DOI: 10.1534/genetics.109.109736

Coalescent Simulation of Intracodon Recombination

Miguel Arenas and David Posada[1]

Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, 36310 Vigo, Spain

Manuscript received September 14, 2009
Accepted for publication November 14, 2009

ABSTRACT

The coalescent with recombination is a very useful tool in molecular population genetics. Under this framework, genealogies often represent the evolution of the substitution unit, and because of this, the few coalescent algorithms implemented for the simulation of coding sequences force recombination to occur only between codons. However, it is clear that recombination is expected to occur most often within codons. Here we have developed an algorithm that can evolve coding sequences under an ancestral recombination graph that represents the genealogies at each nucleotide site, thereby allowing for intracodon recombination. The algorithm is a modification of Hudson's coalescent in which, in addition to keeping track of events occurring in the ancestral material that reaches the sample, we need to keep track of events occurring in ancestral material that does not reach the sample but that is produced by intracodon recombination. We are able to show that at typical substitution rates the number of nonsynonymous changes induced by intracodon recombination is small and that intracodon recombination does not generally result in inflated estimates of the overall nonsynonymous/synonymous substitution ratio (ω). On the other hand, recombination can bias the estimation of ω at particular codons, resulting in apparent rate variation among sites and in the spurious identification of positively selected sites. Importantly, in this case, allowing for variable synonymous rates across sites greatly reduces the false-positive rate and recovers statistical power. Finally, coalescent simulations with intracodon recombination could be used to better represent the evolution of nuclear coding genes or fast-evolving pathogens such as HIV-1.We have implemented this algorithm in a computer program called *NetRecodon*, freely available at http://darwin.uvigo.es.

THE coalescent (Kingman 1982; Hudson 1990) provides an efficient sampling of genealogical histories from a theoretical population evolving under a neutral Wright–Fisher model (Ewens 1979; Kingman 1982; Hudson 1990). Coalescent simulations are commonly used in molecular population genetics to understand the behavior and interactions among evolutionary processes under different scenarios (Innan *et al.* 2005), such as hypothesis testing (DeChaine and Martin 2006), evaluation and comparison of different analytical methods (Carvajal-Rodriguez *et al.* 2006), or estimation of population genetic parameters (Beaumont *et al.* 2002). Indeed, to obtain meaningful biological inferences from these simulations, it is very important that the underlying model is as realistic as possible. In this regard, a number of models have been developed during the last decade that consider different evolutionary processes such as recombination (Simonsen and Churchill 1997; Wiuf and Posada 2003), gene conversion (Wiuf and Hein 2000), selection (Hudson and Kaplan 1988, 1995), and gene flow or demographic history (Slatkin 1987; Pybus and Rambaut 2002).

Despite these advances, and in the face of a plethora of coalescent simulators (Excoffier *et al.* 2000; Hudson 2002; Posada and Wiuf 2003; Spencer and Coop 2004; Mailund *et al.* 2005; Schaffner *et al.* 2005; Marjoram and Wall 2006; Arenas and Posada 2007; Hellenthal and Stephens 2007; Liang *et al.* 2007), it was not possible until very recently to simulate recombining protein-coding DNA sequences within this framework (Anisimova *et al.* 2003; Arenas and Posada 2007). Importantly, to our knowledge, the algorithms described or implemented so far allow recombination only between codons, not within them. The reason for this unrealistic constraint is that standard codon models describe the probabilities of change along a lineage from one codon to another (Yang 2006), whereas recombination can occur between any two nucleotides, potentially resulting in one or more lineages not being shared by all the positions of the codon. In other words, although the unit for substitution in coding sequences is the codon, the unit for recombination in these sequences is still the nucleotide. Here we describe a new algorithm that overcomes this limitation by allowing for the evolution of different positions of the same

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.109736/DC1.

[1] *Corresponding author:* Departamento de Bioquímica, Genética eInmunología, Universidad de Vigo, 36310 Vigo, Spain.
E-mail: dposada@uvigo.es

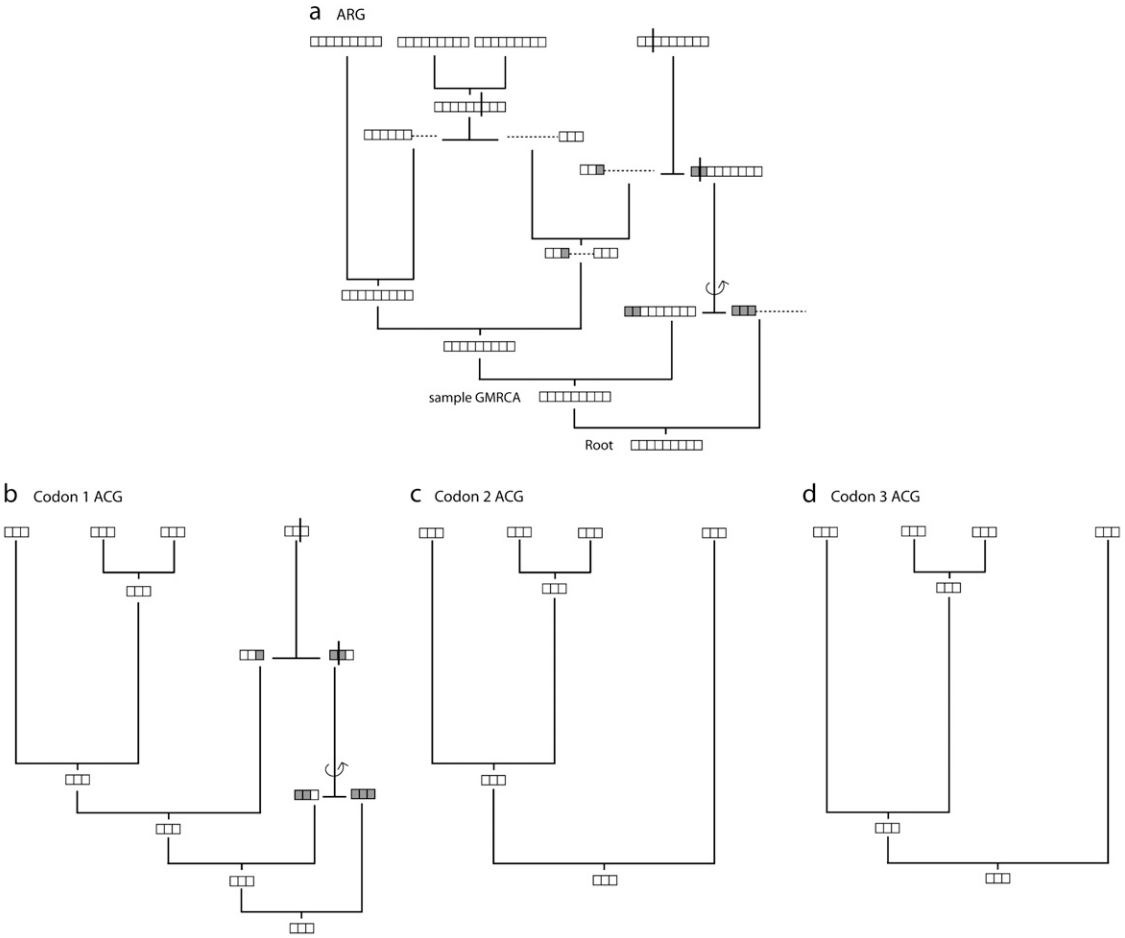

Figure 1.—Generation of ACGs for a coding sequence with three codons. (a) ARG for the whole sequence. Note that the GMRCA of the sample is younger than the root. Inside each node we can see the "sampled" ancestral material (open blocks), the "unsampled" ancestral material (shaded blocks), and the non-ancestral material (dotted lines). Vertical lines across the segments indicate recombination breakpoints. Three recombination breakpoints occur in the ARG: after the first and second positions of the first codon and between the second and third codon. The two intracodon recombination events result in a reticulated ACG for codon 1 (b), while for codons 2 (c) and 3 (d), the ACG are binary trees.

codon in distinct genealogies. Furthermore, we use this algorithm to evaluate the effect of intracodon recombination on the generation of nonsynonymous (NS) diversity and on the estimation of the ratio of nonsynonymous-to-synonymous substitution rates (ω or d_N/d_S) (Li and Gojobori 1983) and the hypotheses derived from it.

