DOI: 10.1534/genetics.104.032144

Detecting Amino Acid Sites Under Positive Selection and Purifying Selection

Tim Massingham[1] and Nick Goldman

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom

Manuscript received June 8, 2004
Accepted for publication December 8, 2004

ABSTRACT

An excess of nonsynonymous over synonymous substitution at individual amino acid sites is an important indicator that positive selection has affected the evolution of a protein between the extant sequences under study and their most recent common ancestor. Several methods exist to detect the presence, and sometimes location, of positively selected sites in alignments of protein-coding sequences. This article describes the "sitewise likelihood-ratio" (SLR) method for detecting nonneutral evolution, a statistical test that can identify sites that are unusually conserved as well as those that are unusually variable. We show that the SLR method can be more powerful than currently published methods for detecting the location of positive selection, especially in difficult cases where the strength of selection is low. The increase in power is achieved while relaxing assumptions about how the strength of selection varies over sites and without elevated rates of false-positive results that have been reported with some other methods. We also show that the SLR method performs well even under circumstances where the results from some previous methods can be misleading.

ANALYZING the instantaneous rate of nonsynonymous (amino acid-changing) and synonymous (silent) nucleotide substitutions in protein-coding molecular sequences can give important clues to understanding how they evolved. In particular, the ratio of the rates of nonsynonymous and synonymous fixation has been used to measure the level of selective pressure on proteins (McDonald and Kreitman 1991, for example). Synonymous mutations do not change the encoded protein and so are often assumed to be selectively neutral. If a nonsynonymous mutation does not affect the fitness of a protein, it would become fixed within the population at the same rate as a synonymous mutation, giving a nonsynonymous/synonymous rate ratio (ω) of 1. If a nonsynonymous change makes the protein more or less fit on average then ω will be greater or less than 1, respectively. A significant excess of nonsynonymous over synonymous substitution has been used as evidence for adaptive evolution (for a review, see Yang and Bielawski 2000).

Several methods have been proposed to detect the presence of positive selection in an aligned set of homologous protein sequences, from counting the observed number of synonymous and nonsynonymous differences between pairs of sequences (*e.g.*, Li *et al.* 1985; Nei and Gojobori 1986; McDonald and Kreitman 1991) to more sophisticated techniques that allow for the phylogenetic relationship between many sequences and permit a statistical test of the result (Nielsen and Yang 1998; Suzuki and Gojobori 1999). These last two methods extract more information by considering the evolutionary relationships between the sequences analyzed but differ markedly in their approaches.

If a site has $\omega > 1$, it is unusually variable and is said to have evolved under positive selection. Similarly, a site with $\omega < 1$ is unusually conserved and is said to have been subject to purifying selection. We describe methods based on the discovery of such sites as tests for detecting the location of selection. The Suzuki and Gojobori (SG) method is such a test, assessing each site of an alignment separately for deviations from neutrality. We refer the reader to the original article (Suzuki and Gojobori 1999) for a more detailed description, but in essence the SG method reconstructs ancestral sequences, counts the number of implied synonymous and nonsynonymous changes, and then tests the result for deviation from neutrality. The SG method largely ignores the uncertainty in the ancestral reconstruction and in the evolutionary path taken between ancestral and extant sequences and failing to account for this extra variability may adversely affect the performance of the test.

In contrast, the Nielsen and Yang (NY) method uses the entire sequence to detect whether any part of it has undergone positive selection, allowing information from all sites to be used to estimate those quantities common to all sites (*e.g.*, evolutionary distances) more accurately. We describe such methods as tests for detecting the presence of positive selection. The NY method describes variation in the level of selection along the sequence as

[1]*Corresponding author:* Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom.
E-mail: timm@ebi.ac.uk

if each site were a random draw from some distribution. It is not clear what distribution appropriately describes how ω varies along the sequence and Yang *et al.* (2000a) investigated the relative performance of several different parametric families. By using maximum-likelihood (ML) methods to compare the performance of the chosen distribution with a suitable "null" model that does not allow for positive selection, a likelihood-ratio test (LRT) for the presence of positive selection can be performed (Anisimova *et al.* 2001). The NY method can also be used with a *post hoc* empirical Bayesian analysis to perform a test of the location of positive selection (Nielsen and Yang 1998). This may be done with or without a prior LRT, although we provide evidence below that empirical Bayes analysis of data where there is not a significant indication of the presence of positive selection can result in unacceptably high rates of false-positive results.

Simulation studies have shown that some variants of the NY method permit unexpectedly high rates of false-positive results when analyzing sequences that, in truth, contain large proportions of neutrally, or almost neutrally, evolving sites (Anisimova *et al.* 2002; Suzuki and Nei 2002). Swanson *et al.* (2003) argue that, with sufficient data, applying some variants of the NY method to data consisting of a mixture of unusually conserved and strictly neutral sites may falsely detect the presence of positive selection 50% of the time regardless of how the nominal size of the test is chosen. These elevated high rates of false positives are a consequence of the distributions used to model the variation in ω and not all choices are affected equally; notably the variant proposed by Swanson *et al.* (2003; henceforth denoted Swanson-Nielsen-Yang, SNY) does not have this unfortunate property when testing for the presence of positive selection, although in this article we show that the method still has some undesirable properties when used in conjunction with empirical Bayes techniques to detect the location of positively selected sites.

This article introduces a test for nonneutral selection that combines the statistical foundation and more realistic models of substitution used by the NY method with the sitewise testing approach used by the SG method. This new "sitewise likelihood-ratio" (SLR) method tests each site individually for neutrality but uses the entire alignment to determine quantities common to all sites, such as evolutionary distances. The SLR method is thus a test for the location of selection, devised from the best features of existing phylogeny-based tests for both the presence and the location of positive selection. It can be used to detect either positive or purifying selection and can form the basis of a test for the presence of selection; we discuss both these points further below.

