A counting renaissance: Combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection

Philippe Lemey 1,*, Vladimir N. Minin 2, Filip Bielejec 1, Sergei L. Kosakovsky Pond 3, and Marc A. Suchard 4,5,∗

1 Department of Microbiology and Immunology, Rega Institute, KU Leuven, B-3000 Leuven, Belgium.
2 Department of Statistics, University of Washington Seattle, WA, 98195, USA.
3 Department of Medicine, University of California, San Diego, San Diego, CA, 92103, USA.
4 Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, 90095, USA.
5 Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA, 90095, USA.

ABSTRACT
Motivation: Statistical methods for comparing relative rates of synonymous and nonsynonymous substitutions maintain a central role in detecting positive selection. To identify selection, researchers often estimate the ratio of these relative rates (dN/dS) at individual alignment sites. Fitting a codon substitution model that captures heterogeneity in dN/dS across sites provides a reliable way to perform such estimation, but it remains computationally prohibitive for massive data sets. By employing crude estimates of the numbers of synonymous and nonsynonymous substitutions at each site, counting approaches scale well to large data sets, but they fail to account for ancestral state reconstruction uncertainty and to provide site-specific dN/dS estimates.

Results: We propose a hybrid solution that borrows the computational strength of counting methods, but augments these methods with empirical Bayes modeling to produce a relatively fast and reliable method capable of estimating site-specific dN/dS values in large data sets. Importantly, our hybrid approach, set in a Bayesian framework, integrates over the posterior distribution of phylogenies and ancestral reconstructions to quantify uncertainty about site-specific dN/dS estimates. Simulations demonstrate that this method competes well with these more principled statistical procedures and in some cases even outperforms them. We illustrate the utility of our method using human immunodeficiency virus, feline panleukopenia and canine parvovirus evolution examples.

Availability: Renaissance counting is implemented in the development branch of BEAST, freely available at http://code.google.com/p/beast-mcmc/. The method will be made available in the next public release of the package, including support to set up analyses in BEAuTi.

Contact: philippe.lemey@rega.kuleuven.be

*to whom correspondence should be addressed

1 INTRODUCTION

Quantifying selective pressures on protein-coding genes is central to the goal of characterizing Darwinian processes in evolutionary biology. Among comparative and summary statistic approaches, the relative rate of silent and replacement substitutions represents one of the most popular measures to detect the molecular footprint of selection. Nonsynonymous mutations that offer fitness advantages are expected to become fixed at a higher rate than synonymous mutations, implying that a nonsynonymous/synonymous substitution rate ratio (ω = dN/dS) greater than one provides evidence for diversifying positive selection.

Although estimation of ω has led to the identification of positive selection in several systems (e.g., Hughes and Nei 1988; Messier and Stewart 1997), there are clear boundaries to the conditions under which it can reveal an unambiguous trace of molecular adaptation. First, an excess of nonsynonymous substitutions over synonymous substitutions is almost invariably restricted to a handful of amino acid sites responsible for adaptive evolution. Therefore, sensible ω estimates need to take into account the variation in selection intensity across codon sites (Nielsen and Yang, 1998). Moreover, while divergent sequences can yield considerable information to estimate nonsynonymous and synonymous substitution rates, the ratio of these rates may offer little insight when inferred from segregating polymorphisms within a single population (Kryazhimskiy and Plotkin, 2008). Finally, it is also important to distinguish between different selective regimes underlying molecular adaptation. Diversifying selection maintains amino-acid diversity at a given site and naturally results in elevated ω values, whereas directional selection may operate through a restricted number of amino-acid replacements which has less impact on ω but can lead to rapid fixation of a new allele in the population. The former is ubiquitous in antagonistic systems such as pathogen-host interactions (Yang and Bielawski, 2000). Not surprisingly,
\(d_S/d_S\) estimation methods have been frequently applied to viral gene sequences in order to detect escape from host immune responses or adaptation to novel hosts. Rapidly evolving viruses benefit from the ability to generate adaptive mutations de novo whereas other organisms must rely on pre-existing variation maintained by population structure or balancing selection (Pybus and Rambaut, 2009). The impact of immune pressure on the genetic diversity of viruses can be highly dependent on how potent the immune responses are to different viral variants, a concept that has been put forward within the framework of phylodynamics (Grenfell et al., 2004). For this reason, methods to estimate \(\omega\) have become an integral part of the ‘phylodynamic toolbox’ (Pybus and Rambaut, 2009), which generally refers to recent statistical and computational developments to simultaneously estimate epidemic and spatiotemporal dynamics from viral genes sequences (Lemey et al., 2009, 2010).

