Genet. Res., Camb. (2003), **81**, *pp.* 115–121. *With* 1 *figure.*
DOI: 10.1017/S0016672303006128 Printed in the United Kingdom

Potential impact of recombination on sitewise approaches for detecting positive natural selection

DANIEL SHRINER[1]*, DAVID C. NICKLE[1], MARK A. JENSEN[1] AND JAMES I. MULLINS[1,2]
[1] *Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98195-8070, USA*
[2] *Departments of Medicine and Laboratory Medicine, University of Washington School of Medicine, Seattle, WA 98195-8070, USA*

(*Received 8 July 2002 and in revised form 19 November 2002*)

Summary

Current sitewise methods for detecting positive selection on gene sequences (the *de facto* standard being the CODEML method (Yang *et al*., 2000)) assume no recombination. This paper presents simulation results indicating that violation of this assumption can lead to false positive detection of sites undergoing positive selection. Through the use of population-scaled mutation and recombination rates, simulations can be performed that permit the generation of appropriate null distributions corresponding to neutral expectations in the presence of recombination, thereby allowing for a more accurate estimation of positive selection.

1. Introduction

An active area of current research is the detection of adaptive evolution, as measured by codons that experience positive natural selection. Older approaches to this issue use summary statistics that average measures of genetic polymorphism across sites, as exemplified by Tajima's *D* test (Tajima, 1989). This test of neutrality does not differentiate between directional (positive) and purifying (negative) selection, because both forces result in an excess of low-frequency mutations that are indistinguishably reflected in the test statistic. The methods of Nei & Gojobori (1986), Li *et al*. (1985), Pamilo & Bianchi (1993), Li (1993), Comeron (1995) and Nei & Kumar (2000), as implemented in the Molecular Evolutionary Genetics Analysis software package (MEGA, version 2.1) (Kumar *et al*., 2001), are all designed to differentiate between positive and negative selection. However, these methods assume constant selection pressure across sites, with the consequence that the signal for sites experiencing negative selection could average away the signal for sites experiencing positive selection. Furthermore, if positive selection was detected in a region, it would not be possible to localize the signal to individual sites.

* Corresponding author. Tel: +1 206 732 6149. Fax: +1 206 732 6167. e-mail: dshriner@u.washington.edu

In an attempt to improve power, more recent approaches consider the data sitewise by reconstructing ancestral sequences at the interior nodes of the sample's phylogeny, thereby allowing for a separate consideration of inferred mutational events that have occurred at each site. These sitewise approaches include the CODEML method implemented within the Phylogenetic Analysis by Maximum Likelihood software package (PAML, version 3.1) (Nielsen & Yang, 1998; Yang *et al*., 2000) and the parsimony method of Gojobori (Suzuki & Gojobori, 1999; Yamaguchi-Kabata & Gojobori, 2000). The maximum-likelihood methods of Yang *et al*. have become a *de facto* standard, having been used to study a wide range of organisms, including conifers (Kusumi *et al*., 2002), vertebrates (Yang *et al*., 2000; Kao & Lee, 2002), rickettsiae (Allsopp *et al*., 2001), viruses (Nielsen & Yang, 1998; Yang *et al*., 2000; Haydon *et al*., 2001; Woelk *et al*., 2001; Brault *et al*., 2002; DeFilippis *et al*., 2002), and mammals (Swanson *et al*., 2001; Baum *et al*., 2002). Some analyses of HIV evolution have also used CODEML to identify sites experiencing positive selection (Nielsen & Yang, 1998; Zanotto *et al*., 1999; Yang *et al*., 2000; Ross & Rodrigo, 2002).

Both of these sitewise approaches assume that there is complete linkage between collinear sites. However, processes that disrupt linkage are present across many evolutionary lineages, including viruses,

bacteria, yeast, *Drosophila* and humans, whether the processes are termed lateral or horizontal gene transfer, conjugation, transformation, hybridization, meiotic crossing over, gene conversion, strand transfer, reassortment, chimerization or recombination. For simplicity, we refer to all such processes as recombination. Recombination does not affect the mean of D in Tajima's D test but it decreases the variance, making the test conservative (Wall, 1999; Schierup & Hein, 2000). However, the effects of recombination on sitewise approaches are not known.

