Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data

David Posada

Department of Zoology, Brigham Young University, Provo, Utah

The performance of 14 different recombination detection methods was evaluated by analyzing several empirical data sets where the presence of recombination has been suggested or where recombination is assumed to be absent. In general, recombination methods seem to be more powerful with increasing levels of divergence, but different methods showed distinct performance. Substitution methods using summary statistics gave more accurate inferences than most phylogenetic methods. However, definitive conclusions about the presence of recombination should not be derived on the basis of a single method. Performance patterns observed from the analysis of real data sets coincided very well with previous computer simulation results. Previous recombination inferences from some of the data sets analyzed here should be reconsidered. In particular, recombination in HIV-1 seems to be much more widespread than previously thought. This finding might have serious implications on vaccine development and on the reliability of previous inferences of HIV-1 evolutionary history and dynamics.

Introduction

The detection of recombination from DNA sequences is relevant to the understanding of evolutionary and molecular genetics. Not surprisingly, a plethora of methods have been developed during the last 15 years to detect the presence of recombination in sequence alignments. However, the performance of these methods has been evaluated only recently (Maynard Smith 1999; Brown et al. 2001; Posada and Crandall 2001; Wiuf, Christensen, and Hein 2001).

Although computer simulations offer an unlimited range of possibilities for the evaluation of the statistical performance of a given method, results from simulation studies alone are always limited. Models used in simulation studies are simplifications of reality. Recombination among sequences has been simulated using the coalescent with recombination (Brown et al. 2001; Posada and Crandall 2001; Wiuf, Christensen, and Hein 2001) or using a more classical forward approach (Maynard Smith 1999). In these simulations, recombination fragments are defined from a single breakpoint to the end of the sequences, and the exchange is performed in a reciprocal way, so each sequence is a donor and a receptor of a fragment. However, this is not the way that recombination occurs in many organisms, including viruses and bacteria, where recombinational events often include more than one breakpoint, and nonreciprocal exchanges are the norm. In addition, nucleotides are evolved under simple reversible substitution models, which assume independence of sites and a stationary and homogeneous substitution process. Moreover, it is not easy to design a simulation study that does not favor some method(s) (Hillis, Mable, and Moritz 1996). Different methods make different assumptions on the processes that generate data. In a simulation study, methods that make the same assumptions as the model used to generate the data should be, in principle, favored. For example, Brown et al. (2001) use the performance of a likelihood ratio test of recombination based on the coalescent as the reference with which the performance of other recombination detection methods is compared.

Despite their limitations, computer simulations are useful. If a result holds in simulations across a wide variety of conditions, they can suggest that the result is general and should apply in real data sets (Swofford et al. 1996). One obvious way of checking this proposal is through the analysis of real data sets utilizing the same software implementations used in the computer simulations. By comparing the results obtained from computer simulations with the inferences obtained from real data, we simultaneously serve two purposes. First, similar results from both types of studies strengthen conclusions regarding performance of the methods evaluated. Second, the analysis of empirical data provides a check of simulation studies, validating (or not) the utility of the models used in the simulations to evaluate a method's performance.

The analysis of empirical data sets alone may be useful to study the performance of recombination methods, especially when the recombinational events are more or less known (Drouin et al. 1999). However, performance of a method is best evaluated through the combined analysis of empirical and simulated data. Indeed, the analysis of empirical data sets is a very desirable and rare complement to simulation studies (Hillis and Huelsenbeck 1994). Here, several empirical data sets are analyzed for the presence of recombination to compare the inferences with those obtained from previous computer simulations (Posada and Crandall 2001). The objectives of this study are threefold (1) to decipher the presence of recombination in several representative empirical data sets, (2) to obtain a better understanding of the performance of recombination detection methods, and (3) to evaluate the utility of models of reticulated sequence evolution.

Materials and Methods

The methodology employed here consists of the following steps: (1) select two groups of empirical data

Key words: detection of recombination, mtDNA recombination, HIV-1 recombination.

Address for correspondence and reprints: David Posada, Variagenics, 60 Hampshire Street, Cambridge, Massachusetts 01239. E-mail: dposada@variagenics.com.

Mol. Biol. Evol. 19(5):708–717. 2002

sets, where (a) recombination is assumed to be absent and (b) recombination has been suggested; (2) apply 14 different methods to detect the presence of recombination; and (3) for each data set, compare the inferences obtained among different recombination methods.

Methods for Detecting Recombination

For the last 20 years, a number of methods have been developed to detect recombination from a set of aligned DNA sequences. David Robertson (Zoology, University of Oxford) provides a web site with links to many of these methods at http://grinch.zoo.ox.ac.uk/RAP_links.html. In general, recombination methods could be tentatively classified as described in the subsequent paragraphs.

(a) Distance Methods

Distance methods search for inversions along the alignment of the distance patterns among sequences (Weiller 1998). In general, they use a sliding window approach and estimate statistics based on genetic distances among sequences. Because the phylogeny does not need to be known, these methods are typically fast.

(b) Phylogenetic Methods

Several methods infer recombination when phylogenies from different parts of the genome result in discordant topologies. When comparisons of adjacent sequences yield different branching patterns, there is reason to suspect recombinational events. If the consequence of such changes results in reconciling different sequence phylogenies to a single phylogeny, then the existence of such events becomes a reasonable hypothesis (Hein 1990, 1993; Grassly and Holmes 1997; Holmes, Worobey, and Rambaut 1999; Lole et al. 1999; Martin and Rybicki 2000). These are the methods most extensively used in the literature.

(c) Compatibility Methods

Compatibility methods are phylogenetic methods that are based on site-by-site analyses (Drouin et al. 1999). These methods define two sites as compatible when their evolutionary history is congruent with the same tree (Sneath, Sackin, and Ambler 1975), and they do not require the phylogeny of the sequences to be known (Jakobsen and Easteal 1996; Jakobsen et al. 1997).