In this paper, we are interested in site-specific estimation of \(d_S/d_S\) ratios. There are two main classes of methods designed for this task. The first class prescribes to build a model of codon evolution, with the \(d_S/d_S\) ratio as a model parameter that is allowed to vary across sites (Nielsen and Yang, 1998; Pond and Muse, 2005). Although we fully endorse these approaches and use them whenever possible, codon-based models become computationally prohibitive when analyzing large data sets that include hundreds, if not thousands, of molecular sequences (Lemey et al., 2007). Because of this computational limitation, researchers often turn to a second class of methods that are based on imputing (counting) unobserved synonymous and nonsynonymous substitutions (Suzuki and Gojobori, 1999). These substitution counts are then used to construct a test statistic to assess the null hypothesis of neutrality (\(H_0 : d_S/d_S = 1\)). Such counting methods enjoy computational efficiency and often work well in practice (Kosakovsky Pond and Frost, 2005), but have two important shortcomings. First, counting approaches experience difficulties in handling uncertainty about the phylogenetic tree and about other nuisance parameters that play a role in counting unobserved synonymous and nonsynonymous substitutions. Zhai et al. (2007) propose to overcome this problem by combining stochastic mapping, introduced by Nielsen (2002), with traditional counting methods’ test statistics as discrepancy measures in a posterior predictive diagnostics framework (Gelman et al., 1996). However, posterior predictive diagnostics, undoubtedly useful for visual exploration of model fit, are difficult to calibrate and to use in a semi-automatic fashion. The second limitation, shared by all approaches, is their inability to provide reliable sites-specific \(d_S/d_S\) estimates and to quantify the relative strengths of selection across amino acid sites.

2 APPROACH

We propose a new counting method that overcomes both of the aforementioned shortcomings. Our solution arises in two steps. In the first step, we follow Nielsen (2002) and Zhai et al. (2007) and produce multiple stochastic mapping-based realizations of synonymous and nonsynonymous counts while integrating over the posterior distribution of the nuisance parameters including the phylogenetics tree. However, to gain computational tractability, we exploit nucleotide-based codon partition models (Yang, 1996) in this step. These models can be fit to data in a fraction of the time it takes to fit even the simplest codon-based evolutionary models first introduced by Muse and Gaut (Muse and Gaut, 1994) and Goldman and Yang (Goldman and Yang, 1994). Although codon partition models do not account for selective pressures at amino acid sites \(d_N/d_S\) is one for all sites under these models), emerging probabilistic counting procedures have been shown to be robust to such gross model misspecification (Minin and Suchard, 2008a; O’Brien et al., 2009; Minin et al., 2011).

The second step of our approach is the main novelty of this paper. We treat each random realization of synonymous and nonsynonymous counts as pseudo-data and shrink these site-specific counts toward their means over all sites. We accomplish this regularization through an empirical Bayes procedure. We then use these regularized synonymous and nonsynonymous counts to form site-specific \(d_S/d_S\) estimates. The end result is a posterior distribution of these ratios for all sites in the alignment. This distribution can be used for estimation (e.g., we report the posterior mean and 95% credible intervals (CIs) of site-specific \(d_S/d_S\) values) and for testing diversifying positive selection (e.g., a site is classified as positively selected if the posterior probability of \(d_S/d_S > 1\) at this site is at least, say, 0.95). Empirical Bayes and, similar in spirit, full hierarchical Bayesian methods have been developed before for site-specific \(d_S/d_S\) estimation and for identifying sites under diversifying positive selection (Huelsenbeck and Dyer, 2004; Yang et al., 2005). However, to our knowledge, we are the first to combine empirical Bayes philosophy and counting approaches, providing a revival of these simple methods to confront a formidable inferential problem. By performing the empirical Bayes regularization on the imputed count data, we avoid fitting computationally expensive codon-based models, making our ‘renaissance counting’ method practical for analyzing large data sets.

We assess our method using a simulation study almost identical to the one conducted and validated by Kosakovsky Pond and Frost (2005). The simulations demonstrate that our method’s site-specific \(d_S/d_S\) estimates and positively selected site identification are remarkably comparable to the estimates produced by the state-of-the-art codon-based models. We also apply our new renaissance approach to two empirical data sets. First, we re-analyze serially sampled feline panleukopenia and canine parvovirus VP2 capsid sequences. This analysis shows that renaissance counting produces sensible site-specific \(d_S/d_S\) estimates, simultaneously accounting for uncertainty in all model parameters, and highlights important differences between our new and codon-based models. We conclude by using our method in tandem with conventional approaches to examine differences in adaptive evolution in HIV pol genes in treated and drug naïve patient populations.

3 METHODS

Probabilistic counting under a codon partition model

Let \(y = \{y_{ij}\}\) be a codon alignment matrix, where \(i\) spans \(n\) sequences, \(l\) runs over \(L\) sites in the alignment, and each \(y_{ij}\) is one of 64 nucleotide triplets (codons). We assume that each site \(y_i\) follows a codon partition model, meaning that each nucleotide position \(s = 1, 2, 3\) within a codon evolves independently along a phylogenetic tree \(\tau\) according to one of the standard nucleotide substitution models (Yang, 1996). In all our examples and without loss of generality, we use an HKY model (Hasegawa et al., 1985) for each nucleotide position, with position-specific transition/transversion ratio \(r_s\), substitution rate \(\mu_s\), and stationary distribution \(\pi_s = (\pi_{1s}, \ldots, \pi_{Ts})\).

2
Position rates are normalized for identifiability for contemporaneously sampled sequences, but remain free parameters in the case of serially sampled data.

