BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 5 2005, pages 589–595
doi:10.1093/bioinformatics/bti040

Sequence analysis

Stepwise detection of recombination breakpoints in sequence alignments

Jinko Graham[1,*], Brad McNeney[1] and Françoise Seillier-Moiseiwitsch[2]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada V5A 1S6 and [2]Division of Biostatistics and Bioinformatics, Lombardi Cancer Center, Georgetown University, Washington, DC, WA 20057, USA

Received on May 18, 2004; revised on July 29, 2004; accepted on September 3, 2004
Advance Access publication September 23, 2004

ABSTRACT

Motivation: We propose a stepwise approach to identify recombination breakpoints in a sequence alignment. The approach can be applied to any recombination detection method that uses a permutation test and provides estimates of breakpoints.

Results: We illustrate the approach by analyses of a simulated dataset and alignments of real data from HIV-1 and human chromosome 7. The presented simulation results compare the statistical properties of one-step and two-step procedures. More breakpoints are found with a two-step procedure than with a single application of a given method, particularly for higher recombination rates. At higher recombination rates, the additional breakpoints were located at the cost of only a slight increase in the number of falsely declared breakpoints. However, a large proportion of breakpoints still go undetected.

Availability: A makefile and C source code for phylogenetic profiling and the maximum χ^2 method, tested with the gcc compiler on Linux and WindowsXP, are available at http://stat-db.stat.sfu.ca/stepwise/

Contact: jgraham@stat.sfu.ca

INTRODUCTION

Recombination leads to different evolutionary histories for different sites within samples of sequences from a population. The multiple correlated histories that result provide more evolutionary information than a single common history. Thus, the presence of recombination can improve the estimation and testing of genetic parameters in population biology. For example, genomic regions with recombination are preferred for detecting geographic subdivision when migration between subpopulations is relatively low (Hudson *et al*., 1992).

Per se, locating recombination breakpoints plays a role in understanding gene genealogies (e.g. DuBose *et al*., 1988) and haplotype structure within populations (e.g. Daly *et al*., 2001). Locating breakpoints is also essential to assessing the possibility of an individual being infected by two genetically diverse viral strains (that have subsequently recombined). For instance, in the case of HIV-1, there is an evidence of recombination of strains from the same subtype (e.g. Groenink *et al*., 1992) and different subtypes (e.g. Leitner *et al*., 1995; Fang *et al*., 2004).

The strength of signal left by a recombination event varies and is affected by factors such as the mutation rate, the level of divergence of the parental sequences that gave rise to the recombinant, how far back in time the recombination event occurred and the relative numbers of descendants of the recombinant and parental sequences in the alignment (e.g. Weiller, 1998; Posada *et al*., 2002). Many recombination events have little or no impact on the data and so are difficult or impossible to detect (Hudson and Kaplan, 1985; Myers and Griffiths, 2003). Likelihood methods for inference of recombination rates (e.g. Griffiths and Marjoram, 1996; Kuhner *et al*., 2000; Nielsen, 2000; Fearnhead and Donnelly, 2001) can take into account such undetectable events.

A variety of methods have been developed to detect recombination within alignments. Posada and Crandall (2001) provide a review and comparison (see also Brown *et al*., 2001; Wiuf *et al*., 2001). Several of these methods also estimate the location of breakpoints within the alignment and are therefore useful for locating breakpoints not proposed in advance.

Since some recombination events would leave stronger signals than others, conditioning on previously found breakpoints can reduce the unexplained variability in the data and improve a method's ability to find further breakpoints. We introduce such a stepwise approach. The approach can be applied with any permutation-based method for detecting recombination, which also identifies breakpoint locations. Examples of such methods include phylogenetic profiling (Phylpro) (Weiller, 1998) and the maximum χ^2 (MaxChi) method (Smith, 1992), as implemented by Posada and Crandall (2001) and Wiuf *et al*. (2001), Chimaera (Posada, 2002) and the GENECONV method (Sawyer, 1989). We illustrate the approach with analyses of a simulated dataset and alignments of HIV-1 *env* gene sequences and single nucleotide polymorphisms (SNPs) in a 150 kb region of human chromosome 7. Following this, we present simulation results comparing statistical properties of the one-step and two-step procedures.

SYSTEMS AND METHODS

For detecting recombination breakpoints that are not proposed in advance, several methods may be used in conjunction with permutation tests. Loosely speaking, each possible breakpoint or fragment with different ancestry within the alignment is considered, and the strength of its recombination signal is summarized by some site- or fragment-specific measure. The set of rank-ordered measures may then be considered in permutation tests. Assuming that sites have independent mutation processes with identically distributed outcomes, their permutations are equally likely outcomes of the same random evolutionary process, under the null hypothesis that all sites share the same ancestry (no recombination). A null distribution for the rank-ordered measures can thus be obtained by permuting sites in the alignment. The

*To whom correspondence should be addressed.

permutation approach makes fewer assumptions than approaches based on models of sequence evolution, thus providing a potentially more robust method.

