

Recombination Estimation Under Complex Evolutionary Models with the Coalescent Composite-Likelihood Method

Antonio Carvajal-Rodríguez,*† Keith A. Crandall,* and David Posada†

*Department of Microbiology and Molecular Biology, Brigham Young University, Provo, Utah; and †Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, Vigo, Spain

The composite-likelihood estimator (CLE) of the population recombination rate considers only sites with exactly two alleles under a finite-sites mutation model (McVean, G. A. T., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**:1231–1241). While in such a model the identity of alleles is not considered, the CLE has been shown to be robust to *minor* misspecification of the underlying mutational model. However, there are many situations where the putative mutation and demographic history can be quite complex. One good example is rapidly evolving pathogens, like HIV-1. First we evaluated the performance of the CLE and the likelihood permutation test (LPT) under more complex, realistic models, including a general time reversible (GTR) substitution model, rate heterogeneity among sites (Γ), positive selection, population growth, population structure, and noncontemporaneous sampling. Second, we relaxed some of the assumptions of the CLE allowing for a four-allele, GTR+ Γ model in an attempt to use the data more efficiently. Through simulations and the analysis of real data, we concluded that the CLE is robust to severe misspecifications of the substitution model, but underestimates the recombination rate in the presence of exponential growth, population mixture, selection, or noncontemporaneous sampling. In such cases, the use of more complex models slightly increases performance in some occasions, especially in the case of the LPT. Thus, our results provide for a more robust application of the estimation of recombination rates.

Introduction

The population recombination rate is a fundamental parameter for evolutionary biology and medical population genetics. Not only is recombination a key force shaping the architecture of genomes, but its distribution across genomic regions is essential for association studies of genetic diseases (Weiss and Clark 2002) and to understand the evolution of pathogens and drug resistance (Posada, Crandall, and Holmes 2002; Awadalla 2003; Stumpf and McVean 2003; Rambaut et al. 2004). Recombination can also generate new allelic variants within a population or improve the estimation of population genetic parameters like gene flow (Hudson, Boos, and Kaplan 1992; Hudson, Slatkin, and Maddison 1992). On the other hand, ignoring the presence of recombination can have misleading effects when estimating other population genetic parameters (Schierup and Hein 2000) or phylogenetic relationships (Schierup and Hein 2000; Posada and Crandall 2002).

Clearly, the estimation of the population recombination rate is not an easy task (Hudson 2001). Recently, several coalescent methods have been proposed to estimate this parameter that use all the information contained in the data (Griffiths and Marjoram 1996; Kuhner, Yamato, and Felsenstein 2000; Fearnhead and Donnelly 2001). However, these full-likelihood methods are very computationally intensive, becoming impractical for many real data sets. To avoid this difficulty, several methods were in turn developed to approximate, instead of to compute, the full-likelihood surface (Stumpf and McVean 2003). Coalescent methods assume neutral evolution, random mating, and constant population size. However, the pseudolikelihood methods make some extra assumptions (see below) in order to reduce computation time.

Among the latter class of methods, Hudson (2001) proposed a composite-likelihood estimator (CLE) of the population recombination rate that combines the coalescent likelihoods of all pairwise comparisons for segregating sites. McVean, Awadalla, and Fearnhead (2002) extended Hudson's method to allow for a finite-sites mutation model, introducing also a likelihood permutation test (LPT). They assume a two-allele model with reversible, symmetric mutation, with a mutation rate per generation homogeneous across sites. This implies that, during the estimation procedure, only sites with exactly two alleles are considered, and that the identity of these alleles (A, C, G, or T) is lost. McVean, Awadalla, and Fearnhead (2002) tested their method with data simulated under a range of models of sequence evolution, concluding that their model was robust to *minor* misspecification of the underlying mutation model.

The model complexity explored by McVean, Awadalla, and Fearnhead (2002) was limited relative to those available that better fit biological reality. For example, models of substitution in HIV-1 can be quite complex (Posada and Crandall 2001a). In fact, the use of a general time reversible (GTR) model of nucleotide substitution (Rodríguez et al. 1990) coupled with a Γ distribution for rate variation among sites has been suggested as a reasonable model of HIV sequence evolution (Leitner, Kumar, and Albert 1997; Anderson et al. 2001). Moreover, HIV-1 population size is far from being constant (Wilson, Falush, and McVean 2005) and undergoes important selection pressures (Nielsen and Yang 1998; Crandall et al. 1999; Templeton et al. 2004). Indeed, HIV-1 achieves high levels of diversity through high replication and mutation rates, as well as recombination and selection (Anderson et al. 2001).

With this study we have two major goals. First, we want to evaluate the performance of McVean et al.'s estimator and recombination test under more complex, realistic models. Second, we relaxed some of the assumptions of this estimator in an attempt to use the data more efficiently and increase the estimator's robustness. Namely, we developed a four-allele mutation model capable of using all sites in an

Key words: recombination, complex models, coalescent, composite likelihood, HIV-1.

E-mail: ac549@email.byu.edu.

Mol. Biol. Evol. 23(4):817–827. 2006

doi:10.1093/molbev/msj102

Advance Access publication February 1, 2006

alignment of DNA sequences, which considers all possible six reversible substitutions between nucleotides and rate variation among sites (i.e., a GTR+ Γ model) during the estimation procedure. We then performed several simulations in which the data were generated under more general models of nucleotide substitution such as GTR with rate heterogeneity, and with different schemes of population growth, selection, and population subdivision. We also explored the effect of including noncontemporary sequences in the same sample, as is often the case with HIV-1 samples (Shankarappa et al. 1998). In addition, we explored the behavior of both estimators, the original and our extension, with real data.

Material and Methods

McVean et al.'s Estimator (CLE)

The CLE is implemented in the program *Pairwise*, included in the *LDhat* package, and freely available at <http://www.stats.ox.ac.uk/~mcvean/LDhat/>. We used the source code and binary application included in the *LDhat* 1.0 distribution. *Pairwise* requires a set of aligned segregating sites, their location, and the specification of the population mutation rate, $\theta = 4N\mu$ (where N is the effective population size and μ is the mutation rate per site per generation), for which the default value is a finite-sites version of the Watterson estimate (Watterson 1975). Considering only sites with two alleles, *Pairwise* estimates the coalescent likelihood of each pair of segregating sites, treating them as independent. The recombination rate ($\rho = 4Nrl$, where N is effective population size, r is recombination rate per site per generation, and l is the total alignment length) for the entire alignment is then estimated over a grid defined by the user. As part of the algorithm, pairs of segregating sites are classified into equivalence sets, reducing the total number of different patterns in order to speed up the computation. In addition, because the number of possible combinations of allelic states in a two-locus model can be easily enumerated, these can be tabulated and then consulted without reference to the data. Consequently, the estimation method is computationally efficient. On the other hand, meaningful confidence intervals for the recombination estimate can be obtained only by extensive simulation (Hudson 2001). Furthermore, and building upon their estimator, McVean, Awadalla, and Fearnhead (2002) also proposed a LPT for the occurrence (yes/no) of recombination. This test is based upon the fact that under a model of no recombination, sites are exchangeable. However, if recombination is present in the data the sites are no longer exchangeable and the likelihood of the data with exchanged sites must be lower than that of the original set.

