Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm

Sergei L. Kosakovsky Pond, David Posada,† Michael B. Gravenor,‡ Christopher H. Woelk,* and Simon D. W. Frost**

*Department of Pathology, University of California San Diego; †University of Vigo, Vigo, Spain; and ‡School of Medicine, University of Swansea, Swansea, Wales, United Kingdom

The evolution of homologous sequences affected by recombination or gene conversion cannot be adequately explained by a single phylogenetic tree. Many tree-based methods for sequence analysis, for example, those used for detecting sites evolving nonneutrally, have been shown to fail if such phylogenetic incongruity is ignored. However, it may be possible to propose several phylogenies that can correctly model the evolution of nonrecombinant fragments. We propose a model-based framework that uses a genetic algorithm to search a multiple-sequence alignment for putative recombination break points, quantifies the level of support for their locations, and identifies sequences or clades involved in putative recombination events. The software implementation can be run quickly and efficiently in a distributed computing environment, and various components of the methods can be chosen for computational expediency or statistical rigor. We evaluate the performance of the new method on simulated alignments and on an array of published benchmark data sets. Finally, we demonstrate that prescreening alignments with our method allows one to analyze recombinant sequences for positive selection.

Introduction

Point mutation and recombination are 2 major evolutionary mechanisms driving diversity and adaptation, but their relative contribution varies greatly between genes and organisms (Worobey and Holmes 1999; Awadalla 2003). In many retroviruses, for example, HIV, the rate of recombination may rival or exceed that of point mutation (Zhuang et al. 2002). Conversely, although it has long been thought that animal mitochondrial genomes exhibit low recombination rates, a recent study (Tsaousis et al. 2005) presents evidence to the contrary.

Over the last 2 decades, there has been an explosion of stochastic models for studying evolution due to point mutations (Felsenstein 1981; Muse and Gaut 1994; Felsenstein and Churchill 1996; Savill et al. 2001). These advances led to rapid development of popular phylogeny-based inference methods, such as ancestral dating based on molecular clocks (Korber et al. 2000) and methods for detecting nonneutral evolution at the level of individual codons (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Kosakovsky Pond and Frost 2005b). Recombination can mislead the phylogenetic estimation procedure (Posada and Crandall 2002) and distort subsequent inferences based on inferred phylogenies (Schierup and Hein 2000a, 2000b). Furthermore, the likelihood methods for quantifying selection pressure on codon alignments developed by Nielsen and Yang (1998) may suffer from high rates of false positives when the sequences being analyzed have undergone recombination (Anisimova et al. 2003; Shriner et al. 2003). This is intuitively clear because the evolution of homologous recombinant sequences must be modeled by several phylogenies—one for each nonrecombinant fragment in the alignment. Consequently, an essential step in any phylogeny-based analysis is to screen for and quantify evidence of recombination.

There exist numerous algorithms and software tools geared toward detection and analysis of recombination. Posada and colleagues have compared the performance of different methods on simulated (Posada and Crandall 2001) and biological data (Posada 2002) and found that they can yield vastly different results, necessitating the use of a consensus method approach to obtain reliable inference. However, it may be excessively laborious and not necessarily illuminating to apply multiple methods to an alignment and attempt to integrate the results. For instance, some methods may claim that an alignment is recombination free, whereas others find many recombination events. Moreover, when the goal is not only to merely test for recombination, but also to identify break points and recombinant sequences and to establish statistical support for the inferences, many methods are incapable of this level of detail. Lastly, when the ultimate objective is to apply a phylogeny-based method to a data set with evidence of recombination, it is imperative to determine which segments of the alignment are nonrecombinant and infer an appropriate phylogeny for each segment. This procedure may be preferable to discarding sequences that may have undergone recombination or assuming a single phylogenetic history for the entire alignment.

With this in mind, we propose a pragmatic approach—Genetic Algorithm Recombination Detection—or GARD for short, to rapidly screen multiple-sequence alignments for recombination. The method is designed from the outset to search for evidence of segment-specific phylogenies. Given the maximum number of break points (B, this number can also be inferred), the method will search the space of all possible locations for B or fewer break points in the alignment, inferring phylogenies for each putative nonrecombinant fragment, and assess goodness of fit by an information-based criterion—such as small sample Akaike Information Criterion (AIC) (Sugiura 1978) ($\mathrm{AIC_c}$)—derived from a maximum likelihood model fit to each segment. For $B = 1$, it is practical to quickly screen all possible locations of the break point. This simple approach is shown to perform at least as well as any of the 14 methods examined by Posada and Crandall (2001) on the same simulated data. When $B > 1$, it is often infeasible to perform a "brute-force" search for long sequences. We propose a genetic algorithm (GA) heuristic to quickly explore such a large-state space. Drawing upon the standards of multimodel inference, we

Key words: recombination, phylogenetic incongruence, model selection, genetic algorithms, multimodel inference.

E-mail: spond@ucsd.edu.

Mol. Biol. Evol. 23(10):1891–1901. 2006
doi:10.1093/molbev/msl051
Advance Access publication July 3, 2006

combine the information from all fitted models and assign a level of support to the placement of break points and support for different phylogenies among inferred nonrecombinant segments. We reanalyze 2 collections of previously published biological sequence alignments (Posada 2002; Chare et al. 2003) and compare our findings with those presented originally. Based on several simulation scenarios, we show that GARD has good power and accuracy to detect recombination and identify nonrecombinant sequence fragments. Lastly, we demonstrate how screening for nonrecombinant sequence fragments helps reduce false-positive error rates in the fixed effects likelihood (FEL) phylogenetic method (Kosakovsky Pond and Frost 2005b) used for selection analysis.

Materials and Methods

Rapid Screening for Recombination Using a Single Break Point

Consider an alignment of S sequences with N characters each. Sequences could consist of nucleotides, amino acids, codons, or characters from other alphabets; however, in this article, we apply the GARD method to nucleotide data. If none of the sequences are recombinant, a single phylogeny should fit the data well; otherwise, different regions of sequences may yield different phylogenies if analyzed separately. We assume that the process of point mutation can be adequately modeled by an appropriate time-reversible model of nucleotide substitution, up to the general reversible model (Tavaré 1986), with site-to-site rate variation accounted for with the β–Γ distribution (Kosakovsky Pond and Frost 2005c). Because we can only resolve the location of break points up to the nearest variable site, we adopt the convention that break points must coincide with variable sites, whose number is denoted by $V \in [2, N]$. Actual break points may reside somewhere between variable sites, but phylogenetic search procedures cannot be used to identify exactly where because invariable sites contain no phylogenetic signal. For computational expediency, we employ the Neighbor-Joining (NJ) method (Saitou and Nei 1987) with the TN93 distance metric (Tamura and Nei 1993) to reconstruct the tree topology on each putative nonrecombinant fragment. The NJ method has been shown to perform reasonably well on reconstructing trees from simulated alignments, including large alignments (Tamura et al. 2004), and if additional accuracy is desired, a more computationally demanding method can be invoked. Branch lengths and other model parameters are fitted using the maximum likelihood framework (Felsenstein 1981).