The CODEML method uses codon-based models that allow for variable selection intensities among sites within protein-coding DNA sequences. Selection intensities are measured by ω, the ratio of nonsynonymous mutations per potential nonsynonymous site (dn) to synonymous mutations per potential synonymous site (ds). Models differ in how sites are distributed into categories of different ω values. Model parameters are estimated in a maximum-likelihood framework, which allows for likelihood-ratio testing of nested models. An empirical Bayesian approach is used to identify positively selected sites. Recent work on this method addressed the accuracy and the power of the likelihood-ratio test for choosing between nested models (Anisimova *et al*., 2001) and the accuracy and power of the Bayesian prediction (Anisimova *et al*., 2002). Implicit in this analytical program, however, is the assumption that there is no recombination. We therefore performed a simulation study to examine the possible effects of unacknowledged recombination on the CODEML method.

2. Simulation results

To investigate the potential impact of recombination on the ability of the CODEML method to detect positive natural selection, we performed the following simulations using the algorithm of Hudson (1983) as implemented in the program 'hudson' (Schierup & Hein, 2000). 100 independent realizations of the evolutionary process were generated under a neutral coalescent model with a per-locus population-scaled mutation rate $\theta = 42$ for each of the following per-locus population-scaled recombination rates: (1) the no-recombination case with $\rho = 0$; (2) $\rho = 18$, an intermediate rate corresponding to values estimated from HIV evolution (Shriner *et al*., unpublished); and (3) $\rho = 105$, a high rate corresponding to the average ratio ρ/θ from 24 loci from *Drosophila melanogaster* (Andolfatto & Przeworski, 2000). Nucleotide sequences (ten sequences of 700 bp) were then generated under a Jukes–Cantor model of substitution. These sample size and sequence length values corresponded to typical values in studies of HIV *env* gene evolution (e.g. Shankarappa *et al*., 1999). The maximum-likelihood phylogeny was reconstructed for each data set (Swofford, 2002), with the understanding that no single tree might accurately reflect the history of all sites in the presence of recombination. The data sets and trees were then input into CODEML.

Table 1. *False positive error rate for neutral data with and without recombination. The parameters for the coalescent simulation were: population-scaled mutation rate* $\theta = 42$ *per locus; population-scaled recombination rate* $\rho = 0$, *18 or 105 per locus; locus length of 700 bp; and sample size of ten sequences, under a Jukes–Cantor model of substitution. For all 300 replicate data sets, stop codons were stripped and sequences were truncated to a final length of 639 bp*

	M0 vs M3	M7 vs M8
$\rho = 0$	0%	1%
$\rho = 18$	7%	68%
$\rho = 105$	4%	98%

Following Anisimova *et al*. (2001), we tested the same four models (models 0 vs 3 and 7 vs 8) using the same likelihood-ratio test procedure in order to assess the false-positive error rate when data are analysed with unacknowledged recombination. Model 0 assumes one ω ratio for all sites, whereas model 3 allows a user-defined number of ω-ratio categories with estimated proportions of sites in each category and freely estimated ω values for each category (we used the default value of three categories) (Yang *et al*., 2000). Model 7 assumes all sites have an ω value that follows a β distribution in which ω is bounded between 0 and 1 (i.e. there is no positive selection) (Yang *et al*., 2000). Model 8 adds to model 7 one additional category of sites with a freely estimated ω, referred to as ω_1 (Yang *et al*., 2000). Models 0 and 3 are frequently compared (Swanson *et al*., 2001; Woelk *et al*., 2001; Brault *et al*., 2002; Kusumi *et al*., 2002), which, according to Anisimova *et al*. (2001), is more a test of rate heterogeneity than of positive selection. Models 7 and 8 are also frequently compared (Haydon *et al*., 2001; Swanson *et al*., 2001; Woelk *et al*., 2001; DeFilippis *et al*., 2002; Kusumi *et al*., 2002), which, according to Anisimova *et al*. (2001), is a strict test of positive selection (if $\omega_1 > 1$). Because all of the data we examined were simulated under conditions corresponding to a constant ω across all sites with an expected mean of 1, we performed both model comparisons using the same data, where the data conformed to the null hypotheses of both models 0 and 7.