(d) Substitution Distribution

Nucleotide substitution distribution methods examine the sequences either for a significant clustering of substitutions or for a fit to an expected statistical distribution (Stephens 1985; DuBose, Dykhuizen, and Hartl 1988; Sawyer 1989; Maynard Smith 1992; Takahata 1994; Sneath 1995; Maynard Smith and Smith 1998; Sawyer 1999; Worobey 2001).

Implementation of Recombination Detection Methods

Fourteen recombination methods were implemented (table 1). The details of each of these methods have been described elsewhere (see also Posada and Crandall 2001). Particular details of the implementations are described subsequently.

(1) Bootscanning (Salminen et al. 1996)

The windows program Simplot (Lole et al. 1999) (http://www.med.jhu.edu/deptmed/sray/download/) was modified by Stuart Ray (Simplot's author) to implement the bootscanning of every sequence in the alignment against the rest. A sliding window size of 200 bp and step size of 10 nt were used. Neighbor-Joining trees were estimated using F84 distances (Felsenstein 1984, 1991), and bootstrap values were obtained from 100 replicates. Several bootstrapping thresholds for assignment of parenthood were explored (70%, 90%, and 95%), but because only the 95% threshold provided a false positive rate below 10% in computer simulations, this was the threshold used here.

(2) Geneconv (Sawyer 1999)

The program Geneconv 1.81 (http://www.math.wustl.edu/~sawyer/geneconv/index.html) was employed. The global permutation *P*-values based on BLAST-like global scores (10,000 replicates) smaller than 0.05 were considered as evidence of recombination. A multiple comparison correction is already built into these *P*-values. The default value of the parameter *gscale* (*gscale* = 0) was used.

(3) Homoplasy Test (Maynard Smith and Smith 1998)

Two qbasic programs written by J. Maynard Smith (http://www.biols.susx.ac.uk/home/John_Maynard_Smith/) were translated into a single C program. No outgroup was used, and the number of effective sites, *Se*, was taken to be 0.6 × the total number of sites. For coding sequences, only third positions were included in the analysis.

(4) Informative Sites Test (Worobey 2001)

The program Pist (Worobey 2001) (http://evolve.zoo.ox.ac.uk/software/PIST/PIST.html) was used. Maximum likelihood trees and substitution estimates were obtained in PAUP* (Swofford 2000) for the best-fit model selected by Modeltest 3.06 (Posada and Crandall 1998). In the case of coding sequences, only third positions were used. For the parametric simulation of the null distribution of the test statistic, 100 replicates were used.

(5) Maximum Chi-Square (Maynard Smith 1992)

The computer program MaxChi2 was written in C, implementing a modification of the maximum chi-square method (Maynard Smith 1992) suggested by Wiuf, Christensen, and Hein (2001). The statistic employed was the maximum chi-square in the original

Table 1
Methods for Detecting Recombination Evaluated for Performance

Method	Implementation	Reference	Category	Inference of Recombination[a]	Recombinant Identification[b]	Breakpoint Identification[c]	Amount of Recombination[d]	Minimum Number of Sequences[e]
Bootscanning	*Simplot*	(Salminen et al. 1996; Lole et al. 1999)	Phylogenetic	Yes/No	Yes	Yes	No	4
Geneconv	*Geneconv*	(Sawyer 1999)	Substitution	Permutation of sites	Yes	Yes	No	3
Homoplasy test	*Homoplasy Test*[f]	(Maynard Smith and Smith 1998)	Substitution	Permutation of sites	No	No	Yes	4
Informative sites test	*Pist*	Worobey 2001)	Substitution	Monte Carlo	No	No	Yes	4
Maximum chi-square	*MaxChi*[2f]	(Maynard Smith 1992)	Substitution	Permutation of sites	Yes	Yes	No	2
Maximum mismatch chi-square	*Chimaera*	(Posada and Crandall 2001)	Substitution	Permutation of sites	Yes	Yes	No	3
Phylogenetic profile	*PhyPro*[f]	(Weiller 1998)	Distance	Permutation of sites	Yes	Yes	No	3
Partial likelihood	*Plato*	(Grassly and Holmes 1997)	Phylogenetic	Monte Carlo	No	Yes	No	4
Rdp	*Rdp*	(Martin and Rybicki 2000)	Phylogenetic	Binomial distribution	Yes	Yes	No	3
Recombination parsimony	*RecPars*	(Hein 1990)	Phylogenetic	Yes/No	Yes	Yes	No	4
Reticulate	*Reticulate*	(Jakobsen and Easteal 1996)	Compatibility	Permutation of sites	No	No	Yes	4
Runs Test	*Runs Test*[f]	(Takahata 1994)	Substitution	Geometric distribution	No	No	No	2
Sneath Test	*Sneath Test*[f]	(Sneath 1995)	Substitution	Normal distribution	Yes	No	No	2
Triple	*Triple*[f]	(Kuhner et al. 1991)	Phylogenetic	Derived distribution	Yes	Yes	No	3

[a] The presence of recombination was identified by a Yes or No answer or by a *P*-value that was obtained from permutations of the data, Monte Carlo, or from other statistical distributions.
[b] Gives direct information on the sequences involved in the recombinational event (recombinant and parental[s]).
[c] Explicitly identifies the recombination breakpoints in the alignments or in the individual sequences.
[d] Explicitly measure the amount of recombination.
[e] Minimum number of sequences needed.
[f] Program written in C available from the author.

alignment. For each pair of sequences, this statistic was calculated on a sliding window that moved one nucleotide at a time and included only variable sites. The width of this window was arbitrarily set to the total number of variable sites divided by 1.5. The *P*-value for the null hypothesis of no recombination was estimated as the number of times the maximum chi-square was smaller than the maximum chi-square out of 1,000 permuted alignments (obtained by randomizing the position of the columns in the alignment).