Treating many parameters as common to all sites allows the SLR method to extract more information from the data without having to make strong assumptions about how selection varies along the sequence. The sitewise nature of the SLR test means that there is no need to specify a model of how ω varies along the sequence and so it is not susceptible to the elevated rates of false-positive results that some variants of the NY method permit. The weaker assumptions underlying the SLR method mean that it is more generally applicable to real data and more robust to potential errors when data violate the models of variation that the NY method assumes.

The behavior of the SLR method is studied using data simulated under a variety of conditions similar to those previously used to investigate the strengths and weaknesses of the NY method (Anisimova *et al.* 2001; Suzuki and Nei 2002) and for two real data sets that provide interesting case studies. In the simulations, data simulating both strictly neutral evolution and evolution with a proportion of sites under positive selection were used. Looking at the performance of the method on neutrally evolving data allows the actual size of the test to be checked, confirming that the rate of false positives is controllable and that the test is neither liberal nor unduly conservative. Simulated data containing sites evolving under positive selection are used to compare the power (ability to detect positive selection) of the SLR and SNY tests. In particular, detecting sites under weak positive selection is extremely difficult and so, while such sites may not be of the most interest when analyzing real data, comparing the ability to find such sites is a good way to discriminate between different methods.

THEORY AND METHODS

Substitution model: The SLR method is based on the same probabilistic model of sitewise evolution as the NY method, which assumes that substitutions at a given codon site occur independently of every other site and that the process can be modeled as a continuous-time Markov process. This model of substitution may be thought of as a two-stage process, describing background mutation and subsequent selection within a population. At each codon position, instantaneous nucleotide mutations are assumed to occur in a fashion similar to the HKY85 model of evolution (Hasegawa *et al.* 1985). These mutations then become fixed in the population with some probability: 0 if the mutation involves the creation of a stop codon, p_S if the mutation is synonymous, or p_N if the mutation is nonsynonymous. Splitting the substitution process into these two stages gives a biological interpretation to the parameterization of the model of codon substitution proposed by Muse and Gaut (1994): their α and β parameters are proportional to the probabilities of fixation given that a synonymous or nonsynonymous mutation has occurred, respectively. Splitting the process in this manner also highlights that the model assumes that all observed mutations are fixed within the population. Consequently, care should be taken when

applying the methods in situations where this may not be true: for example, if the data contain SNPs.

The rate matrix for the Markov chain therefore has entries

$$q_{ij} = \mu\pi_j \times \begin{cases} 0 & (j \text{ a stop codon or } i \to j \\ & \quad \text{not a single-nucleotide mutation}) \\ p_S & (i \to j \text{ a synonymous transversion}) \\ p_N & (i \to j \text{ a nonsynonymous transversion}) \\ \kappa p_S & (i \to j \text{ a synonymous transition}) \\ \kappa p_N & (i \to j \text{ a nonsynonymous transition}), \end{cases}$$

where q_{ij} is the instantaneous rate of substitution from codon i to codon j, κ is the transition/transversion rate ratio, π_j is the equilibrium frequency of codon j, and μ is a parameter confounded with time that describes the rate at which mutations occur. By setting $\omega = p_N/p_S$, constant over all codons, the model "M0" described by Yang *et al.* (2000a; see also Goldman and Yang 1994) is recovered. This is equivalent to assuming that every site has been subject to the same strength of selection. Alternatively, the NY method describes variation in the strength of selection along the sequence as if the value of ω at each site was independently drawn from some distribution, with all of the other parameters common to all sites. By using distributions in this fashion, the variation in the strength of selection can be described using only a few parameters, although its actual value at a particular site remains unknown. In contrast, the SLR method models each site separately with ω taking the value ω_i at site i, using more parameters but directly estimating the strength at each site and making no assumptions about the overall distribution.

For the NY method, the quantity μ is chosen so that, in unit time, the expected number of nucleotide substitutions per codon site is 1. Models of evolution that allow for some distribution of values of ω along the sequence have a consequent variation in the rate of evolution along the sequence; these models set μ so that the average rate across all sites is 1, even though each individual site may not be evolving at this rate (Nielsen and Yang 1998; Yang *et al.* 2000a). In contrast, for the SLR method the rate of evolution could potentially be different at every site as the level of selection changes and it is appropriate to choose μ so that one nucleotide change is expected per unit time for neutrally evolving data. This means that sites under purifying selection (with $p_N < p_S$ or $\omega_i < 1$) appear to evolve more slowly than neutral sites ($\omega_i = 1$), which in turn evolve slower than positively selected sites ($\omega_i > 1$), agreeing with biological intuition.

The SLR test consists of performing a likelihood-ratio test on a sitewise basis, testing the null model (neutrality, $\omega_i = 1$) against an alternative model ($\omega_i \neq 1$; or see below for alternatives).

While revising this article we were made aware of the maximum-likelihood method for detecting positive selection of Suzuki (2004), which shares many similarities with the SLR method. The main differences between the methods are in how the common parameters are estimated, although it is not clear what process Suzuki recommends. Suzuki (2004) does not analyze the size and power of the test presented but the results in this article will be largely transferable if similar methods to estimate common parameters are used.

Parameter estimation and likelihood calculation: All inferences and tests are performed using likelihood calculations and optimizations standard in phylogenetics (see Felsenstein 2003, for example). In the SLR method, the tree topology, branch lengths, equilibrium codon frequencies, and the transition/transversion rate ratio are considered to be common to all sites in an alignment and so are estimated using information from every site.

The null model for the SLR test at site i is that $\omega_i = 1$ while all other parameters, including the strength of selection at other sites (ω_j for all $j \neq i$), are free to vary. Similar techniques to those described in Yang (1996) can be used to hold appropriate parameters common to all sites while allowing others to vary on a "site-by-site" basis, the contribution of each site toward the log-likelihood being calculated using the familiar pruning algorithm (Felsenstein 2003). The alternative model for site i allows all the parameters to vary freely (still including the strength of selection at other sites). Consequently, the alternative model parameter estimates and maximum log-likelihood are identical at every site i, and one high-dimensional optimization (involving all the common parameters and one parameter ω_i for each site i) has to be performed to find them.