As shown in Table 1, for data simulated under the neutral coalescent model without recombination, the false-positive error rate was 0% for the comparison of models 0 and 3, and 1% for the comparison of models 7 and 8 (the significance level α was 5%). These findings are in agreement with those of Anisimova *et al*. (2001) that the likelihood-ratio test is

https://doi.org/10.1017/S0016672303006128 Published online by Cambridge University Press

Table 2. *Estimates of the numbers of segregating sites,* S*, and the recombination rate,* R*, and the standard errors of the means* (*SEM*). *Estimates of* S *and* R *were obtained from DnaSP version 3.53* (*Rozas & Rozas, 1999*). R *is the estimate from Hudson* (*1987*) *and is equivalent to our* ρ

	S	SEM	*R*	SEM
$\rho = 0$				
False positive ($n = 1$)	208		0	
True negative ($n = 99$)	85·53	3·59	12·14	1·90
$\rho = 18$				
False positive ($n = 68$)	97·12	2·28	41·33	3·95
True negative ($n = 32$)	87·66	4·15	21·55	2·47
$\rho = 105$				
False positive ($n = 98$)	94·59	1·40	205·17	25·95
True negative ($n = 2$)	94·00	9·00	107·20	72·80

conservative. By contrast, for data simulated under the neutral coalescent model with an intermediate recombination rate, the false-positive error rate was 7% when comparing model 0 to model 3 and 68% when comparing model 7 to model 8. For data simulated under the neutral coalescent model with a high recombination rate, the false-positive error rate was 4% when comparing model 0 to model 3 and 98% when comparing model 7 to model 8. Thus, unacknowledged recombination inflated the false-positive error rate, leading to erroneous rejection of neutrality and false acceptance of the presence of positive natural selection, for the model 7 vs model 8 test of positive selection. By contrast, the model 0 vs model 3 test of rate heterogeneity was relatively robust to violation of the assumption of no recombination.

In Table 2, we report the number of segregating sites, *S*, and the per-locus population-scaled recombination rate *R*, as estimated using DnaSP (Rozas & Rozas, 1999). We partitioned the estimates by false positives and true negatives. As expected, *S* is unaffected by the recombination rate, whereas *R* is known to be biased upwards (Hey & Wakeley, 1997). This problem notwithstanding, false-positive cases had larger estimates of *R*, suggesting that datasets that experienced more recombination events had a greater tendency to be false-positive cases.

In Table 3, we report the means and 95% confidence intervals (CIs) of ω from the three sets of simulations. ω is reported as the average value across the entire sequence length. Recombination increased the mean ω for both models 0 and 3 by approximately equal amounts. By contrast, recombination decreased the mean ω for model 7 and increased the mean ω for model 8. These patterns explain the different effects of unacknowledged recombination on the accuracies of the two likelihood-ratio tests. All mean values of ω in the presence of recombination were significantly different from the expected value of 1. The 95% CIs illustrate the large variance around the expected mean of 1 under the conditions of our simulations. Recombination had, at best, a marginal effect on variance. We also found that the mean ω was biased even in the absence of recombination, as was previously noted (Yang & Nielsen, 2000).

In Table 4, we report the average p_1 value (the average proportion of sites in the category of sites with a freely estimated ω), the average ω_1 value (the average ω value for the category of sites with a freely estimated ω), the average $\bar{\omega}$ value (the average ω value across all sites) and the tree length for the three simulations, separated by false positives and true negatives. For all false-positive cases, ω_1 was greater than 1, thereby suggesting the presence of positively selected sites. In the presence of recombination, the average $\bar{\omega}$ was significantly greater in the false-positive cases than in the true-negative cases ($P < 0{\cdot}0001$ for $\rho = 18$ and $P = 0{\cdot}0193$ for $\rho = 105$, Wilcoxon rank sums test). The total tree length, as measured by the number of steps, was significantly greater in the false-positive cases for $\rho = 18$ ($P = 0{\cdot}0003$, Wilcoxon rank sums test). Furthermore, the total tree length increased on average with an increasing recombination rate (87 with $\rho = 0$ vs 116 with $\rho = 18$ vs 143 with $\rho = 105$, $P < 0{\cdot}0001$ for all three pairwise comparisons, Wilcoxon rank sums test), suggesting that more recombination led to more inferred homoplasies.