(6) Maximum Match Chi-Square (Posada and Crandall 2001)

The computer program Chimaera was written in C implementing the maximum mismatch chi-square method. This method is a modification of the maximum chi-square (see previously). The only differences are that a different test statistic is used (in this case the statistic is the maximum match chi-square), and that the test statistic is calculated for each possible triplet of sequences (instead of for each pair). For each triplet, each sequence was treated alternatively as the potential recombinant, and the maximum match chi-square statistic was calculated. The maximum match chi-square is the chi-square statistic from the contingency table built with the number of sites in the putative recombinant that match each one of the parental, before and after an arbitrary breakpoint.

(7) Phylogenetic Profile (Weiller 1998)

The computer program PhyPro was written in C extending the Phylogenetic Profile method, which in its original form does not provide statistical significance. The test statistic employed was the minimum distance vector correlation in the original alignment. For each data set, this statistic was calculated on a sliding window that moved one nucleotide at a time and included only variable sites. The width of this window was arbitrarily set to the total number of variable sites divided by 1.5. The *P*-value for the null hypothesis of no recombination was estimated as the number of times the minimum distance vector correlation was smaller than the minimum distance vector correlation out of 1,000 permuted alignments (obtained by randomizing the position of the columns in the alignment).

(8) Partial Likelihood Assessed Through Optimization (Grassly and Holmes 1997)

The program Plato (Grassly and Holmes 1997) (http://evolve.zoo.ox.ac.uk/software/Plato/Plato2.html) was used. Maximum likelihood trees and substitution model estimates were obtained in PAUP* (Swofford 2000) for the best-fit model selected by Modeltest 3.06 (Posada and Crandall 1998). For the simulation of the null distribution, 100 Monte Carlo replicates were used. Default window settings were used (minimum size = 5, step = 1).

(9) Recombination Detection Program (Martin and Rybicki 2000)

The Windows program Rdp (Martin and Rybicki 2000) (ftp://ftp.uct.ac.za/pub/data/geminivirus/recomb.htm) was generously modified by D. Martin for the simulations. The reference sequences used were internal and external. The window size was set to 10 nt.

(10) Recombination Parsimony (Hein 1990, 1993)

The C program RecPars (K. Fisker; ftp://ftp.daimi.aau.dk/pub/empl/kfisker/programs/RecPars) was used. In order to specify the substitution costs between nucleotides, a step matrix was estimated upon the maximum likelihood tree for each data set using MacClade4.0 (Maddison and Maddison 2000). A recombination cost of three times the maximum substitution cost ($d = 3 \times s$) was used. This particular recombination cost was chosen because it performed the best in previous computer simulations. Recombination was inferred when more than one history was suggested for a given data set.

(11) Reticulate (Jakobsen and Easteal 1996)

The C program Reticulate (Jakobsen and Easteal 1996) (http://jcsmr.anu.edu.au/dmm/humgen/ingrid/reticulate.htm) was modified for the simulations. The test statistic used was the neighbor similarity score, and the number of permutations was set to 1,000.

(12) Runs Test (Takahata 1994)

A C program was written implementing the Runs Test proposed by Takahata (1994).

(13) Sneath Test (Sneath 1995, 1998)

A program written by P. Sneath in qbasic (ftp://ftp.ebi.ac.uk/pub/software/dos/) was translated into C for the simulations.

(14) Triple (Kuhner et al. 1991)

A program written in Fortran by Mary Kuhner (Kuhner et al. 1991) that implements an extension to the method of Stephens (1985) was translated into C. On the basis of the results from computer simulations, the Bonferroni correction was applied to each individual test with a family alpha level of 0.01 to obtain an approximate false positive rate of 5%.

Performance Evaluation

The question addressed in this study is whether recombination methods detect the presence of recombination. Although some methods provide a qualitative answer for the presence of recombination (yes or no), most methods calculate a *P*-value (table 1). In the latter case, recombination was inferred when the provided *P*-value (calculated using some statistical distribution or using permutations) was smaller than 0.05.

Table 2
Empirical Data Sets Evaluated for the Presence of Recombination (Available from the Author)

Data Set	Reference	Gene	Homology	Group	Taxonomic Level
BoletalesATP6	(Kretzer and Bruns 1999)	ATP6	Orthologs	Fungi	Order
Candidula16S	M. Pfenninger and D. Posada (unpublished data)	16S	Orthologs	Snail	Species
DaphniaCO1	(Schwenk, Posada, and Hebert 2000)	CO1	Orthologs	Cladoceran	Genus
DmelCytB	(Ballard and Kreitman 1994)	CytB	Orthologs	Fly	Species
GymnND4	Pellegrino et al (personal communication)	ND4	Orthologs	Lizard	Species
HIV(B)EnvNR	Los Alamos HIV database[a]	Env	Orthologs	Virus	Subtype
HIVEnvNR	Los Alamos HIV database[a]	Env	Orthologs	Virus	Group
HumanHRVI	HVR database[b]	HRVI	Orthologs	Human	Species
InsectaCOII	M. Whiting (personal communication)	COII	Orthologs	Insecta	Class
Perom12S	(Sullivan, Holsinger, and Simon 1995)	12S	Orthologs	Rodent	Subfamily
VertebCOI	(Cunningham 1997)	COI	Orthologs	Vertebrate	Above Superclass
WolfCR	(Vilá et al. 1999)	Control region	Orthologs	Wolf	Species
Armillaria-mtDNA	(Saville, Kohli, and Anderson 1998)	mtDNA	Orthologs	Fungi	Species
Candida-mtDNA	(Anderson et al. 2001)	mtDNA[c]	Orthologs	Fungi	Species
Fusarium3	(O'Donnell et al. 2000)	Al-Tri101-Pp[d]	Orthologs	Fungi	Species
FusariumTri101	(O'Donnell et al. 2000)	Tril01[e]	Orthologs	Fungi	Species
HGVgenome	(Worobey and Holmes 2001)	Genome	Orthologs	Virus	Species
HIVEnv	Los Alamos HIV database[b]	Env	Orthologs	Virus	Group
HumanDRB1	(Klein and Schonbach 1993)	DRB1	Orthologs	Human	Species
MaizACT	(Moniz de Sá and Drouin 1996)	Actin	Paralogs	Plant	Species
MammPDH	(Fitzgerald et al. 1996)	Pdh	Paralogs	Mammal	Class
MammPGK	(Fitzgerald et al. 1996)	Pgk	Paralogs	Mammal	Class
NeisseriaArgF	(Zhou and Spratt 1992)	ArgF	Orthologs	Bacteria	Genus
PetuniaS-RNase	(Wang et al. 2001)	S-RNase	Orthologs	Plant	Species