A valid test of significance of the maximum of the observed site- or fragment-specific measures would involve comparing this maximum to the distribution of maxima over permutation replicates. Similarly, the significance of the second-largest observed measure might be assessed against the distribution of second-largest values over permutation replicates, etc. To avoid keeping track of the null joint distribution of the ordered measures, we use the null distribution of the maximum as the reference distribution for all observed measures. Hence, significance of each of the observed measures is assessed against the permutation null distribution of the maximum recombination signal over the alignment. A similar idea has been used in analysis-of-variance contexts for *post hoc* tests of the significance of pairwise differences between means, and is also the basis of the global P-values of GENECONV. Our procedure for detecting recombination breakpoints differs from the associated procedure for detecting the occurrence of recombination in the key aspect of the choice of the observed test statistic. For detecting the occurrence of recombination, the test statistic is the extreme value of the site-specific measures of recombination signal. In contrast, the test statistic for detecting breakpoints is the set of ordered site-specific measures. However, the reference distribution used in either case is the same.

A new stepwise procedure

Permutation tests assume that sites with the same ancestry are independent outcomes of the same random evolutionary process. Known breakpoints define segments of the alignment whose ancestries may differ. Hence, permutation of sites is appropriate within these segments but not between them. The idea of permuting sites within segments may be used in a stepwise procedure that, at each step, conditions on breakpoints declared at earlier stages of the analysis. Specifically, in the first step, the null hypothesis is that there are no recombination breakpoints and the permutation null distribution is obtained by permuting all sites in the alignment. If any breakpoints are declared, we proceed to a second step in which the null hypothesis is that there are no additional recombination breakpoints. At the second step, the null distribution is constructed by permuting sites within segments of the alignment with common ancestries given the breakpoints declared at the first stage. Conditioning on previously found breakpoints reduces variability of the test statistic which, in principle, increases the ability to detect additional breakpoints. The stepwise analysis continues until a prespecified number of steps is reached or until no additional breakpoints are declared.

To illustrate, consider a Phylpro-based procedure with a moving window of fixed width. In the first step, the permutation null distribution would be built from the minimum correlation, over all test sequences and over all polymorphic sites outside a window half-width of the ends of the alignment. All polymorphic sites in the alignment would be permuted. In subsequent steps, the permutation null distribution would be built from the minimum correlation, over all test sequences and over all polymorphic sites outside a window half-width of the ends of the alignment or previously declared breakpoints. Permutation of sites would be restricted to within the same segments of the alignment defined by breakpoints declared in previous steps.

As pointed out by a reviewer, McGraw *et al*. (1999) described a related approach in which segments of sequences defined by previously proposed breakpoints are partitioned at the location showing strongest recombination signal. The partitioning continues recursively for a fixed number of steps to produce a list of proposed breakpoints. The significance of breakpoints is then assessed by using a Monte Carlo procedure. Although each step involves proposing breakpoints, the significance of these breakpoints is not assessed until all steps are completed. In calculating the reference distribution for the test statistics, the null hypothesis is that there are no recombination breakpoints. In contrast, we assess significance at each step, conditional on breakpoints declared in the previous steps. At each step of our approach, the reference distribution is based on a null hypothesis that there are no breakpoints other than those declared previously.

Implementation

We implemented the stepwise approach with Phylpro, MaxChi and GENECONV, using a Monte Carlo permutation test to assess statistical significance of potential recombination locations. For all analyses, P-values for each statistical test were approximated from 1000 permutation replicates and significance was assessed at the 5% level.

Even when analysis of the HIV alignment is restricted to polymorphic sites within highly variable regions, heterogeneity in selection pressure along the HIV genome makes the assumption of an independent and identically distributed mutation process potentially problematic. In an attempt to reduce the effect of varying selection pressure in the HIV alignment, a second set of permutation tests was performed in which third-position sites within codons were treated separately from first- and second-position sites.

Moving-window approaches such as MaxChi and Phylpro cannot identify breakpoints within a window half-width of previously declared breakpoints. To minimize the resulting loss of resolution at each step for these moving-window approaches, we chose to identify breakpoints simultaneously rather than one-at-a-time at each step. However, a single true breakpoint may lead to recombination signal that exceeds the critical value not only at the true breakpoint location but also at several nearby polymorphic sites. We therefore attributed declared breakpoints clustered in nearby polymorphic sites to a single underlying true breakpoint. Specifically, adjacent polymorphic sites at which the site- or fragment-specific measures exceeded the critical value were blocked together. In simulations, true breakpoints tend to be associated with longer runs of significant sites punctuated by a few marginally significant sites. In contrast, false-positive signals were relatively infrequent and associated with shorter runs of significant sites. On the basis of these observations, blocks of adjacent polymorphic sites were expanded by $\sim$1% of the total number of polymorphic sites at each end and merged if they overlapped. For GENECONV, blocks containing the first or last site in the alignment were discarded. For all analyses, a single breakpoint was declared from the most extreme measure within a block.