METHODS

Simulation of intracodon recombination under the coalescent: The simulation of intracodon recombination occurs in two independent steps: the construction of the ancestral recombination graph (ARG) (Griffiths 1991; Griffiths and Marjoram 1996, 1997) and the evolution of the coding sequences. The first step is an extension of the standard coalescent *m*-loci continuous-time model with recombination (Kaplan and Hudson 1985). The novelty comes from the fact that intracodon recombination, apart from breaking the ancestral material in two segments as in the case of intercodon recombination, also results in ancestral material that never reaches the sample (Figure 1). Therefore, we distinguish between "sampled" and "unsampled" ancestral material, while in the standard coalescent all the ancestral material appears in the sample. Because sampled and unsampled ancestral material can meet in the same codon, we need to be sensitive to recombination and substitution events occurring not only in the sampled ancestral material, as usual, but also in the unsampled ancestral material. We also tried a shortcut in which recombination events were allowed only in the sampled material and obtained almost identical results as with the full algorithm. However, this simplified algorithm was discarded because it did not reduce the computational costs much because, in practice, recombination events within unsampled material were very rare. Because we keep track of the relationships between recombinant segments containing ancestral material, we are able to define at the end the exact genealogy for each codon site, which we call

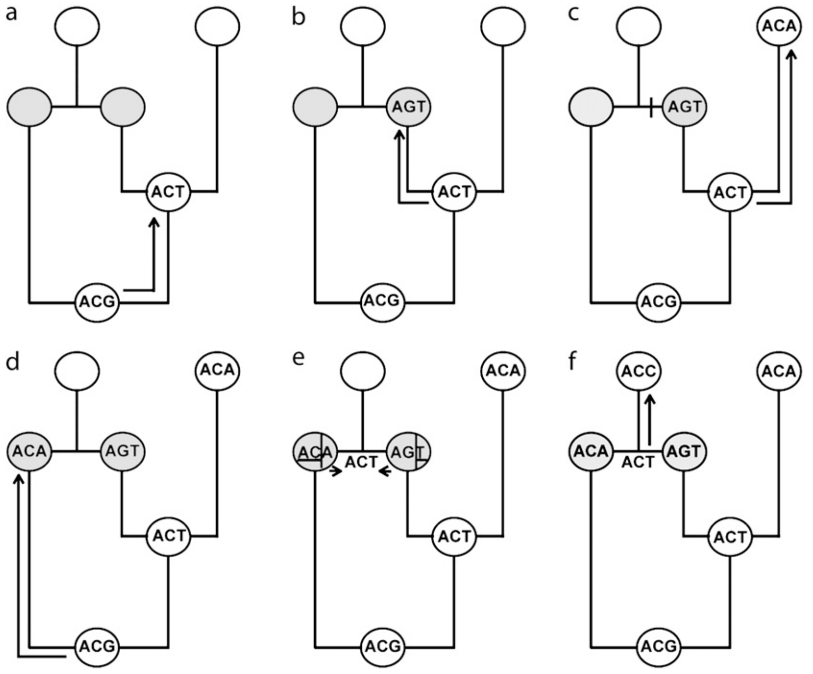

FIGURE 2.—An example of codon evolution along the ACG. Open and shaded circles correspond to coalescence and parental nodes, respectively. (a) Starting from the GMRCA, the codon is evolved between nodes according to the probabilities specified by the codon model and the branch length. (b) The process then encounters a parental node, and because the other parental node has not been assigned a codon yet, it waits there. (c) The algorithm continues its recursion toward the present. (d) The process encounters a parental node, and because the other parental node has already been assigned a codon, (e) it combines the two codons according to the recombination breakpoint. (f) Finally, the resulting recombinant codon (ACT) is evolved.

the ancestral codon graph (ACG) (Figure 1). Note that, in the presence of recombination, the ACGs along the alignment can be different from each other. Moreover, for codons that contain recombination breakpoints, the ACGs are reticulated graphs or networks, while, for the nonrecombining codons, the ACGs are binary trees. The algorithm for building the ARG is as follows:

1. Start with k nodes $= n$ sampled sequences. Each node contains a single segment encompassing all the codons considered.
2. Sample the time back to the next event from an exponential distribution with the parameter $k(k-1)/2 + 2Nrg$, where N is the effective population size, r is the recombination rate per site per generation, and g is the number of possible breakpoint sites summed over all sequences in the sample. A breakpoint site is a potential recombining site only when it has non-MRCA (most recent common ancestor) ancestral material (sampled or unsampled; see below) before and after it (see WIUF and HEIN 1999).
3. Choose the type of event. It will be a coalescent event (CA) with probability $[k(k-1)/2)]/[k(k-1)/2 + 2Nrg]$ and a recombination event (RE) with probability $2Nrg/[k(k-1)/2 + 2Nrg]$.
4. Complete the event. If it is a CA event, select two nodes at random and merge them into a new ancestral node inheriting all the segments from both coalescing and descendant nodes and set $k = k - 1$. If it is a RE event, draw a random site among all possible breakpoints across all segments in all nodes. Cut all the segments in the (recombinant) node that contains the chosen (breakpoint) site into left and right segments. If the event occurs between codons, create two ancestral parental nodes that will inherit either the left or the right segments and set $k = k + 1$. If the event occurs within a codon, create two ancestral parental nodes that will inherit the left and right segments (ancestral material that does reach the sample or "sampled material") and the left and right site(s) flanking the breakpoint in the same codon (ancestral material that never reaches the sample or "unsampled material"), set $g = g + 3$ and $k = k + 1$. In both cases, keep the location and ancestral relationships of every segment.
5. If $k = 1$, label this node as the root and end the process; otherwise, go to step 2.

Importantly, note that intracodon recombination increases the amount of (unsampled) ancestral material by three nucleotides.

In a second step, each codon is evolved independently along its ACG, starting at the root, which contains all the sampled and unsampled ancestral material (Figure 1). Note that in the presence of intracodon recombination the grand most recent common ancestor (GMRCA) (GRIFFITHS and MARJORAM 1996) of the sample is not necessarily the root node. Given branch lengths and a Markov model of codon evolution, it is straightforward to calculate the probabilities of change between any two nodes and to use them to evolve the codons along a nonreticulated ACG (*i.e.*, a tree) (YANG 2006). If the ACG is reticulated, the process is slightly more involved. The general recursion proceeds as follows, independently for each codon position (Figure 2):

1. Starting at the root, choose a random codon by sampling it from the equilibrium distribution.

2. Evolve this codon to produce new codons at the descendant nodes.
3. For every node with an assigned codon that has not been evolved yet:
    a. If the node is a tip, do nothing else.
    b. If the node is a coalescent node, evolve its codon to produce new codons at the two descendant nodes.
    c. If the node is a parental node, check whether both parentals involved in the same recombination event have been already assigned a codon. If not, wait until this assignment is made. Once both parentals have been assigned a codon, combine both codons according to the breakpoint location and evolve the resulting codon to produce a new codon at the descendant recombinant node. If a codon stop is generated, erase all codons in the ACG and go back to step 1.
4. If all the nodes in the ACG have been assigned a codon, stop.