NOTE.—Plain font indicates that recombination is assumed to be absent; bold font indicates that the presence of recombination has been inferred in previous studies.
[a] http://hiv-web.lanl.gov. A partial alignment was used.
[b] http://db.eva.mpg.de/hvrbase/index.html.
[c] Seven mtDNA fragments.
[d] Ammonia ligase + Tril01 + Phosphate permase.
[e] 3-*O*-acetyltransferase.

Selection of Empirical Data Sets

A total of 24 nucleotide data sets were selected from the literature and public databases. In 12 of these data sets recombination is assumed to be absent, whereas in the other 12 data sets the presence of recombination has been suggested in the literature (table 2). Data sets were selected to represent different conditions (levels of genetic diversity, taxonomic level, number of sequences, number of sites, rate heterogeneity among sites) (table 3). Already aligned data sets (available from the author) were obtained. In some cases, regions with abundant missing data were removed from the alignment. Nucleotide substitution parameters (required by some recombination methods) for each data set were estimated in PAUP* (Swofford 2000) for the best-fit model of nucleotide substitution (Posada and Crandall 1998).

Results

Recombination Inference

Different methods for detecting recombination disagreed regarding the presence or absence of recombination for several data sets. In some cases, inference was not possible because of the requirements of a given method (e.g., more than three sequences). In only one case was recombination not inferred by all methods (DmelCytB). There were no data sets for which all methods detected recombination. Notice that in figure 1 the data sets are quite scrambled in terms of the preconceived notions of recombination (data sets in bold vs. not data sets not in bold). However, four of five data sets with the least recombination detected were previously thought to be recombination free and all five data sets with the strongest inference of recombination were previously thought to contain recombinants. The behavior of each recombination method is briefly summarized subsequently.

MaxChi2 and Rdp seem to identify quite well when recombination is present (or likely to be present) without claiming that recombination has occurred when it seems unlikely.

Chimaera could be inferring potential false positives (MammPGK) and missing some likely recombinant data sets (HIV(B)EnvNR).

Reticulate and Geneconv seem to wrongly infer recombination in some divergent data sets (DaphniaCO1 and InsectaCOII), but they do recognize well the occurrence of recombination where it is most likely.

PhyPro does not infer many potential false positives, but it seems to fail to identify some likely recombinant data sets (HIV(B)EnvNR and MaizACT).

The Homoplasy Test seems to incorrectly infer recombination when rate heterogeneity among sites is high (WolfCR and Armillaria-mtDNA). Also, it did not seem able to detect recombination in highly diverse data sets (HIV(B)EnvNR, HIVEnvNR, BoletalesATP6, and HIVEnv). Interestingly, this is exactly the same behavior

Table 3
Genetic Variation

Data set	Number of Sequences[a]	Number of Sites	Segregating Sites	Informative Sites	Diversity	α
BoletalesATP6	31	639	372	265	0.1704	0.3611
Candidula16S	46	326	37	17	0.0187	1.4693
DaphniaCO1	19 (18)	466	214	183	0.1630	0.0095
DmelCytB	17 (6)	1,137	8	1	0.0009	∞
GymnND4	14	803	358	270	0.1837	1.3078
HIV(B)EnvNR	15	2,454	831	407	0.0888	0.6789
HIVEnvNR	11	2,724	1,646	945	0.2456	0.5193
HumanHRVI	21 (12)	428	22	12	0.0110	∞
InsectaCOII	7	609	303	178	0.2480	6.6827
Perom12S	9	757	93	48	0.0452	0.0912
VertebCOI	5	1,506	586	276	0.2292	1.2810
WolfCR	34	230	28	16	0.0243	0.0007
Armillaria-mtDNA	23	2,234	7	5	0.0022	0.0018
Candida-mtDNA	49 (37)	2,553	62	58	0.0070	0.1523
Fusarium3	28 (16)	1,336	220	118	0.0083	∞
FusariumTril01	28 (24)	4,146	64	36	0.0105	∞
HGVgenome	16	8,508	2,413	1,513	0.0936	0.5088
HIVEnv	20	2,724	1,748	1,185	0.2121	0.5660
HumanDRB1	3	153	24	0	0.1046	∞
MaizACT	8	1,008	363	230	0.1708	0.2801
MammPDH	5	1,104	399	130	0.1889	0.4411
MammPGK	6	1,248	584	231	0.2294	0.4725
NeisseriaArgF	9	787	234	122	0.1152	0.6862
PetuniaS-RNase	14	504	391	300	0.3604	2.3846

NOTE.—Plain font indicates that recombination is assumed to be absent; bold font indicates that recombination has been inferred in previous studies. Diversity is calculated as the average pairwise number of observed differences per site. α is the shape of the gamma distribution and represents the strength of rate variation among sites (Yang 1993). Small values of α indicate strong rate variation among sites, whereas $\alpha = \infty$ indicates that there is no rate variation among sites.

[a] The number of haplotypes is shown in parentheses.

observed in previous computer simulations (Posada and Crandall 2001).

Pist showed a tendency to wrongly detect recombination in divergent data sets (e.g., GymnND4, DaphniaCO1, and VertebCOI), although it did not clearly detect recombination in the most divergent data set (PetuniaS-RNase). In the simulations, however, the amount of diversity did not seem to induce false positives for Pist, except when high rate variation was present in the data.

Plato seems to erroneously infer recombination in two data sets with high rate variation (WolfCR and Armillaria-mtDNA), even though rate variation was included in the model used for the calculation of the likelihoods. In addition, when rate variation is included in the model, recombination was not detected in argF; this data set was used as an example of recombination in the original paper.