Our algorithm consists of 4 steps:

1. Infer a NJ tree for the entire alignment and obtain the $\mathrm{AIC_c}$ (A_0) score for a given nucleotide model using maximum likelihood to estimate rate parameters and branch lengths. $\mathrm{AIC_c}$ score of a model with p parameters fitted to a sample of size N is defined as $\mathrm{AIC_c} = -2\log L(\hat{\theta}|\mathrm{data}) + 2p\left(\frac{N}{N-p-1}\right)$. $\mathrm{AIC_c}$ is a second-order correction to the standard AIC, and its use has been advocated when the number of samples is not much larger (40× or fewer) than the number of parameters (Burnham and Anderson 2003, p. 66). Note that the use of $\mathrm{AIC_c}$ sensibly requires that there be more observations (alignment columns) than the number of estimated model parameters. The formal requirement for this setting is, consequently, $N > 2(2S-3) + b_\mathrm{p}$, where $2(2S-3)$ counts the number of branches in 2 trees fitted by the GA and b_p refers to the number of rate and frequency parameters in the evolutionary model.

   We hold the estimates of base frequencies and substitution bias parameters at the values obtained from this step. Indeed, it is reasonable to expect that the stationary base distribution and the parameters of the point substitution process on each nonrecombinant fragment are not strongly affected by recombination.
2. We consider all $V-1$ partitions of N sites into 2 continuous blocks, where each block contains at least one variable site and each break point coincides with a variable site.
3. If the break point is placed at site i, we infer a NJ tree individually for each block and compute the $\mathrm{AIC_c}$ score (A_i) of the model that fits branch lengths to each partition independently, holding other parameters fitted in step 1 constant. The single–break point model will have $2S-3$ more estimable parameters than the single-partition model. This step is repeated for all $V-1$ possible locations of the break point.
4. If $A_i < A_0$ for at least one i, then we deduce that some of the sequences in the alignment are recombinant. We equate the relative support for having the break point at site i to its Akaike (1983) weight: $w_i = \frac{\exp(-\Delta_i/2)}{\sum_r \exp(-\Delta_r/2)}$, where $\Delta_i = \mathrm{AIC}^i_\mathrm{c} - \min \mathrm{AIC_c}$, AIC^i_c is score for the model placing the break point at the i-th variable site, and r indexes possible locations of the break point.

Searching for Multiple Break Points Using a GA

When multiple break points are introduced, a brute-force approach rapidly becomes impractical. Even with the simplifying assumption that break points are restricted to V variable sites, there are $\binom{V}{B} \sim \mathrm{O}(V^B)$ possible combinations of $B \geq 1$ break points, when $B \ll V$. For example, if one were to examine the entire HIV-1 genome (~10 kb) for recombination and assume that a quarter of the sites were variable, there would be approximately 10^9 models with 3 break points to consider.

Consequently, we utilize an aggressive population-based hill climber—the CHC GA (Eshelman 1991; Kosakovsky Pond and Frost 2005a)—to search the space of candidate models. A candidate model for B break points is represented by the ordered vector $\mathbf{b} = (v_1, v_2, \ldots, v_B)$, where $1 \leq v_1 \leq v_2 \leq, \ldots, \leq v_B \leq V$ represent the locations of break points in the coordinates of variable sites, ordered left to right. When 2 coordinates are equal, the model collapses to $B-1$ break points. The GA operates on the binary representation of this vector, with single bits serving as units of evolution.

The parameter space for this optimization problem has 2 components: a discrete allocation of possible positions in the sequence to B break points and a vector of real valued parameters corresponding to branch lengths. The CHC algorithm is employed to search through the discrete component of the parameter space, and conventional numerical optimization techniques are used to find maximum likelihood estimates of all other model parameters, given vector **b**. The fitness of every model is measured by its $\mathrm{AIC_c}$ score. Individuals are chosen for mating with probabilities proportional to their fitness.

The CHC always retains the most fit individual from the previous generation and performs 2 basic operations on individuals currently in the population:

- Mating with free recombination: When 2 individuals $\mathbf{b}_1$ and $\mathbf{b}_2$ are picked to mate, their offspring, $\mathbf{b}_O$, is equally likely to inherit bit b_i from either parent.
- Hypermutation: If the diversity of the sample (measured by the range of $\mathrm{AIC_c}$ scores normalized by the score of the best individual) falls below a fixed threshold—0.1% in our implementation—then all individuals in the population, excluding the most fit one, have 15% of randomly selected bits toggled.

Before the new individuals generated by these operations are placed into the population, it may be necessary to re-sort the break points in ascending order to avoid equivalent representations of the same model. The algorithm terminates if the best $\mathrm{AIC_c}$ score remains unchanged over 100 consecutive generations. To increase the proportion of all possible models examined by the GA, a master list of all fitted models is maintained, and if a previously examined model is generated, the algorithm will randomly mutate such an individual (one position at a time), until a new model has been proposed, provided there are any remaining. A typical GA run considers 10^3–10^4 models, hence it is practical to maintain such a list in memory.

For computational expediency, we make the same simplifications as in the single–break point case: NJ trees are reconstructed for each fragment and parameters of the substitution model are estimated first from the entire alignment and held constant for the entire GA run. For each data set, we start with $B = 0$ break points and increase B by 1 for subsequent GA runs, until the $\mathrm{AIC_c}$ score of the best model stops decreasing with increasing B. We note that such incremental changing of B may underestimate the correct number of break points. A more careful search procedure might investigate a fixed range of B (e.g., $B = 1, \ldots, 20$) but incur greater computational costs. Ideally, B would also be a parameter to be determined automatically during the search, but it is challenging to implement a GA that can properly search a parameter space whose dimension may change at run time.

Model-Averaged Break Point Locations

Having fitted M models with B break points each using the GA and computed their corresponding Akaike weights, w_i, where $i = 1, \ldots, M$, we can compute model-averaged support probability P^j_n that the j-th break point rests on nucleotide position n in the alignment $P^j_n = \sum_{i \in M^j_n} w_i$. Here M^j_n denotes the set of models which place their j-th break point (ordered by increasing nucleotide position of the break point) on site n. It is easy to see that $\sum_{n=1}^{N} P^j_n = 1$, for all $1 \leq j \leq B$.