CODEML overlays an empirical Bayesian approach on the maximum-likelihood estimates in order to calculate the posterior probabilities of any given site belonging to the category of positively selected sites (Nielsen & Yang, 1998). Figure 1 depicts the numbers of sites identified with a 95% or higher posterior probability of belonging to the category of positively selected sites, as well as the ω_1 value for the category of positively selected sites. Fig. 1a shows the results from the 68 false-positive model 8 cases with $\rho = 18$. ω_1 ranged from 1·95 to 32·34 (values that were all outside the 95% CIs of $\bar{\omega}$ shown in Table 3), whereas our expectation was $\omega = 1$. The number of positively selected sites ranged from 1 to 88, corresponding to ~41% of sites, whereas our expectation was 0. Fig. 1b shows the results from the 98 false-positive model 8 cases with $\rho = 105$. ω_1 ranged from 1·87 to 9·78, and the number of positively selected sites ranged from 0 to 80.

3. Discussion

We investigated the effects of unacknowledged recombination on the ability of the CODEML method to detect sites experiencing positive selection. Using values of θ and ρ corresponding to *in vivo* estimates from HIV-1 and *Drosophila*, an increase in the false-positive rate was observed for both pairs of model

https://doi.org/10.1017/S0016672303006128 Published online by Cambridge University Press

Table 3. *Effects of recombination on means and 95% confidence intervals (CIs) of* ω. P *values resulted from a Wilcoxon signed-rank test against an expected mean of 1. Values significant at the 5% level are indicated with asterisks*

		$\rho=0$	$\rho=18$	$\rho=105$
Model 0	Mean ω	1·0468 ($P=0{\cdot}451$)	1·1104 ($P=0{\cdot}002$*)	1·0890 ($P=0{\cdot}032$*)
	95% CIs	0·6103, 1·8967	0·7153, 1·8004	0·6236, 1·9163
Model 3	Mean ω	1·0468 ($P=0{\cdot}451$)	1·1166 ($P=0{\cdot}002$*)	1·0925 ($P=0{\cdot}026$*)
	95% CIs	0·6103, 1·8967	0·7152, 1·8004	0·6236, 1·9163
Model 7	Mean ω	0·8924 ($P<0{\cdot}001$*)	0·7385 ($P<0{\cdot}001$*)	0·5167 ($P<0{\cdot}001$*)
	95% CIs	0·5906, 1·0000	0·5000, 1·0000	0·3538, 0·8935
Model 8	Mean ω	1·0466 ($P=0{\cdot}424$)	1·1517 ($P<0{\cdot}001$*)	1·1652 ($P<0{\cdot}001$*)
	95% CIs	0·5423, 1·9081	0·7231, 1·8691	0·6071, 2·0764

Table 4. *Effects of recombination on Model 8 parameter estimates.* p_1 *indicates the average of the proportion of sites in the category of sites with a freely estimated* ω *value.* ω_1 *indicates the average for the category of sites with the freely estimated* ω *value.* $\bar{\omega}$ *indicates the average across all sites. 'Steps' indicates the average tree length in the parsimony sense in which all inferred mutational events are equally weighted*

	p_1	ω_1	$\bar{\omega}$	Steps
$\rho=0$				
False positive ($n=1$)	0·07	3·73	0·91	206
True negative ($n=99$)			1·05	85
$\rho=18$				
False positive ($n=68$)	0·30	4·72	1·24	124
True negative ($n=32$)			0·95	101
$\rho=105$				
False positive ($n=98$)	0·28	3·98	1·18	143
True negative ($n=2$)			0·65	142

comparisons in the presence of unacknowledged recombination. The reason why the CODEML method errs in the presence of recombination appears to lie within the phylogenetic nature of the approach. Conceptually, nonsynonymous and synonymous changes are counted along all of the branches of the phylogeny for each individual site. Highly variable sites experiencing multiple different nonsynonymous changes are interpreted as signifying divergent evolution, whereas highly variable sites experiencing multiple identical nonsynonymous changes are interpreted as signifying parallel evolution. If recombination is unaccounted for by the evolutionary model, the homoplasies induced by recombination are treated as real mutational events and, as a consequence, the reconstructed phylogeny becomes longer. Notice that the determination of whether a homoplasy is nonsynonymous or synonymous only depends upon the initial mutational event. If early and/or frequent recombination events were randomly to affect an asymmetric distribution of nonsynonymous mutations over synonymous mutations then an excess of nonsynonymous homoplasies would be propagated throughout the tree (thereby increasing ω_1), leading to the false appearance of highly variable sites with multiple nonsynonymous changes. We expected 74·5% potential nonsynonymous sites and 25·5% potential synonymous sites in all of our neutral simulations, so that our neutral expectation was for an excess in the absolute number of nonsynonymous mutations.