RecPars did not detect recombination in clear recombinant data sets, such as Fusarium3 or Candida-mtDNA. This could be because of the somehow low number of parsimony informative sites in these data sets (118 and 58, respectively). This may imply that RecPars needs a minimum sequence divergence (>1%) to detect recombination.

Runs Test performance does not fit an obvious pattern, and it does not detect recombination in data sets where recombination most likely has occurred. However, it detects recombination in data sets where recombination seems unlikely. Interestingly, the Runs Test showed the worst performance of all methods in previous computer simulations (Posada and Crandall 2001).

The bootscanning implementation, Simplot, seems very conservative when a 95% threshold is used. When the bootscanning threshold was dropped from 95% to 70%, it detected recombination in the HIV data sets (data not shown). However, in simulations it was observed that even a bootscanning threshold of 90% resulted in high false positive rates. In addition, it infers recombination in the VertebCOI data set, which may be a false positive.

The Sneath Test seems to work well, except for the likely wrong detection of recombination in Perom12S and InsectaCOII. Because the Sneath Test works in a pairwise manner, this inference could reflect some specific pattern not recognized in the data as a whole.

Finally, Triple does infer what seem to be false positives (e.g., Perom12S), while failing to reject what seem to be clear recombinant data sets (e.g., HGVgenome).

Discussion

Recombination Detection Patterns

The different recombination methods explored in this study showed different behavior. Conclusions here depend on a good knowledge of which data sets are recombinant and which ones are not. Obviously, we do not know the true recombination status of these data sets, but it seems reasonable to assume that in some of

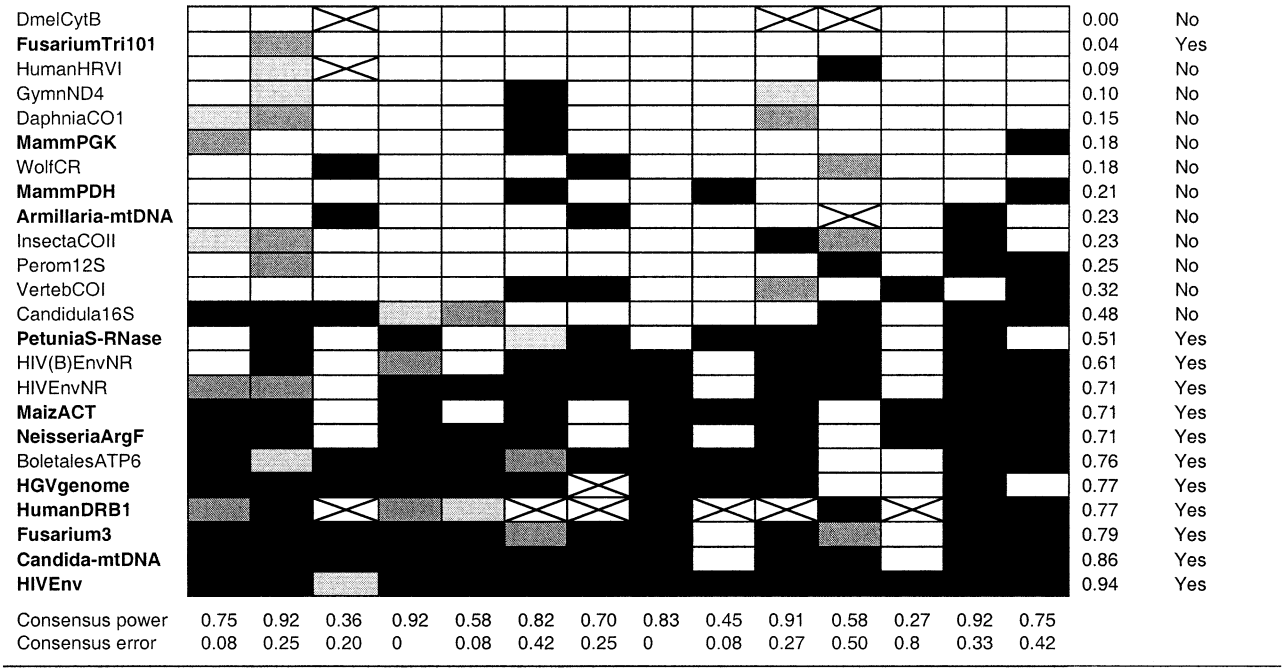

FIG. 1.—Recombination inference. Data sets in plain font indicate that recombination was assumed to be absent. Data sets in bold font indicate that recombination has been suggested. Cells in black indicate recombination was detected. Dark gray indicates that recombination was detected but with a *P*-value marginally significant ($0.05 > P\text{-value} > 0.01$). Light gray indicates recombination was not detected but the *P*-value was marginally nonsignificant ($0.10 > P\text{-value} > 0.05$). Cells in white indicate that recombination was not detected. Crossed cells indicate that inference was not possible. Data sets were ordered by the strength of the recombination inference, from least to most, using an arbitrarily defined recombination score (*R*-score). The *R*-score, the 50% consensus, and the consensus power and false positive error rates are described in the text. Methods abbreviations are—Chi: Chimaera; Gen: Geneconv; Hom: Homoplasy Test; Max: MaxChi2; Phy: PhyPro; Pis: Pist; Pla: Plato; Rdp: Rdp; Rec: RecPars; Ret: Reticulate; Run: Runs Test; Sim: Simplot; Sne: Sneath Test; Tri: Triple.

the data sets recombination has occurred, whereas in others it is absent.

