEME. Having computed $P\left[\beta_b = \beta^{+} | D_s, \hat{\Theta}\right]$ for each branch b, we evaluate the empirical Bayes factor for the event of observing positive selection at each branch:

$$EB\left[\beta_b = \beta^{+} | D_s, \hat{\Theta}\right] = \frac{P\left[\beta_b = \beta^{+} | D_s, \hat{\Theta}\right] / P\left[\beta_b = \beta^{-} | D_s, \hat{\Theta}\right]}{(1 - q^{-})/q^{-}}.$$

When $EB > K > 1$, sequence data increase the prior odds of observing selection at the branch. We do not recommend using this type of inference other than for the purposes of data exploration, even for large values of K (e.g. 100). Intuitively, all the information contributing to the estimate of EB is derived from observing the evolution along a single branch at a single site (i.e. from a sample with size ≈ 1). To quantify this supposition, we simulated sequence data using the vertebrate rhodopsin phylogeny and branch lengths, applied positive selection of varying strength to five branches in the tree selected *a priori* (see Text S1), and applied the EB procedure to infer the identity of selected branches.

Results

Model assessment

To assess the performance of MEME on both simulated and empirical data, we selected the fixed effects likelihood method (FEL [24]) as the most appropriate reference test for pervasive diversifying selection, because FEL most closely matches the assumptions made by MEME (see Text S1). We simulated data sets under a number of scenarios: refer to Text S1 for details of simulation strategies.

Assessing the rates of false positives. Under the scenario where each site was evolved under the worst-case null hypothesis of constant $\omega = 1$, MEME had well controlled rates of false positives at test p-value of 0.05 (Figure S1, also see Text S1 for the empirical derivation of the asymptotic distribution of the test statistic for this hypothesis). MEME appears to be conservative for smaller sample sizes (numbers of sequences, N), but not for larger samples. The rates of false positives were < 0.01 ($N = 8$), 0.01 ($N = 16$), 0.03 ($N = 32$), 0.04 ($N = 64$), and 0.05 ($N = 128$ and 256). We also analyzed simulations based on seven large ($N = 517 - 640$) phylogenies downloaded from TreeBase (http://www.treebase.org). The rate of false positives remained well controlled ($0.047 - 0.053$) at a nominal p-value of 0.05, suggesting that further increasing the number of taxa does not lead to a degradation of Type I error rates.

A further analysis using 36 trees from a variety of published studies downloaded from TreeBase, to simulate 10 replicates from each tree (see Text S1 and Tables S1 and S2 for details), revealed that MEME is generally conservative for alignments of with low pairwise divergence (e.g. < 0.1 nucleotide substitutions per site), nominal for those with medium to high pairwise divergence ($0.1 - 0.4$ nucleotide substitutions per site), and nominal to slightly anti-conservative for higher pairwise divergence (> 0.4 nucleotide substitutions per site), although this relationship is influenced by other factors. Overall, we conclude that false positive rates of MEME, are well controlled in the setting of the most pessimistic (strict neutral) null.

Constant selection pressure at individual sites. At nominal $p = 0.05$ MEME consistently tracked FEL on sequence alignments simulated under the lineage-constant model assumed by FEL (Table S3), losing several percentage points of power because of its more conservative test statistic. Because each simulated alignment contained a subset of sites generated under the null (neutral model), we could derive empirical estimates of the size of the test and set the nominal p-value to achieve a Type I error rate of 5%. When calibrated to deliver a 5% Type I error rate, MEME held a small edge in power. This finding is not surprising, because at a fixed Type I rate, MEME should find every site found by FEL, and resolve FEL borderline cases affected by stochastic variation in ω throughout the tree.

Variable selection pressure at individual sites. The difference in power between MEME and FEL became stark when selection at individual sites varied among lineages, with each branch evolving under positive selection (ω^{+}) with probability q^{+}, and negative selection (ω^{-}) with complimentary probability $1 - q^{+}$. For every combination of independent simulation parameters ($\omega^{-}, \omega^{+}, q^{+}$), MEME had more power to detect sites under episodic diversifying selection (Table 1). Both methods gained power with an increasing proportion of positively selected lineages and/or a greater degree of diversification. The largest differences between MEME and FEL were observed when a small proportion of lineages ($q^{+} = 0.1$) were subjected to diversifying selection. Regardless of the strength of background purifying selection, FEL was effectively powerless (power $0 - 10\%$) to detect episodes of positive selection under any of the three phylogenetic simulation scenarios, whereas MEME achieved low ($4 - 53\%$ when $\omega^{+} = 4$), modest ($15 - 95\%$ when $\omega^{+} = 12$), and excellent ($37 - 100\%$ when $\omega^{+} = 36$) power. Under these conditions, the power of MEME increased with the alignment size, whereas the power of FEL remained very low. Although FEL gained appreciable power when 25% (or 50%) of the lineages were subject to diversification, its power was on average only $\approx 24\%$ ($\approx 67\%$) of that realized by MEME.