Maximizing the log-likelihood of the null model requires a similar high-dimensional optimization at each site and so may require considerable computing resources. Instead of attempting to perform these optimizations, two approximations are made: (a) the estimates of common parameters by M0 are unbiased and consistent, and (b) the effect of a single site on the ML values of the common parameters is negligible. The first approximation allows the common parameters to be estimated using M0 and, in conjunction with the second approximation, held fixed. Given fixed common parameters, the distribution at each site is independent of every other site and so the ML estimate for selective pressure at site i ($\hat{\omega}_i$) can be found via a one-dimensional optimization. In all, these approximations mean that the two high-dimensional optimizations required to find the maximum-likelihood estimates of all parameters can be reduced to a comparatively low-dimensional optimization to estimate the common parameters and a one-dimensional optimization at each site to estimate the strength of selection.

There is some evidence (Yang 2000) that branch lengths are underestimated when applying M0 to data

with rate variation, and so approximation a may not be realistic. Reanalyzing Yang's "small" data set, we find the correlation between synonymous branch length estimates under M8 (dnM8) and those under M0 (dsM0) to be dsM8 = 1.018 × dsM0 − 1.195*e*-07 with an R^2 value of 0.9995. The underestimation is significant and consistent across branches, which may cause an increase in the number of false positives reported by the SLR method. The simulations presented in this article take the underestimation into account and it is not found to make an appreciable difference to the false-positive rate. If the number of sites is large, the contribution from any one site toward the common parameters will be swamped by that from all the others and so approximation b is reasonable.

In all cases, the codon frequencies π_j were estimated from the empirical counts in the observed data rather than using the procedure outlined above.

SLR test and distribution of test statistic: The sitewise LRT statistic for nonneutral evolution at site i, Λ_i, is twice the difference in log-likelihood between the null and alternative models, individually maximized. Approximation b, above, means that Λ_i is in fact simply twice the difference between the contributions of site i to the log-likelihoods under the null and alternative models. Writing $l_i(\omega_i)$ for the contribution of site i to the log-likelihood given common parameter values and selective pressure ω_i, $l_i(1)$ and $l_i(\hat{\omega}_i)$ are the contributions toward the maximum log-likelihood of the null and alternative models, respectively, and

$$\Lambda_i = 2(l_i(\hat{\omega}_i) - l_i(1)).$$

Note that $\Lambda_i \geq 0$, necessarily. Treating all parameters except ω_i as fixed, the usual asymptotic theory (*e.g.*, Garthwaite *et al.* 2002) states that Λ_i should be compared to a χ^2_1 distribution for a statistical test of neutrality ($\omega_i = 1$). The alternative hypothesis is that the site has evolved under purifying or positive selection ($\omega_i \neq 1$).

Often only one of two possible deviations from neutrality may be of interest (for example, detecting positively selected sites), and the power of the test can be improved by considering the likelihood-ratio analog of a one-tailed test. By placing a boundary at $\omega_i = 1$, and finding the maximum likelihood over $\omega_i \geq 1$ (positive selection) or $\omega_i \leq 1$ (purifying selection), only one of the possible alternatives is considered by the test. When the null hypothesis occurs at a boundary of the alternative hypothesis like this, the likelihood-ratio test statistic is asymptotically distributed as an equal mixture of a point mass on 0 and a χ^2_1 distribution (Self and Liang 1987). For this distribution, *P*-values can easily be calculated by halving those obtained from using a χ^2_1 distribution. Following Goldman and Whelan (2000), who provide a table of relevant critical values, this mixture distribution is denoted $\overline{\chi}^2_1$.

As with most parametric tests, the asymptotic distribution of the test statistic is only an approximation to its actual distribution for a finite number of observations and so *P*-values calculated using this asymptote may not be strictly correct. An alternative to approximating the test distribution by its asymptote is to estimate it by a parametric bootstrap, as described by Goldman (1993). Under the null hypothesis $\omega_i = 1$, with all the other model parameters at their previously estimated values, the distribution of all possible observations is completely determined and pseudo-replicates of the data can be generated by Monte Carlo simulation. The observed values of the test statistic for these replicates form an estimate of its actual distribution, from which *P*-values can be derived. In the SLR test, the null model is the same at every site and therefore the bootstrap estimate of the distribution needs to be made only once. The bootstrap estimate is dependent on the tree topology and other estimated parameters, so it is not valid to reuse it to analyze different data—a new bootstrap estimate must be made.

SNY test: The variant of the NY method introduced in Swanson *et al.* (2003) describes the distribution of selective pressure along the sequence as a mixture of a two-parameter beta distribution and a point mass. The beta distribution describes the variation in conservation at sites under purifying selection and the point mass describes neutrally evolving or positively selected sites. The point mass is fixed at neutral ($\omega = 1$) under the null model (model M8A of Swanson *et al.* 2003; see also Wong *et al.* 2004) and the presence of positive selection is tested for by performing a LRT of $\omega = 1$ against $\omega \geq 1$ [a restricted variant of model M8 ("M8B"): Yang *et al.* 2000a; Swanson *et al.* 2003; Wong *et al.* 2004], comparing the LRT statistic to a $\overline{\chi}^2_1$ distribution. This test does not conform to typical statistical procedure since (a) the parameter of interest is on a boundary under the null model, (b) the parameters involved in specifying the shape of the beta distribution can disappear when all the probability of the mixture distribution is placed on the point mass, and (c) the parameter specifying the position of the point mass can disappear when all the probability is placed on the beta distribution. The consequences for performing LRTs are considered below.

The NY method can infer the location of positively selected sites by using empirical Bayes techniques (Nielsen and Yang 1998). We write simply "the SNY test" to denote the SNY test for the location of positive selection using the model M8B with the NY method, estimating parameters by ML and inferring location by using empirical Bayes with some specified cutoff. "SNY + LRT_x" explicitly denotes the use of the SNY test only if a prior LRT (with significance level x%) between M8A and M8B indicates significant positive selection.