Extended Estimator

Aiming to make a more efficient use of the data and increase the robustness of the method, we relaxed some of the assumptions made in the program *Pairwise*:

1. We used all segregating sites, and not only those with two alleles.
2. We assumed a GTR substitution model that allows for six different rates between all four nucleotides (A, C, G,

and T), instead of a single rate between two undifferentiated states (0, 1).

3. We allowed for rate variation among sites instead of assuming that all sites evolve at the same rate. To model this variation we use a gamma distribution (Γ) (Yang 1993).

We implemented these extensions in the program *Kpairwise*, which is based on the original source code of *Pairwise* (*Kpairwise* is freely distributed from <http://darwin.uvigo.es/software/kpairwise.html>). However, our program is considerably slower than *Pairwise* for several reasons. In *Pairwise*, the number of possible combinations of allelic states in a two-locus biallelic model can be easily enumerated, and these can be tabulated and calculated without reference to the data. On the other hand, in *Kpairwise* we use a more complex substitution model which is no longer parent-independent because the type of a mutant depends on the type of its parent (Stephens and Donnelly 2000). Consequently, the number of allele combinations is much higher and the enumeration in tables becomes much less efficient. In any case, *Kpairwise* incorporates a hash table system that keeps track of likelihoods for a given theta, nucleotide substitution parameters, number of sequences, and a grid of ρ values. In every run, the program looks for particular entries in this table. If the entry exists, the program uses the stored value; otherwise it proceeds to calculate a new likelihood that is stored in the table afterward. *Kpairwise* includes also McVean et al.'s LPT.

Estimation Models

We estimated the recombination rate and implemented the permutation test under six different models.

JC2: assumes the JC model and uses only sites with two alleles. This is the model assumed in *Pairwise*.

JCall: assumes the JC model and uses all sites.

GTR2: assumes the GTR model and uses only sites with two alleles.

GTRall: assumes the GTR model and uses all sites.

GTRall+ Γ : assumes the GTR model, uses all sites, and allows different pairs of sites to evolve at different rates. Only in this case the estimation model is not misspecified.

To model rate variation across pairs of sites, we use a discrete gamma approximation (Γ) with four categories (Yang 1994). Given a specific gamma shape, which we estimate from the data, we draw four numbers from the gamma distribution that correspond to the average for each rate category, and use them to obtain four-scaled θ s. For each of these, we obtain a likelihood surface for the recombination rate and average them to obtain a single likelihood surface. Indeed, these calculations imply that computation time is significantly increased.

Data Simulation

To evaluate the performance of *Pairwise* and *Kpairwise* under complex models we simulated DNA alignments using the coalescent with recombination (Hudson 1983), a model of codon substitution (Goldman and Yang 1994),

and a modified forward simulator (Liu et al. 2004) under a range of different conditions.

- (a) Coalescent with recombination (in-house software: source code available upon request from the authors)
- (i) Number of sequences = 20
 - (ii) Number of replicates = 100
 - (iii) Sequence length = 100, 200, or 1,000 bp
 - (iv) Effective population size = 1,000
 - (v) Nucleotide substitution model: a fully parameterized GTR substitution model with rate variation among sites following a Γ distribution (freq. A = 0.35; freq. C = 0.17, freq. G = 0.23, freq. T = 0.25; rate AC = 3.0, rate AG = 5.0, rate AT = 0.9, rate CG = 1.3, rate CT = 5.3, rate GT = 1.0; $\alpha = 0.7$). These values are typical for HIV-1 (Posada and Crandall 2001a).
 - (vi) Population recombination parameter $\rho = 4Nrl = 0, 2, 10, 50,$ and 100
 - (vii) Population mutation parameter $\theta = 4N\mu = 0.1$
 - (viii) Exponential growth rate $G = Ng = 0, 20, 40, 80$, where g is the growth rate per individual per generation.
- The values above are typical for fast-evolving pathogens like HIV-1 (Pybus, Holmes, and Harvey 1999; Posada and Crandall 2001a; Seo et al. 2002; Shriner et al. 2004).
- (b) Codon model (implemented in the package *evolver* and included in the Paml 3.14 program [Yang 1997])
- (i) Number of sequences = 20
 - (ii) Number of replicates = 100
 - (iii) Sequence length = 900 bp
 - (iv) One random tree with height = 0.5
 - (v) Population recombination parameter, $\rho = 4Nrl = 0$
 - (vi) Ratio of nonsynonymous/synonymous substitution rates, $\omega = d_N/d_S = 0.2, 1,$ and 5
- (c) Forward simulations (software modified from that of Liu et al. [2004])
- (i) Number of sequences = 20
 - (ii) Number of replicates = 100
 - (iii) Sequence length = 1,000 bp
 - (iv) Effective population size = 1,000
 - (v) Nucleotide substitution model: Jukes-Cantor (Jukes and Cantor 1969)
 - (vi) Population recombination parameter, $\rho = 4Nrl = 0$ and 10
 - (vii) Population mutation parameter $\theta = 4N\mu = 0.1$
 - (viii) Selection:
 - a. Directional and weak selection: 10 selective sites with selection coefficient $s = 0.1$. Maximum fitness is twice the initial one.
 - b. Divergent and strong selection: 100 selective sites with selection coefficient $s = 0.1$ or 0.01. Maximum fitness is four times the initial one.
 - (ix) Population subdivision without migration
 - (x) Noncontemporaneous samples

Selection Model

Our selection model is an extension of that of Liu et al. (2004). It consists of haploid organisms with constant pop-

ulation size, and each individual is represented by a DNA sequence. In every sequence, there are a number of sites that undergo positive selection. Fitness is evaluated at the fecundity level, so that the offspring of each individual is proportional to its fitness value. We introduced recombination in this model by selecting pairs of parents during reproduction. After mutation, recombination occurs between each pair of sites with probability $r = 2.5 \times 10^{-6}$. A recombination event results in a pair of recombinant sequences, but in order to maintain a haploid model, only one is randomly chosen. The nucleotide substitution process followed a Jukes-Cantor model. Samples of 20 sequences were taken from a population of 1,000 individuals after 1,000 generations.