Taken together, the constant and variable selection pressure simulations demonstrate the uniform superiority of MEME over a standard test for diversifying positive selection. MEME has well controlled rates of false positives, has power comparable to FEL when selective forces are uniform at individual sites, and gains a large power advantage when these forces are variable, as is undoubtedly the case in most biological data sets.

Power and accuracy of the empirical Bayes procedure to identify branches subject to diversifying selection at a single site. Our exploratory simulations (see Figure S2) suggest that it is difficult to accurately identify individual positively selected branches at an individual site. We restricted the analysis to only those sites, which were found to be under episodic diversifying selection by MEME ($p \leq 0.05$) and set the threshold of 20 for the empirical Bayes factor to call an individual branch selected. The best results are achieved when selected branches are placed in the background of strongly conserved lineages ($\omega = 0.1$) – an individual branch is correctly detected in approximately 25% of cases, while *at least* one selected branch is found in 89.8% of cases (see Figure S3). However, while none of the negatively selected background branches are reported in more than 5% of cases, in

Table 1. Comparative performance of FEL and MEME on simulated data where ω varies along phylogenetic lineages.

		Japanese encephalitis virus *env*			Vertebrate rhodopsin			Camelid VHH		
ω^-	q^+	**$\omega^+=4$**	**$\omega^+=12$**	**$\omega^+=36$**	**$\omega^+=4$**	**$\omega^+=12$**	**$\omega^+=36$**	**$\omega^+=4$**	**$\omega^+=12$**	**$\omega^+=36$**
0	0.1	0.00 **0.06**	0.01 **0.25**	0.03 **0.50**	0.00 **0.21**	0.00 **0.53**	0.02 **0.81**	0.00 **0.53**	0.00 **0.95**	0.04 **0.99**
0	0.25	0.01 **0.12**	0.06 **0.32**	0.12 **0.51**	0.01 **0.30**	0.04 **0.68**	0.15 **0.88**	0.00 **0.66**	0.14 **0.98**	0.56 **1.00**
0	0.5	0.06 **0.12**	0.19 **0.29**	0.34 **0.45**	0.09 **0.28**	0.34 **0.59**	0.54 **0.82**	0.23 **0.77**	0.85 **0.98**	0.96 **0.98**
0.2	0.1	0.00 **0.05**	0.01 **0.21**	0.02 **0.41**	0.00 **0.09**	0.01 **0.35**	0.02 **0.67**	0.00 **0.16**	0.01 **0.87**	0.04 **0.98**
0.2	0.25	0.02 **0.08**	0.07 **0.27**	0.14 **0.48**	0.03 **0.17**	0.09 **0.55**	0.17 **0.84**	0.01 **0.42**	0.27 **0.96**	0.62 **0.99**
0.2	0.5	0.05 **0.11**	0.18 **0.29**	0.36 **0.49**	0.13 **0.25**	0.36 **0.60**	0.55 **0.76**	0.30 **0.72**	0.84 **0.99**	0.90 **0.99**
0.4	0.1	0.00 **0.04**	0.01 **0.15**	0.03 **0.37**	0.01 **0.07**	0.02 **0.30**	0.03 **0.57**	0.01 **0.10**	0.04 **0.78**	0.10 **0.97**
0.4	0.25	0.02 **0.06**	0.09 **0.27**	0.15 **0.45**	0.04 **0.16**	0.09 **0.49**	0.21 **0.78**	0.03 **0.32**	0.33 **0.97**	0.63 **0.99**
0.4	0.5	0.07 **0.10**	0.17 **0.26**	0.33 **0.46**	0.17 **0.28**	0.39 **0.58**	0.51 **0.76**	0.40 **0.62**	0.82 **0.94**	0.96 **1.00**

Power to detect sites under selection ($p=0.05$) are reported for FEL and MEME (in **boldface**) for each unique combination of negative selection (ω^-), positive selection (ω^+), and proportion of branches under positive selection (q^+) parameters.

doi:10.1371/journal.pgen.1002764.t001

55% of cases *at least* one background branch was falsely detected as positively selected. In a more difficult case of neutrally evolving background, the EB procedure performs considerably worse: at least one select branch is found in 55.6% of cases, whereas at least one background branch is detected in 86.5% instances. 18 background neutral branches are reported as selected at over 5% frequency, while the 5 positively selected branches are identified at $3.4-26\%$ of selected sites.

Empirical data

To gauge the comparative performance of MEME and FEL when identifying sites subject to pervasive diversifying selection, we used a collection of 16 protein-coding alignments, representing a diverse array of taxa, genes subject to differing levels of conservation, and a range of data set sizes (Table 2). In 12/16 alignments analyzed, MEME identified all the sites inferred by FEL to be under diversifying positive selection and found between

Table 2. Comparative performance of MEME and FEL on 16 empirical alignments (see Results and Text S1 for an extended discussion of each individual case).