For simulated data, a declared breakpoint was classified as an accurate call of location if a true breakpoint was within the block. Each block was said to accurately call only a single true breakpoint even if more than one true breakpoint was present within the block. Another possible choice would be to count all true breakpoints within a block as accurate calls. However, in practice, any method would have difficulty distinguishing nearby breakpoints. Hence, such a convention would tend to inflate the proportion of accurately called breakpoints. In contrast, the scoring scheme we have adopted is expected to under report the proportion of accurately called breakpoints, particularly for higher recombination rates. The rate of falsely declared breakpoints could also be decreased. However, any downward bias in false-positive rates is expected to be minimal given that false-positive signals are relatively infrequent and associated with shorter runs of significant sites. For moving window approaches, a further scoring issue relates to the counting of true breakpoints within a window half-width of the ends of the alignment or of previously declared breakpoints. Such breakpoints will necessarily go undetected and are therefore counted as true breakpoints that are not detected.

Statistical properties

In a stepwise procedure, the possibility of multiple applications of a given recombination detection method raises the issue of multiple testing. Multiple testing is often addressed by controlling the experiment-wise type I error rate, which is defined as the chance of rejecting any of the null hypotheses given that all hold. In a procedure taking up to n steps, the null hypotheses at each step are,

- H_{01}: no breakpoints exist in the alignment,
- H_{02}: no breakpoints exist other than those declared at the first step,
- H_{03}: no breakpoints exist other than those declared in the first and second steps, and so on up to
- H_{0n}: no breakpoints exist other than those declared in the first $n-1$ steps.

If all null hypotheses hold, the event that any of them is rejected is equivalent to the event that the first (H_{01}) is rejected. To see why, consider the event that H_{01} is rejected. It then follows that at least one of the hypotheses is rejected. Conversely, if at least one of the hypotheses is rejected, it must be the case that H_{01} was rejected because of the stepwise nature of the procedure. The experiment-wise type I error rate for detecting recombination is therefore the chance that H_{01} is incorrectly rejected which, in turn, is the size of the test at the first step.

However, the experiment-wise type I error rate reflects errors in detecting the presence of recombination rather than in identifying the location of recombination breakpoints. In fact, investigators may be more concerned that any declared breakpoint reflect a nearby real breakpoint. Falsely declared breakpoints may correspond to true breakpoints whose locations are inaccurately called, or they may have no correspondence to any true breakpoint. In either case, declared breakpoints that are far from any true breakpoint can be considered as false-positive results.

Statistical properties of the one-step and two-step versions of Phylpro and MaxChi were compared in a simulation study since these methods were relatively straightforward to automate as stepwise procedures. To assess statistical properties with respect to detecting the presence of recombination, we summarized the empirical size and power of the one-step and two-step procedures (for H_{01} and H_{02}, respectively). To assess statistical properties with respect to identifying the location of breakpoints, we summarized the proportion of all declared breakpoints that were inaccurately called (i.e. the rate of falsely declared breakpoints), the proportion of true breakpoints that were accurately called, and the mean number of true and falsely declared breakpoints found per simulated alignment for the one-step and two-step procedures. We also summarized the means and selected percentiles of the histograms of the number of additional true and falsely declared breakpoints per simulated alignment in the second step of the two-step procedure.

DATA

We examined simulated sequence alignments, for which the location of recombination breakpoints was known, and alignments of HIV-1 and human chromosome 7 sequences, for which the location of recombination breakpoints was unknown.

Simulated alignments

Evolutionary histories were simulated for random samples of 30 sequences of length 1000. Simulation parameters were chosen to roughly mimic evolution of the *env* gene in a population of HIV-1 particles within an infected individual. The simulation programs Treevolve version 1.3 (Grassly and Holmes, 1997) and Hudson (Schierup and Hein, 2000) were used to generate data under a coalescent simulation with recombination (Hudson, 1983; Griffiths and Marjoram, 1997), assuming a constant effective population size of $N = 2000$ (Leigh Brown, 1997). The coalescent with recombination assumes that each recombination event involves a single breakpoint and that sequences evolve neutrally. Therefore, the model does not describe the possibility of multiple breakpoints per recombination event and the selection observed in viral evolution. Nonetheless, coalescent simulation is a useful starting point for understanding properties of recombination detection methods (Posada, 2002). The alignment used to illustrate the approach had a recombination rate of $r = 4 \times 10^{-3}$ per sequence per generation (i.e. replication cycle), leading to a compound recombination parameter of $2Nr = 16$. To evaluate statistical properties in the simulation study, $2Nr$ was set to be 0, 2, 4, 8, 16 and 28 when using Treevolve and to 32, 64 and 128 when using Hudson. Under values of $2Nr > 28$, simulation of multiple datasets with Treevolve was computationally prohibitive. For $2Nr = 0$, 10 000 alignments were generated. For values of $2Nr > 0$, a total of 1000 alignments with at least one recombination breakpoint were generated at each setting. For $2Nr = 2$, only 1002 simulated alignments were required to obtain 1000 with recombination. For each value of $2Nr > 2$, all 1000 alignments simulated had recombination. The overall per-site mutation rate was chosen to be $\mu = 3.4 \times 10^{-5}$ per replication cycle, a value consistent with the estimated mutation rate for the HIV-1 envelope gene (Mansky and Temin, 1995). In Treevolve simulations, the general time-reversible substitution model (Tavaré, 1986; Rodrígues *et al*., 1990) was used, with a gamma distribution of rates. Random rates were used to model heterogeneity in selection pressure. Base frequencies and relative substitution rates were taken from the estimates of Anderson *et al*. (2000) for a region of the HIV-1 *env* gene (region 6) close to the V1–V2 and V3 subregions. Nucleotide frequencies were 0.3820 for A, 0.1758 for C, 0.1846 for G and 0.2576 for T; the gamma shape parameter was 0.54366. Relative rates of nucleotides are reported in Table 1. The Hudson program does not currently support simulation of sequences under the general time-reversible model. Therefore, for $2Nr = 32, 64$ and 128, the Kimura two-parameter substitution model with a gamma distribution of rates was used. Rates of transition and transversion for the Kimura two-parameter model were set equal to the mean rates of transition and transversion in the corresponding general time-reversible model.