Population Subdivision

The model above was extended to allow for population subdivision. We simulated two populations, each with 1,000 individuals. These populations evolved independently (no migration), with or without recombination. Samples of 20 sequences (10 sequences from each population) were taken after 1,000 generations.

Longitudinal and Noncontemporaneous Samples

Using the models described above, we obtained longitudinal samples at generations 250, 500, 750, and 1,000. From these same samples we also built at random 100 noncontemporaneous data sets, sampling each time five sequences from each time point.

Empirical Data

We selected two subjects (P1 and P2) from a nine-individual HIV-1 longitudinal study of the C2-V5 region of the *env* gene (Shankarappa et al. 1999). P1 is a short-term progressor, from whom 10 different time points were sampled. P2 is a long-term progressor, and in this case 14 longitudinal samples were available. The average time between each sample was 8 months. In order to study the effect of noncontemporary sequences on the recombination estimate, we also analyzed 10 samples of P1 consisting of a mixture of two randomly chosen sequences from each time point.

Estimation Procedure

For each data set, we estimated the parameters of the substitution model (GTR) and the parameter alpha of the gamma distribution when necessary using PAUP* (Swofford 2002). To estimate recombination we used the "crossover" recombination model from *Pairwise* (see *Discussion*). The allowed range of recombination rates during estimation was always $0 \leq 4Nrl \leq 100$, using a grid of 101 points for the importance sampling method of Fearnhead and Donnelly (2001). For each set of conditions, we report the median estimate across replicates and define success as the fraction of the time that estimates are within a factor of 2 from the generating, true ρ value (Wall 2000). Power of the permutation test was defined as the number of times it detects the presence of recombination when $\rho > 0$.

Table 1
Recombination Detection and Estimation for 20–25 Segregating Sites

ρ	S(S2)	Substitution Models				
		JC2	JCall	GTR2	GTRall	GTRall + Γ
0	23 (20)	0 [51] (3)	1 [45] (2)	0 [51] (3)	0 [51] (5)	2 [45] (2)
2	25 (22)	3 [47] (32)	4 [43] (41)	3 [50] (35)	3 [43] (41)	4 [36] (43)
10	23 (20)	9 [75] (73)	12 [69] (86)	10 [69] (72)	11 [76] (87)	11 [71] (88)
50	23 (20)	46 [88] (89)	57 [96] (97)	53 [92] (92)	54 [94] (97)	55 [100] (96)
100	23 (20)	92 [83] (78)	100 [86] (94)	99 [84] (79)	96 [83] (95)	100 [87] (96)

NOTE.—Data were simulated under the neutral coalescent with recombination and constant population size. ρ indicates the simulated recombination rate. S(S2) indicates the total number of segregating sites and in parenthesis, the number of sites with only two alleles. In the remaining columns, we indicate the median recombination rate estimate, success (the fraction of the time that estimates are within a factor of 2 from the simulated ρ value) in brackets, and the percentage of recombination detection ($P < 0.05$) of the LPT in parenthesis. Substitution models are explained in the text.

Estimation Repeatability

Because the method to estimate the likelihoods is based on an importance sampling scheme (Fearhead and Donnelly 2001), different runs of the program using the same data could result in different estimates of the recombination rate. To understand the magnitude of this variation, we measured the SD associated to the estimates by repeating the estimation process 10 times in several of the empirical data sets and in one of the simulated data sets (100 replicates, GTR+ Γ , $\rho = 50$, $l = 200$).

Comparison with Other Recombination Detection Methods

We also compared the performance of the LPTs with the best result obtained by any of the 14 recombination detection methods evaluated in Posada and Crandall (2001b), using the same simulated data sets. We restricted our comparison to the JC and JC+ Γ models, with $\theta = 0.01$ and 0.05 , $\rho = 0, 1, 4, 16$, and 64 , and $\alpha = \infty, 2, 0.5$, and 0.05 .

Results

Coalescent Simulations Under GTR+ Γ

When data were simulated under GTR+ Γ , and sequence length was 100 bp (resulting in 20–30 segregating sites), estimates of ρ were quite accurate, independent of the substitution model assumed (table 1). Success values increased with higher recombination rates, from 40%–50% to 80%–90%. Similar results were obtained when sequence

length was 200 bp (encompassing 40–50 segregating sites) (table 2).

The LPT for the presence of recombination showed false-positive rates around 5% for all estimation models (table 1, first row). Power of the test increased with increasing values of ρ , from 30%–40% at $\rho = 2$ (7 recombination events expected on average) to 89%–97% when $\rho = 100$ (355 recombination events expected), although for the two-allele models power seemed higher for $\rho = 50$ than for $\rho = 100$. Using the “all”-allele models slightly increased power, but incorporating rate variation among sites (Γ) did not have a significant effect. Results were very similar when sequence length was 200 bp, although power increased 10%–15% at low levels of recombination (table 2).

Simulations Under Codon Models

Simulating data under a codon model without recombination did not result in any case of increased false-positive rate for the LPT. False-positive rates for sequences simulated under $\omega = 0.2, 1$, and 5 were only 6%, 3%, and 2%, respectively.

Simulations with Exponential Growth

Exponential growth resulted in a consistent underestimation of the recombination rate (table 3). Increasing values of the population growth parameter G were associated with decreasing estimates. At the highest growth rate ($G = 80$), the median estimates were 0, 0.5, and 4, for the JC2, JCall, and GTRall models, respectively, when the true value of $4Nrl$ was 10. The power of the LPT was also strongly

Table 2
Recombination Estimation and Detection for 40–50 Segregating Sites

ρ	S(S2)	Substitution Models			
		JC2	JCall	GTR2	GTRall
0	49 (42)	0 [71] (3)	1 [48] (5)	0 [54] (3)	0 [55] (5)
2	48 (41)	3 [51] (56)	4 [44] (65)	3 [49] (65)	4 [45] (65)
10	49 (42)	10 [74] (93)	12 [77] (94)	11 [78] (92)	11 [78] (94)
50	48 (41)	50 [89] (100)	61 [94] (100)	55 [92] (100)	54 [92] (100)
100	47 (40)	96 [94] (100)	100 [96] (100)	100 [95] (100)	100 [96] (100)

NOTE.—Data were simulated under the neutral coalescent with recombination and constant population size. ρ indicates the simulated recombination rate. S(S2) indicates the total number of segregating sites and in parenthesis, the number of sites with only two alleles. In the remaining columns, we indicate the median recombination rate estimate, success (the fraction of the time that estimates are within a factor of 2 from the simulated ρ value) in brackets, and the percentage of recombination detection ($P < 0.05$) of the LPT in parenthesis. Substitution models are explained in the text.