Data set	N	S	Mean	Classes of sites detected at $p \leq 0.05$				Mean q^+		Sites where
			Div.	M^+F^0	M^+F^+	M^+F^-	M^-F^+	M^+F^{0-}	M^+F^+	MEME>FEL at $p=0.05$
Abalone sperm lysin	25	134	0.43	17	9	0	1 (0.04/0.05)	0.17	0.35	19
Camelid VHH	212	96	0.27	22	6	2	0 (n/a)	0.11	0.50	26
Diatom SIT	97	300	0.54	12	0	36	0 (n/a)	0.05	n/a	82
Drosophila *adh*	23	254	0.26	9	1	0	0 (n/a)	0.09	0.19	7
Echinoderm H3	37	111	0.33	0	0	1	0 (n/a)	0.02	n/a	3
Flavivirus NS5	18	342	0.48	3	0	1	0 (n/a)	0.16	n/a	7
Hepatitis D virus Ag	33	196	0.29	13	7	0	1 (0.05/0.07)	0.08	0.37	10
HIV-1 *rt*	476	335	0.08	12	10	7	0 (n/a)	0.04	0.69	27
HIV-1 *vif*	29	192	0.08	5	2	0	7 (0.04/0.06)	0.11	0.59	3
IAV H3N2 HA	349	329	0.04	7	11	2	3 (0.04/0.06)	0.04	0.73	8
JEV *env*	23	500	0.13	2	1	1	0 (n/a)	0.11	1.00	3
Mamallian β-globin	17	144	0.38	10	2	0	0 (n/a)	0.20	0.31	11
Primate *COXI*	21	510	0.36	3	0	1	0 (n/a)	0.18	n/a	4
Salmonella *recA*	42	353	0.04	1	0	0	0 (n/a)	0.02	n/a	0
Vertebrate rhodopsin	38	330	0.34	13	1	5	0 (n/a)	0.11	0.74	39
West Nile virus NS3	19	619	0.13	1	1	0	0 (n/a)	0.04	1.00	2
Total/Mean				130	51	56	12	0.10	0.59	

N (S) reports the number of sequences (codons) in the alignment. M^+ (M^-) refers sites found by MEME to be positively (negatively) selected ($p \leq 0.05$). F^+ (F^-) denote sites found by FEL to be positively (negatively) selected ($p \leq 0.05$). F^0 references sites that are classified as neutrally evolving by FEL. Values in parentheses for the M^-F^+ column show the mean p-values for FEL and MEME on this set of sites, respectively. Values reported in the rightmost column count the number of sites where MEME fits significantly better than FEL, based on a 2-degrees of freedom LRT ($p \leq 0.05$). Abbreviations: IAV = Influenza A virus, JEV = Japanese encephalitis virus.

doi:10.1371/journal.pgen.1002764.t002

1 (e.g. West Nile virus NS3) and 48 (Diatom SIT) additional sites that were subject to episodic diversifying selection (Table 2). In four data sets, $1-7$ sites identified by FEL with p-values close to 0.05 were missed by MEME. Note that MEME p-values for these sites remained in the $0.05-0.07$ range (Table 2), i.e. marginally significant.

Sites identified by both methods tended to have a greater average proportion of lineages under selection (0.59, measured by the mean of MLE estimates of q^+); sites found only by MEME experienced more episodic selection (0.10). In 9 data sets (Table 2), sites that FEL inferred to be under purifying selection are instead identified by MEME as likely to have been subjected to episodic diversifying selection. Almost universally (Tables S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19), such sites had a smaller estimated proportion of positively selected lineages ($<10\%$). This behavior is consistent with the relative performance of the two tests on simulated data and corroborates the expectation that MEME has greater power to identify sites when only a proportion of lineages evolved under positive selection. Vertebrate rhodopsin, Japanese encephalitis virus *env*, and Camelid VHH are investigated in detail below; for a discussion other genes, see Text S1.

Vertebrate rhodopsin

The vertebrate rhodopsin (a low-light vision protein) data set was previously experimentally investigated for the substitutions that modulate the wavelength of the light absorbed by the molecule (λ_{max}, [18]). The authors asserted that, because none of the 12 sites that they had determined as affecting λ_{max} by site-directed mutagenesis were detected by site-level computational methods, "statistical tests of positive selection can be misleading without experimental support." Other authors reanalyzed the same data set more comprehensively and went even further, questioning the utility of ω-based methods for detecting experimentally validated sites, because "most of the current statistical methods are designed to identify codon sites with high ω values, which may not have anything to do with functional changes. The codon sites showing functional changes generally do not show a high ω value" [29]. The validity of this generalization has been correctly questioned with a simple counter-argument that the sites detected by computational methods may also be functionally important, because the change in λ_{max} is unlikely to be the sole determinant of adaptation [17].