From the trends revealed by the simulations, we can formulate expectations for the extreme case of free recombination. For the comparison of models 0 and 3, it appears that the likelihood-ratio test becomes less conservative as the false-positive rate approaches the designated significance level. For the comparison of models 7 and 8, the false-positive rate approaches 100%. The comparison of models 7 and 8 revealed a greater false-positive rate than the comparison of models 0 and 3. One possible explanation for this finding is that models 0, 3 and 8 allow for sites with $\omega>1$, whereas model 7 does not. Model 7 assumes a strict boundary between neutrality, where $\omega=1$, and positive selection, where $\omega>1$. Given the large variance (Table 3, Fig. 1), perhaps model 7 would be better specified as having all sites with ω bounded between 0 and 1·8, for example, and then specify model 8 as having some proportion of sites with ω bounded between 0 and 1·8, and the rest of the sites with $\omega>1{\cdot}8$.

As an alternative to the current procedure, it has been suggested (Urwin *et al.*, 2002) that one possible way to handle the explicit requirement of a phylogeny by CODEML in the presence of recombination is to input a star phylogeny. In a star phylogeny, all lineages diverge from the root node, such that all mutations are independently derived. This approach was intended to remove the effects of the phylogeny on CODEML's estimations and to determine the extent of false-positive signal in the extreme case of no phylogenetic history. Urwin *et al.* stated that 'The fact that positive selection was still observed at these same sites after removing the effect of phylogenetic

https://doi.org/10.1017/S0016672303006128 Published online by Cambridge University Press