Table 3
Recombination Estimation and Detection in Fluctuating Populations

G	S(S2)	Substitution Models		
		JC2	JCall	GTRall
0	238 (203)	8.5 [81] (99)	12 [81] (100)	10.5 [78] (100)
20	43 (42)	7 [52] (21)	6 [53] (25)	4 [47] (30)
40	29 (29)	2 [9] (12)	4 [41] (11)	3 [37] (8)
80	19 (18)	0 [12] (4)	0.5 [9] (7)	4 [32] (8)

NOTE.—Data were simulated under the neutral coalescent with recombination and exponential growth. ρ was always 10. $G = Ng$, where g is the growth rate per individual per generation. N was always 1,000. S(S2) indicates the total number of segregating sites and in parenthesis, the number of sites with only two alleles. In the remaining columns, we indicate the median recombination rate estimate, success (the fraction of the time that estimates are within a factor of 2 from the simulated ρ value) in brackets, and the percentage of recombination detection ($P < 0.05$) of the LPT in parenthesis. Substitution models are explained in the text.

diminished by growth. When $G = 20$, power was only 21%–30%, decreasing to 8%–12% at the highest growth rate. The different estimation models did not have a clear effect on the recombination rate estimates, although for $G = 40$ and 80 using all alleles resulted in better estimates.

Simulations with Recombination and Selection

The ability to estimate or detect recombination decreased with the presence and intensity of selection. In the absence of recombination, the estimated recombination rate was zero or close to zero even in the presence of selection (table 4). Also, selection did not affect the otherwise conservative false-positive rate of the recombination test. When $\rho = 10$, selection reduced considerably the accuracy and success of the estimator, especially when selection was strong and only biallelic sites were taken into account. In this case, selection clearly reduced the power of the recombination test, from 90% in the neutral case, to almost 70% and 50% for the weak and strong selection cases, respectively. All calculations above used the specific Watterson's estimate of $4N\mu$ for each replicate. Interestingly, when the simulated (parametric) value of $4N\mu = 0.1$ was used instead, the impact of selection on success was diminished (data not shown).

Table 4
Recombination Estimation and Detection in the Presence of Selection

Selective Regime	ρ	S(S2)	Substitution Models	
			JC2	JCall
Neutral	0	88 (85)	0 [90] (4)	0 [84] (5)
	10	82 (80)	8 [72] (90)	10 [72] (92)
Weak	0	86 (84)	0 [96] (2)	0 [97] (3)
	10	91 (88)	6 [50] (69)	8 [62] (73)
Strong	0	57 (56)	3 [22] (4)	0 [99] (1)
	10	58 (57)	3 [38] (46)	12 [39] (48)

NOTE.—Data were simulated under a forward selection models with constant population size. ρ indicates the simulated recombination rate. S(S2) indicates the total number of segregating sites, and in parenthesis the number of sites with only two alleles. In the remaining columns, we indicate the median recombination rate estimate, success (the fraction of the time that estimates are within a factor of 2 from the simulated ρ value) in brackets, and the percentage of recombination detection ($P < 0.05$) of the LPT in parenthesis.

Table 5
Recombination Estimation and Detection in Subdivided Populations

Population Structure	ρ	S(S2)	Substitution Models	
			JC2	JCall
Not subdivided	0	88 (85)	0 [90] (4)	0 [84] (5)
	10	82 (80)	8 [72] (90)	10 [72] (92)
Subdivided	0/0	181 (170)	0 [81] (10)	0 [86] (6)
	10/10	180 (170)	4 [39] (82)	8 [94] (95)
	0/10	181 (171)	4 [15] (74)	7 [90] (86)

NOTE.—Data were simulated under a forward neutral model with constant population sizes. ρ indicates the simulated recombination rate in one or two populations. S(S2) indicates the total number of segregating sites, and in parenthesis the number of sites with only two alleles. In the remaining columns, we indicate the median recombination rate estimate, success (the fraction of the time that estimates are within a factor of 2 from the simulated ρ value) in brackets, and the percentage of recombination detection ($P < 0.05$) of the LPT in parenthesis. Substitution models are explained in the text.

Simulations Under Population Subdivision

When data were simulated under a population subdivision model with no migration, but this subdivision was ignored during the estimation procedure, there was a tendency to underestimate the simulated value of the recombination rate. This was especially true when only sites with two alleles were used. For example, when there was recombination in just one population ($\rho = 10$), the median estimate decreased from 8 to 4, and success from 72% to 15%, when only biallelic sites were considered. A similar effect was observed for the recombination permutation test. False positives increased and power decreased with population subdivision when only sites with two alleles were used. (table 5).

Simulations with Longitudinal Sampling

Sampling time had an effect on the estimation of the recombination rate. When sampling occurred as early as generation 250, the estimated recombination rate was 0 although the simulated value was 10 (table 6), even though there were already 61 segregating sites in the sample. From

Table 6
Recombination Estimation and Detection in Longitudinal Samples

Sampling Time	S(S2)	Substitution Models	
		JC2	JCall
250	61 (60)	0 [23] (21)	0 [9] (18)
500	90 (88)	6 [53] (65)	10 [79] (70)
750	106 (103)	8 [69] (80)	8 [74] (89)
1000	120 (115)	7 [70] (93)	10 [81] (94)
Noncontemporaneous	134 (128)	12 [75] (88)	12 [79] (96)

NOTE.—Data were simulated under a forward neutral model with constant population size. The simulated ρ was 10. Sampling time is in generations passed since the beginning of the simulations. Noncontemporaneous sequences include a mixture of different time points. S(S2) indicates the total number of segregating sites, and in parenthesis the number of sites with only two alleles. In the remaining columns, we indicate the median recombination rate estimate, success (the fraction of the time that estimates are within a factor of 2 from the simulated ρ value) in brackets, and the percentage of recombination detection ($P < 0.05$) of the LPT in parenthesis. Substitution models are explained in the text.