Table 1. Average relative substitution rates

	A	C	G	T
A	—	1.7254	4.3979	0.6212
C	1.7254	—	0.9954	4.1248
G	4.3979	0.9954	—	1
T	0.6212	4.1248	1	—

We chose a single alignment at random from those generated under $2Nr = 16$ to illustrate the approach. This alignment consisted of 28 unique sequences with 239 polymorphic sites. As summarized in Table 2, there were 26 potentially detectable breakpoints; that is, breakpoints associated with recombination events which changed the coalescent tree (topology and/or branch lengths) for polymorphic sites.

HIV-1 alignment

The HIV-1 alignment consisted of 72 unique, ungapped sequences of length 294 from the V1–V2 and V3 regions of the envelope gene. There were 63 polymorphic sites. Sequences were derived from proviral DNA of peripheral blood mononuclear cells (PBMCs) and cervical secretions, and from plasma viral RNA. Only hypervariable regions were analyzed since selection varies across the HIV genome. Samples were collected from a Kenyan woman infected with a clade A virus six times over a period of 17 months post seroconversion, from August 1993 to January 1995. As noted by Poss *et al*. (1998), the patient had two distinct viral subpopulations at seroconversion.

Chromosome 7 alignment

We considered 135 SNPs with a minor allele frequency of at least 5% in a densely genotyped 150 kb region of chromosome 7 (base positions 126224043–126374042) from release 7 of the International

Table 2. Recombinations in simulated alignment

Interval[a]	Next break[b]	When[c]	Parental divergence[c]	Expected difference (%)[d]
93, 96	2	128	286	2.9
98, 100	6	711	1738	15.5
120, 126	2	1938	2569	21.5
127, 139	3	141	267	2.7
152, 154	41	859	1411	12.9
300, 305	8	59	473	4.7
335, 342	1	1505	766	7.4
342, 347	2	17	1227	11.4
350, 366	5	955	380	3.8
377, 392	1	364	880	8.4
392, 401	12	111	177	1.8
453, 463	2	65	468	4.6
470, 483	2	19	2252	19.3
486, 496	8	48	1512	13.7
549, 556	9	9	1550	14.0
588, 604	1	197	2073	18.0
604, 616	7	43	340	3.4
637, 644	15	1096	463	4.6
712, 720	20	48	649	6.3
776, 782	2	248	2023	17.6
784, 790	4	895	1376	12.6
802, 805	10	2216	1809	16.0
827, 828	5	374	3652	28.2
851, 852	9	3183	843	8.1
883, 884	5	1129	2896	23.6
896, 908	—	29	503	4.9

[a]Breakpoint between polymorphic sites.
[b]Number of polymorphic sites to next downstream breakpoint.
[c]One time-unit is 1 generation back.
[d]Expected percentage of sites at which parents differ.

HapMap Project (http://www.hapmap.org/). Haplotypes for 60 parents in the 30 available Caucasian parent–child trios were inferred using PHASE version 2.0.2 (Stephens *et al*., 2001; Stephens and Donnelly, 2003). Only haplotypes for the 47 individuals for whom phase could be determined with posterior probability ≥98% were considered. The resulting alignment contained 25 unique haplotypes and 135 polymorphic sites.

PARAMETERS

The MaxChi method makes pairwise comparisons between sequences. Briefly, for each pair, a χ^2-statistic is computed from a 2×2 table with counts of matches and mismatches upstream and downstream of the proposed breakpoint. Large values of the maximum χ^2-statistic over all pairs are taken as evidence for recombination at the breakpoint. In Phylpro, pairwise genetic distances between a target sequence and all other sequences in the alignment are contrasted upstream and downstream of a proposed breakpoint. Discordance between upstream and downstream distance vectors may be measured by the sample correlation. The minimum correlation over all sequences in the alignment, each regarded in turn as a target sequence, is taken as evidence for recombination at the proposed breakpoint.