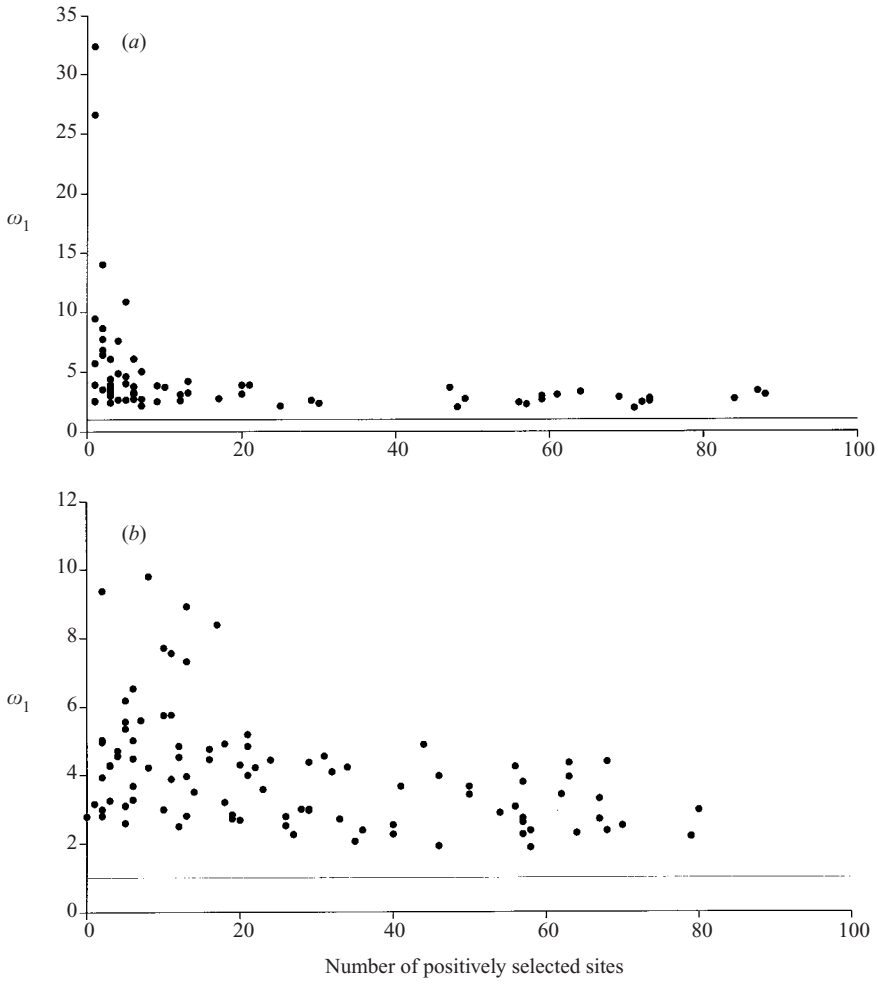

Fig. 1. Bayesian prediction of positively selected sites for false-positive model 8 cases with $\rho = 18$ (a) and $\rho = 105$ (b). The numbers of sites in each category with $\omega > 1$ with posterior probabilities greater than 95% are plotted. The horizontal line indicates the neutral expectation of $\omega = 1$.

history by assuming a star phylogeny indicated that the analysis was unlikely to have been greatly biased by recombination.' However, this approach is flawed. There is phylogenetic history, even in the presence of a high recombination rate. The optimization principle of phylogenetic reconstruction ensures that a phylogeny with fewer steps than the star phylogeny will be reconstructed, except for cases in which the data were derived from a star phylogeny, as would be the case for rapid population growth. In other words, the number of steps on a star phylogeny will usually overestimate the true number of steps. Therefore, the inferred number of mutational events will be overestimated. Consequently, the false-positive error rate might be overestimated. Furthermore, this approach only considers the extreme case of a star phylogeny. As such, it does not allow for testing against intermediate levels of recombination. Finally, no phylogeny will have more steps than a star phylogeny. Thus, the false-positive distribution generated from a star phylogeny maximally delimits the proportion of positively selected sites and their ω value, ultimately yielding a conservative likelihood-ratio test. However, the identification of sites that might have experienced positive selection should require them to be outside the false-positive distribution, not within it. That is, the false-positive distribution should account for the sites potentially affected by recombination. Then, exclusion of these sites should leave the sites potentially affected by positive selection.

In consideration of the large confidence intervals around the mean ω value, estimates of the variance should accompany attempts to interpret ω. To this end, we propose as an alternative procedure that parametric bootstrapping methods, such as those used in this work, be used to generate appropriate null distributions. This would involve first estimating the population-scaled mutation and recombination rates from sequence data using a program such as RECOMBINE (Kuhner *et al.*, 2000), which provides maximum-likelihood estimates of these two parameters. Then, neutral data with recombination could be simulated under the coalescent conditioned upon these rate estimates. These simulated data can then be

https://doi.org/10.1017/S0016672303006128 Published online by Cambridge University Press

analysed by CODEML to determine the proper null distributions.

There are two caveats about this procedure. First, RECOMBINE explicitly assumes no selection. Therefore, we suggest excluding nonsynonymous mutations from this analysis and performing the rate estimates on just synonymous mutations. Second, because this procedure involves estimating unknown θ and ρ values from real data, uncertainties in the point estimates are potentially problematic. For example, RECOMBINE's estimate of the recombination rate is biased upwards, particularly when θ and ρ are small (Kuhner *et al*., 2000). Conditioning upon the upper 95% confidence limits of the estimated rates maximizes the possible contribution of recombination. Consequently, exceeding these rates would be expected to yield a conservative test for positive selection. Conversely, exceeding the lower 95% confidence limits would be expected to yield a liberal test for positive selection. Although it seems reasonable that conditioning upon maximum-likelihood point estimates should provide a valid hypothesis test, simulations should be carried out on a case-by-case basis to investigate the extent of this problem.

Caution is clearly warranted when interpreting the results of CODEML analyses. Previous studies that reported evidence for positive selection should be revisited, particularly for data derived from sources reasonably expected to have experienced recombination. With respect to HIV studies, sites identified as experiencing positive selection may be inferred to reflect previously known epitopes and to identify previously unknown epitopes under strong immune pressure (Suzuki & Gojobori, 1999; Yamaguchi-Kabata & Gojobori, 2000). The identification of HIV-1 cytotoxic T-lymphocyte epitopes is thought to be important for vaccine development (Borrow *et al*., 1997; Korber *et al*., 2001); hence, the misidentification of sites experiencing positive selection, but which are actually experiencing recombination, is of great concern.

We thank Brian Charlesworth for helpful advice and the two anonymous reviewers for their comments. This work was supported by grants from the US Public Health Services including support from the University of Washington Center for AIDS Research. DS was a Howard Hughes Medical Institute Predoctoral Fellow.