Note that, in step 3c, if a codon stop is created, we start the substitution process again from the root, keeping the simulated genealogy to afford computational costs and to avoid favoring smaller genealogies because larger genealogies tend to have more recombination events and thus a higher chance of generating stop codons.

Algorithm development and validation: The algorithm was implemented in C in a program called *NetRecodon*, which is a major redraft of *Recodon* (Arenas and Posada 2007). To validate it, we compared the results obtained with this algorithm with the theoretical expectations for the mean and variances of different simulation statistics, such as the number of recombination events or the time to the most recent common ancestor (Hudson 1990). We also checked whether these summary statistics agreed with those obtained with the *ms* program (Hudson 2002) under different evolutionary scenarios. In addition, substitution and codon model parameters were estimated from the simulated data using HYPHY (Kosakovsky Pond *et al.* 2005) and PAUP* (Swofford 2000). These estimates agreed very well with the expected values from the simulations. Moreover, we implemented additional features that correspond to a variety of real scenarios, such as exponential growth, migration, longitudinal samples (dated tips), haploid/diploid populations, and a broad set of codon models that allow ω to change along the sequences according to different distribution (see Yang *et al.* 2000).

Simulation of protein-coding sequences with recombination: *Global* ω*:* We simulated coding sequences under different values of the population mutation parameter [$\theta = 4N\mu l = 10, 20, 50, 100$, and 200, where N is the effective (diploid) population size, μ is the substitution rate per codon, and l is the number of codons], recombination rates ($\rho = 4Nrl = 0, 1, 4, 16, 64$, and 128; where r is the recombination rate per nucleotide), and d_N/d_S ratios (ω = 0.2, 1.0, and 5.0). The number of sequences in the sample ($n = 10$), alignment length (l = 333 codons), and effective population size (N = 1000) was constant. The codon model used was GY94 (Goldman and Yang 1994) with a transition/transversion ratio of 0.5, and under the standard genetic code. For every combination of parameters ($5 \times 6 \times 3 = 90$ combinations), we simulated recombination in two different ways. In one setting, we used the algorithm introduced above, where recombination was free to occur within and between codons. In addition, we also repeated the simulations but forced recombination to occur only between codons because this setting had been previously used to understand the impact of recombination on the detection of positive selection (Anisimova *et al.* 2003; Shriner *et al.* 2003). Indeed, we made sure that the total recombination rate, ρ, was the same in both situations. For every scenario, we simulated 200 alignments.

The global ω was estimated from the simulated data using the Nei and Gojobori method (NG86) (Nei and Gojobori 1986) as implemented in SNAP (Korber 2000) and maximum likelihood under the GY94 model as implemented in HYPHY. The NG86 is a very simple method, commonly used for closely related sequences, which does not use phylogenetic information. The GY94 requires a phylogeny, which was estimated with the neighbor-joining (NJ) algorithm (Saitou and Nei 1987).

ω *per site:* To understand the effect of recombination on a site-by-site basis, we also performed some simulations parameterized according to a real data set—an HIV-1 *env* 2007 subtype reference alignment (A–K, without recombinants)—downloaded from the Los Alamos National Laboratory HIV sequence database (http://hiv-web.lanl.gov) with 40 sequences and 956 codons. In particular, we studied the effect of (inter- and intra-) codon recombination on the M0 (one rate) and M1 (neutral) likelihood-ratio test (LRT) for homogeneity of ω across sites (see Yang *et al.* 2000) and on the identification of positively selected sites (PSS) ($\omega > 1$).

The estimated nucleotide diversity was 0.16, which roughly corresponds to θ = 250 under our settings. According to PAML (Yang 2007) under the M0, M1, and M8 models, $\omega_{M0} = 0.5$ (but we also studied cases with $\omega_{M0} = 1.0$ and $\omega_{M0} = 2.0$), $\omega_{M1} = 0.1$, $p0_{M1} = 0.6$, $p0_{M8} = 0.9$, $p_{M8} = 0.2$, $q_{M8} = 0.3$, and $\omega_{M8} = 3.8$—the values used in the simulations. As before, we assumed that $\rho = 0, 1, 4, 16, 64$, and 128, $l = 999$ nt, and N=1000, but we increased the sample size ($n = 30$) because these models are more complex. The number of replicates was 200 for M0 and M1 and 2000 for M8. For each data set, maximum-likelihood phylogenetic trees were estimated using PHYML (Guindon and Gascuel 2003)

TABLE 1

Effect of the substitution rate on the number of nonsynonymous changes induced by recombination

		Type of recombination events (%)				
		Intercodon	Intracodon			
θ	Observed nucleotide diversity (%)	0 NS	0 NS	1 NS	2 NS	% NS changes induced by recombination
10	1	32.6	66.0	1.4	0.0	4.6
20	2	32.5	64.5	2.9	0.1	6.1
50	5	32.5	60.5	6.7	0.3	7.0
100	9	32.6	54.7	11.9	0.8	7.1
200	16	32.6	47.3	17.9	2.2	6.3
500	36	32.7	33.8	27.3	6.2	4.9

Results shown correspond to $\rho = 64$ (~170 observed recombination events per replicate) and to $\omega = 1.0$ for different values of the substitution rate (θ). The last column shows the proportion of the total nonsynonymous (NS) changes observed in the history of the sample that have been produced by intracodon recombination.

and fed into PAML to obtain model likelihoods. Results for the M0 and M1 LRTs were classified as true negatives (TN) or as false positives (FP; *P*-value < 0.05) when data were simulated under M0 and as true positives (TP) or as false negatives (FN) when data were simulated under M1. We used this classification to calculate the false-positive rate [$\text{FPR} = \text{FP}/(\text{TN} + \text{FP})$] and the power [$\text{TP}/(\text{TP} + \text{FN})$] of the LRTs. To identify PSS, we used fixed-effects likelihood (FEL) model (Kosakovsky Pond and Frost 2005b) assuming a "one-rate" (FEL-1R; d_S is held constant across sites) and a "two-rate" (FEL-2R; d_S is adjusted across sites) model, as implemented in HYPHY upon a NJ tree. False-positive rates and power were calculated as before, but on a site-by-site basis in the context of the M8 model.

SIMULATION RESULTS

Nonsynonymous recombination: In theory, intracodon recombination can generate new codons that have 0, 1, or 2 NS changes when compared to the parental codons. In the simulations, recombination within codons was indeed twice as common as intercodon recombination, and only at high substitution rates did the proportion of total recombination events that resulted in NS changes reach a significant proportion (12–34%) (Table 1). Nonetheless, the proportion of the total NS changes observed in the history of the sample due to intracodon recombination was always small. This proportion indeed depended on the recombination and substitution rates and reached 10–12% when $\rho = 128$ (~230 recombination events in the history of the sample) (Table 2).