We fit the above codon partition model in a Bayesian framework and use Markov chain Monte Carlo (MCMC) integration to obtain a sample from the posterior distribution of model parameters \( \Pr(\theta \mid y) \), where \( \theta = (\tau, \kappa_1, \kappa_2, \kappa_3, \kappa_4, \tau_1, \tau_2, \pi_1, \pi_2, \pi_3) \). Let \( \lambda_j \) be parameter values at MCMC iteration \( j \). At each iteration \( j \), we use stochastic mapping to impute the full evolutionary history of each nucleotide position within each codon site in our alignment (Nielsen, 2002). We capitalize on a uniformization method, a modification of the original approach by Nielsen (2002), to draw realizations of the HKY continuous-time Markov chains (CTMCs) conditional on the codon data at the tips of the phylogeny \( \tau_j \) (Lartillot, 2006; Rodrigue et al., 2008). See Hobolth and Stone (2009) for an excellent review of different stochastic mapping algorithms. We also use simple forward CTMC simulation to draw the full evolutionary history of each site without conditioning on the data for normalization. For each stochastic mapping realization, conditional (C) and unconditional (U) on the data, we record the numbers of synonymous (S) substitutions \( C_{jl}^{(SC)} \), \( C_{jl}^{(SC)} \), \( C_{jl}^{(NC)} \), \( C_{jl}^{(NC)} \) and nonsynonymous (N) substitutions \( C_{jl}^{(NC)} \), \( C_{jl}^{(NC)} \) for each site \( l = 1, \ldots, L \). Computing the ratio \( (C_{jl}^{(NC)} / C_{jl}^{(SC)}) / (C_{jl}^{(NC)} / C_{jl}^{(SC)}) \) yields one straightforward way of converting these site-specific synonymous and nonsynonymous counts into site-specific \( dS/d\bar{d} \) ratios. However, this naive method produces highly unstable estimates due to the high variance of site-specific substitution counts. Moreover, in many cases some of the substitutions counts are zero, resulting in a \( dS/d\bar{d} \) estimate of 0 or \( \infty \). To circumvent these problems, we apply an empirical Bayes regularization procedure to each of the four substitution counts, producing regularized rate estimates \( \lambda_{jl}^{(SC)}, \lambda_{jl}^{(SC)}, \lambda_{jl}^{(NC)}, \lambda_{jl}^{(NC)} \), and \( \lambda_{jl}^{(NC)}, \lambda_{jl}^{(NC)} \), with the regularized rate estimates at hand, we form our \( dS/d\bar{d} \) estimates

\[
\omega_{jl}^{SC} = \left( \lambda_{jl}^{(SC)} / \lambda_{jl}^{(SC)} \right) / \left( \lambda_{jl}^{(NC)} / \lambda_{jl}^{(NC)} \right).
\]

We explain our empirical Bayes regularization procedure in detail in the next section.

**Regularized site-specific \( dS/d\bar{d} \) ratios**

Within each MCMC iteration, we view the realized values of synonymous and nonsynonymous substitutions as pseudo-data. Since our empirical Bayes regularization procedure is identical for all four types of imputed counts, we use the generic notation \( C_l \) to denote one of the four possible count types, \( C_{jl}^{(SC)}, C_{jl}^{(SC)}, C_{jl}^{(NC)}, C_{jl}^{(NC)} \), at site \( l \). Each of these counts represents the number of specifically labelled jumps of the codon model CTMC. In general, such random variables follow nonstandard distributions that reduce to a Poisson distribution only in special cases (Minin and Suchard, 2008b). However, in the phylogenetic context, a Poisson distribution approximates to a Poisson distribution only in special cases (Minin and Suchard, 2008b).

To execute our empirical Bayes approach, we start with sample means and variances

\[
\hat{\mu} = \frac{1}{L} \sum_{l=1}^{L} C_l \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{L} \sum_{l=1}^{L} (C_l - \hat{\mu})^2,
\]

and match these empirical quantities with the theoretical mean and variance of a gamma-Poisson distribution with parameters \( \alpha \) and \( \beta \). This method of moments produces the following hyperparameter estimates

\[
\hat{\alpha} = \frac{Q^2}{\hat{\sigma}^2 - \hat{\mu}} \quad \text{and} \quad \hat{\beta} = \frac{\hat{\mu}}{\hat{\sigma}^2 - \hat{\mu}}.
\]

Given these estimated values, we have

\[
\lambda_l \mid C ~ \text{Gamma}(C_l + \hat{\alpha}, 1 + \hat{\beta}),
\]

from which we arrive at empirical Bayes estimates

\[
\hat{\lambda}_l = \mathbb{E}(\lambda_l \mid C) = \left( \frac{C_l + \hat{\alpha}}{1 + \hat{\beta}} \right) \quad \text{if} \quad \hat{\sigma}^2 > \hat{\mu} \quad \text{and} \quad \hat{\mu} \quad \text{otherwise}.
\]

In summary, to obtain a sample from the posterior distribution of the \( dS/d\bar{d} \) ratios, we

1. Draw \( M \) samples from the posterior distribution of model parameters \( \Pr(\theta \mid y) \) under the simple nucleotide-based codon partition model, and then

2. Postprocess the MCMC output:

   a. For each iteration \( j = 1, \ldots, M \), impute the numbers of synonymous and nonsynonymous substitutions, \( C_{jl}^{(SC)}, C_{jl}^{(SC)}, C_{jl}^{(NC)}, C_{jl}^{(NC)} \), and \( C_{jl}^{(NC)} \), with the help of stochastic mapping.

   b. Use empirical Bayes regularization to arrive at rates of synonymous and nonsynonymous substitutions, \( \hat{\lambda}_{jl}^{(SC)}, \hat{\lambda}_{jl}^{(SC)}, \hat{\lambda}_{jl}^{(NC)}, \hat{\lambda}_{jl}^{(NC)} \), and \( \hat{\lambda}_{jl}^{(NC)} \), and finally

   c. Form site-specific \( dS/d\bar{d} \) ratios \( \omega_{jl}^{SC} \) using Equation (1).