Result Verification

GARD does not explicitly require that tree topologies be different among partitions. For example, if the alignment exhibits strong spatially localized changes in diversity or heterotachy, then a model that fits the same topology but differing branch lengths to segments of the alignment might outperform a single-partition model. Optionally, to verify whether the "topologies" were significantly different between adjacent partitions, we performed a posteriori incongruence tests between all the tree topologies derived from adjacent sequence segments. We used the Shimodaira and Hasegawa (1999) test (SH test) and required that at least 1 pair of the adjacent segments show a statistically significant ($P < 0.01$, when corrected for multiple tests) difference in tree topologies. However, this requirement may be too restrictive because not all recombination events give rise to discordant topologies between sequence fragments (only Type 3 recombination events do, using the notation of Wiuf et al. 2001). When only the improvement in the $\mathrm{AIC_c}$ score is used to detect recombination, other types of events may be identified, and 3-sequence alignments, for which there is a single possible tree topology, can also be handled.

Implementation

All sequence analyses and model fitting were performed using the HyPhy (Kosakovsky Pond et al. 2005) software on a P-node message passing interface cluster. $P - 1$ slave nodes were used to fit various models, and a single master node dispatched the jobs and assembled the results. The size of CHC population was set to $2P - 2$ individuals. We set $P = 17$ for the analyses in this article. A single run of the GA algorithm required from several minutes to several hours, based on the size of the alignment and the number of break points. A Web-based interface for GARD is available at http://www.datamonkey.org/GARD/.

Sequence Alignments

We consider 2 collections of alignments previously analyzed for recombination: 24 mixed data sets from Posada (2002) and 78 viral data sets from Chare et al. (2003). These alignments span a range of size and diversity levels and include a large number of both recombinant and nonrecombinant cases. Furthermore, comparing the performance of the new method to existing ones on previously analyzed alignments provides a direct measure of agreement with the tools currently at the disposal of researchers and helps elucidate conditions under which our approach reaches a different conclusion. Both biological and simulated sequence alignments used in this study can be downloaded from http://www.hyphy.org/pubs/GARD/.

Simulations

Scenario 1 (fig. 1) consisted of 8 nonrecombinant sequences and a single recombinant with 2 break points. The length of the sequences, divergence levels, and base

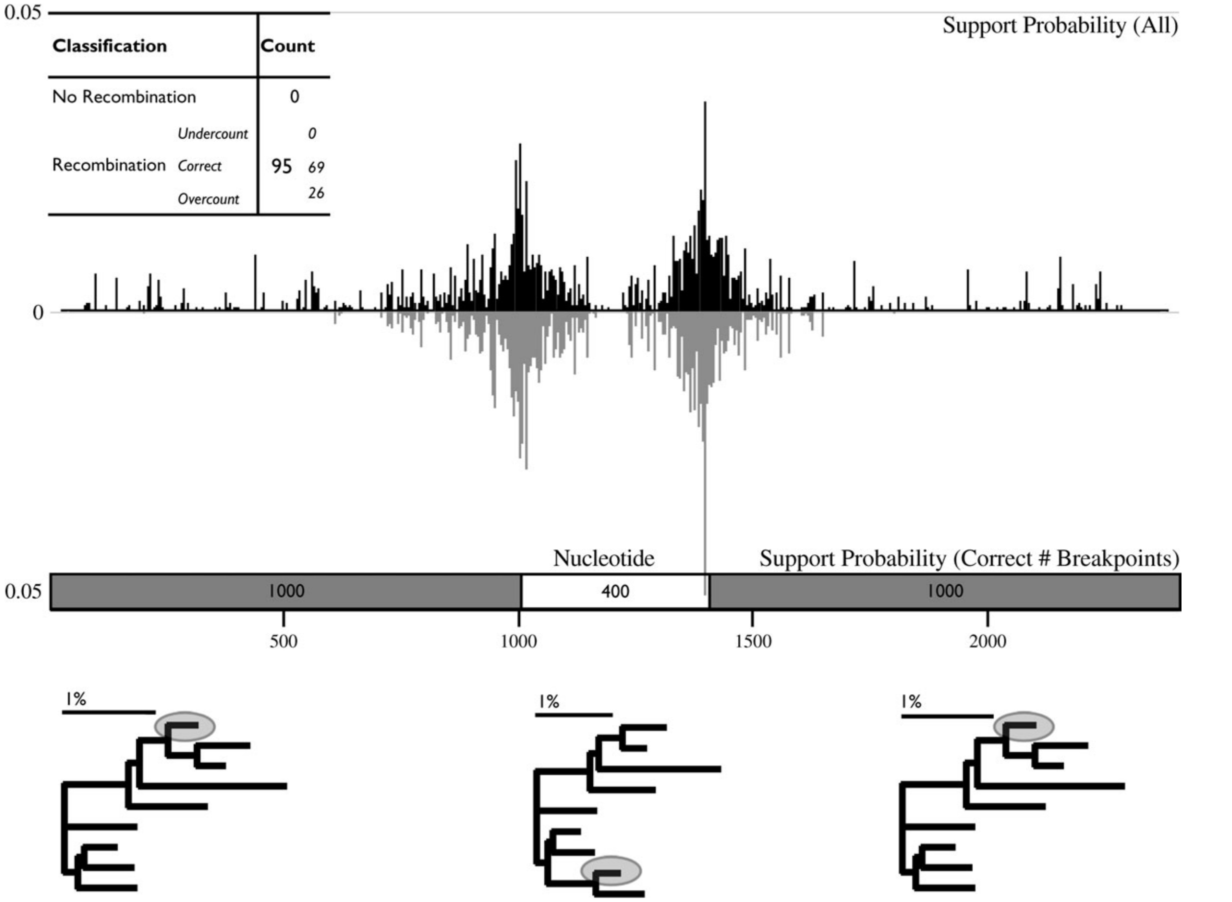
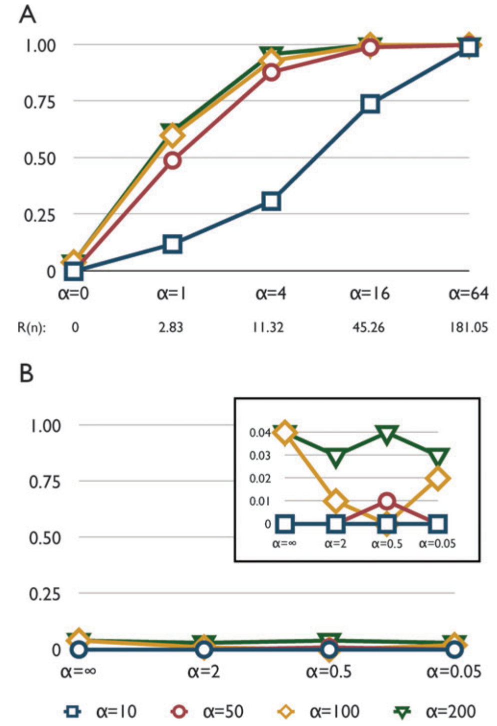