Effect of intracodon recombination on the estimation of ω: We did not observe a significant effect of (intra- or inter-) recombination on the estimation of the global ω (supporting information, Table S1, Table S2, and Table S3). However, at high substitution rates, some significant instances were detected in which increasing recombination rates resulted in estimates of ω closer to 1, regardless of its simulated (true) value. Both methods employed to estimate ω worked better in the presence of elevated substitution rates, although the phylogenetic GY94 model showed a slight tendency to overestimate ω compared to NG86, especially when θ was large.

Effect of intracodon recombination on the LRT for ω variation across sites: Recombination clearly biased the M0 and M1 LRTs toward the rejection of the null hypothesis of homogeneity of ω across sites (Figure 3). Increasing (intra-or inter-) recombination rates ele-

TABLE 2

Effect of the recombination rate on nonsynonymous change

	% NS changes induced by recombination		
ρ	$\omega = 0.2$	$\omega = 1.0$	$\omega = 5.0$
		$\theta = 50$	
0	0.0	0.0	0.0
1	0.2	0.2	0.2
4	0.6	0.5	0.5
16	2.1	2.0	2.0
64	7.3	7.0	6.9
128	13.6	12.9	12.5
		$\theta = 100$	
0	0.0	0.0	0.0
1	0.2	0.2	0.1
4	0.6	0.6	0.5
16	2.2	2.1	2.0
64	7.5	7.1	6.8
128	13.7	12.7	12.2
		$\theta = 200$	
0	0.0	0.0	0.0
1	0.2	0.2	0.1
4	0.6	0.5	0.5
16	2.2	1.9	1.8
64	7.4	6.3	6.0
128	13.5	11.7	10.8

Values correspond to the proportion of the total nonsynonymous (NS) changes observed in the history of the sample produced by intracodon recombination for different substitution (θ), recombination (ρ), and ω values.

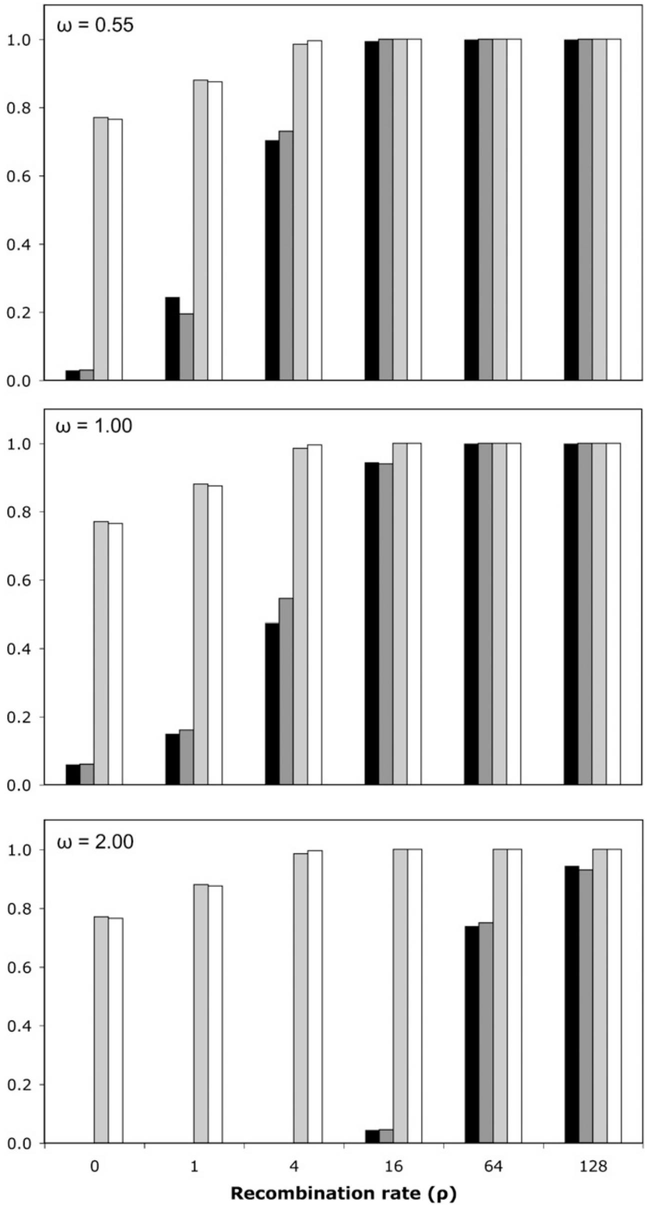

Figure 3.—Performance of the likelihood-ratio test for homogeneous selection pressure across sites in the presence of recombination. Solid and darkly shaded bars indicate the M0 and M1 LRT false-positive rate when data were simulated without/with intracodon recombination, respectively. Lightly shaded and open bars correspond to the power of the LRT for the same two scenarios.

vated both the false-positive rate and power, making these LRTs nonconservative.

Effect of intracodon recombination on the identification of PSS: Recombination significantly increased the number of sites erroneously identified as positively selected by FEL-1R, although the errors were also evident in the absence of recombination. At high recombination rates ($\rho = 64$, 128), the number of false PSSs reached 30–35 (of 333 codons) at the 95% significance level (Figure S1). When the PSSs were identified with FEL-2R, the number of false positives clearly decreased. In this case, the number of false PSSs was tiny when the recombination rate was 0 or very small and increased slowly with higher rates. In all cases, allowing for intracodon recombination did not make a difference. Consequently, the false-positive rate was more or less constant regardless of the recombination rate in the case of FEL-1R and smaller and much more correlated with the recombination rate for FEL-2R (Figure 4). Remarkably, the power to detect PSS was four times higher for FEL-2R than for FEL-1R. In both cases, the recombination rate did not affect the results, except for the fact that the standard errors per replicate decreased at larger recombination rates.

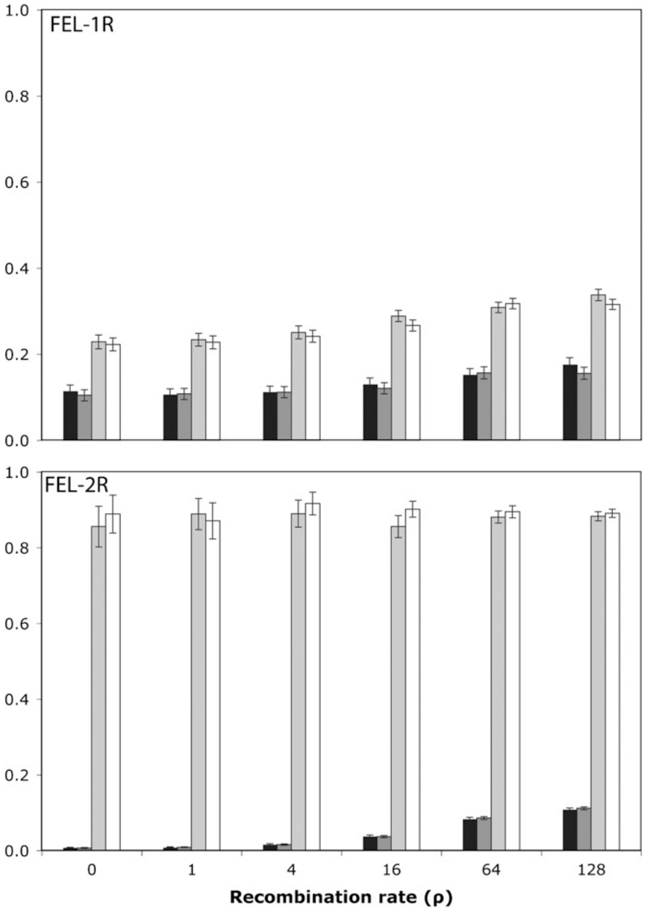

Figure 4.—Performance of the FEL estimator of ω (per site) in the presence of recombination. Data were simulated under a M8 model. Solid and darkly shaded bars indicate the FPR when data were simulated without/with intracodon recombination, respectively. Lightly shaded and open bars correspond to the power for the same two scenarios. (Top) FPR and power per replicate for FEL-1R (2000 replicates). (Bottom) FPR and power per replicate for FEL-2R (200 replicates). Sites identified as PSSs were those with $\omega > 1$ and a *P*-value < 0.05. Error bars indicate 95% confidence intervals per replicate.