**Computational considerations**

Our site-specific \( dS/d\bar{d} \) estimation method has two main steps. This first and the most time consuming step involves obtaining a posterior MCMC sample under the nucleotide-based codon partition model. An analogous step for the simplest codon models with no rate heterogeneity (Muse and Gaut, 1994; Goldman and Yang, 1994) takes at best 6/1/4/3/3 \( \approx 80 \) times as long, but in practice, performs worse since the computational work of matrix exponentiation for the codon model becomes non-negligible. In the second step, we use stochastic mapping coupled with empirical Bayes regularization. Our empirical Bayes regularization of the imputed substitution counts is accomplished almost instantaneously, because this estimation procedure does not involve computationally intensive calculations. The imputation step is potentially time consuming, but we draw stochastic mapping realizations for a small number of saved MCMC iterations \( (10^3 - 10^5) \), which is typically orders of magnitude smaller than the length of the entire Markov chain \( (10^6 - 10^7) \). Therefore, the running time of our two-step procedure is dominated by the first step, making our approach significantly faster than Bayesian site-specific \( dS/d\bar{d} \) estimation under codon-based models.

**Identifying sites under selection**

To classify sites as negatively selected, neutrally evolving or positively selected, we use for each site \( l \) the posterior sample \( \{\omega_{jl}^{SC}, \ldots, \omega_{jl}^{SC}\} \), to estimate the 95% CI \( (\omega_{jL}^{SC} - 2.5\% \omega_{jL}^{SC}, \omega_{jL}^{SC} + 2.5\% \omega_{jL}^{SC}) \). Alternatively, we could use a 95% highest posterior density region. We declare site \( l \) to be under diversifying positive selection if \( 1 < \omega_{jL}^{SC} < \omega_{jL}^{SC} + 2.5\% \omega_{jL}^{SC} \), neutral if \( \omega_{jL}^{SC} < 1 < \omega_{jL}^{SC} - 4\% \omega_{jL}^{SC} \), and under negative selection if \( 1 > \omega_{jL}^{SC} + 4\% \omega_{jL}^{SC} \). Notice that we can not say anything about frequentist properties of this classification procedure. In other words,
We conduct an extended simulation study to compare the relative CIs, including their relative rankings and the degree of overlap. To compare our method is not guaranteed to classify a site correctly 95% of the time during repeated replications of the evolutionary process. However, we point out that our new method offers more than merely a classification algorithm. In addition to classifying sites into the three categories above, we can assess the relative strengths of selection at individual sites by comparing site-specific CIs, including their relative rankings and the degree of overlap. To compare the use of a per-site \( \frac{dS}{dN} \) posterior distribution with a frequentist counting approach, we also explore a test procedure that only conditions on the expectations for the conditional and unconditional counts, \( C_{jl}^{(S|C)} \), \( C_{jl}^{(S|U)} \), \( C_{jl}^{(N|C)} \) and \( C_{jl}^{(N|U)} \), for each site \( l \) to compute \( p \)-values using the extended binomial (EBin) distribution, as discussed in Kosakovsky Pond and Frost (2005).

### 4 RESULTS

#### Performance

We conduct an extended simulation study to compare the relative performance of the renaissance approach using both nonsynonymous and synonymous rate variation across sites (DUAL REL) and the fixed effects likelihood (FEL) model approach implemented in HyPhy (Kosakovsky Pond et al., 2005). We re-analyze the data sets simulated by (Kosakovsky Pond and Frost, 2005). Briefly, the simulation procedure originally considers symmetric bifurcating trees with 8, 16, 32 and 64 tips and generates alignments encompassing 375 codons with a complex distribution of nonsynonymous and synonymous substitution rates (resulting in 75 neutral, 355 negatively selected and 75 positively selected sites) (Kosakovsky Pond and Frost, 2005). We further extend this study to the 128-tip case and focus on the more realistic 16, 32, 64 and 128 range. Codon-based model substitution parameters were inspired by HIV polymerase sequence data and 50 different replicates are generated for each collection of parameters.

We first focus on the binary classification problem of identifying sites under diversifying positive selection and map the proportion of misidentified sites (false positive rate) to the proportion of correctly identified sites (true positive rate) for various nominal \( \alpha \) levels (\( p \)-values, empirical Bayes factors based on the ratio of posterior and prior odds of having \( \omega \neq 1 \), or posterior quantiles) of the test procedures we compare. The resulting receiver operating characteristic (ROC) curves for renaissance counting, DUAL REL and FEL are presented in Figure 1. For renaissance counting, we focus on the results for the CI coverage test approach (Supplementary Figure 1 shows the EBin test has a very similar performance). While the DUAL REL approach performs marginally better for a lower number of taxa, FEL benefits more from a large number of taxa. In general, the renaissance approach competes well with the codon model approaches with a roughly intermediate performance across the different simulation scenarios. While the renaissance counting ROC curves are similar for CI coverage test and the EBin test (Supplementary Figure 1), it is interesting to note that reasonably low false positive rates do not require very low EBin \( p \)-values but they do require relative high posterior quantiles in the coverage approach (Supplementary Table 1). The EBin critical \( p \)-values also clearly decreases with the number of taxa for the same false positive rate whereas a similar trend is not so obvious for the posterior quantiles. We anticipate that cut off values sensitive to the number of taxa will also be affected by sequence divergence and only data-specific simulation provides an objective way to select such values.