FIG. 1.—Performance of the GARD in detecting recombination for Scenario 1. The trees and sequence fragments used to simulate the data are shown, and recombinant sequences are highlighted. The summary table classifies the results of recombination inference based on 100 data replicates. Probability plots display model-averaged probabilities of finding a break point at a given position in the alignment, averaged either over those runs that detected some recombination (95 runs, top) or over the runs that correctly identified 2 break points (69 runs bottom).

frequencies were derived from HIV-1 subtype B/D recombinants. Data were generated parametrically using the HKY85 (Hasegawa et al. 1985) model of sequence evolution with constant rates across sites and the transition/transversion ratio set to 3. Scenario 2 (fig. S1, Supplementary Material online) extended Scenario 1 to 3 recombination break points involving a clade (ancient recombination) and a single sequence (recent recombination). These 2 scenarios are able to measure the performance of the method over multiple replicates of the same evolutionary process with fixed break points and recombinant sequences.

Additionally, we utilized 8 coalescent-based recombination simulations designed to sample from multiple realizations of the evolutionary process with fixed meta-parameters, such as the number of recombination events and mutation rates. Hundred alignments with 8 sequences with 3,000 nt each were simulated for 2 levels of diversity: $\approx$5% (low) and $\approx$25% (high), selected to reflect the level of divergence within—and between—subtype HIV-1 sequences. Each alignment contained 0, 1, 2, 4, or 8 recombination events. General Time Reversible (GTR) + Γ ($\alpha = 0.5$) model was used to describe the process of character substitution. Base frequencies ($\pi_A = 0.35$, $\pi_C = 0.19$, $\pi_G = 0.22$, and $\pi_T = 0.24$) and substitution rates ($r_{AC} = 2$, $r_{AG} = 5$, $r_{AT} = 0.7$, $r_{CG} = 0.8$, $r_{CT} = 4$, and $r_{GT} = 1$) were chosen to reflect parameters similar to those found in HIV-1 sequences.

Finally, we considered a simulation scenario (Neutral Scenario), in which 32-codon sequences were evolved using 2 randomly generated trees (one for 400 codons and another for 100 codons) and then concatenated. Hundred replicates using 2 fixed trees (one on each partition) were generated. This rather extreme scenario was chosen to model a fixed recombination hot spot that sustains high recombination rates. Phylogenetic signals to the left and the right of the hot spot were effectively independent of one another, but there was strong consistent phylogenetic signal within each region. Each fragment was evolved under a neutral ($dN = dS = 1$) $MG94 \times REV$ codon model (Kosakovsky Pond and Frost 2005c) estimated from an alignment of HIV-1 reverse transcriptase sequences. The concatenated alignment was next analyzed for site-by-site selection with FEL (Kosakovsky Pond and Frost 2005b) based on the NJ tree inferred from the entire alignment. Following a recombination screen with the single–break point model, we next applied FEL to each of the inferred nonrecombinant fragments with NJ trees derived from each segment independently. Additionally, we performed FEL analyses on biological data sets showing strong evidence



C

α	α	α	Best 14	LPT-JC2/LPT-JCall	Single Breakpoint AIC$_c$
10	*0*	∞	*4*	*1/6*	*0*
10	*0*	*2*	*5*	*3/3*	*0*
10	*0*	*0.5*	*5*	*7/3*	*0*
10	*0*	*0.05*	*4*	*3/1*	*0*
10	1	∞	17	22/23	12
10	4	∞	39	58/52	31
10	16	∞	80	79/83	74
10	64	∞	100	97/96	99
50	*0*	∞	*5*	*8/3*	*0*
50	*0*	*2*	*5*	*2/3*	*0*
50	*0*	*0.5*	*5*	*4/4*	*1*
50	*0*	*0.05*	*5*	*9/7*	*0*
50	1	∞	35	30/39	49
50	4	∞	77	73/71	88
50	16	∞	100	98/99	99
50	64	∞	100	100/100	100

FIG. 2.—Single–break point method performance on simulated data from Posada and Crandall (2001). Panels *A* and *B* show the proportion of 100 data replicates that the test classified as recombinant. "θ" denotes the scaled mutation rate. In panel *A*, ρ denotes the scaled recombination rate and *R*(*n*), the expected number of recombination events per replicate. In panel *B*, alignments were simulated without recombination but with gamma-distributed site-to-site substitution rate variation (mean 1, variance of 1/α). These plots are directly comparable with figure 1 in Posada and Crandall (2001). Panel *C* shows the number of data sets (out of 100) that were classified as recombinant by: the best performing method of the 14 examined by Posada and Crandall (2001); the composite likelihood permutation test (McVean et al. 2002), using benchmark results from Carvajal-Rodriguez et al. (forthcoming) based on the Jukes–Cantor model (Jukes and Cantor 1969) of nucleotide substitution and either all sites (JCall) or only those with exactly 2 alleles (JC2); and our single–break point recombination scan.

for recombination, with and without splitting the alignments into nonrecombinant fragments to explore how recombination can affect the inference of codons under selection.

Results

Single–Break Point Screen for Recombination

Using the single–break point screening procedure (see Materials and Methods), we reanalyzed nucleotide data (10 sequences with 1,000 bp) simulated under varying levels of divergence and recombination originally presented in Posada and Crandall (2001). The results for detecting whether recombination has acted on an alignment are summarized in figure 2. Single–break point scanning outperforms all the 14 methods presented in the original study and a more recent composite likelihood permutation test (McVean et al. 2002), both in terms of false positives and power, except for the cases of low sequence divergence, when the performance is comparable to the best of the other methods (fig. 2*C*). This finding is quite remarkable because the assumption of a single break point is likely violated for a vast majority of simulations (fig. 2*A*), where recombination did occur. When the recombination rate is high ($\rho = 64$), there are over 180 recombination events per alignment, on average, and one could expect that all phylogenetic signal is lost. Nonetheless, our method can reliably (>95%) detect recombination even if the level of divergence is low. Because we explicitly model site-to-site rate variation, it is not surprising that the rate of false positives for nonrecombinant data simulated with variable rates is low (fig. 2*B*). The software implementation of our method runs very quickly—for example, 100 alignments with 10 sequences, 1,000 bp long, were screened in about 5 min on 24 cluster nodes. Hence, our approach can be recommended as a rapid recombination-screening tool.