The MEME analysis of this gene suggests another obvious alternative, also expounded by previous studies [17]: the failure of the original computational analysis [18] to identify functionally important sites results from the fact that these sites have been subjected to episodic selection, which is masked by predominantly purifying selection elsewhere in the tree. Indeed, among three sites that alter λ_{max} found by MEME (96, 183 and 195, versus none found by FEL), no more than 13% of the branches exhibited $\omega > 1$ (Table S17); at these sites, the average ω is less than 1. We note that, because adaptive evolution will not always adhere to a single, simple scenario of episodic diversifying selection, we do not expect MEME to find all 12 sites experimentally confirmed to alter λ_{max}. For example, three of the nine missed sites (83,194,292) appear to have been subjected to partial selective sweeps and have been detected using a specialized model of directional evolution [29].

Three sites from this alignment can be used to illustrate how the inclusion of lineage variability modifies inference of selection (Figure 2). Site 54 was inferred to have experienced pervasive non-synonymous substitutions throughout its evolutionary history. Both FEL and MEME detect this site as positively selected ($p = 0.02$). Sixty three percent of the lineages at this site evolved with $\beta^+ > \alpha$, whereas the remainder were conserved ($\alpha = \beta^- = 0$), according to MEME. The log-likelihood of the site is only marginally higher for MEME, which suggests that MEME behaves like FEL at sites with

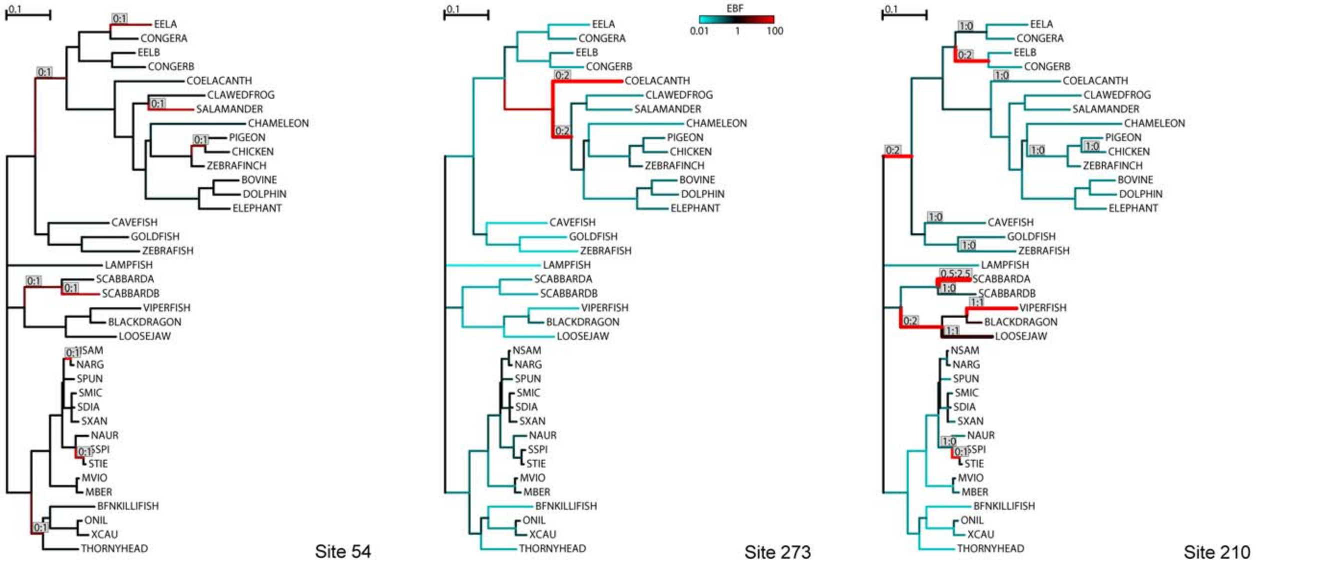

Figure 2. Individual sites of the vertebrate rhodopsin alignment used to illustrate similarities and differences between FEL and MEME. Branches that have experienced substitutions, based on most likely joint maximum likelihood ancestral reconstructions at a given site, are labeled as count of synonymous substitutions:count of non-synonymous substitutions. The thickness of each branch is proportional to the minimal number of single nucleotide substitutions mapped to the branch. Branches are colored according to the magnitude of the empirical Bayes factor (EBF) for the event of positive selection: red – evidence for positive selection, teal – evidence for neutral evolution or negative selection, black –Ê no information. See Methods for more detail. All three sites were identified as experiencing positive diversifying selection by MEME. FEL reported site 54 as positively selected, site 273 as neutral, and site 210 as negatively selected.
doi:10.1371/journal.pgen.1002764.g002

"canonical" patterns of diversifying selection, corroborating the simulation results.