Next, we investigate the ability of the renaissance approach to estimate site-specific \( \frac{dS}{dN} \) and \( \frac{dS}{dN} \) values. In Figure 2, we compare the mean relative errors of such site-specific estimation produced by our method, DUAL REL, and FEL during the analysis of the 128-taxon simulated data set. Since the total number of simulations is
Table 1. Comparison of sites under positive diversifying selection as identified using renaissance counting, two random effects likelihood (REL) models, the two-rate FEL approach, and the SLAC method applied to the feline and carnivore parvovirus data set. One REL model incorporates nonsynonymous substitution rate variation (NS) and the other includes both synonymous and nonsynonymous rate variation (DUAL). Underlined sites were identified as positively selected in the original study (Shackleton et al., 2005). Starred and double starred intervals correspond to the 95% and 99% highest posterior density intervals respectively that exclude $\omega_{NC}^S=1$. $\omega_M$ and $\omega_D$ represent the rate class of the general discrete distribution with three rate classes to accommodate variation in $d_S$ and $d_S$ rates among sites respectively.

<table>
<thead>
<tr>
<th>Site</th>
<th>$\omega$ (95% CI)*</th>
<th>p-value</th>
<th>$\omega_{NC}^S$ log(10)</th>
<th>$\omega_M$ log(10)</th>
<th>$\omega_{NC}^D$ log(10)</th>
<th>p-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC</td>
<td>0.55 (0.47-0.63)</td>
<td>0.01</td>
<td>1.86</td>
<td>1.85</td>
<td>1.83</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>87</td>
<td>2.20 (1.35-3.44)*</td>
<td>0.00</td>
<td>2.22</td>
<td>0.90</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>277</td>
<td>1.18 (0.75-1.93)</td>
<td>0.00</td>
<td>1.21</td>
<td>0.90</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>450</td>
<td>1.03 (0.78-1.72)*</td>
<td>0.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>232</td>
<td>2.29 (1.35-3.44)*</td>
<td>0.00</td>
<td>1.33</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>491</td>
<td>1.18 (0.75-1.72)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>240</td>
<td>1.11 (0.74-1.72)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>206</td>
<td>1.12 (0.74-1.72)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>426</td>
<td>1.16 (0.74-1.72)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>80</td>
<td>1.11 (0.28-1.72)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>69</td>
<td>1.11 (0.28-1.72)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

fairly small, owing to the high computational costs of DUAL REL and FEL under phylogenetic tree uncertainty, we also report 95% Monte Carlo error (grey shaded areas). Although the renaissance approach tends to overestimate small $d_S/d_S$ values, our method performs remarkably well when the true $d_S/d_S$ values are above or equal to one, the situation in which practitioners are most interested. This error pattern is reversed for the REL method, outperforming the renaissance method on small $d_S/d_S$ values, but loses to our approach on sites under positive selection. The relative errors of the FEL method are difficult to quantify due to the fact that this method occasionally produces unrealistically high $d_S/d_S$ estimates. If we remove infinite $d_S/d_S$ estimates during the relative error calculations, FEL still produces unrealistically high $d_S/d_S$ estimates on some of the remaining simulations (bottom left plot in Figure 2). Only when we remove $d_S/d_S$ estimates that are above 10,000 do we arrive at reasonable (but now biased) relative errors for the FEL method. Although after this filtering the FEL approach outperforms both the renaissance and REL methods, the FEL approach remains problematic because for some of the sites it fails to provide reasonable $d_S/d_S$ estimates in as many as 11 simulations (22% of all simulations).