DISCUSSION

Simulation is indeed a powerful tool in population genetics, with a rich variety of applications, but most of its benefits rely on biologically meaningful models. The algorithm described here facilitates the generation of more realistic protein-coding sequence samples in the presence of recombination and, importantly, allows for the estimation of the nonsynonymous substitutions induced by recombination, relative to mutation. Here, a necessary assumption is that the genealogy is independent of ω (as in Anisimova *et al.* 2003; Shriner *et al.* 2003; Scheffler *et al.* 2006; Wilson and McVean 2006). Our results suggest that recombination does not have a strong overall effect on the generation of nonsynonymous changes, although this does not mean that it cannot mislead positive selection analyses, especially those based on phylogenies (Anisimova *et al.* 2003). Other authors have studied the impact of recombination on the estimation of (the global) ω. Shriner *et al.* (2003) studied the effect of recombination on the maximum-likelihood phylogenetic methods implemented in PAML (Yang 1997) for the characterization of molecular adaptation. They used the program *ms* (Hudson 2002) to simulate the data, which does not allow for intracodon recombination. Although they found that recombination leads to false-positive detection of sites undergoing positive selection, the effect of recombination on the estimate of ω across the entire sequence length was unclear. This is because, although point estimates were reported as significantly higher than the expected value of 1, in fact the 95% confidence intervals included 1 in most cases. Similarly, Anisimova *et al.* (2003), also ignoring intracodon recombination, did not find a strong effect of recombination on PAML's estimation of ω, although there was some inflation of the number of sites identified as positively selected ($\omega > 1$, PSS). In addition, Kosakovsky Pond *et al.* (2006) found that a single recombination hotspot could increase the number of PSSs detected by the FEL analysis (Kosakovsky Pond and Frost 2005a), but they did not investigate its effect on the estimation of ω. More recently, Kosakovsky Pond *et al.* (2008) did not find a significant effect of a single recombination event on the estimation of ω, although they did find that the PSSs identified were different when the recombination breakpoint was explicitly taken into account. Here we show that considering intracodon recombination does not make a difference in the assessment of the impact of recombination on the estimation of global ω and that, as previously shown, this impact is low.

A different issue is the effect of recombination on the estimation of ω per site. We show that recombination can give the impression that the selection pressure varies along the sequence when in fact it is constant and that some sites appear to be under positive selection when they are not. Obviously, this is especially relevant for those studies trying to understand positive selection across recombining genomes, such as studies of Drosophila, humans, or HIV-1 (Nielsen and Yang 1998; Zanotto *et al.* 1999; De Oliveira *et al.* 2004; Bustamante *et al.* 2005; Clark *et al.* 2007; Nielsen *et al.* 2007). In theory, recombination could inflate the value of ω (global or per site) through two different mechanisms. One mechanism could be the generation of nonsynonymous changes, but we have shown that the number of nonsynonymous changes generated by recombination is low compared to substitution. Indeed, many recombination events occur between identical codons, especially at low substitution rates. The other possible mechanism operates to confound phylogenetic estimation and therefore also those methods that use a phylogeny to estimate this parameter (so, in theory, the NG86 method should not be affected by this bias). Our results suggest that recombination affects selection analyses mainly because it confounds the phylogenies used in those analyses. Indeed, in our simulations, the consideration of intracodon recombination does not change the effects of recombination when only intercodon events, which cannot result in nonsynonymous changes, are allowed. Previously, Anisimova *et al.* (2003) found that just using an incorrect phylogeny can have a severe effect on the comparison of codon models. Indeed, phylogenetic variation across sites can be wrongly interpreted by the LRTs as variation in synonymous and nonsynonymous substitution rates. This can happen because different trees along the alignment can have different heights (see Figure 1, for example), and when d_S is held constant (as it is in most popular codon models), this variation in tree length can be interpreted as variation in d_N and therefore as variation of the d_N/d_S ratios (Anisimova *et al.* 2003). The fact that a model that does not hold d_S constant (FEL-2R) showed much better properties in terms of false-positive rate and power (and provides a better statistical fit; data not shown) in the presence of recombination supports this idea. However, some bias still persisted, and this might be due to recombination misleading phylogenetic inference and the derived LRTs (Anisimova *et al.* 2003). Indeed, we have previously shown (Posada and Crandall 2002) that the trees inferred by ignoring recombination can be quite different from the underlying true phylogenies. In any case, using a "two-rate" or "dual" model such as FEL-2R seems highly recommended if recombination is suspected to have occurred in the history of the sample under study.

By chance, 2/3 of the recombination events that occur within coding sequences should happen within codons. We have analyzed recombination breakpoints for the circulant recombinant forms of HIV-1 reported at the Los Alamos HIV database (http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/breakpoints.html). Note that these breakpoints are very gross estimates and do not constitute a random sample. Among the 290 breakpoint locations listed, only 28% occur within codons, which might suggest that intracodon recombination events are selected against and/or are more difficult to detect (especially if the detection is carried out at the amino acid level). More reliable data are needed to better understand this question.

To our knowledge, only one method has been developed to co-estimate ρ and ω (Wilson and McVean 2006). This method showed very good performance in simulations, but these did not allow recombination within codons. One of the potential uses of our algorithm is a more meaningful evaluation of these kinds of methods. The algorithm described here has been implemented in a computer program called *NetRecodon*, freely available from the software section at http://darwin.uvigo.es, which also simulates migration, demographic periods, and dated tips. The program is reasonably fast and can produce large alignments (100 sequences with 1000 codons will take 2 min). Note that the execution time depends directly on the recombination and substitution rates. Conveniently, *NetRecodon* can also run in parallel (using the Message Passing Interface libraries). This algorithm could be used to more realistically model the evolution of nuclear genes and fast-evolving pathogens such as HIV-1 or the estimation of genetic parameters using approximate Bayesian computation (Beaumont *et al.* 2002; Tallmon *et al.* 2004; Excoffier *et al.* 2005; Tanaka *et al.* 2006). Certainly, in coding sequences, recombination occurs more often within codons than between them, and therefore recombination needs to be taken into account.

We thank two anonymous reviewers for their excellent suggestions. In particular, one of the reviewers suggested some important improvements to the original algorithm. This work was supported by the Spanish Ministry of Science and Education (grant no. BIO2007-61411 to D.P., Formación de Personal Investigador (FPI) fellowship BES-2005-9151 to M.A.).

LITERATURE CITED

Anisimova, M., R. Nielsen and Z. Yang, 2003 Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164:** 1229–1236.