Table 7
Recombination Estimation and Detection from Empirical Data Sets

Time	Patient 1						Patient 2					
	Size	S(S2)	JC2	JCall	GTRall	GTRall + Γ	Size	S(S2)	JC2	JCall	GTRall	GTRall + Γ
1	7	7 (7)	6	0	0	0	10	18 (17)	6	0	10	56
2	10	23 (22)	16	12	12	100	11	24 (24)	27	50	95	71
3	9	41 (41)	15*	8*	9*	8*	10	37 (37)	10	12	11	13
4	8	34 (33)	14*	29*	28*	20*	11	50 (50)	38*	40*	30*	32*
5	10	52 (51)	93*	53*	42*	55*	8	49 (45)	19	15*	13*	20*
6	6	57 (55)	15*	12*	13*	13*	8	68 (64)	26*	82*	55*	67*
7	10	55 (50)	22*	42*	30*	50*	9	44 (42)	0	0	0	0
8	8	45 (43)	5	5	4	7	10	77 (71)	56*	65*	52*	64*
9	9	40 (36)	19*	14*	11*	13*	10	76 (72)	10*	14*	12*	12*
10	10	48 (46)	25*	37*	31*	8*	8	38 (37)	23	99*	67*	71*
11	—	—	—	—	—	—	8	69 (67)	21	18	17	34
12	—	—	—	—	—	—	9	57 (51)	23*	28*	33*	41*
13	—	—	—	—	—	—	9	61 (61)	17*	22*	21*	24*
14	—	—	—	—	—	—	11	70 (64)	46*	42*	47*	61*

NOTE.—P1 and P2 are the correspondent subjects in Shankarappa et al. (1999). Size is the number of sequences. S(S2) indicates the total number of segregating sites, and in parenthesis the number of sites with only two alleles. Each cell displays the estimated recombination rate. The symbol * indicates that the LPT was significant ($P < 0.05$).

generation 500 onward the estimation improved, especially if all sites were considered. The power of the recombination test also increased with increasing number of generations completed before sampling.

Noncontemporaneous Sequences

When sequences sampled at different time points were considered as contemporaneous and treated as a single sample, the median estimate seemed to slightly overestimate the actual value of the recombination rate (table 6). Estimator success and power of the recombination test were a little higher when all alleles were considered than when only biallelic sites were used.

Empirical Data

Estimates of recombination from patient 1 (P1) changed between different time points, with a single peak at time point 5 (table 7). Most estimates were significantly different from zero (i.e., the LPT was significant), except for time points 1 and 2. In most cases, the model assumed did not have a strong effect. In patient 2 (P2), there were several nonsignificant time points interspersed between different recombination peaks. In this case the model had a stronger effect. For example, at time points 5 and 10, recombination was not detected when considering only biallelic sites, but it was inferred when considering all sites, even resulting in large estimates.

When we analyzed 10 mixed samples including two randomly chosen P1 sequences from each time point, we obtained consistent estimates of the recombination rate (mean estimate JC2 = 8.4, SD JC2 = 4), but well below most of the estimates previously obtained at each point. In 8 out of the 10 mixed samples examined, the LPT was significant.

Variability Between Runs

The SD associated with 10 repetitions of the recombination rate estimate from sample time 5 of P1 was very high

using the JC2 (SD = 24.88) or JCall (SD = 18.51) models, and significantly smaller ($P < 0.01$) with the GTRall model (SD = 5.87) (table 8). When we repeated this experiment with samples 10 and 14 from individual P2, we obtained similar results. On the other hand, the results of the LPT seem to be very consistent across different runs (table 8). However, the same experiment performed with simulated data (GTR + Γ , $\rho = 50$, 40–50 segregating sites) resulted in much better repeatability. For JC2, the average over 100 replicates of the SD associated with 10 repetitions was 3.4 ± 1.9 .

Comparison with Other Recombination Detection Methods

With low variation, and especially at low recombination rates, the LPT was significantly more powerful than any of the 14 methods evaluated by Posada and Crandall (2001b) (fig. 1A). At higher variation levels the LPT was tied with the best methods (fig. 1B). At the same time,

Table 8
Repeatability of the Estimation Procedure

Run	Substitution Models		
	JC2	JCall	GTRall
1	96*	35*	46*
2	90*	48*	54*
3	47*	50*	48*
4	49*	98*	42*
5	34*	47*	46*
6	41*	52*	50*
7	46*	56*	39*
8	72*	66*	43*
9	31*	51*	59*
10	23*	31*	49*
Average	52.9	53.4	47.6
SD	24.88	18.51	5.87
Range	23–96	31–98	42–54

NOTE.—Recombination was estimated 10 times (starting from random seeds) from sample 5 in individual P1. Each cell displays the estimated recombination rate. The symbol * indicates that the LPT was significant ($P < 0.05$).