**Empirical Data**

We analyze a feline and carnivore parvovirus data set to compare renaissance counting implemented in BEAST (Drummond et al., 2012) to REL codon-based model approaches. This data set includes 91 VP2 capsid sequences sampled from various geographical regions over a 42-year period and was previously used to demonstrate the rapid adaptation of feline panleukopenia parvovirus to canine hosts (Shackleton et al., 2005). Table 1 lists the positively selected sites identified using renaissance counting and REL approaches and Supplementary Figure 2 plots the site-specific DUAL REL $d_S/d_S$ estimates against the renaissance counting $d_S/d_S$ estimates and their credible intervals. The REL approaches include nonsynonymous (NS REL) and both nonsynonymous and synonymous (DUAL REL) rate variation across sites. To model this rate variation, we used generalized discrete distributions with three rate categories. The DUAL REL model provides a significantly better fit when compared to the NS REL model (p < 0.001), indicating significant variation in $d_S$ across sites. Shackleton at al (Shackleton et al., 2005) originally employ a NS REL approach as implemented in PAML (Yang, 1997) and relied on fairly conservative criteria to identify positively selected sites (underlined in Table 1). If we use the criterion of $\omega_{NC}^S > 1$ for renaissance counting and a log Bayes factor (for $d_S/d_S > 1$) > 4.0 for the REL models, which is indeed the cut off that corresponds to a false positive rate of 5% according to data-specific simulations (data not shown), then these three approaches provide support for roughly the same sites as being positively selected with few exceptions. The FEL approach however, does not yield support for any site under positive diversifying selection unless one is willing to accept high critical p-values as evidence. In Table 1, we also report results of a single-likelihood ancestor counting (SLAC) method (Kosakovsky Pond and Frost, 2005), a more advanced version of the original counting method of Suzuki and Gojobori (1999), which performs similarly to the FEL model. In addition, the same issues that emerged in the simulation analyses also complicate the site-specific estimation of $d_S/d_S$ for this data set, and for this reason, (Kosakovsky Pond and Frost, 2005) suggested to use scaled $d_S$ minus $d_S$ as a measure of selection to compare FEL with other methods. Similar to the FEL and SLAC p-values, the EBIN p-values derived from the average counts for each site are of little use compared to a test approach that draws information from the posterior distribution of regularized site-specific $d_S/d_S$ ratios. The poor performance of these approaches may be attributed to the relatively low diversity in this data set (mean pairwise diversity of 0.0086 substitutions per site). Between the two clear exceptions to the consistency of renaissance counting and REL results, site 101 is only identified by NS REL to be under positive selection. When also considering rate variation for synonymous substitutions, this site falls under a class with high synonymous substitution rates in DUAL REL, resulting also in a considerably lower $d_S/d_S$. So, it appears that despite conservative assignment, site 101 may be falsely identified as being positively selected when synonymous rate variation across sites is ignored. In this respect, renaissance counting is in better agreement with DUAL REL, and in general, the renaissance estimates are in good agreement with the posterior $d_S/d_S$ expectations estimated by DUAL REL. Both REL approaches find support for site 491 being under positive diversifying selection despite the observation that the site also experiences a high synonymous substitution rate ($S$ rate class = 2 in DUAL REL), whereas renaissance counting classifies the site as neutrally evolving. It is interesting to note that for this site, only two unrelated sequences have a different amino acid (histidine and arginine) compared to the wild-type amino acid (glutamine). These amino acid substitutions on external branches could therefore represent (slightly) deleterious substitutions that have subsequently experienced purifying selection (Pond et al., 2006; Lemey et al., 2007). We note that for renaissance counting, similar to other counting and FEL approaches, we can restrict ourselves to summarizing the substitution history on internal branches only to avoid the impact of such transient mutations.

To appreciate the computational efficiency that renaissance counting affords, we first recall that direct run-time comparison with the existing implementation of the REL and FEL models is flawed because these implementations condition on a fixed tree. In this paper, we overcome this implementation limitation by
avoids the problem of having to consider the entire posterior distribution of \( \omega_{n} \) for each site. Instead, we can compute the posterior probability of a site being under positive selection, \( \pi \), and then compare \( \pi \) across sites to determine which sites are most likely to have undergone adaptive evolution.

Note that the posterior probability of a site being under positive selection, \( \pi \), is not directly comparable to the posterior probability of \( \omega_{n} \) being greater than 1, \( \pi(\omega_{n} > 1) \), because the former is a function of the latter and the current state of the data. However, the posterior probability of \( \omega_{n} \) being greater than 1, \( \pi(\omega_{n} > 1) \), is a useful measure of the strength of the evidence for adaptive evolution at each site. A site with \( \pi(\omega_{n} > 1) > 0.5 \) is likely to have undergone adaptive evolution, while a site with \( \pi(\omega_{n} > 1) < 0.5 \) is unlikely to have undergone adaptive evolution.