Biological Data Analyses

Table 1 and table S1 (Supplementary Material online) summarize GARD results based on a collection of biological sequence data. For the data sets from Posada (2002), we find high levels of agreements between our results and those achieved with the 50% consensus of 14 recombination detection methods. When recombination is detected, there is invariably a strong level of statistical support for multiple segments, both in terms of goodness of fit and phylogenetic discordance between adjacent partitions. GARD found between 1 and 9 recombination break points, with a wide spectrum of lengths of nonrecombinant fragments. It is worth noting that, in most cases, recombination events appear to follow a complex pattern involving multiple sequences, as evidenced by the low proportion of phylogenetic splits shared among the trees derived from each fragment.

Table 1
GARD Results Based on the Alignments from Posada (2002)

		Nucleotides/	Diversity	Recombination		Mean Tree Splits	Significance	
Gene	*S*	Variable Sites	(%)	Previous	Fragments (bp)	Identity %	$\Delta\mathrm{AIC}_c$	SH $\hat{P} \leq 0.01$
Maize Actin	8	1,008/367	34.3	Yes	111, 183, 162, 424, 128	20.0	203.2	1/4
Neiserria ArgF	9	787/234	25.6	Yes	139, 55, 311, 33, 249	7.0	187.4	1/4
Candida mtDNA	18	2,553/126	1.4	Yes	2,102, 28, 423	11.1	137.5	2/2
Fusarium3	24	4,146/225	1.1	Yes	2,751, 1,006, 389	17.8	203.6	2/2
HGV Genome	16	8,508/2,413	21.1	Yes	69, 1,113, 384, 771, 1,861 274, 116, 689, 2,781, 450	0.9	587.2	3/9
HIV B *env* NR	15	2,234/753	11.7	Yes	40, 802, 53, 102, 70 254, 266, 58, 57, 632	2.5	249.0	2/9
HIV *env*	20	2,205/1,396	44.3	Yes	81, 665, 60, 457 94, 99, 359, 241, 149	4.1	950.9	6/8
HIV *env* NR	11	2,352/1,444	59.8	Yes	37, 762, 160, 74 209, 349, 374, 314, 73	2.9	290.6	1/8
Human DRB1	3	153/24	11.6	Yes	59, 41, 53	100	11.2	N/A[a]
Insecta CO II	7	549/276	51.8	No	138, 125, 173, 89, 24	0	32.2	0/4
Mammalian PDH	5	1,104/400	32.7	No	159, 219, 54, 265, 158, 249	6.7	27.5	0/5
Mammalian PGK	6	1,249/1,014	N/A[b]	No	630, 125, 494	21.5	5.2	0/2
Perom12S	9	750/93	5.8	No	85, 134, 134, 397	5.6	14.6	0/3
PetuniaS-RNase	14	504/405	80.6	Yes	167, 133, 204	40	64.2	0/2
Vertebrate CO I	5	1,506/586	53.2	No	422, 444, 568, 8, 10, 53	3.3	85.2	0/5
Armillaria mtDNA	18	2,234/17	0.1	No	None			
Candidula 16S	44	326/43	1.9	No	None			
Daphnia CO1	18	466/180	53.3	No	None			
Dmel CytB	17	1,137/8	0.09	No	None			
Fusarium Tri 101	16	1,336/64	0.9	No	None			
Gymn ND4	14	663/301	36.7	No	None			
Human HRVI	12	428/42	1.3	No	None			
Wolf CR	34	230/32	3.8	No	None			

NOTE.—Column labels are as follows—S: number of unique haplotypes in the data set; nucleotides/variable sites: total nucleotides and the number of variable sites; diversity, %: mean pairwise distance between sequences derived using a NJ tree built from the entire alignment; previous: 50% method consensus call from Posada (2002); fragments: lengths of nonrecombinant fragments inferred by GARD; mean tree splits identity, %: proportion of tree splits shared by trees, averaged over all pairs; $\Delta\mathrm{AIC}_c$: the improvement in AIC_c of the best model compared to the single tree model; SH $\hat{P} \leq 0.01$: proportion of break points found significant by the SH test on flanking trees, at $P = 0.01$ using Bonferroni correction for multiple comparisons. Data sets are grouped by strength of support for recombination: top group has both AIC_c and SH support, middle group, only AIC_c support, and bottom group has no support. Alignments in bold were thought to have undergone recombination in the original publications (Posada 2002).

[a] Test is not directly applicable to a 3-sequence alignment.

[b] Substitutions along one of the branches were saturated, resulting in a branch length of numerical ∞.

GARD reconfirmed all 5 genes found to have undergone recombination by the phylogenetic incongruence test in Chare et al. (2003), with very similar locations of break points (Table S1, Supplementary Material online). Two of these genes were found to contain multiple break points. Interestingly, our approach found 11 additional genes with putative recombinant sequences, both based on AIC_c goodness of fit and the SH test for phylogenetic incongruence. Thirteen other genes were classified as recombinant by AIC_c alone, indicating the possibility of Type 1 or 2 recombination events or perhaps the inadequacy of the model for character substitution but did not have strong evidence of phylogenetic discordance. This suggests that another process (such as space-localized selection or substitution rate variation) could be affecting branch lengths of the phylogeny along the sequence. Some of the phylogenetic incongruence signal (e.g., Measles M gene) is not likely to be a result of recombination but rather the effect of adenosine to inosine hypermutation events from cases of subacute sclerosing panencephalitis, which result in phylogenetic patterns resembling convergent evolution (Woelk et al. 2002). GARD does not rely on manual identification of sequences "migrating" along the tree between different sequence fragments, as does the Chare et al. (2003) method, thus it is not surprising that it appears to be more sensitive. Indeed, all genes with evidence of recombination as detected by at least two of the three methods used by Chare et al. (2003) are also detected by our approach.

Simulated Data Analyses

When the model used to simulate recombinant sequences matches the underlying assumptions (i.e., there is phylogenetic incongruence between 2 or more sequence fragments, but the evolutionary process is the same for the entire sequence) of our detection methods, as is the case for simulation Scenarios 1 and 2, GARD performed well both in detecting the number of break points and their location in the sequence (fig. 1 and fig. S1, Supplementary Material online). Recombination was detected reliably (95/100 and 100/100 cases, respectively). The correct number of break points was inferred in 69/100 and 82/100 cases. GARD had a slight tendency to overcount the number of break points (26 times for Scenario 1 and 11 times for Scenario 2). When break point counts were inferred correctly, their location was found reasonably accurately, even though confidence intervals (CIs) for the locations were fairly wide (fig. 1 and fig. S1, Supplementary Material online).