At codon 273, FEL obtained a maximum likelihood estimate of $\beta > \alpha$, but failed to infer positive selection, as the signal was not statistically significant ($p = 0.70$). MEME, on the other hand, allocated 0.04 (0.013–0.10: 95% confidence interval obtained by latin hypercube sampling importance resampling [30]) of branches to a rate class with $\alpha = 0.0, \beta^+ = 9.49$ (2.94–6726) and inferred positive selection ($p = 0.03$). The difference in log-likelihoods between MEME and FEL is 4.9 points: MEME fits significantly better, based on a 2-degrees of freedom likelihood ratio test ($p = 0.007$). The maximum likelihood estimates of individual model parameters have large associated errors (although in all posterior samples we obtained $\beta^+ > \alpha$), as is expected for inference based on a single site. This has also been noted by Yang and dos Reis [17]. The point estimates themselves, however, are immaterial for inferring whether or not a site is positively selected, since the likelihood ratio test is used for that purpose.

Perhaps the most dramatic example of the added power of MEME is illustrated by site 210. At this site, the evolutionary history is replete with non-synonymous substitutions along deep lineages followed by extensive synonymous evolution, indicative of purifying selection. There is also a small clade with repeated synonymous and nonsynonymous substitutions. Averaging over all branches, FEL determined that the site, overall, is under negative selection ($p = 0.01$). MEME reported that 89% of the branches were under a very strong selective constraint ($\alpha = 2.13, \beta^- = 0.0$), but that the remaining $11\%(5.5 - 18.6\%)$ were under strong diversifying selection ($\beta^+ = 26.5(10.1 - 6519)$). The log-likelihood improvement is now 13.4 at the cost of two parameters, which is highly significant ($p < 0.001$). Site 210 is the ideal illustration of why it is undesirable to average ω over all lineages: bursts of diversification followed by conservation will most likely be missed by traditional site methods.

Japanese encephalitis virus *env*

No evidence for selection was found in this envelope gene in previous analyses [28], and FEL found only one site under positive selection. Despite the low levels of divergence among a relatively small number of taxa (23 isolates), MEME found episodic selection at sites called negatively selected by FEL (Table S12). Two of these sites fall within a beta-barrel epitope known to be involved in escape from neutralizing antibodies [31]. Sites 33 and 242 showed evidence of repeated toggling at terminal lineages. Remarkably, site 33 – likely a part of a neutralizing antibody epitope [32] – changed from isoleucine to leucine on 6 terminal lineages; site 242 changed from phenylalanine to serine on 5 terminal lineages. These substitutions co-occur on three terminal lineages. Evidence of recombination was detected in this alignment, and corrected for using a partitioning approach (details on how MEME can correct for recombination are in Text S1).

Camelid VHH

The camelid VHH data set comprises partial variable domain sequences (germline alleles) of llama and dromedary heavy chain only antibodies (Table S3). 11 of 16 sites in the variable complementarity determining regions (CDR) 1 (sites 26–33) and 2 (sites 51–58) were found to be under diversifying selection by MEME (2/16 were detected by FEL and 2 more were marginally significant). Because CDR regions are driven to diversify in order to provide a broad basis of antigen recognition, positive selection is expected to be commonplace in the CDRs [33]. MEME was able to uncover selective signatures at a majority of those sites. Of the remaining 19 sites classified by MEME as positively selected, six were associated with VHH family differentiation [34]. Unlike standard antibodies, which must maintain relatively conserved framework regions (FR) involved in binding heavy and light chains to form functional tetramers, VHH antibodies are free of such functional constraints. A previous analysis of camelid VHH for evidence of positive selection using counting methods [35] reported evidence for positive selection at a single site (14) in FR1 (sites 1–25 in Table S3), but this analysis could find no clear evidence of positive or negative selection at 49 FR sites. In contrast, MEME inferred episodic selection at six sites in FR1, six sites in FR2 (sites 34–50), and 7 sites in FR3 (sites $59 - 96$). The well-known lack of power of counting methods to detect even pervasive selection [17] likely hampered the previous study.

Effect of sequence sampling

Although a previous analysis of 38 vertebrate rhodopsin sequences found no sites under selection at posterior probability $\geq 95\%$ [18], the same authors found 7 selected sites in the subset of 11 squirrelfish sequences, and 2 selected sites when the subset of 28 fish sequences was analyzed. These results run counter to the expectation that more data should provide greater power to detect selection. MEME, on the other hand, detects more selected sites when more sequences are included. One site is identified in the squirrelfish alignment, 9 in the fish alignment, and 19 in the complete rhodopsin alignment. All but 5 sites detected in the subset alignments are also identified in the full alignment (Table S20). Allowing ω to vary over branches at least partially mitigates the pathology of constant-ω models which effectively rely on an average ω for inferring selection. A similar pattern is seen in the analysis of the influenza A virus H3N2 hemagglutinin sequences, where site-level methods also appear to be sensitive to sequence sampling ([19], see Text S1 and Table 23).

Discussion

We have presented a mixed effects model of evolution, MEME, and a statistical test for detecting the signal of past episodic positive selection from molecular sequence data. Our model corrects the biologically unrealistic assumption that selective pressure, as measured by the ω ratio, remains constant over lineages. Based on comprehensive simulations and empirical analysis of an array of taxonomically diverse genes, MEME can be recommended as a replacement for existing site models. MEME matches the performance of older approaches when natural selection is pervasive, but possesses greater power to identify sites where episodes of positive selection are confined to a small subset of branches in a phylogenetic tree.