### Table 2.

<table>
<thead>
<tr>
<th>Site</th>
<th>( \omega_{n} ) (95%CPD)</th>
<th>( p )-value</th>
<th>( \omega_{M1} )</th>
<th>( \omega_{M2} )</th>
<th>log(BF)</th>
<th>( \omega )</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.16 (1.37-3.36)**</td>
<td>0.14</td>
<td>8.55</td>
<td>1</td>
<td>2</td>
<td>3.11</td>
<td>( \infty )</td>
</tr>
<tr>
<td>35</td>
<td>1.79 (1.10-2.82)*</td>
<td>0.08</td>
<td>2.3</td>
<td>2</td>
<td>3</td>
<td>7.25</td>
<td>2.62</td>
</tr>
<tr>
<td>79</td>
<td>1.55 (0.99-2.43)</td>
<td>0.25</td>
<td>7.05</td>
<td>1</td>
<td>2</td>
<td>3.68</td>
<td>( \infty )</td>
</tr>
<tr>
<td>60</td>
<td>1.53 (0.89-2.40)</td>
<td>0.08</td>
<td>6.13</td>
<td>1</td>
<td>2</td>
<td>3.88</td>
<td>2.84</td>
</tr>
<tr>
<td>64</td>
<td>1.3 (0.85-1.99)</td>
<td>0.31</td>
<td>8.5</td>
<td>1</td>
<td>2</td>
<td>6.66</td>
<td>( \infty )</td>
</tr>
<tr>
<td>69</td>
<td>3.13 (2.05-4.71)**</td>
<td>0.06</td>
<td>34.89</td>
<td>1</td>
<td>3</td>
<td>10.00</td>
<td>( \infty )</td>
</tr>
<tr>
<td>83</td>
<td>1.24 (0.67-2.09)</td>
<td>0.18</td>
<td>2.02</td>
<td>2</td>
<td>3</td>
<td>6.02</td>
<td>1.58</td>
</tr>
<tr>
<td>102</td>
<td>0.68 (0.44-1.09)</td>
<td>0.56</td>
<td>8.28</td>
<td>1</td>
<td>2</td>
<td>6.12</td>
<td>( \infty )</td>
</tr>
<tr>
<td>123</td>
<td>1.97 (1.11-3.31)*</td>
<td>0.19</td>
<td>1.69</td>
<td>2</td>
<td>3</td>
<td>4.47</td>
<td>0.83</td>
</tr>
<tr>
<td>155</td>
<td>2.27 (1.30-3.80)*</td>
<td>0.11</td>
<td>1.88</td>
<td>2</td>
<td>3</td>
<td>4.39</td>
<td>1.48</td>
</tr>
<tr>
<td>178</td>
<td>2.59 (1.31-4.59)*</td>
<td>0.12</td>
<td>9.96</td>
<td>1</td>
<td>3</td>
<td>10.00</td>
<td>( \infty )</td>
</tr>
<tr>
<td>200</td>
<td>4.61 (2.92-7.27)**</td>
<td>0.02</td>
<td>34.57</td>
<td>1</td>
<td>3</td>
<td>11.14</td>
<td>( \infty )</td>
</tr>
<tr>
<td>207</td>
<td>1.92 (1.15-3.15)*</td>
<td>0.19</td>
<td>1.72</td>
<td>2</td>
<td>3</td>
<td>4.50</td>
<td>3.18</td>
</tr>
<tr>
<td>211</td>
<td>1.16 (0.72-1.85)</td>
<td>0.59</td>
<td>1.8</td>
<td>2</td>
<td>3</td>
<td>7.98</td>
<td>2.54</td>
</tr>
<tr>
<td>215</td>
<td>3.44 (2.17-5.32)**</td>
<td>0.05</td>
<td>39.75</td>
<td>1</td>
<td>3</td>
<td>15.28</td>
<td>( \infty )</td>
</tr>
</tbody>
</table>

**Note:** Sites with \( \pi(\omega_{n} > 1) > 0.5 \) are considered to be under positive selection.

### Table 3.

<table>
<thead>
<tr>
<th>Site</th>
<th>( \omega_{n} ) (95%CPD)</th>
<th>( p )-value</th>
<th>( \omega_{M1} )</th>
<th>( \omega_{M2} )</th>
<th>log(BF)</th>
<th>( \omega )</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>0.98 (1.30-3.00)**</td>
<td>0.06</td>
<td>2.04</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>68</td>
<td>1.78 (1.21-2.51)*</td>
<td>0.13</td>
<td>0.72</td>
<td>1</td>
<td>2</td>
<td>&lt; 1</td>
<td>1.43</td>
</tr>
<tr>
<td>83</td>
<td>0.7 (0.49-0.97)</td>
<td>0.87</td>
<td>1.4</td>
<td>2</td>
<td>3</td>
<td>7.42</td>
<td>0.71</td>
</tr>
<tr>
<td>102</td>
<td>1.68 (1.20-2.33)*</td>
<td>0.14</td>
<td>0.75</td>
<td>2</td>
<td>3</td>
<td>&lt; 1</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>123</td>
<td>0.91 (0.65-1.25)</td>
<td>0.68</td>
<td>1.11</td>
<td>2</td>
<td>3</td>
<td>6.34</td>
<td>0.88</td>
</tr>
<tr>
<td>135</td>
<td>2.31 (1.64-3.20)**</td>
<td>0.01</td>
<td>1.26</td>
<td>2</td>
<td>3</td>
<td>6.63</td>
<td>1.45</td>
</tr>
<tr>
<td>142</td>
<td>0.64 (0.47-0.86)</td>
<td>1</td>
<td>1.11</td>
<td>2</td>
<td>3</td>
<td>6.62</td>
<td>0.56</td>
</tr>
<tr>
<td>177</td>
<td>0.81 (0.60-1.15)</td>
<td>0.77</td>
<td>1.16</td>
<td>2</td>
<td>3</td>
<td>4.39</td>
<td>1.00</td>
</tr>
<tr>
<td>178</td>
<td>4.92 (2.78-7.63)**</td>
<td>&lt; 0.01</td>
<td>2.48</td>
<td>1</td>
<td>3</td>
<td>9.11</td>
<td>1.98</td>
</tr>
<tr>
<td>200</td>
<td>2.64 (1.83-3.72)**</td>
<td>0.01</td>
<td>2.91</td>
<td>1</td>
<td>3</td>
<td>17.85</td>
<td>3.67</td>
</tr>
<tr>
<td>202</td>
<td>4.68 (3.30-6.37)**</td>
<td>&lt; 0.01</td>
<td>0.72</td>
<td>1</td>
<td>2</td>
<td>&lt; 1</td>
<td>&lt; 0.13</td>
</tr>
<tr>
<td>211</td>
<td>1.06 (0.77-1.42)</td>
<td>0.47</td>
<td>2.48</td>
<td>1</td>
<td>3</td>
<td>25.83</td>
<td>3.63</td>
</tr>
</tbody>
</table>
5 DISCUSSION

We present a novel and efficient Bayesian method for detecting positive diversifying selection in molecular sequences. Our method combines the computational speed of counting methods and statistical efficiency of empirical Bayes approaches. A further strength of the method is its simplicity of implementation thanks to recent algorithmic advances in stochastic mapping (Nielsen, 2002; Lartillot, 2006). Stochastic mapping has been used before to speed up MCMC-based Bayesian estimation of codon-based models (Rodrigue et al., 2008), but this and subsequent approaches still operate within the codon-based modeling framework. In contrast, our method uses stochastic mapping to quickly approximate codon-based models that account for rate variation of \( d_N/d_S \) over sites. Our analysis of simulated and real data demonstrates that the proposed approach competes well with more computationally demanding methods in recovering site-specific \( d_N/d_S \) ratios and in accurately identifying sites under positive diversifying selection.