Arenas, M., and D. Posada, 2007 Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography. BMC Bioinformatics **8:** 458.

Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153–1157.

Carvajal-Rodriguez, A., K. A. Crandall and D. Posada, 2006 Recombination estimation under complex evolutionary models with the coalescent composite-likelihood method. Mol. Biol. Evol. **23:** 817–827.

Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver *et al.*, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203–218.

DeChaine, E. G., and A. P. Martin, 2006 Using coalescent simulations to test the impact of quaternary climate cycles on divergence in an alpine plant-insect association. Evolution **60:** 1004–1013.

de Oliveira, T., M. Salemi, M. Gordon, A. M. Vandamme, E. J. van Rensburg *et al.*, 2004 Mapping sites of positive selection and amino acid diversification in the HIV genome: An alternative approach to vaccine design? Genetics **167:** 1047–1058.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Excoffier, L., J. Novembre and S. Schneider, 2000 SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J. Hered. **91:** 506–509.

Excoffier, L., A. Estoup and J. M. Cornuet, 2005 Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics **169:** 1727–1738.

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

Griffiths, R. C., 1991 The two-locus ancestral graph, pp. 100–117 in *Selected Proceedings on the Symposium on Applied Probability* (IMS Lecture Notes, Monograph Series, Vol. 18), edited by I. V. Basawa and R. L. Taylor. Institute of Mathematical Statistics, Hayward, CA.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3:** 479–502.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution*, edited by P. Donelly and S. Tavaré. Springer-Verlag, Berlin.

Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:** 696–704.

Hellenthal, G., and M. Stephens, 2007 msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics **23:** 520–521.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Hudson, R. R., and N. L. Kaplan, 1995 The coalescent process and background selection. Philos. Trans. R. Soc. Lond. B Biol. Sci. **349:** 19–23.

Innan, H., K. Zhang, P. Marjoram, S. Tavare and N. A. Rosenberg, 2005 Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics **169:** 1763–1777.

Kaplan, N., and R. R. Hudson, 1985 The use of sample genealogies for studying a selectively neutral m-loci model with recombination. Theor. Popul. Biol. **28:** 382–396.

Kingman, J. F. C., 1982 The coalescent. Stochastic Processes and their Applications **13:** 235–248.

Korber, B., 2000 HIV signature and sequence variation analysis, pp. 55–72 in *Computational Analysis of HIV Molecular Sequences*, edited by A. G. Rodrigo and G. H. Learn. Kluwer Academic Publishers, Dordrecht, Netherlands.

Kosakovsky Pond, S. L., and S. D. Frost, 2005a Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics **21:** 2531–2533.

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

Kosakovsky Pond, S. L., S. D. Frost and S. V. Muse, 2005 HYPHY: hypothesis testing using phylogenies. Bioinformatics **21:** 676–679.

Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk and S. D. Frost, 2006 Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. **23:** 1891–1901.

Kosakovsky Pond, S. L., A. F. Y. Poon, S. Zárate, D. M. Smith, S. J. Little *et al.*, 2008 Estimating selection pressures on HIV-1 using phylogenetic likelihood models. Stat. Med. **27:** 4779–4789.

Li, W. H., and T. Gojobori, 1983 Rapid evolution of goat and sheep globin genes following gene duplication. Mol. Biol. Evol. **1:** 94–108.

Liang, L., S. Zollner and G. R. Abecasis, 2007 GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics **23:** 1565–1567.

Mailund, T., M. H. Schierup, C. N. Pedersen, P. J. Mechlenborg, J. N. Madsen *et al.*, 2005 CoaSim: a flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics **6:** 252.

Marjoram, P., and J. D. Wall, 2006 Fast "coalescent" simulation. BMC Genet. **7:** 16.

Nei, M., and T. Gojobori, 1986 Simple method 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.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante and A. G. Clark, 2007 Recent and ongoing selection in the human genome. Nat. Rev. Genet. **8:** 857–868.

Posada, D., and K. A. Crandall, 2002 The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. **54:** 396–402.

Posada, D., and C. Wiuf, 2003 Simulating haplotype blocks in the human genome. Bioinformatics **19:** 289–290.

Pybus, O. G., and A. Rambaut, 2002 GENIE: estimating demographic history from molecular phylogenies. Bioinformatics **18:** 1404–1405.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. **15:** 1576–1583.

Scheffler, K., D. P. Martin and C. Seoighe, 2006 Robust inference of positive selection from recombining coding sequences. Bioinformatics **22:** 2493–2499.

Shriner, D., D. C. Nickle, M. A. Jensen and J. I. Mullins, 2003 Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. **81:** 115–121.

Simonsen, K. L., and G. A. Churchill, 1997 A Markov chain model of coalescence with recombination. Theor. Popul. Biol. **52:** 43–59.

Slatkin, M., 1987 Gene flow and the geographic structure of natural populations. Science **236:** 787–792.

Spencer, C. C., and G. Coop, 2004 SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics **20:** 3673–3675.

Swofford, D. L., 2000 *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

Tallmon, D. A., G. Luikart and M. A. Beaumont, 2004 Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. Genetics **167:** 977–988.

Tanaka, M. M., A. R. Francis, F. Luciani and S. A. Sisson, 2006 Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics **173:** 1511–1520.

Wilson, D. J., and G. McVean, 2006 Estimating diversifying selection and functional constraint in the presence of recombination. Genetics **172:** 1411–1425.

Wiuf, C., and J. Hein, 1999 The ancestry of a sample of sequences subject to recombination. Genetics **151:** 1217–1228.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Wiuf, C., and D. Posada, 2003 A coalescent model of recombination hotspots. Genetics **164:** 407–417.

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

Yang, Z., 2006 *Computational Molecular Evolution*. Oxford University Press, Oxford.

Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586–1591.