Phylpro and MaxChi were applied with moving windows, which have fixed rather than variable numbers of sites in order to avoid a bias toward proposing recombination breakpoints at the ends of the alignment, where test statistics are more variable. For the simulated alignments, window half-widths of 30 polymorphic sites were selected. The half-width of 30 was approximately 1/10 of the polymorphic sites in the simulated alignments and permitted good resolution of breakpoints while controlling variability in computed correlations (Phylpro) and MaxChi statistics. Since the HIV-1 and chromomsome 7 alignments had only 63 and 135 polymorphic sites, respectively, we reduced the window half-width from 30 to 20 in order to achieve better resolution of breakpoints. A window half-width of 20 allowed consideration of 24 and 96 possible breakpoint locations for the HIV-1 and chromosome 7 alignments, respectively.

GENECONV, like MaxChi, makes pairwise comparisons between sequences but uses fragment scores to assess the evidence for recombination or gene conversion events. A fragment may be viewed as the maximal segment on which a pair of sequences is similar and its score measures similarity over sites within the segment relative to similarity over all sites. The GSCALE parameter of GENECONV determines the mismatch penalty for fragment scores. Smaller values allow for more discrepancies between sequence pairs and tend to produce longer fragments than larger values. Smaller GSCALE values also make GENECONV more sensitive to fragments from more distant recombination events, such as those in the histories of the simulated alignments. To strike a balance between detection of distant versus recent recombination events, we applied $\texttt{GSCALE} = 2$ throughout. For the HIV-1 alignment, GENECONV was also applied with $\texttt{GSCALE} = 3, 4$ and ∞, leading to similar results.

RESULTS

Simulated alignment

The results of the analysis are summarized in Figure 1. In the first step, two breakpoints were declared with Phylpro, three with MaxChi and eight with GENECONV. In step two, one additional breakpoint was declared with both Phylpro and MaxChi. In step three, one further breakpoint was declared with Phylpro.

Three of the eight breakpoints declared by GENECONV were classified as false-positive results because they are within blocks that do not overlap a true breakpoint location. All breakpoints declared by both Phylpro and MaxChi were classified as true-positive results. The most recent breakpoint in the alignment, between polymorphic sites 549 and 556, also involved relatively divergent parental sequences and is marked by an asterisk in Figure 1. Not suprisingly, this breakpoint had a consistently strong signal across all three methods. For any given method, true breakpoints declared after the first step appeared to derive from either older recombination events or from less divergent parental sequences (or both) than breakpoints found in the first step. The associated recombination events would be expected to leave weaker recombination signal in the dataset.

HIV-1 alignment

The Results of the HIV analysis are summarized in Figure 2. In the first step, one breakpoint was declared with Phylpro and GENECONV, while two were declared with MaxChi. In the second step, one additional breakpoint was declared with both Phylpro and GENECONV.

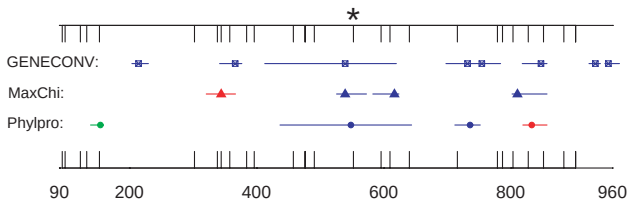

Fig. 1. Graphic summarization of results from the analysis of the simulated alignment. Clustered polymorphic sites with significant recombination signal have been collected into blocks indicated by horizontal lines. A single breakpoint is declared from the strongest signal within each block and is marked by a circle for Phylpro, a triangle for MaxChi and a square for GENECONV. Declared breakpoints from the first, second and third steps are colored in blue, red and green, respectively. The vertical lines on the upper and lower horizontal axis give the location of true breakpoints. The breakpoint for the most recent recombination event is marked with an asterisk.

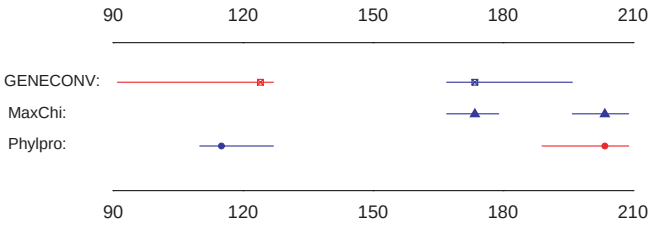

Fig. 2. Graphic summarization of results from the analysis of the HIV-1 alignment. Declared breakpoints from the first step are colored in blue and those from the second step are colored in red.

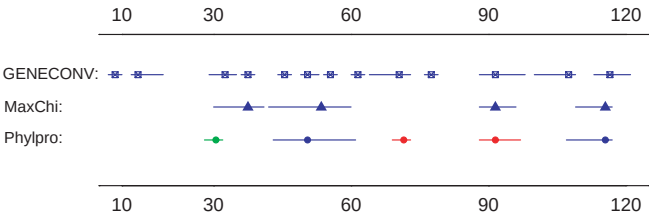

Fig. 3. Graphic summarization of results from the analysis of the chromosome 7 alignment. Declared breakpoints from the first, second and third steps are colored in blue, red and green, respectively.

The same results were obtained when third codon positions were permuted separately from first and second codon positions.

Chromosome 7 alignment

Figure 3 summarizes the results from the chromosome 7 analysis. In the first step, 2 breakpoints were declared with Phylpro, 4 were declared with MaxChi and 13 were declared with GENECONV. In the second step, two additional breakpoints were declared with Phylpro. In the third step, Phylpro declared one further breakpoint.