References

Allsopp, M. T. E. P., Dorfling, C. M., Maillard, J. C., Bensaid, A., Haydon, D. T., van Heerden, H. & Allsopp, B. A. (2001). *Ehrlichia ruminantium* major antigenic protein gene (*map1*) variants are not geographically constrained and show no evidence of having evolved under positive selection pressure. *Journal of Clinical Microbiology* **39**, 4200–4203.

Andolfatto, P. & Przeworski, M. (2000). A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156**, 257–268.

Anisimova, M., Bielawski, J. P. & Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution* **18**, 1585–1592.

Anisimova, M., Bielawski, J. P. & Yang, Z. (2002). Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution* **19**, 950–958.

Baum, J., Ward, R. H. & Conway, D. J. (2002). Natural selection on the erythrocyte surface. *Molecular Biology and Evolution* **19**, 223–229.

Borrow, P., Lewicki, H., Wei, X., Horwitz, M. S., Peffer, N., Meyers, H., Nelson, J. A., Gairin, J. E., Hahn, B. H., Oldstone, M. B. A. & Shaw, G. M. (1997). Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nature Medicine* **3**, 205–211.

Brault, A. C., Powers, A. M., Holmes, E. C., Woelk, C. H. & Weaver, S. C. (2002). Positively charged amino acid substitutions in the E2 envelope glycoprotein are associated with the emergence of Venezuelan equine encephalitis virus. *Journal of Virology* **76**, 1718–1730.

Comeron, J. M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution* **41**, 1152–1159.

DeFilippis, V. R., Ayala, F. J. & Villareal, L. P. (2002). Evidence for diversifying selection in human papillomavirus type 16 E6 but not E7 oncogenes. *Journal of Molecular Evolution* **55**, 491–499.

Haydon, D. T., Bastos, A. D., Knowles, N. J. & Samuel, A. R. (2001). Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics* **157**, 7–15.

Hey, J. & Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research* **50**, 245–250.

Kao, H.-W. & Lee, S.-C. (2002). Phosphoglucose isomerases of hagfish, zebrafish, gray mullet, toad, and snake, with reference to the evolution of the genes in vertebrates. *Molecular Biology and Evolution* **19**, 367–374.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C. & Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin* **58**, 19–42.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.

Kusumi, J., Tsumura, Y., Yoshimaru, H. & Tachida, H. (2002). Molecular evolution of nuclear genes in Cupressacea, a group of conifer trees. *Molecular Biology and Evolution* **19**, 736–747.

Li, W.-H., Wu, C.-I. & Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**, 150–174.

Li, W.-H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**, 96–99.

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.

Nei, M. & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

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

Pamilo, P. & Bianchi, N. O. (1993). Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Molecular Biology and Evolution* **10**, 271–281.

Ross, H. A. & Rodrigo, A. G. (2002). Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *Journal of Virology* **76**, 11715–11720.

Rozas, J. & Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175.

Schierup, M. H. & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879–891.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L. & Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology* **73**, 10489–10502.

Suzuki, Y. & Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **16**, 1315–1328.

Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. (2001). Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proceedings of the National Academy of Sciences of the USA* **98**, 2509–2514.

Swofford, D. (2002). PAUP*: *Phylogenetic Analysis Using Parsimony* (*and Other Methods), version 4.0b10. Sinauer Associates, Sunderland, MA.

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

Urwin, R., Holmes, E. C., Fox, A. J., Derrick, J. P. & Maiden, M. C. J. (2002). Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB. *Molecular Biology and Evolution* **19**, 1686–1694.

Wall, J. D. (1999). Recombination and the power of statistical tests of neutrality. *Genetical Research* **74**, 65–79.

Woelk, C. H., Jin, L., Holmes, E. C. & Brown, D. W. G. (2001). Immune and artificial selection in the haemagglutinin (H) glycoprotein of measles virus. *Journal of General Virology* **82**, 2463–2474.

Yamaguchi-Kabata, Y. & Gojobori, T. (2000). Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *Journal of Virology* **74**, 4335–4350.

Yang, Z. & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32–43.

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

Zanotto, P. M., Kallas, E. G., de Souza, R. F. & Holmes, E. C. (1999). Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* **153**, 1077–1089.