Our results suggest that it may be necessary to revise previous estimates of the proportion of sites under positive selection in many genes. Using the FEL method, which assumes constant selective pressure at a site, we are able to detect 63 sites across all 16 empirical alignments. MEME identifies 51 of these sites (the remaining 12 are borderline significant) and 186 additional sites – nearly 4 times as many as FEL. For individual data sets (e.g. Drosophila *adh* and Diatom SIT, Table 2), the differences may be even more dramatic. The greater power of MEME indicates that selection acting at individual sites is considerably more widespread than constant ω models would suggest. It also suggests that natural selection is predominantly episodic, with transient periods of adaptive evolution masked by the prevalence of purifying or neutral selection on other branches. We emphasize that MEME is not just a quantitative improvement over existing models: for 56 sites in our empirical analyses, we obtain qualitatively different conclusions. FEL asserts that these sites evolved under significant purifying selection, but MEME is able to identify the signature of positive selection on some branches. Furthermore, MEME is less sensitive to sampling effects that plague existing positive selection detection tools [18,19]. Variable levels of purifying selection pressure across different lineages prevented these older methods from detecting instances of episodic positive selection; MEME is able to peer through the fog of purifying selection.

It is important to bear in mind that the mixture χ^2 statistic used to calculate the p-values reported here is based on a null model under which all sites are evolving neutrally. This, however, is not biologically realistic: the null hypothesis against which sites ideally ought to be screened is one under which sites are evolving *either* neutrally *or* under purifying selection. But the proportion of sites evolving under negative selection and the strength of this selection are unknown and vary from case to case, which means that such a null hypothesis would be very sensitive to modeling assumptions that cannot be justified in general. Instead, the neutral null hypothesis represents a worst case scenario for our inference, so that the p-values we obtain are upper bounds of the true p-values. This ensures that our inference is conservative. Even in the worst (and biologically unrealistic) case for MEME, namely when selective pressures are constant throughout the phylogeny, the loss of power compared to FEL is minimal: a site with FEL p-values between 0.0346 and 0.05 will be missed by MEME, since its p-value will be >0.05 for the same ranges of the likelihood ratio test statistic (LRT). In our simulation scenarios under the assumption of constant ω, this translates to no more a 5% loss in power (Table S3).

Our inference is performed in a site-wise rather than an alignment-wide manner, and we therefore control the site-wise rather than the family-wise error rate. We do not recommend combining the results of multiple site-wise inferences to perform alignment-wide inference. To aid interpretation of the results while taking account of multiple testing, we calculate the false discovery rate [36]; the resulting q-value upper bounds are reported alongside their corresponding p-value upper bounds in Tables S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19. This gives an upper bound on how many of the reported sites can be expected to be false discoveries: for instance, of the 30 sites reported in Table S5 we expect no more than 5 (14%) to be false, and probably far fewer because of the conservative choice of null model. We emphasize that q-values are usually much larger than their corresponding p-values and caution that p-values (regardless of whether they have been corrected for multiple testing) cannot be used to estimate an expected number of false discoveries in the same way.

MEME is a conceptual advance over the first generation of random effects models designed to detect episodic selection (called "branch-site models" in the literature [17]). MEME does not require *a priori* designation of, or an exhaustive search for, the branches under selection, and it allows each site to have its own selective history. Whereas branch-site models make restrictive *a priori* assumptions about how ω values are distributed across the tree – sometimes leading to very poor statistical performance [20] – MEME treats the selective class on each branch as a random effect that is marginalized over in the likelihood calculation.

For computational tractability, MEME assumes that the value taken by ω on each branch is independent of that on any other branch, i.e. selective pressures between branches are uncorrelated. This assumption could potentially be violated: for example, if ω changes very slowly across the phylogeny, then ω values on neighboring branches will be correlated. Further research is needed to understand how inference of selection would be affected if these correlations were directly accounted for, and whether the additional model and computational complexity would be justified. In practice, MEME could be combined with models of directional selection to improve power, e.g. [15,16]. Unlike covarion models [37,13], MEME does not allow ω to change in the middle of a tree branch. The effect of this restriction is unclear, but it could be tested by implementing a mixed effects covarion model, where switching rates and proportion of time spent under $\omega > 1$ are estimated at an individual site.

The ability of MEME, or similar substitution model-based methods, to accurately infer the identity of individual branches subject to diversifying selection at a given site seems unavoidably limited. Most of the information that such inference might be based on is limited to character substitutions along a single branch at a single site, i.e. one realization of the Markov substitution process. Selection along terminal branches in the context of negatively selected background can be detected more reliably than selection along interior branches among neutrally evolving background lineages. However, we caution that despite obvious interest in identifying specific branch-site combinations subject to diversifying selection, such inference is based on very limited data (the evolution of one codon along one branch), and cannot be recommended for purposes other than data exploration and result visualization. This observation could be codified as the "selection inference uncertainty principle" – one cannot simultaneously infer both the site and the branch subject to diversifying selection. In this manuscript, we describe how to infer the location of sites, pooling information over branches; previously [20] we have outlined a complementary approach to find selected branches by pooling information over sites.