Simulations: The probabilistic model that underlies both the SNY and SLR methods means that sample data sets can be simulated under known conditions, for example, with 5% of sites subject to a given level of

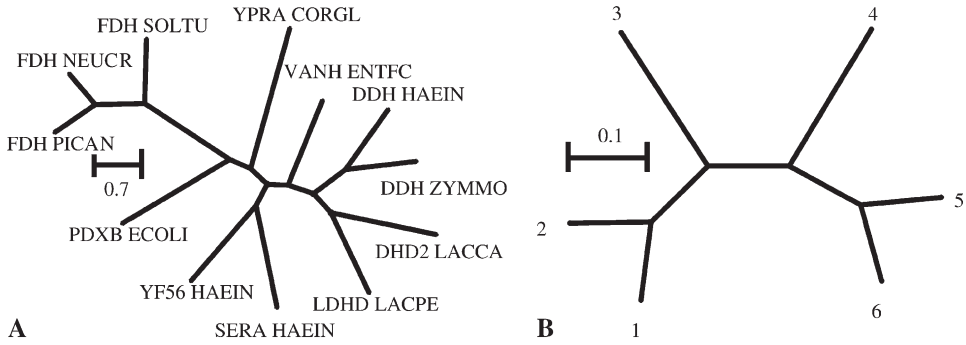

FIGURE 1.—Phylogenetic trees used for simulation and analyses presented in this article. Tree A is derived from 12 sequences of 2-hydroxyacid dehydrogenase (accession no. PF00389) taken from the Pandit database (WHELAN *et al.* 2003). Tree B is an artificial tree used solely for simulations and has previously been used in studies of positive selection tests by ANISIMOVA *et al.* (2001).

positive selection. Neutrally evolving data can be generated to check whether there is any significant difference between the actual distribution of a test statistic and its asymptote and so ensure that critical points obtained from the asymptotic approximation are accurate and the size of the test (probability of type I error or false-positive results) is properly controlled. If data are generated with sites under purifying or positive selection, the location of these sites in the sequence is known and so the number of the sites a method correctly detects can be determined. Multiple methods can be compared by looking at the number of sites each of them correctly (or incorrectly) detects when analyzing the same data. In this article, the situation we study in detail is that of trying to detect positive selection since we expect this variant of the SLR test to be of most interest.

The power of two methods can be fairly compared only if their respective probabilities of a false-positive result are the same. This poses a problem when comparing the SNY and SLR tests since the SNY test produces a Bayesian posterior probability of positive selection rather than a *P*-value, and the SLR produces a *P*-value assuming strict neutrality (the hardest case to distinguish from positive selection) and so is an upper bound on the true probability that the result is a false positive. For this reason, receiver operator characteristic (ROC) curves—plots of the number of correctly identified sites against the number of false positives as the cutoff value for the test is lowered—are useful. These curves allow the power of the methods to be compared as if the probability of a false positive could be chosen perfectly, a situation that is not possible for either test.

Testing for the presence of positive selection: As multiple tests will generally have been performed (*e.g.*, one at each codon site), the detection by the SLR test of one (or more) sites under positive selection may not be sufficient evidence for inferring the presence of positive selection in the sequence as a whole. The statistics must be adjusted for the number of tests performed, using standard techniques for multiple comparisons (HSU 1996). Application of corrections for multiple comparisons to the SG method has been investigated by WONG *et al.* (2004) and these tests are equally applicable to the SLR method.

Standard corrections for multiple comparisons are likely to lead to conservative tests in the SLR method since they assume that all sites have the same probability of falsely indicating positive selection as a neutral site whereas in reality many sites are likely to be under purifying selection and therefore have a lower probability of giving a false-positive result. In addition, the question of interest is whether sites are present under positive selection, rather than identifying particular sites, which is a weaker criterion than assumed by most multiple-comparison corrections. An unusually large number of sites with some evidence for positive selection could be just as convincing as one site with overwhelming evidence.

Sites that exist only in a single species, perhaps as a result of a recent insertion, contain no information about the substitution process and are defined to have $\omega_i = 1$. Such sites cannot be detected as being under positive selection and should not count toward the number of tests performed for the purposes of multiple-comparison corrections.

RESULTS

Distribution of the SLR test statistic: We present two simulations representative of the range of results we have observed in a greater number of experiments. In the first, 25 alignments of 12 sequences, each 200 codons long, were generated under neutral evolution with $\kappa = 2$ and codon frequencies taken from a set of 25 abalone species' sperm lysin sequences (YANG *et al.* 2000b; data also distributed with YANG 1997) on the tree in Figure 1A, using the program *evolver* from the PAML package (YANG 1997). The generated alignments were then analyzed on the correct tree using the SLR method, all common parameters estimated using only data from each alignment. The 5000 sitewise test statistics form a Monte Carlo estimate of their test distribution, allowing for variation in the estimates of common parameters, which was compared to the $\overline{\chi}^2_1$ distribution using a chi-square goodness-of-fit test (SOKAL and ROHLF 1995) and is shown in Figure 2. The *P*-value of the fit was 0.24 and we conclude that, for all practical purposes, there is no difference between the actual distribution of the test statistic and its asymptote in this case.