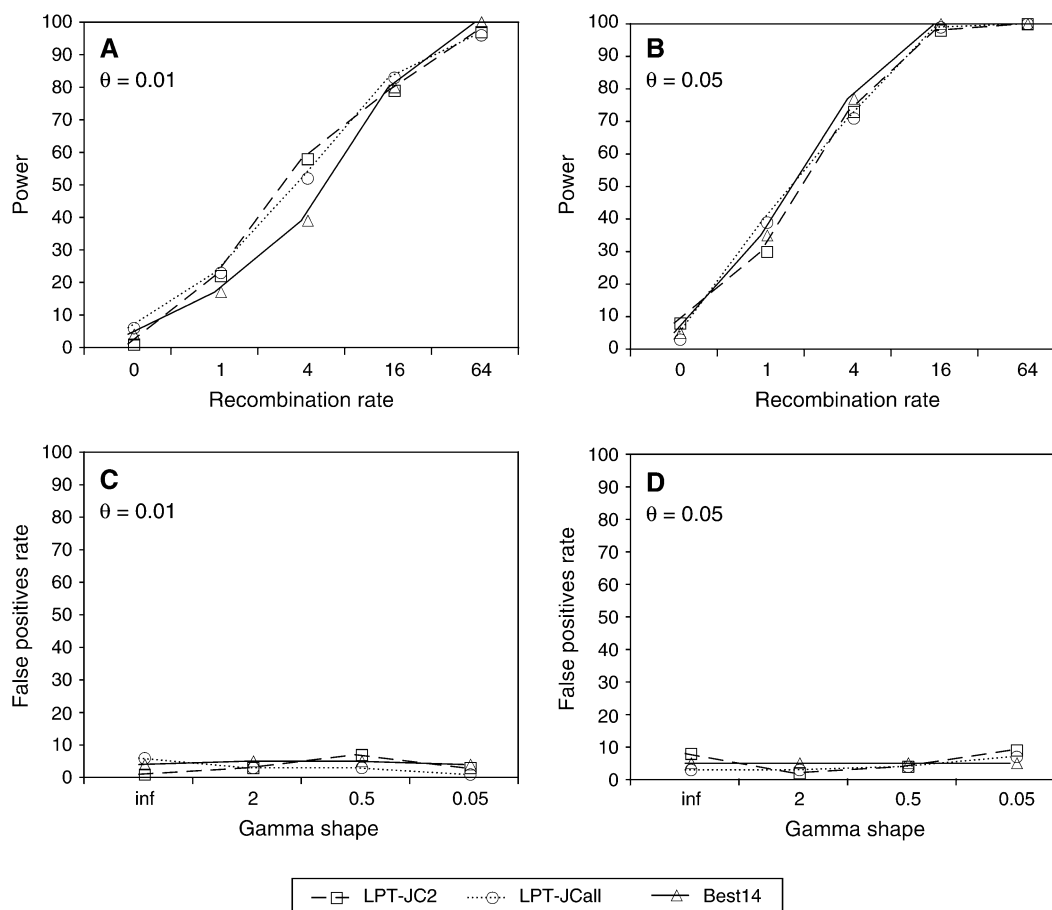


FIG. 1.—Relative power and false-positive rate of the LPT. Performance of the LPT under the JC2 and JCall is compared to the best results obtained by any of the 14 recombination detection methods evaluated by Posada and Crandall (2001*b*). (A) Power to detect recombination when $\theta = 0.01$ (B) Power to detect recombination when $\theta = 0.05$ (C) False-positive rate when $\theta = 0.01$ (D) False-positive rate when $\theta = 0.05$. *Gamma shape* is the shape of the gamma distribution for rate variation among sites (*inf* = infinity).

the likelihood permutation showed a false-positive rate around 5%, and it did not seem to be influenced by increasing levels of rate heterogeneity (fig. 1C and D), whereas many of the other methods had false-positive rates above 5%. These results were independent of the estimation model used (JC2 or JCall).

Discussion

Recombination Estimation Under Complex Substitution Models

We have shown that the CLE of the recombination rate (McVean, Awadalla, and Fearnhead 2002) is robust not only to minor misspecification of the substitution model (McVean, Awadalla, and Fearnhead 2002; Stumpf and McVean 2003), but also to other cases where the substitution model is grossly misspecified. This is the case when we assume a two-allele Jukes-Cantor model to estimate the recombination rate from data simulated under a fully parameterized four-allele GTR + Γ model. Under these conditions, the CLE method still performs quite well, and relaxing several of its assumptions during the estimation procedure (two alleles, equal transition probabilities between nucleotides, and rate homogeneity across sites) does not seem to significantly increase its performance. It seems that the major

limitation of the CLE method comes from the inherent loss of information due to the use of pairwise comparisons, which, in any case, does not seem very important.

Importantly we have used the *Pairwise* crossover (L) model instead of the “gene conversion” (C) one to estimate recombination. However, we performed some estimations under coalescent simulations using also the C model. The performance under the L model was always much better than under the C model (not shown). A possible explanation could be the necessity for the extra parameter “tract length” in the C model. We have tried using the recommended setting for viruses (McVean, Awadalla, and Fearnhead 2002), a tract length of 100 bp, but this seemed to be inadequate for the sequence lengths assayed (100 and 200 bp). Thus we finally decided to use the L model during the estimation procedure.

On the other hand, the detection of the presence of recombination with the LPT, which was derived from the CLE, is more sensitive to model misspecification. In these cases, accounting for more complex models during the permutation test does increase statistical power. For example, when all polymorphic sites are used we obtained an average increase of 12% on the ability to detect recombination, despite the fact that sites with more than two alleles

represented only 12%–15% of the total number of segregating sites. However, incorporating rate variation among sites during the estimation procedure does not result in significant improvement.

The LPT showed correct false-positive rates under complex substitution models. Although recombination can generate apparent rate heterogeneity (Schierup and Hein 2000; Worobey 2001), increasing levels of rate heterogeneity did not result in inflated false-positive rates above the nominal 5% level. Not surprisingly, low false-positive rates were also obtained when data were simulated with codon models (Yang 1997) including nonsynonymous/synonymous substitution rate ratios expected under negative and positive selection.

Growing Populations

The CLE consistently underestimates the recombination rate, and the power to detect recombination decreases, when the population has been growing exponentially. This seems to be true independent of the estimation model assumed. Recombination was virtually not detected with $G = 80$, which can be considered a high growth rate, but a plausible one, for example, for HIV-1 (Pybus, Holmes, and Harvey 1999). The disguising effect of growth on the detection of recombination was previously shown by Wiuf, Christensen, and Hein (2001), although using an extreme value of $G = 5,000$ in their simulations.

In our simulations it was clear that the number of segregating sites rapidly decreased with increasing growth rate. However, this effect does not completely explain the underestimation effect because when we estimated the recombination rate from data sets without growth but also with very few segregating sites, we obtained much better estimates (see table 1). Indeed, under population growth we expect long external branches and short internal ones in the underlying sample genealogy, implying that we will have many mutations at low frequency and only a few common ones (Slatkin and Hudson 1991). In growing populations fewer recombination events will influence the history of a pair of randomly chosen alleles (McVean 2002). Indeed, when genealogies are star-like, sequences contain less information about the topology, making recombination harder to detect (Wiuf, Christensen, and Hein 2001). Clearly, not only recombination can produce apparent growth rate (Schierup and Hein 2000), but population growth can also obscure the signal for recombination.

Positive Selection

Although different types of selection have been shown to have somehow little effect on the shape of gene genealogies (Golding 1997; Neuhauser and Krone 1997; Przeworski, Charlesworth, and Wall 1999; Barton and Etheridge 2004), in our simulations positive selection resulted in the underestimation of the recombination rate and in a decrease of the detection power. This could be the case because our selective regime consists of multiple selective sweeps and our samples are not collected at equilibrium. Different causes can contribute to this underestimation. First, the estimated population mutation param-

eter tends to decrease because Watterson's formulas for the estimation of the scaled mutation parameter are based on the number of segregating sites. However, under selection, the number of segregating sites no longer informs us about neutral mutation rate. Second, if selection affects fertility, the effective population size also decreases (Crow and Kimura 1970), as does θ and ρ . Third, selective sweeps result in an uneven distribution of allele frequencies, with many mutations at low frequency. Certainly, the assumptions of the standard neutral coalescent are violated by realistic forms of natural selection (Seo et al. 2002).