To summarize inferences obtained by the different methods (see fig. 1), data sets were ordered by the strength of the recombination inference, from least to most, using an arbitrary recombination score (*R*-score) that ranges from 0 to 1, defined as *R*-score = ([number of methods inferring recombination with a *P*-value lesser than 0.01×6 + number of methods inferring recombination with a *P*-value greater than 0.01 but lesser than 0.05×3 + number of methods inferring recombination with a *P*-value greater than 0.05 but lesser than 0.10]/[24×6]). For each data set, a 50% consensus decision (yes or no) about the presence of recombination was obtained by considering whether the *R*-score was greater than 0.5. Because recombination was inferred in Fusarium3 by most methods, the 50% consensus for FusariumTriq101, which is a subset of the Fusarium3 data set, including the putative breakpoint, was also scored as yes. In addition, to assess the accuracy and false rate for each recombination method, a consensus power was calculated as the percentage of time a method inferred recombination for a data set with an *R*-score $>$ 0.5 (i.e., the inference consensus for the data set was ''yes'' and the method inferred ''yes''). A consensus false rate was also calculated as the percentage of time a method inferred recombination for a data set with an *R*-score $<$ 0.5 (i.e., the inference consensus for the data set was ''no'' and the method inferred ''yes'').

Given the statistics described previously, Geneconv, MaxChi2, Reticulate, and Sneath Test, were the most powerful methods, followed by Chimaera, Pist, Rdp, and Triple. Among the methods with least power were Homoplasy Test, RecPars, and Simplot. Regarding false positives, Chimaera, MaxChi2, PhyPro, Rdp, RecPars, and Simplot showed the lowest values ($<$0.10), whereas Pist, Runs Test, and Triple showed the highest false positive rates ($>$0.4). These results suggest that substitution and compatibility methods seem to be more powerful than the phylogenetic or the distance methods. Among all methods, MaxChi2 performed the best.

Recombination seems to be easier to detect, in general, with increasing levels of divergence. However, recombination was inferred in Candida-mtDNA by most recombination methods, even though there is not much variation in this data set ($<$1%). No relationship was found between the number of sequences, sites, segregating sites or informative sites, and power to detect recombination, but this is most likely a product of the low number of data sets evaluated. Nevertheless, there is a clear indication that the number of sites influences

the power to detect recombination. A recombination event was suggested to have occurred in FusariumTri101 data set (O'Donnell et al. 2000), but all methods, except one (Geneconv) did not detect recombination in this data set. However, when two other genes (in which there is no evidence of recombination) situated on both sides of the *Tri101* gene are added to build the Fusarium3 data set, most methods detect recombination. In fact, some methods identify the breakpoint in the *Tri101* gene only in the Fusarium3 data set.

The detection patterns observed here are in good agreement with those observed in the computer simulations. The best performing methods here (considering power and false positive error rates) were also the best in previous computer simulations (Posada and Crandall 2001). Also, the worst performing method here (Runs Test) was the worst performing method in the computer simulations. Likewise, recombination methods seem to be more powerful with increased levels of divergence, both with simulated and real data sets. In addition, methods confounded by the presence of high rate variation among sites in simulations here inferred recombination from putative nonrecombinant data sets that showed high levels of rate among sites (e.g., Homoplasy Test with WolfCR).

Reevaluation of Previous Recombination Inferences

The results of the recombination analyses carried out in this study do not completely agree with previous inferences. In some data sets where recombination has been claimed (PetuniaS-RNase) or assumed to be absent (Candidula16S), inferences were very ambiguous, and the disagreement with previous studies cannot be easily elucidated. For other data sets, however, inferences here are congruent enough to suggest that previous beliefs should be reconsidered.

Although it has been claimed that recombination has occurred in MammPGK and MammPDH (Fitzgerald et al. 1996) on the basis of phylogenetic incongruence, the evidence for recombination in those data sets is very weak here, and it is only strongly suggested by a few methods (Pist, Plato, RecPars, Triple). Interestingly, all these methods are phylogenetic methods. For divergent data sets with some rate variation among sites (as in MammPGK and MammPDH), these phylogenetic methods showed some power without inferring false positives in the computer simulations (Posada and Crandall 2001). It is possible that the recombination events in these data sets are very old, and that the substitution pattern has been obscured since by repeated mutation; therefore, they are no longer identifiable by the substitution methods. However, the inference in the original paper has some potential problems. In particular, the authors measured phylogenetic incongruence by considering bootstrap support for different clades in different regions of the alignment. Bootstrap values are not designed to compare phylogenetic hypotheses (trees); for this purpose, adequate methods exist (Shimodaira and Hasegawa 1999). Using such methods in a maximum likelihood framework, the PDH trees regarded as different by Fitzgerald et al. (1996) do not seem to be significantly different. There is thus reasonable doubt about the presence of recombination in these data sets.

Recombination has also been claimed to be present in the Armillaria-mtDNA data set (Saville, Yoell, and Anderson 1996; Saville, Kohli, and Anderson 1998), whereas here only the Homoplasy Test, Plato, and the Sneath Test infer its presence. Armillaria-mtDNA is a low-divergence data set ($<1\%$) with extreme rate variation among sites. It is precisely in this situation where the Homoplasy Test shows 90% false positives (Posada and Crandall 2001). However, Plato and the Sneath Test do not show an excess of false positives in computer simulations under these conditions, but they do infer recombination in some mitochondrial data sets where recombination seems very unlikely (e.g., Perom12S, InsectaCOII, VertebCOI). Of course, there may be some aspects of real data sets that may induce false positives that were not explored in the computer simulations, like nonindependence among sites. The presence of recombination in this data does not seem conclusive but cannot be discarded.

The analysis here suggests that recombination has occurred in the mitochondria of Boletales (Boletales-ATP6). This finding is even more interesting when we consider that this data set is composed of divergent species. Recombination in the mitochondria of fungi also seems to be very evident in the Candida-mtDNA data set.

Although the HIV(B)EnvNR (only subtype B sequences) and HIVEnvNR (several subtypes represented) data sets do not include known circulating recombinant forms (http://hiv-web.lanl.gov/CRFs/CRFs.html), this analysis indicates that recombination is prevalent in those sequences. This finding suggests that the frequency of recombination in HIV-1 is currently underestimated, especially within subtypes. Indeed, this result has very serious implications for vaccine development and HIV-1 dynamics, and it implies that coinfection might be more common than currently thought. In addition, many past inferences based on HIV-1 evolutionary history assumed no recombination, especially when dealing with within-subtype sequences. A thorough analysis of the known HIV-1 sequences is imperative to evaluate the impact of recombination on our current understanding of this virus.