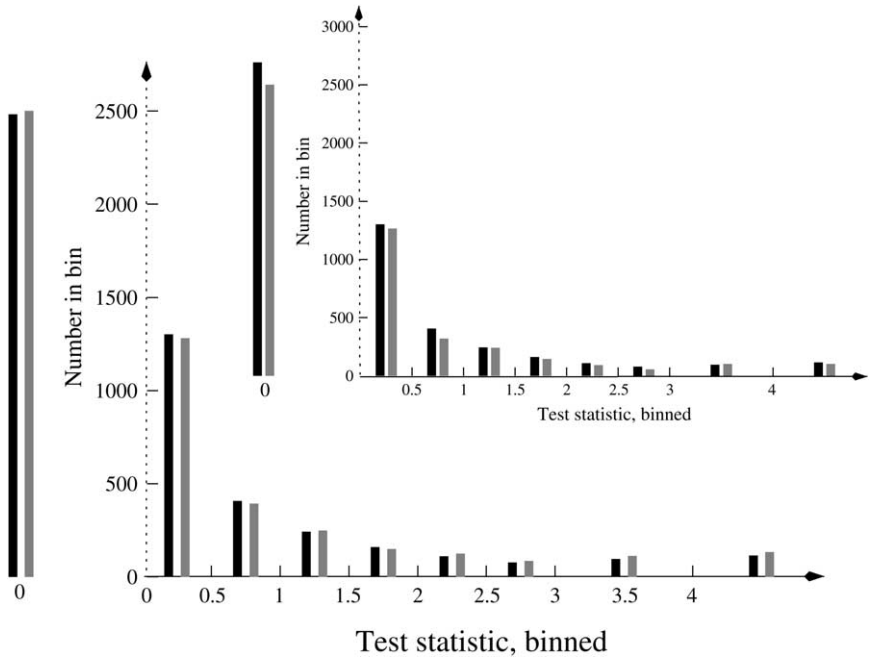

FIGURE 2.—Comparison of actual and assumed distribution of test statistic for SLR. Bar charts show an estimate of the actual distribution of the test statistic (solid bars) against its $\overline{\chi}^2_1$ asymptote (shaded bars). The bars to the left of the *y*-axis represent values where the test statistic is exactly 0. The main graph shows a good fit from data generated on the 12-species tree; the inset is the poorer fit from the 6-species tree.

In the second experiment, the analysis was again performed using the parameter values derived from the sperm lysin alignment but now the tree shown in Figure 1B was used. This tree is short and has few sequences, and hence each site contains relatively little information about its evolution. The distribution of the test statistic is also shown in Figure 2, and the *P*-value for the fit is 1×10^{-7}, indicating a significant difference between the actual and asymptotic distributions. However, the 95 and 99% critical values of the asymptotic $\overline{\chi}^2_1$ distribution, 2.71 and 5.41, correspond to the 95.5 and 99.1% points of the parametric bootstrap distribution, respectively, and so there is little difference in the tail of the distribution. This is the region that is important for all the tests we describe here, and so for the purposes of hypothesis testing in this article the asymptotic critical points are taken as correct.

This similarity between the simulated data and theory agrees with the results of KNUDSEN and MIYAMOTO (2001) for a similar test to detect sitewise changes in evolutionary rate. They also observed that the fit appeared especially good in the tail of the distribution.

False positives and size of SLR test: Table 1 summarizes the performance of the SLR and SNY tests on neutrally evolving sequences ($\omega_i = 1$ for all sites), the most difficult case to distinguish from positive selection for either method. One hundred sets of data, each consisting of 300 codons, were simulated under a neutral model of evolution on the tree shown in SUZUKI and NEI (2002, Figure 1C), using the same branch lengths (all 0.1). SUZUKI and NEI (2002) claim this tree to be a difficult case, in which they report the NY method can give high rates of false positives. As shown in Table 1, the size for the SLR test is approximately correct for these data.

This contrasts markedly with the results for the SNY test used without first performing a LRT for the presence of positive selection, which show a false positive rate >30% even when the empirical Bayes posterior probability cutoff is 0.99. Once the LRT is carried out, the size of the SNY test is reduced to more appropriate values; we note, however, that the size is affected more by the significance level chosen for the LRT than by the posterior probability cutoff chosen for the empirical Bayes analysis. For the analysis of strictly neutral data, the cutoff for empirical Bayes analysis provides no meaningful control over the rate of positive results. As also suggested by WONG *et al.* (2004), these results strongly support the requirement to confirm that there is significant evidence for the presence of positive selection before using empirical Bayes techniques to determine its location. Even though the overall false-positive rate is reasonable for the SNY + LRT_x tests, consideration of the range of the number of false-positive results per data set (from 0 to 100% of sites falsely identified) shows that the behavior of the method is still extreme: when one site is wrongly inferred to be under positive selection, many more false positives are also likely to occur and there is an incorrect inference of apparently pervasive positive selection.

Power of the SLR test: The simulations have shown that the size of the SLR method is properly controlled and that it is reasonable to approximate critical points from the asymptotic $\overline{\chi}^2_1$ distribution of the test statistic. Having established that the method is well behaved, it is meaningful to investigate its power.

The power of the SLR method was compared to that of the SNY test in three situations that differ in the distribution from which the value of ω at each site is drawn. These distributions were: (A) the true model for

TABLE 1

Number of sites falsely identified as having evolved under positive selection

Method	Cutoff	Wrong sites	%	Range (%)
SLR	0.95	1529	5	2–43 (1–14)
	0.99	334	1	0–14 (0–5)
SNY	0.95	10646	35	0–300 (0–100)
	0.99	9540	32	0–300 (0–100)
SNY + LRT_{95} (17)	0.95	2808	9	0–300 (0–100)
	0.99	2586	9	0–300 (0–100)
SNY + LRT_{99} (5)	0.95	1171	4	0–300 (0–100)
	0.99	1106	4	0–300 (0–100)

One hundred data sets, each 300 codons long, were simulated under a model of neutral evolution and analyzed. "SNY + LRT_x (n)" refers to detecting sites using empirical Bayes on data sets that pass a statistical test for the presence of positive selection using significance level x%, the number n in parentheses being the number of such data sets (out of 100) that passed the test. "Cutoff" refers to the significance level used in LRTs for the SLR test and to the empirical Bayes posterior probability cutoff level used in the SNY test—note that these numbers are not directly comparable (see text for further details). "Range" shows the minimum and maximum numbers of falsely identified sites observed within the 100 individual data sets.

the SNY test, M8B with $p_0 = 0.9432$, $\omega = 2.081$, $p = 0.572$, and $q = 2.172$; (B) a mixture distribution taking the value $\omega = 0.5$ with probability 0.75 or else $\omega = 1.5$; and (C) a distribution derived empirically by fitting a large number of categories to an alignment of D-mannose-binding lectin sequences (Pandit accession no. PF01453), seven of which had nonzero weight and one of which was slightly positively selected ($\omega = 1.72$, $p = 0.039$). All other parameters took the same values as in the simulations described above. Cases A and C represent realistic examples of positive selection; in case A the alternative hypothesis model of the SNY test is true, giving that test a particularly good chance of succeeding. Case B was selected to be a difficult problem for both methods. For each model of ω variation, data sets 200 codons long were generated on the tree shown in Figure 1B. The short length of this tree, along with the limited number of sequences, means there is little information per site for a method to detect positive selection and so it is a good example for comparing the power of methods.