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

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

Communicating editor: J. Wakeley

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.109.109736/DC1

Coalescent Simulation of Intracodon Recombination

Miguel Arenas and David Posada

DOI: 10.1534/genetics.109.109736

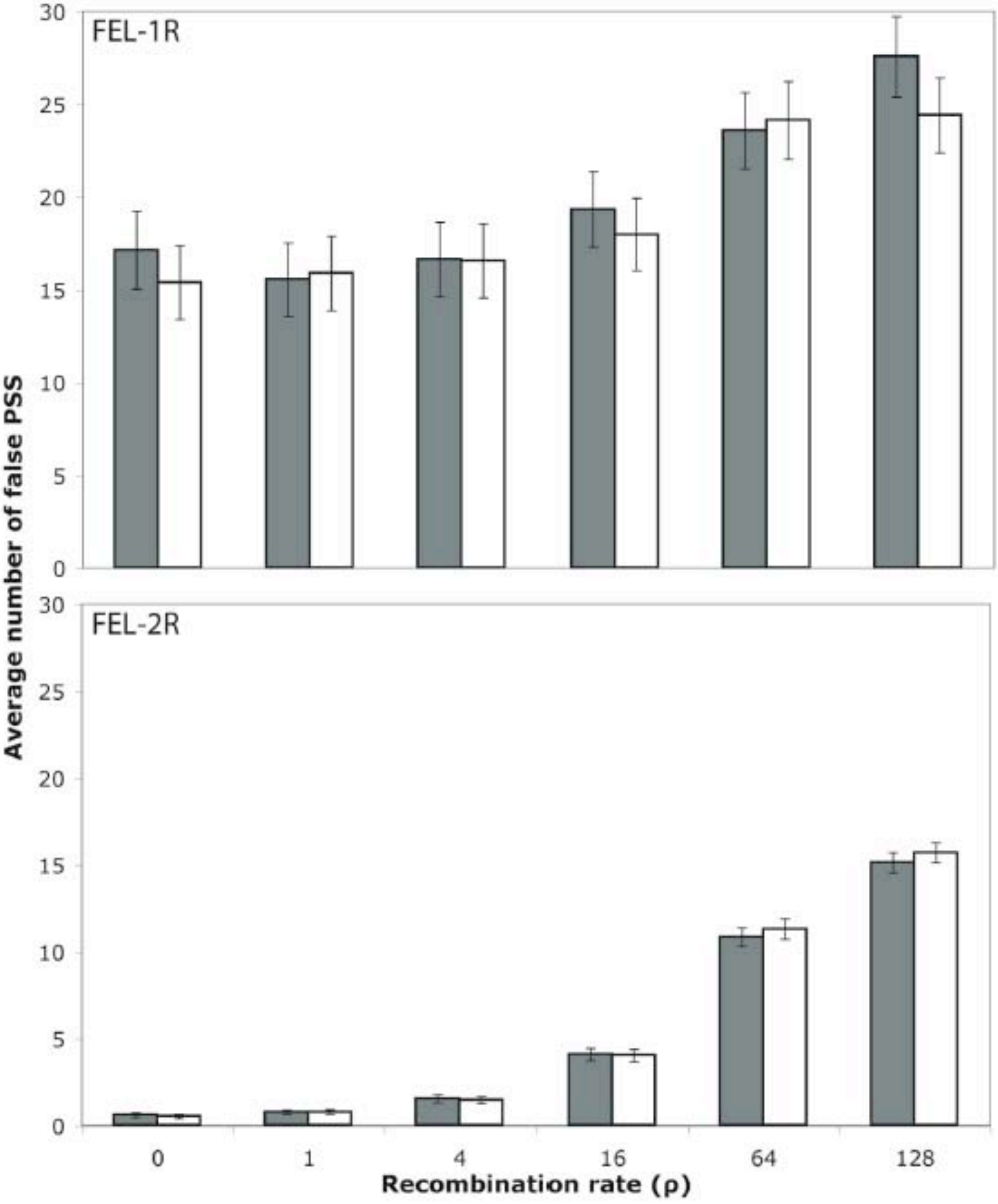

FIGURE S1.—False PSS estimated by FEL-1R and FEL-2R in the presence of recombination. Data was simulated under an M8 model with (white bars) and without (grey bars) recombination breakpoints within codons. The panel shows the average number of false PSS ($\omega > 1$; p-value < 0.05) for FEL-1R (upper; 2000 replicates) and FEL-2R (lower; 200 replicates). Error bars indicate 95% confidence intervals.

TABLE S1

Effect of recombination on the estimation of the global ω when the simulated value was ω = 0.2

		Intercodon and intracodon recombination		*Only intercodon recombination*	
θ	ρ	*NG86*	*GY94*	*NG86*	*GY94*
10	0	0.25 ± 0.02	0.24 ± 0.02	0.25 ± 0.01	0.22 ± 0.01
	1	0.25 ± 0.02	0.23 ± 0.02	0.25 ± 0.02	0.23 ± 0.01
	4	0.23 ± 0.02	0.21 ± 0.01	0.26 ± 0.02	0.24 ± 0.02
	16	0.25 ± 0.02	0.23 ± 0.01	0.26 ± 0.02	0.23 ± 0.02
	64	0.24 ± 0.02	0.23 ± 0.02	0.25 ± 0.02	0.23 ± 0.02
	128	0.25 ± 0.02	0.23 ± 0.02	0.24 ± 0.02	0.22 ± 0.02
20	0	0.23 ± 0.01	0.22 ± 0.01	0.22 ± 0.01	0.21 ± 0.01
	1	0.23 ± 0.01	0.22 ± 0.01	0.23 ± 0.01	0.21 ± 0.01
	4	0.23 ± 0.01	0.22 ± 0.01	0.23 ± 0.01	0.22 ± 0.01
	16	0.22 ± 0.01	0.22 ± 0.01	0.22 ± 0.01	0.22 ± 0.01
	64	0.21 ± 0.01	0.21 ± 0.01	0.22 ± 0.01	0.22 ± 0.01
	128	0.22 ± 0.01	0.22 ± 0.01	0.21 ± 0.01	0.21 ± 0.01
50	0	0.21 ± 0.01	0.21 ± 0.01	0.21 ± 0.00	0.21 ± 0.00
	1	0.22 ± 0.01	0.21 ± 0.00	0.22 ± 0.00	0.21 ± 0.00
	4	0.21 ± 0.01	0.22 ± 0.01	0.21 ± 0.01	0.21 ± 0.00
	16	0.22 ± 0.01	0.22 ± 0.01	0.21 ± 0.00	0.21 ± 0.00
	64	0.21 ± 0.01	0.21 ± 0.01	0.21 ± 0.00	0.22 ± 0.00
	128	0.22 ± 0.01	0.22 ± 0.01	0.21 ± 0.00	0.22 ± 0.00
100	0	0.20 ± 0.00	0.21 ± 0.00	0.21 ± 0.00	0.21 ± 0.00
	1	0.21 ± 0.01	0.21 ± 0.00	0.21 ± 0.00	0.21 ± 0.00
	4	0.21 ± 0.00	0.21 ± 0.00	0.21 ± 0.00	0.22 ± 0.00
	16	0.22 ± 0.00	0.22 ± 0.00	0.21 ± 0.00	0.22 ± 0.00
	64	0.23 ± 0.00	0.23 ± 0.00	0.22 ± 0.00	0.23 ± 0.00
	128	0.22 ± 0.00	0.23 ± 0.00	0.22 ± 0.00	0.23 ± 0.00
200	0	0.20 ± 0.00	0.20 ± 0.00	0.20 ± 0.00	0.20 ± 0.00
	1	0.20 ± 0.00	0.21 ± 0.00	0.21 ± 0.00	0.21 ± 0.00
	4	0.22 ± 0.00	0.22 ± 0.00	0.21 ± 0.00	0.22 ± 0.00
	16	0.23 ± 0.00	0.24 ± 0.00	0.22 ± 0.00	0.23 ± 0.00
	64	0.24 ± 0.00	0.25 ± 0.00	0.24 ± 0.00	0.25 ± 0.00
	128	0.24 ± 0.00	0.26 ± 0.00	0.24 ± 0.00	0.25 ± 0.00

θ is the population mutation rate and ρ is the population recombination rate.ω was estimated using the Nei and Gojobori method (NG86) in SNAP, and the ML phylogenetic framework under the GY94 codon model in HYPHY. Error bars indicated approximate 95% confidence intervals.