Evaluating Methods for Detecting Recombination

The performance of methods for detecting recombination could be evaluated using several criteria. Here the only criterion employed—and the most basic—was the ability to detect, qualitatively, the presence of recombination. Other possible performance criteria relate to the characterization of the recombination events, like the identification of the parental and recombinant sequences involved or the estimation of the recombinational breakpoints. However, these criteria are less adequate for the purpose of comparing these 14 recombination methods. First, there are not many well-characterized recombination events (but see Drouin et al.

1999). Second, several of the data sets studied here presumably have been subject to numerous, and overlapping, recombination events, making the characterization of each recombination event cumbersome. Third, not all methods evaluated here are designed to characterize the recombination events (see table 1).

Recombination methods detect recombination either directly (Yes-No answer; e.g., Bootscanning) or calculate a *P*-value for the null hypothesis of no recombination (*P*-value obtained from permutations of the data, from Monte Carlo simulation, or from a statistical distribution). To evaluate these methods I have used the criteria offered by each program, trying to reproduce the behavior of a researcher using these same programs or methods. In the case of RecPars, an arbitrary decision was made. The output of RecPars consists of one tree or more for different regions of the alignment. Here, recombination was inferred when more than one tree was estimated for a given alignment. This kind of statistic (0/1) is not very powerful, and it could explain why RecPars, as implemented here, did not detect recombination in the Candida-mtDNA or Fusarium3 data sets. It is possible that permutations of the data would be a reasonable approach to build a proper null distribution of this test statistic. However, the program RecPars does not currently provide this possibility. A potential problem with methods that simulate the Monte Carlo *P*-value (Plato and Pist) is that in the generated data sets the amount of diversity is not the same as in the original data set. This might help explain why these methods did not perform among the best.

Choosing a Method to Detect Recombination

A researcher planning to scan his or her data for recombination might select the most appropriate method, according to different criteria. One of the first things to consider is what the purpose of the study is. If the researcher just wants to detect the presence of recombination, any of the methods evaluated here could potentially be used. However, if the characterization of the recombination events (identification of parental and recombinant sequences or recombination breakpoints [or both]) is of interest, only a few methods are appropriate (see table 1). If the interest is to measure the amount of recombination, among those evaluated here only methods like the Homoplasy Test and Pist might be used. Second, the methods contemplated here are able to deal with alignments of homologous sequences. Therefore, whether the sequences analyzed are orthologs from the same locus or members of a multigene family seems irrelevant for the selection of a recombination method. Third, in some cases, a researcher could be interested only in analyzing two sequences where there are some reasons to believe that recombination or gene conversion has occurred. The minimum number of sequences that a method requires is indicated in table 1. Fourth, in order to maximize performance, that is, increase power and decrease false positives, data sets with very low divergence (1%) could be analyzed with the Homoplasy Test (as long as a high level of rate variation among sites does not exist; this conclusion is also based on the computer simulations), MaxChi2, or Rdp. For higher levels of divergence (>1%), the Homoplasy Test is not adequate, and methods like MaxChi2 (the best performing method), Geneconv, Reticulate, or Rdp could be used (see also Posada and Crandall 2001). However, it does not seem that one should rely too much on a single method.

Acknowledgments

Thanks to Keith A. Crandall for discussion. Brandon Gaut and two anonymous reviewers provided very helpful comments. Special thanks to Darren Martin and Stuart Ray for modifying their programs. Thanks to Andrew Rambaut and Mike Worobey for giving access to Pist before it was published and to Mary Kuhner for providing the Triple fortran code. Mike Whiting, Katia Pellegrino, Markus Pfenninger, Mike Worobey, Jim Anderson, and Kerry O'Donnell generously sent their data sets. This work was supported by a Brigham Young University Graduate Studies Award and by a NSF Doctoral Dissertation Improvement Grant (NSF DEB 0073154).

LITERATURE CITED

ANDERSON, J. B., C. WICKENS, M. KHAN, L. E. COWEN, N. FEDERSPIEL, T. JONES, and L. M. KOHN. 2001. Infrequent genetic exchange and recombination in the mitochondrial genome of *Candida albicans*. J. Bacteriol. **183**:865–872.

BALLARD, J. W. O., and M. KREITMAN. 1994. Unraveling selection in the mitochondrial genome of Drosophila. Genetics **138**:757–772.

BROWN, C. J., E. C. GARNER, K. A. DUNKER, and P. JOYCE. 2001. The power to detect recombination using the coalescent. Mol. Biol. Evol. **18**:1421–1424.

CUNNINGHAM, C. W. 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. **14**: 733–740.

DROUIN, G., F. PRAT, M. ELL, and G. D. PAUL CLARK. 1999. Detecting and characterizing gene conversions between multigene family members. Mol. Biol. Evol. **16**:1369–1390.

DUBOSE, R. F., D. E. DYKHUIZEN, and D. L. HARTL. 1988. Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **85**:7036–7040.

FELSENSTEIN, J. 1984. Distance methods for inferring phylogenies: a justification. Evolution **38**:16–24.

———. 1991. PHYLIP (phylogenetic inference package). Version 3.4. University of Washington, Seattle.

FITZGERALD, J., H.-H. M. DAHL, I. B. JAKOBSEN, and S. EASTEAL. 1996. Evolution of mammalian X-linked and autosomal *Pgk* and *Pdh E1alfa* subunit genes. Mol. Biol. Evol. **13**:1023–1031.

GRASSLY, N. C., and E. C. HOLMES. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. **14**:239–247.

HEIN, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98**:185–200.

———. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36**:396–405.