The main distinction between renaissance counting and previous attempts to use stochastic mapping to detect diversifying positive selection (Zhai et al., 2007) is the ability of our renaissance method to directly estimate \( d_N/d_S \) ratios and to quantify uncertainty in these estimates. This advance is important, because all counting methods, including stochastic mapping-based ones, can only produce site-specific \( p \)-values, resulting from testing a null hypothesis of neutrality. However, it is well known that using \( p \)-values has limitations. In particular, since \( p \)-values are always calculated by conditioning on the null hypothesis, it would be inappropriate to use the magnitude of \( p \)-values as evidence in favor of the null (Goodman, 1999). The task of comparing \( p \)-values becomes even more challenging in the method of Zhai et al. (2007), because these authors use notoriously hard to calibrate posterior predictive \( p \)-values (Hjort et al., 2006). Therefore, it is difficult to use previously developed counting methods to quantify relative strength of selection at amino acid sites. Although our method can not possibly resolve the long-standing statistical controversy of quantifying evidence in hypothesis testing, renaissance counting offers a straightforward estimation alternative by producing site-specific posterior distributions of \( d_N/d_S \) ratios. Importantly, this uniquely positions renaissance counting among other positive selection detection methods as a method that also appropriately quantifies the relative strengths of selection at individual sites.

When imputing/counting the unobserved synonymous and nonsynonymous substitutions, we produce substitution counts on each branch of the phylogenetic tree and then sum them up for each site. If a scientifically meaningful partition of branches into groups exists, we can easily sum the substitutions within each group separately and apply our empirical Bayes procedure to the site-specific counts in each group. For applications to within-host HIV evolution for example, we can consider synonymous and nonsynonymous substitutions only on internal branches of the phylogeny. This partition of branch lengths is motivated by a widely accepted hypothesis that many terminal branches of intrahost HIV phylogenies represent lineages that are weeded out by selection (Pond et al., 2006). Therefore, it is reasonable, if not advisable, to avoid the mutational load on these branches and their impact on substitution rates (Lemey et al., 2007), by excluding them when identifying positive diversifying selection.

Concerning the identification site-specific selection patterns, the development of renaissance counting was clearly not to replace, but to complement existing methodology. Even after a comprehensive evaluation of several counting-based and maximum likelihood methods, Kosakovsky Pond and Frost (2005) found it difficult to make definitive recommendations on which methods to use in particular cases. In fact, they concluded that a consensus of several methods, but each accepting relatively high nominal \( \alpha \) levels, would be a reasonable approach to rule out spurious results. In many ways, renaissance counting presents an interesting candidate method to include in such a consensus approach. Embedded in a Bayesian inference framework, renaissance counting accounts for phylogenetic error when estimating \( d_N/d_S \) ratios. Perhaps more importantly, our approach uses a test statistic that is very different from both previous counting methods and maximum likelihood methods. For small or less divergent data sets for example, the test employed in previous counting methods is notoriously conservative (Kosakovsky Pond and Frost, 2005), which is also confirmed by our comparison of the EBin test to the CI coverage test. Empirical Bayesian analysis of random-effects models on the other hand can suffer from large Type I error rates due to excessive errors in the parameter estimates employed (Kosakovsky Pond and Frost, 2005). Renaissance counting avoids the estimation of codon models parameters and estimation uncertainty will be taken into account by the CI coverage test approach. The nucleotide-model approximation also offers a significant speed-up compared to fitting a codon model, but considerable computation time may still need to be invested in integrating over the posterior distribution of phylogenies. In this respect, we emphasize that renaissance counting is implemented in a Bayesian inference framework (BEAST (Drummond et al., 2012)) that provides access to an array of flexible models for estimating genealogical history, divergence times, flexible semi-parametric demographic and phylogeographic histories (Lemey et al., 2009, 2010). Although site-specific selection patterns may not be of prime interest when engaging in such analyses, renaissance counting readily delivers \( d_N/d_S \) ratios for any coding sequence alignment, which may offer a starting point for in-depth characterization of selection pressures involving different methods.

ACKNOWLEDGEMENT

Funding: The research leading to these results has received funding from the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433 and ERC Grant agreement no. 260864, the National Science Foundation (DMS 0856999) and the National Institutes of Health (R01 GM086887 and R01 HG006139). This research has also been supported by the Bioinformatics, Statistical Analysis and Evolutionary Core of the UCSD Center for AIDS Research (5P30AI56214). We acknowledge the support of the National Evolutionary Synthesis Center (NESCent) through a working group (Software for Bayesian Evolutionary Analysis). Further, research was partially completed while the authors were visiting the Institute for Mathematical Sciences, National University of Singapore in 2011.
REFERENCES