Finally, although MEME is considerably more powerful than existing methods at detecting bursts of selection, it still requires that a measurable proportion of lineages ($5-10\%$) experience non-synonymous evolution at a site. When a single substitution modifies an adaptive trait and is subsequently fixed, we expect ω based methods to have very little power. Specialized methods which make use of change in allele frequencies [15,16], or between and within-population diversification patterns [38], will be required in such cases.

Supporting Information

Figure S1 Quantile–Quantile plot of three asymptotic distributions (x-axis) for the MEME LRT test versus the LRT derived by parametric bootstrap (y-axis), limited to the meaningful test p-value range of <0.01. The χ^2_1 distribution is too liberal (lying below the $x=y$ line), the χ^2_2 is too conservative, while the mixture is approximately correct.
(PDF)

Figure S2 Simulation parameters for generating datasets for evaluating the empirical Bayes inference of branch-site combinations under selection. Branches are colored according the the value of ω used to evolve sequences along them; branches simulated under positive selection are also labeled with ω values.
(PDF)

Figure S3 Summary of empirical Bayes inference of branches under selection on data simulated using the selective parameters from Figure S2. Each branch is colored according to the proportion of times it was found to have an empirical Bayes factor of 20 or greater at sites with MEME p-value of 0.05 or less. Branches with $>5\%$ detection rates are also labeled with the values of the rates.
(PDF)

Table S1 False positive rates for data sets simulated under strict neutrality using empirical trees from TreeBase. The entries are sorted in order of increasing mean false positive rate derived from simulated data (10 replicates per tree). Mean divergence between any pair of leaves in a given tree is reported in expected nucleotide substitutions per site. False positive range reports the minimum and maximum values for false positive rates for an individual replicate. 95% confidence intervals are derived from the binomial distribution with the probability of success $p = 0.05$, and the number of trials N equal to the number of codons. This range provides the expected spread of per replicate false positive rates for a test that has the probability of making a false positive error of exactly 0.05 over N tests.
(PDF)

Table S2 False positive rates for three empirical trees from TreeBase when the parameters of the null model are varied: 20% of the branches are simulated with the foreground ω, and the remainder under the background ω. 10 replicates with 300 codons each per tree-ω pair were simulated. The synonymous rate was set to 0.52 for the first 150 codons, 0.9 for the next 100 codons, and 1.58 for the last 50 codons.
(PDF)

Table S3 Comparative performance of FEL and MEME on simulated data where ω does not vary among tree branches. The rate of false positives (FP) and power are reported for a fixed nominal test p-value of 0.05. Power is also shown for the p-value that achieves FP of 0.05, estimated empirically from the distribution of p-values on the subset of sites evolving neutrally.
(PDF)

Table S4 Positively selected sites in abalone sperm lysin. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S5 Positively selected sites in camelid VHH. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S6 Positively selected sites in Diatom silicon transporters found by MEME at $p \leq 0.05$. The FEL result column summarizes the classification obtained by FEL. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S7 Positively selected sites in Drosophila *adh* found by MEME at $p \leq 0.05$. The FEL result column summarizes the classification obtained by FEL. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S8 Positively selected sites in Echinoderm histone H3. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S9 Positively selected sites in Flavivirus NS5. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S10 Positively selected sites in Hepatitis D virus Ag. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S11 Positively selected sites in HIV-1 reverse transcriptase (*rt*). $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S12 Positively selected sites in HIV-1 viral infectivity factor (*vif*). $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S13 Positively selected sites in Influenza A virus hemagglutinin (H3N2 serotype). Superscript letters after the site indicate the epitope in which substitutions can affect phenotype. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S14 Positively selected sites in Japanese encephalitis virus *env*. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S15 Positively selected sites in mammalian β-globin. The FEL result column summarizes the classification obtained by FEL. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S16 Positively selected sites in primate cytochrome c oxidase subunit 1 (*COX1*). $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S17 Positively selected sites in Salmonella *recA*. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites

(FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S18 Positively selected sites in vertebrate rhodopsin. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S19 Positively selected sites in West Nile virus NS3. $+$ stands for a positively selected site and $-$ stands for a negatively selected site (FEL $p > 0.1$). $++$ and $--$ reflect borderline significant sites (FEL p between 0.05 and 0.1). $+++$ and $---$ denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S20 Test p-values for positively selected sites found by MEME in a set of 38 vertebrate rhodopsin sequences analyzed with REL methods in Yokoyama2008fk. Sites with $p \leq 0.05$ are shown in bold. The partial ordering of subsets is as follows: Squirrelfish $\subset$ Fish $\subset$ All, Coelacanth and tetrapods $\subset$ All. Sites found to be under positive selection with posterior probability of $>95\%$ (M8 model) in Yokoyama2008fk in at least one of the subsets are marked with *.
(PDF)