Simulation study

Table 3 reports the empirical size (first row) and power (subsequent rows) of Phylpro and MaxChi at the first and second steps. The first column of the table gives the value of the compound recombination parameter ($2Nr$) and the second reports the expected number of breakpoints (Wiuf *et al*., 2001). The third and fifth columns describe the percentage of alignments going to a second step; i.e. the power to reject H_{01} that there are no breakpoints in the alignment. The fourth and sixth columns describe the percentage of those alignments going to a second step for which further breakpoints were declared; i.e. the power to reject H_{02} that no breakpoints exist other than those declared in the first step. In the first step, the empirical sizes of the tests approximate the nominal 5% level to within simulation error. Power is lower in the second step than in the first because recombination breakpoints with weaker signals are more difficult to detect. In the second step, the empirical size of the Phylpro test (5.1%) approximates the nominal 5% level, but the size of the MaxChi test is lower (1.0%). After the first step, as possible values of the test statistic are limited by the permutation restrictions, test size for statistics whose null distribution is particularly discrete will decrease relative to the nominal level. In the simulations, the possible null values of the MaxChi test statistic were much more limited than those of Phylpro. Consequently, the test for MaxChi is conservative at the second step, as reflected not only in the size of the test at the second step but also in the reduced power relative to Phylpro at the second step for values of $2Nr$ between 2 and 8. On the other hand, in the first step, MaxChi appears to be more powerful than Phylpro.

Table 3. Empirical power (%) at first and second steps

$2Nr$	E(R)[a]	Phylpro		MaxChi	
		Step 1	Step 2	Step 1	Step 2
0	0	5.4	5.1	5.2	1.0
2	7	54.3	28.7	62.1	11.8
4	13	68.4	40.9	84.7	19.4
8	27	81.4	52.2	94.4	40.3
16	53	89.9	57.4	99.2	59.2
28	93	94.4	62.2	99.9	68.0
32	106	95.4	63.0	100	71.5
64	212	94.4	63.1	100	71.9
128	424	91.9	58.7	100	70.6

[a]Expected number of breakpoints.

Table 4 summarizes the percentage of true breakpoints that are accurately called (R_T), the percentage of declared breakpoints that are false or inaccurately called (R_F), and the mean number of true breakpoints detected (m_T) and of falsely declared breakpoints (m_F) per alignment for the one-step and two-step procedures. For the two-step procedure, R_T, R_F, m_T and m_F are cumulative. The table also reports the mean number of additional true (Δ_T) and falsely declared (Δ_F) breakpoints per alignment gained from taking a second step, along with the corresponding 5th (q_5) and 95th (q_{95}) percentiles. The measures R_T, m_T and Δ_T are not meaningful when there are no true breakpoints, and are not reported in the table for $2Nr = 0$. The measures related to falsely declared breakpoints are reported for $2Nr = 0$, including the rate R_F of falsely declared breakpoints, which is necessarily 100%. Generally, both the rates of accurately called and falsely declared breakpoints decrease with increasing recombination rates. The rate of accurately called breakpoints is low because many recombination events leave little trace in the alignment, but also because of the way we define an accurate call (counting only one true breakpoint per block), particularly at higher

Table 4. Rate (%) of accurately called (R_T) and of false or inaccurately called (R_F) breakpoints, the mean number of true breakpoints detected (m_T) and of falsely declared breakpoints (m_F) per alignment for the one-step and two-step procedures, along with the mean number of additional true (Δ_T) and falsely declared (Δ_F) breakpoints identified in the second step with fifth (q_5) and 95th (q_{95}) percentiles