To reflect realistic analysis of data and reduce the false-positive rate for the SNY test, data sets were simulated until there were 100 that passed the LRT test for the presence of positive selection at the 95% level. Only these 100 sets were analyzed; in total 229, 456, and 931 sets of data had to be generated to achieve the required number of passes for cases A, B, and C, respectively.

The ROC curves in Figure 3 show how the numbers of true-positive and false-positive results change for each method as their respective cutoff points are varied. In all three cases, the SLR test dominates the SNY + LRT_{95} test and so is the more powerful discriminator: for any given level of false positives, the SLR method correctly identifies more sites evolving under positive selection. In all cases the nominal 95 and 99% cutoffs are conservative, a result that is expected for the SLR test for reasons similar to those discussed above. (The SNY test does not have to be conservative; here, it probably is so because none of the three models contains any almost neutral sites.) The simulations were repeated using sets of data 1000 codons long, generating data from tree A in Figure 1, or both of these. Using longer sequences makes little difference to the ROC of either method but increasing the number of species does, findings consistent with those of Anisimova *et al.* (2002). The ROC curves for these additional simulations are in the supplementary material (http://www.genetics.org/supplemental/).

The curves in Figure 3 do not tell the entire story, as the SLR method would have correctly detected sites in the many data sets that were discarded because they did not pass the LRT for the presence of positive selection. For the discarded data sets (56, 78, and 89% of data sets in cases A, B, and C, respectively) the SNY + LRT_{95} test has effectively no power, whereas the power of the SLR method does not drop substantially compared to the results shown in Figure 3 (see supplementary material) and so overall it is by far the more powerful of the two tests. As the evidence for the presence of positive selection increases, the proportion of data sets for which the SNY + LRT_{95} test will detect the presence of positive selection will increase and so this difference in performance between it and the SLR test will decrease.

Real examples: By way of comparison on real data, we reanalyze two data sets from Yang *et al.* (2000a), using both the SLR and SNY tests. The vertebrate β-globin and human immunodeficiency virus (HIV)-1 *pol* data sets (D2 and D7 in the original article) were analyzed for positively selected sites with both tests, using the alignments and topologies available from ftp://abacus.gene.ucl.ac.uk/pub/YNGP2000/. The SNY test gives only marginal evidence for the presence of positive selection in β-globin, with a LRT statistic of 5.45, in comparison to the HIV-1 *pol* data for which a value of 60.27 is obtained (*cf.* $\overline{\chi}^2_1$ 95 and 99% critical values of 2.71 and 5.41, respectively). Using criteria consistent with our simulations, there is sufficient evidence of positive selection to carry out empirical Bayes detection of location for both data sets. The results are summarized in Table 2.

The SLR test finds no sites with significant positive selection in the β-globin data, compared to five sites found by the SNY test with a 95% posterior probability cutoff. Correction of the SLR test for multiple comparisons with Hochberg's method (Hsu 1996) gives a *P*-value indistin-

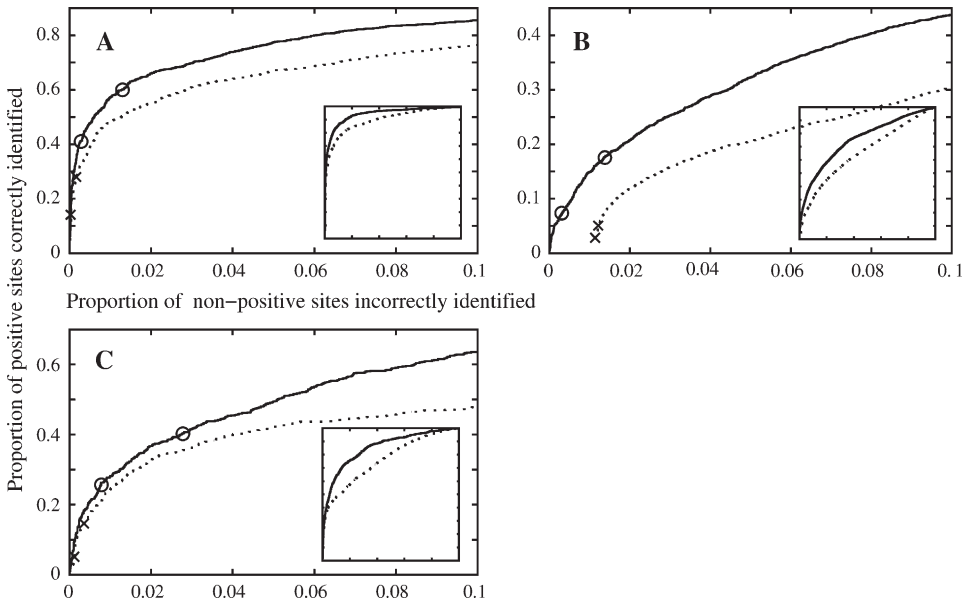

FIGURE 3.—ROC curves comparing the power of the SNY and SLR tests for positive selection. ROC curves are shown for the proportion of sites correctly and incorrectly identified as positively selected in various simulation experiments. The conditions for the simulations (A, B, and C) are described in the text. The main curves show the region with a false-positive rate <10%; the insets are the complete curves (axes unmarked, but both between 0 and 1). Solid lines, SLR test; dotted lines, SNY + LRT_{95} test. The circles (crosses) indicate the results that would be obtained by taking nominal 95 and 99% cutoffs for the SLR (SNY) test. The discontinuity for the SNY test in case B is due to several sites being classified as being under positive selection with posterior probability 1 (possibly due to rounding at the eighth decimal place). A signed version of the SLR test statistic was used when constructing these curves [$\text{sign}(\hat{\omega}_i - 1)\Lambda_i$] so the SLR test could order sites that have no evidence of positive selection.

guishable from 1, again indicating no evidence for the presence of positive selection.