To quantify the performance GARD when the underlying evolutionary model may be different from the one

Table 2
Recombination Inference Results Based on Simulation Scenarios 3 (low diversity) and 4 (high diversity)

		Count by Inferred Break Points $N = 100$									Detectable	
Diversity	Events	0	1	2	3	4	5	6	7	Total > 0	AIC_c	SH
Low	0	90	10							10	N/A	N/A
Low	1	44	39	17						56	31	6
Low	2	26	50	18	4	2				74	46	18
Low	4	16	28	41	11	4				84	55	17
Low	8	3	25	33	24	11	3	1		97	57	0
High	0	94	6							6	N/A	N/A
High	1	24	41	23	8	3	1			76	56	24
High	2	12	33	35	15	5				88	62	25
High	4	1	22	36	25	11	4	1		99	83	43
High	8	2	5	23	30	22	12	5	1	98	86	0

NOTE.—Column labels are as follows—events: number of recombination events simulated in each alignment; count by inferred break points: number of data replicates for which a given number of break points were inferred; total > 0: number of data replicates for which any evidence of recombination was inferred; detectable: how many data replicates showed evidence of phylogenetic incongruence, when break points were placed at the correct (simulated) locations.

assumed by the method, we evaluated 800 coalescent-based simulated data sets. As expected, the ability to detect recombination somewhere in the sequence increases both with the level of divergence and the extent of recombination (table 2), with near-perfect power for alignments with 8 recombination events. However, recombination signal is quickly saturated for small alignments (8 sequences), and the number of break points is often underestimated. Interestingly, this limitation may be due not to the GA search procedure but rather to our limited ability to infer phylogenetic trees from short fragments of small alignments. If we were to use the correct placement of recombination break points and perform the AIC_c or the SH tests as discussed in Materials and Methods, only a small percentage of alignments would contain evidence of recombination (see table 2). The finding is especially striking for scenarios with 8 recombination events per alignment, where not a single replicate contained enough information to statistically support discordant phylogenies. The inability to reliably resolve phylogenies is clearly a fundamental limitation of all tests based on phylogenetic discordance rather than that of GARD alone.

If one is concerned with inferring the location of recombination break points, then the situation is less clear (table 3). The best possible outcome for GARD is to place each inferred break point at a variable site that is nearest to a true recombination break point. This happens about 20% of the time (table 3). More realistically, the model-averaged CI for the location of a given inferred break point should contain at least 1 recombination break point. This is the case about 60% of the time. Overall, 60–70% of inferred break points have a true break point in the 95% CI. Another measure of inference quality is the median distance from each inferred break point to the nearest “true” break point (or more accurately, to the nearest variable site closest to the “true” break point). This quantity, predictably, decreases with the increasing number of break points and level of sequence divergence (table 3). The distribution of distances resembles an exponential form (fig. S2, Supplementary Material online). To quantify the statistical significance of median distances derived with our algorithm, we conducted a simple simulation. We computed median distances to correct break points based on 1,000 random placements of break points in all 100 replicates in every scenario. Random

Table 3
Quality of Break Point Location Inference Based on Simulation Scenarios 3 (low diversity) and 4 (high diversity)

		Inferred Break Points									
		All		Exact		Approximate			Median Distances (bp)		
Diversity	Events	Count	%	Count	%	Count	%	PPV	Inferred	Random	*P* Value
Low	1	73	73	15	20.5	29	39.7	60.3	181	850.75	<0.001
Low	2	106	53	25	23.6	51	48.1	71.7	64	458	<0.001
Low	4	159	39.8	41	25.8	66	41.5	67.3	31	225.5	<0.001
Low	8	228	28.5	80	35.1	100	43.9	78.9	12	116	<0.001
High	1	123	>100	21	16.4	44	34.4	50.8	150.5	899.25	<0.001
High	2	168	84	32	19.0	69	41.1	60.1	36.5	491	<0.001
High	4	239	59.8	60	25.1	92	38.5	63.6	15	240	<0.001
High	8	326	40.8	73	22.4	151	58.0	68.7	15	124	<0.001

NOTE.—Column labels are as follows—events: number of recombination events in each alignment; all inferred break points: the total number (and percentage of all simulated break points) of break points identified across all 100 replicates; exact inferred break points: total number (and percentage of all identified break points) of break points identified to be at the variable site nearest to a true recombination break point; approximate inferred break points: total number (and percentage of all identified break points) of 95% model-averaged CIs that include a true break point; PPV: positive predictive value of having correctly (exactly or within a 95% CI) identified a break point; inferred median distance: median (overall inferred break points) of the distance from the break point to the closest variable site nearest to a true break point; random median distance: median of median distances over 1,000 replicates of random break point placements (see Materials and Methods); *P* value for median distance: proportion of 1,000 random replicates that had median distance from true break points no greater than the inferred median distance.