$2Nr$	Step	Phylpro						MaxChi					
		R_T	R_F	m_T	Δ_T (q_5, q_{95})	m_F	Δ_F (q_5, q_{95})	R_T	R_F	m_T	Δ_T (q_5, q_{95})	m_F	Δ_F (q_5, q_{95})
0	1	—	100	—	—	0.07	—	—	100	—	—	0.06	—
	2	—	100	—	—	0.07	0.0 (0, 0)	—	100	—	—	0.06	0.0 (0, 0)
2	1	9.6	20.3	0.62	—	0.16	—	13.8	13.8	0.89	—	0.13	—
	2	11.4	23.8	0.73	0.1 (0, 1)	0.23	0.1 (0, 1)	14.4	14.8	0.92	0.0 (0, 0)	0.16	0.0 (0, 0)
4	1	8.9	12.5	1.14	—	0.16	—	12.9	9.2	1.66	—	0.17	—
	2	11.2	15.7	1.44	0.3 (0, 2)	0.27	0.1 (0, 1)	13.9	11.8	1.78	0.1 (0, 1)	0.24	0.1 (0, 1)
8	1	7.0	9.6	1.73	—	0.18	—	11.3	6.7	2.77	—	0.20	—
	2	9.1	11.4	2.25	0.5 (0, 2)	0.29	0.1 (0, 1)	12.9	9.2	3.16	0.4 (0, 2)	0.32	0.1 (0, 1)
16	1	5.4	3.5	2.49	—	0.09	—	9.0	3.8	4.16	—	0.16	—
	2	7.3	4.2	3.36	0.9 (0, 3)	0.15	0.1 (0, 1)	10.6	5.3	4.89	0.7 (0, 2)	0.27	0.1 (0, 2)
28	1	4.0	1.5	2.91	—	0.04	—	7.2	1.4	5.28	—	0.07	—
	2	5.5	1.9	4.06	1.1 (0, 3)	0.08	0.0 (0, 0)	8.6	2.0	6.29	1.0 (0, 3)	0.13	0.1 (0, 0)
32	1	3.3	1.0	3.03	—	0.03	—	5.9	0.8	5.50	—	0.04	—
	2	4.5	1.1	4.16	1.1 (0, 3)	0.04	0.0 (0, 0)	7.1	1.1	6.55	1.1 (0, 3)	0.07	0.0 (0, 0)
64	1	2.1	0.2	3.13	—	0.01	—	4.3	0.1	6.26	—	0.01	—
	2	2.9	0.2	4.25	1.1 (0, 3)	0.01	0.0 (0, 0)	5.0	0.2	7.33	1.1 (0, 3)	0.01	0.0 (0, 0)
128	1	1.2	0.0	2.45	—	0.00	—	3.3	0.0	6.79	—	0.00	—
	2	1.7	0.0	3.42	1.0 (0, 3)	0.00	0.0 (0, 0)	3.9	0.0	7.92	1.1 (0, 3)	0.00	0.0 (0, 0)

recombination rates. When comparing a two-step procedure with the corresponding one-step method, the rates of accurately and falsely called breakpoints are always higher. However, the increase per alignment in the mean number Δ_T of accurately called breakpoints from the second step is always greater than the increase Δ_F in the mean number of falsely declared breakpoints. This discrepancy between Δ_T and Δ_F is more pronounced with higher recombination rates. Furthermore, for $2Nr \geq 28$, gains of three or more true breakpoints are plausible from taking a second step. For example, at least $q_{95} = 3$ additional true breakpoints were obtained from a second step in 5% of simulated alignments. On the other hand, $q_{95} = 0$ additional false breakpoints were incurred in 95% of simulated alignments. Nevertheless, the gains are small relative to the number of breakpoints that go undetected.

In summary, the simulation results indicate modest gains from taking a second step that offset the costs of additional falsely declared breakpoints, particularly at higher recombination rates. However, when recombination is pervasive, only a small fraction of breakpoints are expected to be identified, even when an additional step is taken. Still, our results suggest that there is little to lose from taking a second step because the number of additional false-positive calls is minimal.

DISCUSSION AND CONCLUSIONS

We propose a stepwise procedure for identifying recombination breakpoints. The approach can be applied with any permutation-based method for detecting recombination that also provides estimates of breakpoint locations. The approach is illustrated by analyses of a simulated alignment and alignments of HIV-1 sequences from an individual and SNPs from a 150 kb region of human chromosome 7. Stepwise application of phylogenetic profiling, the maximum χ^2 method and the GENECONV method to these datasets detected more breakpoints than a single application.

Methods to identify breakpoints that are based on a permutation test produce extreme values of test statistics when segments of sites in the alignment have different ancestries. The idea motivating the stepwise approach is that conditioning on segments known to have different ancestries will reduce the unexplained variability in the data and lighten the tails of the permutation null distribution. Thus, in subsequent steps, it will be easier to detect breakpoints with weaker recombination signal.

As previously noted, the experiment-wise type I error rate for detecting recombination is the size of the corresponding test at the first step. Our simulation results and those of others (e.g. Posada and Crandall, 2001) indicate that the two methods considered achieve the nominal level. Our simulations also illustrate that more breakpoints are found with a two-step procedure than with a single application of a given method, particularly for higher recombination rates. For example, at $2Nr \geq 28$, a second step of Phylpro and MaxChi each found at least three additional true breakpoints in 5% of the simulated alignments. In contrast, the number of additional false breakpoints from the second step was usually no more than one at any recombination rate. Not surprisingly, recombination breakpoints with weaker signals are more difficult to pinpoint, as reflected by the higher proportion of inaccurate calls at the second step. Accuracy is also expected to fall as the number of previously declared breakpoints increases and more of the alignment lies within a window half-width of a declared break. However, at higher recombination rates, the additional real breakpoints that are found offset the cost of a slight increase in the number of inaccurately called or falsely declared breakpoints. We would therefore argue that a stepwise approach is worthwhile for finding additional recombination breakpoints within an alignment.

Even with these improved methods for detecting recombination breakpoints, a large proportion will still be missed. Ignoring recombination in phylogenetic analysis of population sequence data can lead to important errors in interpretation (Schierup and Hein, 2000). Our results underscore the need to further develop and adopt methods for inference of evolutionary parameters that take into account the large proportion of recombination breakpoints that cannot be identified.