TABLE S2

Effect of recombination on the estimation of the global ω when the simulated value was ω = 1.0

		Intercodon and intracodon recombination		*Only intercodon recombination*	
θ	*ρ*	*NG86*	*GY94*	*NG86*	*GY94*
10	0	1.06 ± 0.06	1.33 ± 0.19	1.11 ± 0.08	1.28 ± 0.12
	1	1.18 ± 0.09	1.38 ± 0.19	1.08 ± 0.08	1.15 ± 0.07
	4	1.09 ± 0.07	1.23 ± 0.13	1.15 ± 0.09	1.23 ± 0.11
	16	1.14 ± 0.07	1.22 ± 0.10	1.19 ± 0.09	1.22 ± 0.10
	64	1.14 ± 0.06	1.17 ± 0.09	1.18 ± 0.07	1.24 ± 0.09
	128	1.29 ± 0.10	1.31 ± 0.14	1.20 ± 0.07	1.25 ± 0.12
20	0	1.15 ± 0.06	1.16 ± 0.08	1.07 ± 0.06	1.08 ± 0.05
	1	1.12 ± 0.07	1.12 ± 0.06	1.09 ± 0.06	1.13 ± 0.06
	4	1.08 ± 0.06	1.11 ± 0.06	1.15 ± 0.06	1.13 ± 0.06
	16	1.14 ± 0.06	1.11 ± 0.06	1.16 ± 0.07	1.11 ± 0.06
	64	1.15 ± 0.07	1.09 ± 0.06	1.21 ± 0.07	1.17 ± 0.08
	128	1.22 ± 0.07	1.16 ± 0.07	1.23 ± 0.07	1.13 ± 0.07
50	0	1.06 ± 0.04	1.06 ± 0.03	1.05 ± 0.04	1.05 ± 0.03
	1	1.05 ± 0.04	1.06 ± 0.03	1.08 ± 0.04	1.06 ± 0.03
	4	1.04 ± 0.03	1.07 ± 0.03	1.05 ± 0.05	1.04 ± 0.04
	16	1.08 ± 0.04	1.08 ± 0.04	1.05 ± 0.04	1.04 ± 0.04
	64	1.04 ± 0.03	1.05 ± 0.03	1.05 ± 0.03	1.04 ± 0.03
	128	1.05 ± 0.03	1.07 ± 0.04	1.03 ± 0.03	1.03 ± 0.03
100	0	1.01 ± 0.02	1.02 ± 0.02	1.04 ± 0.03	1.06 ± 0.02
	1	1.02 ± 0.02	1.03 ± 0.02	1.04 ± 0.03	1.04 ± 0.02
	4	1.00 ± 0.03	1.02 ± 0.02	1.01 ± 0.03	1.03 ± 0.02
	16	1.00 ± 0.02	1.03 ± 0.02	0.99 ± 0.02	1.02 ± 0.02
	64	0.98 ± 0.02	1.02 ± 0.02	0.99 ± 0.02	1.03 ± 0.02
	128	1.00 ± 0.02	1.04 ± 0.02	0.98 ± 0.02	1.03 ± 0.02
200	0	1.01 ± 0.02	1.03 ± 0.02	1.00 ± 0.02	1.03 ± 0.02
	1	0.99 ± 0.02	1.03 ± 0.02	0.99 ± 0.02	1.03 ± 0.02
	4	0.99 ± 0.02	1.03 ± 0.02	0.99 ± 0.02	1.02 ± 0.02
	16	1.00 ± 0.02	1.04 ± 0.02	0.96 ± 0.02	1.00 ± 0.02
	64	0.96 ± 0.02	1.01 ± 0.02	0.98 ± 0.02	1.03 ± 0.02
	128	0.98 ± 0.02	1.04 ± 0.02	0.96 ± 0.01	1.01 ± 0.02

θ is the population mutation rate and ρ is the population recombination rate. ω was estimated using the Nei and Gojobori method (NG86) in SNAP, and the ML phylogenetic framework under the GY94 codon model in HYPHY. Error bars indicated approximate 95% confidence intervals.

TABLE S3

Effect of recombination on the estimation of the global ω when the simulated value was $\omega = 5.0$

		Intercodon and intracodon recombination		*Only intercodon recombination*	
θ	ρ	*NG86*	*GY94*	*NG86*	*GY94*
10	0	3.41 ± 0.33	6.89 ± 0.72	3.34 ± 0.30	6.42 ± 0.62
	1	3.90 ± 0.35	7.55 ± 0.65	3.46 ± 0.42	6.40 ± 0.61
	4	3.46 ± 0.31	6.87 ± 0.64	3.45 ± 0.30	7.10 ± 0.68
	16	3.34 ± 0.23	6.56 ± 0.59	3.23 ± 0.24	6.81 ± 0.62
	64	3.07 ± 0.19	6.23 ± 0.55	3.05 ± 0.20	7.05 ± 0.61
	128	3.04 ± 0.20	6.78 ± 0.58	2.92 ± 0.15	6.80 ± 0.61
20	0	4.98 ± 0.48	6.42 ± 0.55	4.58 ± 0.49	6.03 ± 0.48
	1	4.95 ± 0.44	6.10 ± 0.49	4.85 ± 0.37	6.20 ± 0.47
	4	4.40 ± 0.27	5.52 ± 0.42	4.78 ± 0.40	6.02 ± 0.45
	16	4.64 ± 0.27	5.60 ± 0.44	4.40 ± 0.28	5.70 ± 0.42
	64	4.54 ± 0.25	5.49 ± 0.42	4.74 ± 0.29	5.98 ± 0.46
	128	4.71 ± 0.26	5.61 ± 0.40	4.76 ± 0.26	5.83 ± 0.43
50	0	4.87 ± 0.28	5.52 ± 0.31	5.31 ± 0.36	5.81 ± 0.33
	1	4.76 ± 0.27	5.39 ± 0.29	5.17 ± 0.33	5.47 ± 0.29
	4	5.20 ± 0.35	5.59 ± 0.30	5.19 ± 0.30	5.62 ± 0.33
	16	5.42 ± 0.40	5.62 ± 0.34	5.08 ± 0.30	5.24 ± 0.30
	64	5.59 ± 0.50	5.40 ± 0.32	5.41 ± 0.30	5.28 ± 0.29
	128	5.59 ± 0.35	5.37 ± 0.30	5.41 ± 0.31	5.24 ± 0.27
100	0	4.54 ± 0.21	5.14 ± 0.20	4.92 ± 0.26	5.48 ± 0.25
	1	4.77 ± 0.28	5.22 ± 0.24	4.65 ± 0.22	5.15 ± 0.19
	4	4.44 ± 0.19	4.94 ± 0.20	4.61 ± 0.24	5.05 ± 0.20
	16	4.54 ± 0.18	5.04 ± 0.20	4.54 ± 0.23	5.04 ± 0.23
	64	4.22 ± 0.17	4.84 ± 0.21	4.33 ± 0.17	4.82 ± 0.19
	128	4.27 ± 0.19	4.84 ± 0.21	4.28 ± 0.16	4.83 ± 0.18
200	0	4.26 ± 0.16	5.23 ± 0.18	4.19 ± 0.14	5.15 ± 0.16
	1	4.29 ± 0.14	5.20 ± 0.16	4.13 ± 0.13	5.00 ± 0.14
	4	3.94 ± 0.14	4.76 ± 0.14	3.94 ± 0.12	4.83 ± 0.16
	16	3.73 ± 0.12	4.61 ± 0.15	3.62 ± 0.10	4.45 ± 0.13
	64	3.38 ± 0.08	4.28 ± 0.12	3.50 ± 0.09	4.46 ± 0.13
	128	3.32 ± 0.11	4.22 ± 0.14	3.48 ± 0.08	4.47 ± 0.14

θ is the population mutation rate and ρ is the population recombination rate. ω was estimated using the Nei and Gojobori method (NG86) in SNAP, and the ML phylogenetic framework under the GY94 codon model in HYPHY. Error bars indicated approximate 95% confidence intervals.