For the HIV-1 *pol* data set, the SNY test finds 13 sites with a posterior probability of ≥0.95, of which 6 have posterior probability ≥0.99. The SLR test finds 22 sites with a *P*-value ≤0.05 and 13 with *P*-value ≤0.01; these latter 13 matched exactly the sites found by the SNY test. After Hochberg's correction for multiple comparisons, two sites have adjusted *P*-values ≤0.05. The more extreme of these has a *P*-value of 2×10^{-6}; the other has a *P*-value of 1×10^{-3}. This indicates that the SLR test finds highly significant evidence for the presence of positive selection in the HIV-1 *pol* data, at sites 67 and 347. For this data set, the SLR and SNY tests indicate the presence of significant positive selection and suggest similar locations.

These two data sets show that the SNY and SLR tests have broadly comparable performance on real data, with the SLR test detecting more sites in the HIV-1 *pol* data set at a nominal cutoff although the actual number of false-positive results is unknowable. The SLR test provides no

TABLE 2

Analysis of β-globin and HIV-1 *pol* for positively selected sites

Data set	Model/test	Optimal log-likelihood	Optimal parameters	Positively selected sites
β-Globin	M8A	−3594.55	$\kappa = 1.78$, $p_0 = 0.90$, $p = 0.63$, $q = 3.24$	NA
	M8B	−3591.83	$\kappa = 1.86$, $p_0 = 0.94$, $p = 0.58$, $q = 2.44$, $\omega = 1.62$	7, 50, 67, 85, *123*
	SLR	NA	$\kappa = 1.78$, $\omega = 0.24$	None
HIV-1 *pol*	M8A	−9293.65	$\kappa = 4.87$, $p_0 = 0.87$, $p = 0.70$, $q = 10.92$	NA
	M8B	−9263.52	$\kappa = 5.29$, $p_0 = 0.97$, $p = 0.19$, $q = 1.08$, $\omega = 3.78$	2, 3, 14, 41, *67*, *347*, 379, 459, *478*, *568*, 654, *761*, *779*
	SLR	NA	$\kappa = 4.84$, $\omega = 0.20$	*2*, *3*, 4, *14*, *41*, *67*, 313, *347*, *379*, 388, 431, *459*, 462, *478*, *568*, 570, *654*, 732, *761*, *779*, 782, 890

Results of analysis of β-globin and HIV-1 *pol* data sets are shown, using the SNY test (models M8A and M8B) and the SLR test. For the SLR test, the parameter values given are those optimal under M0. The sites listed are those at which positive selection is detected with a cutoff (significance level or posterior probability, as appropriate to the method used) >95%; those >99% are in italics. The italic underlined sites are those at which there is still evidence for positive selection after correcting the SLR test for multiple comparisons.

evidence for positive selection in β-globin and the SNY test and associated parameter value estimates for M8B suggest that, should any positive selection exist, it is weak. The results for β-globin differ from those obtained by Yang *et al.* (2000a) for two reasons: the comparison of models M8A and M8B in the SNY test permits a direct test for the presence of positive selection, whereas other comparisons used by Yang *et al.* (2000a) may not (Swanson *et al.* 2003), and the codon frequencies were estimated differently (Yang *et al.* 2000a used four nucleotide frequencies for each of the three codon positions, whereas we use 61 codon frequencies).

DISCUSSION

This article has presented a new method to detect both positive and purifying selection. It has been shown that the SLR test can achieve higher power than the SNY test while providing numerous other advantages: the method relaxes some important assumptions about how variation in the level of selection is modeled, the probability of a false-positive result from the SLR method is controllable, and the method is better behaved than the SNY test when analyzing strictly neutral data and does not occasionally give drastically wrong results in such cases.

Having to model how the level of selective pressure varies along a sequence has caused a proliferation of variants of the NY method and appears to be the source of the observed problems with the LRT (Anisimova *et al.* 2001; Suzuki and Nei 2002; Massingham 2003). By making fewer assumptions about how the level of selection varies along the sequence, the SLR method is more generally applicable to data and robust to possible errors when assumptions about the distribution of ω are violated.

The SLR test appears to have excellent control over the levels of false-positive inference of sites evolving under positive selection, with no evidence of the high rates that have been reported for variants of the NY method. In contrast, Table 1 shows that the SNY test can occasionally make large mistakes when analyzing strictly neutral data. When a set of data falsely passes the LRT, empirical Bayes analyses can go badly wrong, often implying that many sites are under positive selection with extremely high confidence—apparently strong evidence of pervasive positive selection when none was actually present.

We believe that this behavior is due to the empirical Bayes analysis taking the estimated parameter values as true and not allowing for their inherent uncertainty. For example, consider the case where the truth consists of most sites evolving under strictly neutral evolution with a small number evolving with rates as if chosen from a beta distribution: even for a large amount of data the estimated location of the point mass in model M8B will place $\omega > 1$ ~50% of the time. This point mass may have a large weight and under these circumstances an empirical Bayes analysis will indicate positive selection with high confidence.

In contrast, a full Bayes analysis (Huelsenbeck and Dyer 2004) would give confidence only of 50%. When analyzing real data, false indications of pervasive positive selection may be differentiated from real ones by looking at the standard errors of the parameter values or by checking whether the sites identified as being under positive selection are distributed randomly along the sequence. An inference of a high proportion of sites with ω close to 1 should be cause for extreme caution.