ACKNOWLEDGEMENTS

We are grateful to Mary Poss for help with the HIV sequence alignment. We thank David Posada and two anonymous reviewers for their useful comments on the manuscript. This work was initiated while J.G. and B.M. were at the National Institute of Statistical Sciences, Research Triangle Park, NC. This research was supported in part by the National Science Foundation (grants DMS-9208758, DMS-9700867 and DMS-9711365 to the National Institute of Statistical Sciences), the Natural Sciences and Engineering Research Council of Canada (grants 222886-99 to J.G. and 227972-00 to B.M.), the American Foundation for AIDS Research (70428-15-RF to F.S.M.) and the National Institutes of Health (R01-AI47068 to F.S.M. and P30-HD37260 to the Center for AIDS Research at the University of North Carolina at Chapel Hill, NC).

REFERENCES

Anderson,J., Rodrigo,A., Learn,G., Madan,A., Delahunty,C., Coon,M., Girard,M., Osmanov,S., Hood,L. and Mullins,J. (2000) Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. *J. Virol.*, **74**, 10752–10765.

Brown,A.J. (1997) Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population. *Proc. Natl Acad. Sci. USA*, **94**, 1862–1865.

Brown,C., Garner,E., Dunker,K. and Joyce,P. (2001) The power to detect recombination using the coalescent. *Mol. Biol. Evol.*, **18**, 1421–1424.

Daly,M., Rioux,J.D., Schnaffner,F., Hudson,T. and Lander,E.S., (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.

DuBose,R., Dykhuisen,D. and Hartl,D. (1988) Genetic exchange among natural isolates of bacteria: recombination within the *phoa* locus of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **85**, 7036–7040.

Fang,G., Weiser,B., Kuiken,C., Philpott,S., Rowland-Jones,S., Plummer,F., Kimani,J., Shi,B., Kaul,R., Bwayo,J., Anzala,O. and Burger,H. (2004) Recombination following superinfection by HIV-1. *AIDS*, **18**, 153–159.

Fearnhead,P. and Donnelly,P. (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.

Grassly,N. and Holmes,E. (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, **14**, 239–247.

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

Griffiths,R. and Marjoram,P. (1997) An ancestral recombination graph. In Donnelly,P. and Tavaré,S., (eds), *Progress in Population Genetics and Human Evolution*. IMA Volumes in Mathematics and its Applications, Springer-Verlag, Berlin, Vol. 87, pp. 257–270.

Groenink,M., Andeweg,A., Fouchier,R., Broersen,S., van der Jagt,R., Schuitemaker,H., de Goede,R., Bosch,M., Huisman,H. and Tersmette,M. (1992) Phenotype-associated *env* gene variation among eight i related human immunodeficiency virus type 1 clones: evidence for *in vivo* recombination and determinants of cytotropism outside the V3 domain. *J. Virol.*, **66**, 6175–6180.

Hudson,R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoret. Popul. Biol.*, **23**, 183–201.

Hudson,R. and Kaplan,N. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111**, 147–164.

Hudson,R., Boos,D. and Kaplan,N. (1992) A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.*, **9**, 138–151.

Kuhner,M., Yamato,J. and Felsenstein,J. (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics*, **156**, 1393–1410.

Leitner,T., Escanilla,D., Marquina,S., Wahlberg,J., Brostrom,C., Hansson,H., Uhlen,M. and Albert,J. (1995) Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. *Virology*, **209**, 136–146.

Mansky,L. and Temin,H. (1995) Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.*, **69**, 5087–5094.

McGraw,E., Li,J., Selander,R. and Whittam,T. (1999) Molecular evolution and mosaic structure of α, β, and γ intimins of pathogenic *Escherichia coli*. *Mol. Biol. Evol.*, **16**, 12–22.

Myers,S. and Griffiths,R. (2003) Bounds on the minimum number of recombination events in a sample history. *Genetics*, **163**, 375–394.

Nielsen,R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Posada,D. (2002) Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Mol. Biol. Evol.*, **19**, 708–717.

Posada,D. and Crandall,K. (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl Acad. Sci. USA*, **98**, 13757–13762.

Posada,D., Crandall,K. and Holmes,E. (2002) Recombination in evolutionary genomics. *Annu. Rev. Genet.*, **36**, 75–97.

Poss,M., Rodrigo,A., Gosink,J., Learn,G., De Vange Panteleeff,D., Martin,H.,Jr. Bwayo,J., Kreiss,J. and Overbaugh,J. (1998) Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1. *J. Virol.*, **72**, 8240–8251.

Rodrígues,F., Oliver,J., Marín,A. and Medina,J. (1990) The general stochastic model of nucleotide substitution. *J. Theoret. Biol.*, **142**, 485–501.

Sawyer,S. (1989) Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, **6**, 526–538.

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

Smith,J.M. (1992) Analyzing the mosaic structure of genes. *J. Mol. Evol.*, **34**, 126–129.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Stephens,M., Smith,N.J. and Donnelly,P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In Miura,R., (ed.), *Lectures on Mathematics in the Life Sciences*. American Mathematical Society, Providence, RI pp. 455–486.

Weiller,G. (1998) Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol.*, **15**, 326–335.

Wiuf,C., Christensen,T. and Hein,J. (2001) A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol.*, **18**, 1929–1939.