HILLIS, D. M., and J. P. HUELSENBECK. 1994. To tree the truth: biological and numerical simulations of phylogeny. Pp. 55–67 *in* D. M. FAMBROUGH, ed. Molecular evolution of phys-

iological processes. Rockefeller University Press, New York.

HILLIS, D. M., B. K. MABLE, and C. MORITZ. 1996. Applications of molecular systematics: the state of the field and a look to the future. Pp. 515–543 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. Sinauer Associates, Sunderland.

HOLMES, E. C., M. WOROBEY, and A. RAMBAUT. 1999. Phylogenetic evidence for recombination in Dengue virus. Mol. Biol. Evol. **16**:405.

JAKOBSEN, I. B., and S. EASTEAL. 1996. A program for calculating and displaying compatibility matrices as an aid to determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12**:291–295.

JAKOBSEN, I. B., S. E. WILSON, and S. EASTEAL. 1997. The partition matrix: exploring variable phylogenetic signals along nucleotide sequences alignments. Mol. Biol. Evol. **14**: 474–484.

KLEIN, J., and C. SCHONBACH. 1993. Origins of MHC diversity. Pp. 16–37 *in* B. G. SOLHEIM, ed. The HLA system in clinical transplantation. Springer, New York.

KRETZER, A. M., and T. D. BRUNS. 1999. Use of atp6 in fungal phylogenetics: an example from the Boletales. Mol. Phylogenet. Evol. **13**:483–492.

KUHNER, M. K., D. A. LAWLOR, P. D. ENNIS, and P. PARHAM. 1991. Gene conversion in the evolution of the human and chimpanzee MHC class I loci. Tissue Antigens **38**:152–164.

LOLE, K. S., R. C. BOLLINGER, R. S. PARANJAPE, D. GADKARI, S. S. KULKARNI, N. G. NOVAK, R. INGERSOLL, H. W. SHEPPARD, and S. C. RAY. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73**:152–160.

MADDISON, D. R., and W. P. MADDISON. 2000. MacClade 4: analysis of phylogeny and character evolution. Version 4.0. Sinauer Associates, Sunderland, Mass.

MARTIN, D., and E. RYBICKI. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics **16**: 562–563.

MAYNARD SMITH, J. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34**:126–129.

———. 1999. The detection and measurement of recombination from sequence data. Genetics **153**:1021–1027.

MAYNARD SMITH, J., and N. H. SMITH. 1998. Detecting recombination from gene trees. Mol. Biol. Evol. **15**:590–599.

MONIZ DE SÁ, M., and G. DROUIN. 1996. Phylogeny and substitutions rates of angiosperm actin genes. Mol. Biol. Evol. **13**:1198–1212.

O'DONNELL, K., H. C. KISTLER, B. K. TACKE, and H. H. CASPER. 2000. Gene genealogies reveal global phylogeographic structure and reproductive isolation among lineages of *Fusarium graminearum*, the fungus causing wheat scab. Proc. Natl. Acad. Sci. USA **97**:7905–7910.

POSADA, D., and K. A. CRANDALL. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **14**:817–818.

———. 2001. Performance of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA **98**:13757–13762.

SALMINEN, M. O., J. K. CARR, D. S. BURKE, and F. E. MCCUTCHAN. 1996. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. AIDS Res. Hum. Retrovir. **11**:1423–1425.

SAVILLE, B. J., Y. KOHLI, and J. B. ANDERSON. 1998. mtDNA recombination in a natural population. Proc. Natl. Acad. Sci. USA **95**:1331–1335.

SAVILLE, B. J., H. YOELL, and J. B. ANDERSON. 1996. Genetic exchange and recombination in populations of the root-infecting fungus *Armillaria gallica*. Mol. Ecol. **5**:485–497.

SAWYER, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.

SAWYER, S. A. 1999. GENECONV: a computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer.

SCHWENK, K., D. POSADA, and P. D. N. HEBERT. 2000. Molecular systematics of European hyalodaphnia: the role of contemporary hybridization in ancient species. Proc. R. Soc. Lond. B **267**:1833–1842.

SHIMODAIRA, H., and M. HASEGAWA. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1234.

SNEATH, P. H. A. 1995. The distribution of the random division of a molecular sequence. Binary **7**:148–152.

———. 1998. The effect of evenly spaced constant sites on the distribution of the random division of a molecular sequence. Bioinformatics **14**:608–616.

SNEATH, P. H. A., M. J. SACKIN, and R. P. AMBLER. 1975. Detecting evolutionary incompatibilities from protein sequences. Syst. Zool. **24**:311–322.

STEPHENS, J. C. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2**:539–556.

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. Mol. Biol. Evol. **12**:988–1001.

SWOFFORD, D. L. 2000. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. Sinauer Associates, Sunderland, Mass.

TAKAHATA, N. 1994. Comments on the detection of reciprocal recombination or gene conversion. Immunogenetics **39**: 146–149.

VILÁ, C., I. R. AMORIM, J. A. LEONARD, D. POSADA, J. CASTROVIEJO, F. PETRUCCCI-FONSECA, K. A. CRANDALL, H. ELLEGREN, and R. K. WAYNE. 1999. Mitochondrial DNA phylogeography and population history of the gray wolf *Canis lupus*. Mol. Ecol. **8**:2089–2103.

WANG, X., A. L. HUGHES, T. TSUKAMOTO, T. ANDO, and T. KAO. 2001. Evidence that intragenic recombination contributes to allelic diversity of the S-RNase gene at the self-incompatibility (S) locus in *Petunia inflata*. Plant Physiol. **125**:1012–1022.

WEILLER, G. F. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombination in homologous sequences. Mol. Biol. Evol. **15**:326–335.

WIUF, C., 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.

WOROBEY, M., and E. C. HOLMES. 2001. Homologous recombination in GB virus C/hepatitis G virus. Mol. Biol. Evol. **18**:254–261.

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

ZHOU, J., and B. G. SPRATT. 1992. Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. Mol. Microbiol. **23**:2135–2146.

BRANDON GAUT, reviewing editor

Accepted January 14, 2002