The observed size of the LRT for the SNY test did not agree with what may have been expected from theory (Self and Liang 1987). The discrepancy may be because the distribution used to generate the data lies on the boundary of parameter space for the null model, making the two parameters describing the beta distribution component of the model inestimable. When parameters cannot be estimated under the null model, even nuisance parameters not directly involved in the test, the regularity conditions necessary for the asymptotic approximations do not hold and there is no reason to expect *P*-values from a $\overline{\chi}^2_1$ distribution to be accurate. In this case, the techniques described in Davies (1987) may be useful in constructing critical values whose nominal *P*-values more accurately reflect the actual size of the test. This discrepancy arises only when the truth is strictly neutral evolution, although bad fit to the asymptotic distribution may also be observed for small to medium sample sizes when the truth is close to strictly neutral (for example, when a large proportion of sites have $\omega \simeq 1$). Whatever the cause, at the moment the SNY test allied with an empirical Bayesian analysis gives no predictable control over the size of the ensuing test for the location of positive selection.

The examples presented in this article suggest that the power of the SLR test to detect the location of positive selection exceeds that of the SNY test, even when considering only those sets of data with significant evidence of the presence of positive selection. In addition, the SLR method does not suffer from uncontrollable rates of false-positive results when analyzing strictly neutrally evolving data and relaxes restrictions about how the variation in the strength of selection is modeled. The apparent paradox of an increase in power accompanying a relaxation of restrictions may be due in part to the ill fit between the assumptions and the data (in some examples) and to the failure of the empirical Bayes analysis to take the variability of parameter estimates into account.

The SLR method assumes that the background pattern of mutation is constant along the sequences being analyzed and that the differences between the observed rates of evolution at each site are solely due to differences in the strength of selection. For data consisting of alignments of single proteins, the genomic environment of each site is similar, as will be the probability of repair and other factors influencing the rate at which synonymous mutations become fixed within a popula-

tion. However, if significant variation in mutation rates is expected along the sequence, extra parameters could be added at each site to describe the rate of mutation at that site—equivalent to allowing μ to vary on a site-by-site basis. Modifying the method in this manner might cause a noticeable drop in power as many more parameters would have to be estimated from the same amount of data. Alternatively, rate variation could be incorporated in a manner common to all sites using techniques similar to those described by Yang (1993, 1994).

Last, we note that the SLR method is a test of whether a given site has undergone selection or not, and the test statistic summarizes the strength of the evidence for selection rather than the strength of the selection itself. The variance of the estimate of ω may be very large and values at different sites or from different data sets should not be compared without reference to confidence intervals. The program used to carry out the SLR test is available on request from TM.

The authors thank Ziheng Yang for his generous distribution of the PAML software package and Simon Whelan for useful discussions and are grateful to Rasmus Nielsen, Wendy Wong, and Ziheng Yang for access to and discussion of unpublished data [and acknowledge their priority in discovering that the actual size of the test for the Swanson modification does not always agree closely with the theory from Self and Liang (1987); this result was published separately as part of Wong *et al.* (2004)]. T.M.'s work was funded by a postdoctoral fellowship from the European Bioinformatics Institute of the European Molecular Biology Laboratory. N.G. is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Research.

LITERATURE CITED

Anisimova, M., J. P. Bielawski and Z. Yang, 2001 Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18:** 1585–1592.

Anisimova, M., J. P. Bielawski and Z. Yang, 2002 Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. **19:** 950–958.

Davies, R. B., 1987 Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **74:** 33–43.

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

Garthwaite, P., I. Jolliffe and B. Jones, 2002 *Statistical Inference*, Ed. 2. Oxford University Press, Oxford.

Goldman, N., 1993 Statistical tests of models of DNA substitution. J. Mol. Evol. **36:** 182–198.

Goldman, N., and S. Whelan, 2000 Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. **17:** 975–978.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

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

Hsu, J. C., 1996 *Multiple Comparisons. Theory and Methods.* Chapman & Hall, London.

Huelsenbeck, J. P., and K. A. Dyer, 2004 Bayesian estimation of positively selected sites. J. Mol. Evol. **58:** 661–672.

Knudsen, B., and M. M. Miyamoto, 2001 A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl Acad. Sci. USA **98:** 14512–14517.

Li, W.-H., C.-I Wu and C.-C. Luo, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

Massingham, T., 2003 *Detecting Positive Selection in Proteins: Models of Evolution and Statistical Tests.* Ph.D. Thesis, University of Cambridge, Cambridge, UK.

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

Muse, S. V., and B. S. Gaut, 1994 A likelihood approach for comparing synonymous and non-synonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11:** 715–724.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

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

Self, S. G., and K.-Y. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82:** 605–610.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*, Ed. 3. W. H. Freeman, New York.

Suzuki, Y., 2004 New methods for detecting positive selection at single amino acid sites. J. Mol. Evol. **59:** 11–19.

Suzuki, Y., and T. Gojobori, 1999 A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16:** 1315–1328.

Suzuki, Y., and M. Nei, 2002 Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **19:** 1865–1869.

Swanson, W. J., R. Nielsen and Q. Yang, 2003 Pervasive adaptive evolution in mammalian fertilization proteins. Mol. Biol. Evol. **20:** 18–20.

Whelan, S., P. I. W. de Bakker and N. Goldman, 2003 Pandit: a database of protein and associated nucleotide domains with inferred trees. Bioinformatics **19:** 1556–1563.

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

Yang, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:** 1396–1401.

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

Yang, Z., 1996 Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. **42:** 587–596.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

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

Yang, Z., and J. P. Bielawski, 2000 Statistical methods for detecting molecular adaptation. TREE **15:** 496–503.

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

Yang, Z., W. J. Swanson and V. D. Vacquier, 2000b Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. **17:** 1446–1455.

Communicating editor: J. Wakeley