Table S21 Test p-values for positively selected sites found by MEME in a set of 86 influenza A virus hemagglutinin sequences (Set 3) and its various subsets, analyzed with REL methods in Chen2011fk. Sites with $p \leq 0.05$ are shown in bold. The partial ordering of subsets is as follows: Set 4 $\subset$ Set 1 $\subset$ Set 3, Set 5 $\subset$ Set 2 $\subset$ Set 3, Set 6 $\subset$ Set 3, Set 7 $\subset$ Set 3. Sites found to be under positive selection with posterior probability of $>95\%$ (M3 model) in Chen2011fk in at least one of the subsets are marked with *.
(PDF)

Text S1 Supplementary methods, results, and discussion.
(PDF)

Author Contributions

Conceived and designed the experiments: BM KS SLKP. Performed the experiments: BM JOW SM TW SLKP. Analyzed the data: BM JOW SM TW SLKP. Contributed reagents/materials/analysis tools: BM SLKP. Wrote the paper: BM JOW KS SLKP.

References

1. 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.
2. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–36.
3. Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. Brief Bioinform 10: 97–109.
4. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26: 255–271.
5. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167–70.
6. Bonhoeffer S, Holmes EC, Nowak MA (1995) Causes of HIV diversity. Nature 376: 125.
7. Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. Nature 385: 151–4.
8. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–6.
9. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170. doi:10.1371/journal.pbio.0030170
10. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–36.
11. 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: 2832–7.
12. Brault AC, Huang CYH, Langevin SA, Kinney RM, Bowen RA, et al. (2007) A single positively selected West Nile viral mutation confers increased virogenesis in American crows. Nat Genet 39: 1162–6.
13. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. Proceedings of the National Academy of Sciences of the United States of America 101: 12957–12962.
14. Delport W, Scheffler K, Seoighe C (2008) Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog 4: e1000242. doi:10.1371/journal.ppat.1000242
15. Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, et al. (2007) A model of directional selection applied to the evolution of drug resistance in HIV-1. Mol Biol Evol 24: 1025–31.
16. 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 inuenza A virus. Mol Biol Evol 25: 1809–24.
17. Yang Z, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28: 1217–28.
18. Yokoyama S, Tada T, Zhang H, Britt L (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci U S A 105: 13480–5.
19. Chen J, Sun Y (2011) Variation in the analysis of positively selected sites using nonsynonymous/synonymous rate ratios: an example using inuenza virus. PLoS ONE 6: e19996. doi:10.1371/journal.pone.0019996
20. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, et al. (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28: 3033–3043.
21. Kosakovsky Pond S, Delport W, Muse SV, Scheffler K (2010) Correcting the bias of empirical frequency parameter estimators in codon models. PLoS ONE 5: e11230. doi:10.1371/journal.pone.0011230
22. Felsenstein J (1981) Evolutionary trees from DNA-sequences – a maximum-likelihood approach. J Mol Evol 17: 368–376.
23. Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human inuenza virus A. Journal of Molecular Evolution 51: 423–432.
24. 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.
25. Wertheim JO, Kosakovsky Pond SL (2011) Purifying selection can obscure the ancient age of viral lineages. Mol Biol Evol.
26. Pond SK, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22: 2375–85.
27. Self SG, Liang KY (1987) Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. J Am Stat Assoc 82: 605–310.
28. Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155: 431–449.
29. Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A 106: 6700–5.
30. Pond SLK, Scheffler K, Gravenor MB, Poon AFY, Frost SDW (2010) Evolutionary fingerprinting of genes. Mol Biol Evol 27: 520–36.
31. Wu KP, Wu CW, Tsao YP, Kuo TW, Lou YC, et al. (2003) Structural basis of a avivirus recognized by its neutralizing antibody: solution structure of the domain III of the Japanese encephalitis virus envelope protein. J Biol Chem 278: 46007–13.
32. Gangwar RS, Shil P, Cherian SS, Gore MM (2011) Delineation of an epitope on domain I of Japanese encephalitis virus Envelope glycoprotein using monoclonal antibodies. Virus Res 158: 179–87.
33. Tanaka T, Nei M (1989) Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol 6: 447–59.
34. Harmsen M, Ruuls R, Nijman I, Niewold T, Frenken L, et al. (2000) Llama heavy-chain V regions consist of at least four distinct subfamilies revealing novel sequence features. Molecular Immunology 37: 579–590.
35. Su C, Nguyen VK, Nei M (2002) Adaptive evolution of variable region genes encoding an unusual type of immunoglobulin in camelids. Mol Biol Evol 19: 205–15.
36. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57: pp. 289–300.
37. Tuffley C, Steel M (1998) Modeling the covarion hypothesis of nucleotide substitution. Mathematical biosciences 147: 63–91.
38. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–4.