Indeed, positive selection occurs in many organisms with high recombination rates, like HIV-1 (Nielsen and Yang 1998). The interaction between recombination and selection can be complicated. In fact, it has been shown that recombination can also lead to the spurious inference of selection (Anisimova, Nielsen, and Yang 2003; Shriener et al. 2003). In any case, independent of the effect of nonsynonymous/synonymous substitution rates, if the assumption of evolutionary neutrality does not hold, then caution is needed when estimating and detecting recombination. Certainly, the interaction of selection and recombination clearly deserves further study.

Population Structure

Population structure or subdivision affects the estimation and detection of recombination. Despite the logical increase of the number of segregating sites, in our simulations the recombination rate is underestimated. Detection power was high, although lower than expected for the observed number of segregating sites. This effect is expected because we are mixing two independent populations with independent gamete frequencies, therefore increasing the levels of linkage disequilibrium and reducing the apparent recombination. Indeed, our simulations represent an extreme case, without migration and with a 50% of admixture in the sample. The effect of more subtle cases, allowing for restricted gene flow between populations, should be studied in the future.

Longitudinal Sampling and Noncontemporaneous Sequences

Sequence data obtained at different sampling points from populations of rapidly evolving pathogens as HIV-1 are increasingly available, as are new estimation approaches (Drummond et al. 2002). Here we have studied the variation of the recombination rate estimates through time. As expected, as time proceeds estimates get better because more information should be available from the data. In our simulations, and in the presence of recombination, after 250 generations the median recombination estimate was still 0, and recombination was not detected most of the time. Five hundred generations were needed to obtain reasonable estimates of the recombination rate. We did not detect a large impact, maybe a slight overestimation, on the estimation of recombination when we mixed sequences sampled at different time points. In this particular case, this is likely due to the fact that after generation 500, recombination estimates were quite homogeneous.

Estimation from Empirical Data

Finally, we briefly worked out an example with real HIV-1 longitudinal data originally from Shankarappa et al. (1999). These authors found that in nine infected individuals the pattern of viral evolution within each patient was very similar. Seo et al. (2002) studied the same nine patients and found a negative correlation between the effective viral population size and the rate of evolution. Furthermore, Ross and Rodrigo (2002) again studied these same nine patients and found that the proportion of sites under positive selection was a statistically significant predictor of disease duration. Strikingly, there have not been many attempts to study recombination in these data except for Buendia and Narasimhan (2004). Although methods to estimate several evolutionary parameters using serially sampled data have been developed (Fu 2001; Drummond et al. 2002; Seo et al. 2002), these parameters do not include the recombination rate.

Here we studied the variation on the recombination rate in two of the nine patients (P1 and P2) using the CLE method. We found a common pattern for both patients, despite the fact that P1 is a short-term progressor and P2 is a long-term progressor. We did not detect recombination in the early samples, where the number of segregating sites is lowest, but recombination appeared quite clearly afterward. Remarkably, in both patients there are one (P1) or two (P2) late points in which the recombination signal fades away. A similar pattern for individual P2 was also found by Buendia and Narasimhan (2004) upon inspection of a phylogenetic network. Those decreases in the recombination estimates might be associated with selective sweeps, although we did not detect a significant correlation ($P > 0.1$) with the proportion of positively selected sites in Ross and Rodrigo (2002).

The effect of mixing noncontemporaneous sequences in the same sample is stronger than in the case of simulated data, in part due to the higher variance associated with each sampling point. In this case mixing sequences resulted in the underestimation of the recombination rate, which resembles the effect of population structure. Although most researchers do not combine on purpose sequences sampled at different time points, in organisms like HIV-1 it is possible that the existence of reservoirs provokes the unintentional sampling of noncontemporaneous sequences.

Estimation Repeatability

The original implementation of CLE is affected by the stochasticity of the estimation process, showing an appreciable variation between runs with empirical data, and less variation with simulated data. This might be explained by the rupture of the standard neutral coalescent assumptions in the empirical data (e.g., selection). Estimation repeatability is significantly improved when the GTRall model is used, independent of the number of segregating sites. Because we have not altered the original importance sampling scheme, the improved estimation repeatability can be attributed to the use of a more realistic model. This implies that a researcher using the model implemented in the program *Pairwise* in *LDhat* should repeat the estimation several

times in order to get a feeling of the stochastic error of the estimate. This is not necessary if the program *Kpairwise* is used. Although *Kpairwise* is slower than *Pairwise*, a single run of *Kpairwise* could be faster than 5–10 runs of *Pairwise*. If one is just interested in detecting the presence of recombination with the LPT, a single run of *Pairwise* should suffice.

Comparison with Other Recombination Detection Methods

The detection of recombination is not an easy problem, and detection methods are not very powerful (Posada and Crandall 2001b). Importantly, for small recombination rates and low diversity, the LPT was more powerful than any of the 14 methods evaluated by Posada and Crandall (2001b). For other conditions it was always among the best. Clearly, the LPT is one of the most powerful methods to detect recombination available.

Conclusions

The main conclusions derived from this study are

1. The CLE of the recombination rate (McVean, Awadalla, and Fearnhead 2002; Stumpf and McVean 2003) performs very well, being robust to severe misspecification of the substitution model.
2. On the other hand, the CLE, as implemented in the program *Pairwise*, shows poor repeatability between runs. The estimation process should be repeated with different seeds in order to take this variability into account. Our program *Kpairwise*, albeit slower, does not show this variation.
3. Relaxing some assumptions of the CLE does not significantly increase performance, although power of the LPT increases when all alleles are taken into account. Nevertheless, the LPT is one of the most powerful tests for the detection of recombination available.
4. The CLE underestimates the recombination rate in the presence of rapid exponential growth, positive selection, population structure, or noncontemporaneous sampling. Such situations are common in the case of rapidly evolving pathogens like HIV-1. Therefore, new recombination estimators are necessary. These new estimators should consider scenarios other than the Wright-Fisher model. Extensions of the standard coalescent model that consider growing population size (Griffiths and Tavaré 1994), selection (Stephens and Donnelly 2003), and longitudinal data (Rodrigo and Felsenstein 1999) currently exist. Improved estimates of recombination will come with the incorporation of these biological realities within the powerful pseudolikelihood framework.