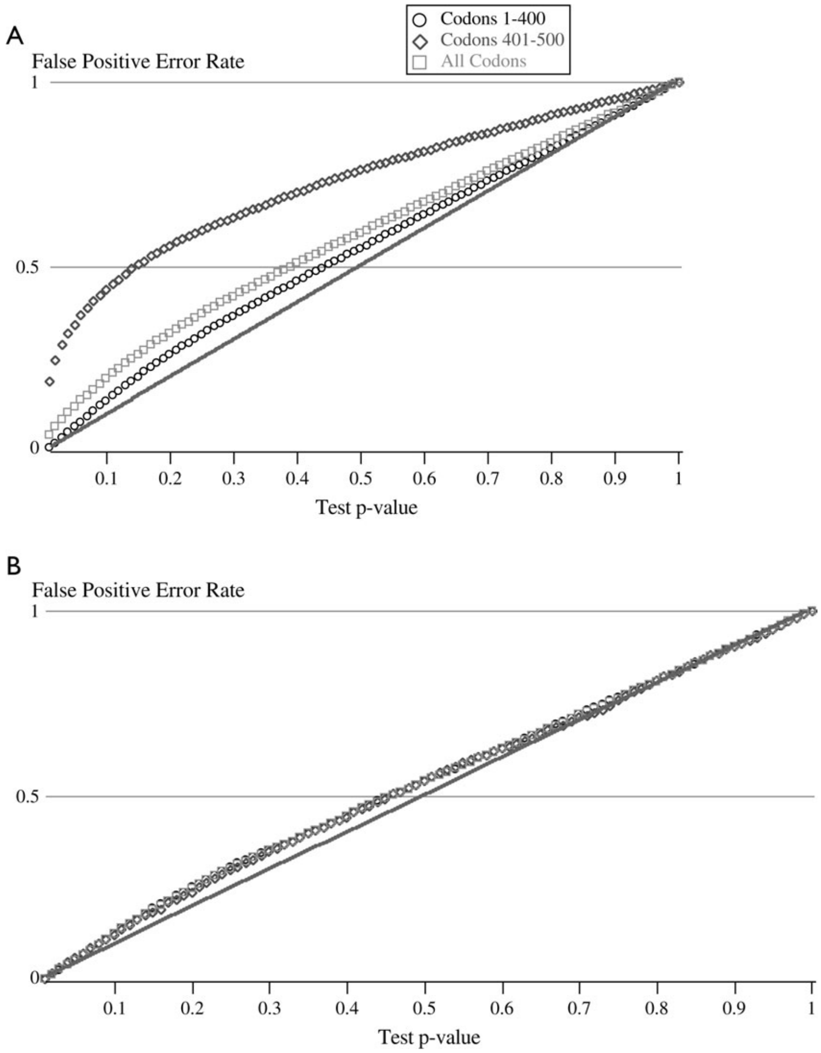


FIG. 3.—False-positive error rates for the FEL test for selected (both positively and negatively) sites based under the Neutral Scenario. Panel *A* shows the error rates for the uncorrected (single partition) FEL and panel *B*, for the corrected (2 partitions) FEL. Solid lines indicate expected error rates, based on the *P* value. Tabulated error rates are presented for the first 400 codons (evolved under one tree), the last 100 codons (evolved under a different tree), and the joint error rate for all 500 codons, averaged over 100 replicates.

break points were placed on variable sites only, and the number of break points allocated to a replicate was randomly drawn from the distribution of the number of inferred break points for that scenario. *P* values of observing smaller median distances to correct break points by chance were computed based on 1,000 replicates. In all cases, the median distance from inferred break points to correct ones was significantly less than that expected by chance.

Effect of Recombination on Site-by-Site Analyses of Selection

False-positive rates of the FEL method were adversely affected (fig. 3) by extensive recombination in the Neutral Scenario (see Materials and Methods). Generally, FEL is expected to have well-controlled rates of false positives, relative to the *P* value of the test (Kosakovsky Pond and Frost 2005b), but in this case, inference on the last 100 codons of each of the alignments was subject to Type I error far in excess of the nominal *P* value (fig. 3*A*), whereas the error rate for the first 400 codons was effectively the same as the *P* value. Intuitively, a topology inferred from all 500 codons is "almost" correct for the first 400 codons and "never" correct for the last 100 codons. A simple corrective procedure, in which we split each of the 100 simulated alignments into 2 fragments, identified by the single–break point scan on that alignment, and then analyzed each segment separately with FEL, restored good statistical properties of FEL (fig. 3*B*).

We also analyzed 14 data sets (see Table S1, Supplementary Material online) that showed both AIC_c and SH support for recombination, using FEL, with and without splitting the alignment into nonrecombinant fragments. The list of sites subject to positive selection can vary substantially between corrected and uncorrected FEL analyses (table 4). This observation reinforces previous findings (Anisimova et al. 2003; Shriner et al. 2003) indicating that

Table 4
Effect of Correcting for Recombination When Using FEL to Detect Positively Selected Sites

	Positively Selected Codons	
Virus and Gene	Uncorrected FEL	Corrected FEL
Cache valley G	212, 516, 546, 551	None
Canine distemper H	158, **179**, **264**, **444**	**179**, **264**, **444**, 548
Crimean Congo hemm. fever NP	**195**	9, **195**
Hantaan G2	None	None
Human parainfluenza (1) HN	37, **91**, **358**, 556	**91**, **358**
Influenza A (human H2N2) HA	**87**, 166, **252**, **358**	**87**, 147, **252**, **358**
Influenza B NA	**42**, **106**, **345**, **436**	**42**, **106**, **345**, **436**
Mumps F	**57**, **480**	**57**, **480**
Mumps HN	399	None
Newcastle disease F	**1**, 4, **5**, **7**, **16**, 18, **108**, 516	**1**, **5**, **7**, **16**, **108**, 493, 505
Newcastle disease HN	**2**, 54, **58**, **228**, **262**, **284**, **306**, **471**	**2**, **58**, **228**, **262**, **284**, **306**, **471**
Newcastle disease N	**425**, **430**, **466**	**425**, **430**, 462, **466**
Newcastle disease P	12, **56**, **65**, **174**, **179**, 188, **189**, **204**, **208**, **213**, 217, **218**, 239, **306**, **332**	**56**, **65**, 146, 153, **174**, **179**, **189**, 193, **204**, **208**, **213**, **218**, 261, **306**, **332**
Puumala NP	79	None

Note.—Test $P < 0.1$ was used to classify sites as selected. Codon sites found under selection by both methods are shown in bold.

recombination can significantly alter the results of selection analyses.

Discussion

Recombination is a major evolutionary force in many organisms and can have a profound impact on evolutionary rates. Not only is recombination of interest in its own right, but also analyses of selection pressure may be confounded by its presence or absence. Maximum likelihood methods of codon substitution, used to estimate selection pressures on sites in terms of the ratio of nonsynonymous to synonymous substitution rates, may generate many false-positive sites when recombination is not taken into account. Results based on Poisson random field models (Sawyer and Hartl 1992), used to infer selection pressures from the frequency spectrum of sites under the assumption of loose linkage between sites, may be misleading in the absence of recombination. Hence, screening of recombination should be an integral part of phylogenetic analyses.

The use of phylogenetic incongruence among fragments of a sequence alignment to detect recombination is not a new approach (Koop et al. 1989; Fitch and Goodman 1991; Salminen et al. 1996; Grassly and Holmes 1997; McGuire and Wright 1998). Many of the existing methods (Holmes et al. 1999; Archibald and Roger 2002) rely on a sliding window approach. However, the length of the sliding window and the way it is moved along the sequence can strongly influence recombination inference. Several methods based on Markov Chain Monte Carlo (Husmeier and McGuire 2002; Suchard et al. 2002; Minin et al. 2005) are free of the sliding window limitations, but due to computational expense, they can only be used to examine small or medium alignments. In contrast, GARD is very intuitive, simple to implement and extend, and runs quickly on a computer cluster. Most importantly, it works very well in multiple scenarios, yielding good power and low rates of false positives.

It is remarkable that even our single–break point method outperforms almost all existing methods when detecting the presence of recombination, even when the true number of break points is extremely high. Given the speed of this approach, it can be recommended if one is only interested in screening for the presence or absence of recombination. The one method (Maynard Smith and Smith 1998) that performs better under some parameter regimes is not robust to rate heterogeneity. Given that it is difficult to tease apart recombination and rate heterogeneity, robustness of results is an important consideration in choosing a method to detect recombination. Our multiple break point model generates a rich set of inferences, on the number and location of break points, sequences involved in the recombination events, and the confidence in these inferences. We have also demonstrated that in certain situations, a simple screen for recombination can be used to correct analyses that detect sites evolving adaptively and to mitigate high rates of false positives incurred by uncorrected analyses of recombinant sequences for selection.