Acknowledgments

The authors want to thank the two anonymous reviewers for discussion and comments on the manuscript. This work was supported by grant R01-GM66276 from the US National Institutes of Health (A.C.-R., K.A.C., and D.P.), grant BFU2004-02700 of the Spanish Ministry

of Education and Science (D.P.), grant PGIDT05P-XIC31001PN of Dirección Xeral de Investigación e Desenvolvemento da Xunta de Galicia and by the Ramón y Cajal initiative of the Spanish government (D.P.).

Literature Cited

- Anderson, J. P., A. G. Rodrigo, G. H. Learn, Y. Wang, H. Weinstock, M. L. Kalish, K. E. Robbins, L. Hood, and J. I. Mullins. 2001. Substitution model of sequence evolution for the human immunodeficiency virus type 1 subtype B gp120 gene over the C2-V5 region. *J. Mol. Evol.* **53**:55–68.
- 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.
- Awadalla, P. 2003. The evolutionary genomics of pathogen recombination. *Nat. Rev. Genet.* **4**:50–60.
- Barton, N. H., and A. M. Etheridge. 2004. The effect of selection on genealogies. *Genetics* **166**:1115–1131.
- Buendía, P., and G. Narasimhan. 2004. MinPD: distance-based phylogenetic analysis and recombination detection of serially-sampled HIV quasispecies. Proceedings of the IEEE Computer Society Bioinformatics Conference, Stanford, Calif.
- Crandall, K. A., C. R. Kelsey, H. Imamichi, C. H. Lane, and N. P. Salzman. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* **16**:372–382.
- Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper & Row, New York.
- Drummond, A., G. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**:1307–1320.
- Fearnhead, P., and P. Donnelly. 2001. Estimating recombination rates from population genetic data. *Genetics* **159**:1299–1318.
- Fu, Y.-X. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol. Biol. Evol.* **18**:620–626.
- Golding, G. B. 1997. The effect of purifying selection on genealogies. Pp. 271–285 in P. Donnelly and S. Tavaré, ed. *Progress in population genetics and human evolution*. Springer-Verlag, New York.
- 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., and P. Marjoram. 1996. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**:479–502.
- Griffiths, R. C., and S. Tavaré. 1994. Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B* **344**:403–410.
- Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**:183–201.
- . 2001. Two-locus sampling distributions and their application. *Genetics* **159**:1805–1817.
- Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* **9**:138–151.
- Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**:583–589.
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 in H. M. Munro, ed. *Mammalian protein metabolism*. Academic Press, New York.
- Kuhner, M. K., J. Yamato, and J. Felsenstein. 2000. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**:1393–1401.
- Leitner, T., S. Kumar, and J. Albert. 1997. Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* **71**:4761–4770.
- Liu, Y., D. C. Nickle, D. Shriner, M. A. Jensen, H. Gerald, J. Learn, J. E. Mittler, and J. I. Mullins. 2004. Molecular clock-like evolution of human immunodeficiency virus type 1. *Virology* **329**:101–108.
- McVean, G. A. T. 2002. A genealogical interpretation of linkage disequilibrium. *Genetics* **162**:987–991.
- McVean, G. A. T., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**:1231–1241.
- Neuhauser, C., and S. M. Krone. 1997. The genealogy of samples with selection. *Genetics* **145**:519–534.
- Nielsen, R., and Z. Yang. 1998. Likelihood methods for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929–936.
- Posada, D., and K. A. Crandall. 2001a. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.* **18**:897–906.
- Posada, D., and K. A. Crandall. 2001b. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA* **98**:13757–13762.
- 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., K. A. Crandall, and E. C. Holmes. 2002. Recombination in evolutionary genomics. *Annu. Rev. Genet.* **36**:75–97.
- Przeworski, M., B. Charlesworth, and J. D. Wall. 1999. Genealogies and weak purifying selection. *Mol. Biol. Evol.* **16**:246–252.
- Pybus, O. G., E. C. Holmes, and P. H. Harvey. 1999. The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history. *Mol. Biol. Evol.* **16**:953–959.
- Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes. 2004. The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5**:52–61.
- Rodrigo, A. G., and J. Felsenstein. 1999. Coalescent approaches to HIV population genetics. Pp. 233–272 in K. A. Crandall, ed. *Evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- Rodríguez, F., J. F. Oliver, A. Marín, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**:485–501.
- Ross, H. A., and A. G. Rodrigo. 2002. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J. Virol.* **76**:11715–11720.
- Schierup, M. H., and J. Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**:879–891.
- Seo, T.-K., J. L. Thorne, M. Hasegawa, and H. Kishino. 2002. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. *Genetics* **160**:1283–1293.
- Shankarappa, R., P. Gupta, Learn, G. H. Jr, A. G. Rodrigo, Rinaldo, C. R. Jr, M. C. Gorry, J. I. Mullins, P. L. Nara, and G. D. Ehrlich. 1998. Evolution of human immunodeficiency virus type 1 envelope sequences in infected individuals with differing disease progression profiles. *Virology* **241**:251–259.
- Shankarappa, R., J. B. Margolick, S. J. Gange et al. (12 co-authors). 1999. Consistent viral evolutionary changes associated with the

- progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**:10489–10502.
- 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.
- Shriner, D., A. G. Rodrigo, D. C. Nickle, and J. I. Mullins. 2004. Pervasive genomic recombination of HIV-1 in vivo. *Genetics* **167**:1573–1583.
- Slatkin, M., and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**:555–562.
- Stephens, M., and P. Donnelly. 2000. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B* **62**:605–655.
- . 2003. Ancestral inference in population genetics models with selection (with discussion). *Aust. N. Z. J. Stat.* **45**:395–430.
- Stumpf, M. P. H., and G. A. T. McVean. 2003. Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.* **4**:959–968.
- Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.
- Templeton, A. R., R. A. Reichert, A. E. Weisstein, X.-F. Yu, and R. B. Markham. 2004. Selection in context: patterns of natural selection in the glycoprotein 120 region of human immunodeficiency virus 1 within infected individuals. *Genetics* **167**:1547–1561.
- Wall, J. D. 2000. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17**:156–163.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**:256–276.
- Weiss, K. M., and A. G. Clark. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**:19–24.
- Wilson, D. J., D. Falush, and G. McVean. 2005. Germs, genomes and genealogies. *Trends Ecol. Evol.* **20**:39–45.
- Wu, C. F., T. Christensen, and J. Hein. 2001. A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol.* **18**:1929–1939.
- Worobey, M. 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Mol. Biol. Evol.* **18**:1425–1434.
- Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**:1396–1401.
- . 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**:306–314.
- . 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**:555–556.

Scott Edwards, Associate Editor

Accepted January 25, 2006