Our method has a number of limitations. Sometimes discordant phylogenetic signal may arise not through recombination but through a region evolving under a different evolutionary model. This may be the case for measles virus, which is thought to be recombination free. Despite our best attempts to automate the process of recombination screening as much as possible, we stress the importance of analyzing the resultant trees for the different segments and making an informed judgment about whether the results make sense or not. Optimistically, we note that GARD can easily accommodate substitution models of arbitrary complexity, and as methodological developments occur in this area, the performance of our approach may also improve. Like all methods, ours cannot detect recombination in regions where there is no genetic diversity.

In conclusion, we have developed a straightforward method for detecting discordant phylogenetic signal in alignments of DNA or protein sequences, which provides estimates of the number and location of break points and segment-specific phylogenetic trees. GARD does not require a nonrecombinant reference alignment (cf. bootscanning, see Salminen et al. 1995; Lole et al. 1999), and recombination between ancestral sequences is also accommodated. GARD outperforms other methods in terms of levels of Type I and Type II error (fig. 1) and can employ arbitrarily complex models of substitution. Furthermore, it can be run in parallel on a cluster of computers, and so is well positioned to screen for recombination in large data sets. We hope that this will encourage researchers to make recombination screening a routine part of their evolutionary analyses.

Supplementary Material

Supplementary Table S1 and Figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This research was supported in part by the National Institutes of Health (AI43638, AI47745, and AI57167), the University of California Universitywide AIDS Research Program (grant number IS02-SD-701), and by a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to S.D.W.F. and S.L.K.P (AI36214). D.P. was supported by grant R01-GM66276 from the US National Institutes of Health, grant BFU2004-02700 of the Spanish Ministry of Education and Science, and the "Ramón y Cajal" programme of the Spanish government. C.H.W. was further assisted by the Veterans Affairs Research Center for HIV and Hepatatis C Virus Infection.

Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health grant no. AI57167.

Literature Cited

Akaike H. 1983. Information measures and model selection. Int Stat Inst 44:139–49.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–36.

Archibald JM, Roger AJ. 2002. Gene conversion and the evolution of euryarchael chaperonins: a maximum likelihood-based method for detecting conflicting phylogenetic signals. J Mol Evol 55:232–45.

Awadalla P. 2003. The evolutionary genomics of pathogen recombination. Nat Rev Genet 4:50–60.

Burnham K, Anderson D. 2003. Model selection and multimodel inference. 2nd ed. New York: Springer.

Carvajal-Rodriguez A, Crandall KA, Posada D. 2006. Recombination estimation under complex evolutionary models with the coalescent composite likelihood method. Mol Biol Evol. Forthcoming.

Chare ER, Gould EA, Holmes EC. 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. J Gen Virol 84:2691–703.

Eshelman LJ. 1991. The CHC adaptive search algorithm: how to do safe search when engaging in nontraditional genetic recombination. In: Rawlins GJE, editor. Foundations of genetic algorithms. San Mateo, CA: Morgan Kaufmann Publishers. p 265–83.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–76.

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104.

Fitch D, Goodman M. 1991. Phylogenetic scanning: a computer assisted algorithm for mapping gene conversion and the recombination events. Comput Appl Biosci 7:207–15.

Grassly N, Holmes E. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol Biol Evol 14:239–47.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Mol Biol Evol 21:160–74.

Holmes EC, Worobey M, Rambaut A. 1999. Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol 16:405–9.

Husmeier D, McGuire G. 2002. Detecting recombination with MCMC. Bioinformatics 18:S345–53.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HM, editor. Mammalian protein metabolism. New York: Academic Press. p 21–132.

Koop B, Siemieniak D, Slightom J, Goodman M, Dunbar J, Wright P, Simons E. 1989. Tarsius delta- and beta-globin genes: conversions, evolution, and systematic implications. J Biol Chem 264:68–79.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789–96.

Kosakovsky Pond SL, Frost SD. 2005a. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22:478–85.

Kosakovsky Pond SL, Frost SD. 2005b. Not so different after all: a comparison of methods for detecting amino-acid sites under selection. Mol Biol Evol 22:1208–22.

Kosakovsky Pond SL, Frost SD. 2005c. A simple hierarchical approach to modeling distributions of substitution rates. Mol Biol Evol 22:223–34.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–9.

Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 73:152–60.

Maynard Smith J, Smith N. 1998. Detecting recombination from gene trees. Mol Biol Evol 15:590–9.

McGuire G, Wright F. 1998. TOPAL: recombination detection in DNA and protein sequences. Bioinformatics 14:219–20.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–41.

Minin VN, Dorman KS, Fang F, Suchard MA. 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics 21:3034–42.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–24.

Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–36.

Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol Biol Evol 19:708–17.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci 98:13757–62.

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

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–25.

Salminen M, Carr J, Burke D, McCutchan F. 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retrovir 11:1423–5.

Salminen M, Carr J, Burke D, McCutchan F. 1996. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. AIDS Res Hum Retrovir 11:1423–5.

Savill NJ, Hoyle DC, Higgs PG. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics 157:399–411.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics 132:1161–76.

Schierup M, Hein J. 2000a. Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–91.

Schierup M, Hein J. 2000b. Recombination and the molecular clock. Mol Biol Evol 17:1578–9.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–6.

Shriner D, Nickle DC, Jensen MA, Mullins J. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res 81:115–21.

Suchard MA, Weiss RE, Dorman KS, Sinsheimer JS. 2002. Oh brother, where art thou? a Bayes factor test for recombination with uncertain heritage. Syst Biol 51:715–28.

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat Theory Meth A7:13–26.

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16: 1315–28.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–26.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci 101:11030–5.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci 17:57–86.

Tsaousis AD, Martin DP, Ladoukakis ED, Posada D, Zouros E. 2005. Widespread recombination in published animal mtDNA sequences. Mol Biol Evol 22:925–33.

Wiuf C, Christensen T, Hein J. 2001. A simulation study of the reliability of recombination detection methods. Mol Biol Evol 18:1929–39.

Woelk C, Pybus O, Li J, Brown D, Holmes E. 2002. Increased positive selection pressure in persistent (SSPE) versus acute measles virus infections. J Gen Virol 83:1419–30.

Worobey M, Holmes EC. 1999. Evolutionary aspects of recombination in RNA viruses. J Gen Virol 80:2535–43.

Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, Preston BD, Dougherty JP. 2002. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J Virol 76:11273–82.

Arndt von Haesler, Associate Editor

Accepted June 27, 2006