

Sequence Note

A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints

D.P. MARTIN,¹ D. POSADA,^{2,3} K.A. CRANDALL,² and C. WILLIAMSON¹

ABSTRACT

We have developed a modified BOOTSCAN algorithm that may be used to screen nucleotide sequence alignments for evidence of recombination without prior identification of nonrecombinant reference sequences. The algorithm is fast and includes a Bonferroni corrected statistical test of recombination to circumvent the multiple testing problems encountered when using the BOOTSCAN method to explore alignments for evidence of recombination. Using both simulated and real datasets we demonstrate that the modified algorithm is more powerful than other phylogenetic recombination detection methods and performs almost as well as one of the best substitution distribution recombination detection methods.

DETECTION OF RECOMBINATION is a central component of HIV nucleotide sequence analyses. Of the over 20 currently published recombination detection methods (for a list of programs implementing most of these look at <http://www.umber.embnnet.org/~robertson/recombination/index.shtml>), the BOOTSCAN¹ and RIP² methods (implemented in the programs SimPlot and Recombination Identification Program, respectively) are most popular for the analysis of HIV sequences. Both methods were developed within the HIV research community to facilitate identification and characterization of intersubtype HIV-1 group M recombinants and both have proven very useful for this purpose.

However, current implementations of BOOTSCAN and RIP have various limitations that seriously restrict their more general utility (e.g., detecting recombination within subtypes). Proper use of both methods is heavily reliant on prior identification of a suitable set of potential parental (or reference) sequences against which putative recombinant (or query) sequences can be scanned. This is a serious problem if, for example, detection of intra-HIV-1 subtype C recombinants is desired, because choosing a suitably representative set of nonrecombinant subtype C reference sequences would itself require testing of candidate sequences for recombination. Conclusions drawn from

analyses where one or more of the selected reference sequences are themselves recombinant could be potentially misleading.

Also, whereas the BOOTSCAN method gives a good quantitative impression of conflicting phylogenetic signals in different parts of an alignment, the RIP method does not. Conversely, the RIP method employs a χ^2 test to verify the significance of potential recombination signals, while the BOOTSCAN method relies entirely on a rather arbitrary bootstrap cutoff. The lack of a statistical test for recombination is a major shortcoming of current implementations of the BOOTSCAN method since evidence of conflicting phylogenetic signal is not necessarily evidence of recombination. The problem is compounded when one considers use of the BOOTSCAN method to explore data for recombination (rather than its use to describe recombination) because there is no obvious way of correcting bootstrap values for multiple comparisons. Without an appropriate multiple comparisons correction the probability of detecting false positives will increase as the number of sequences scanned increases.

Despite these problems we believe that core features of the BOOTSCAN and RIP methods have considerable appeal. We have therefore developed a modified version of the BOOTSCAN algorithm, hereafter referred to as RECSAN, which incorporates some of RIP's features and is also suitable for detection and

¹Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, 7925, South Africa.

²Department of Microbiology and Molecular Biology, Brigham Young University, Provo, Utah 84602.

³Department of Biochemistry, Genetics and Immunology, University of Vigo, 36200 Vigo, Spain.

characterization of recombination events without prior identification of parental reference sequences. We demonstrate here that the recombination detection power of RECSCAN compares very well with that of the best recombination detection methods currently available.

The main difference between RECSCAN and BOOTSCAN/RIP is the use of a triplet scanning approach analogous to that used in the RDP,³ LARD,⁴ and CHIMAERA⁵ recombination detection methods. Unlike BOOTSCAN and RIP, where a query (or potentially recombinant sequence) is scanned against a set of three or more assumed nonrecombinant reference sequences, RECSCAN exhaustively scans every possible set of three sequences in an alignment without categorizing the sequences into query and reference groups. In effect, every sequence is considered a potential recombinant or parent. The rationale behind a triplet scanning approach is that a recombination signal will be clear-

est when the three sequences in a triplet are the recombinant and two parental sequences (or more often two sequences closely related to the real parents).

The obvious problem with using a triplet scanning approach in the BOOTSCAN algorithm is that there is only one possible branching order in an unrooted tree with just three sequences—how then can one determine which of the two sequences in a triplet is more closely related? We have developed two solutions to this problem, both of which massively increase the computational speed of the algorithm. The first solution is to construct bootstrap replicates and trees at every window position during a “scanning phase” of the algorithm (Fig. 1). Given information on the relative positions of every sequence within every tree constructed, sequence triplets are then screened in the “detection phase” of the algorithm (Fig. 1) and the relationships of the three sequences are determined in the context of

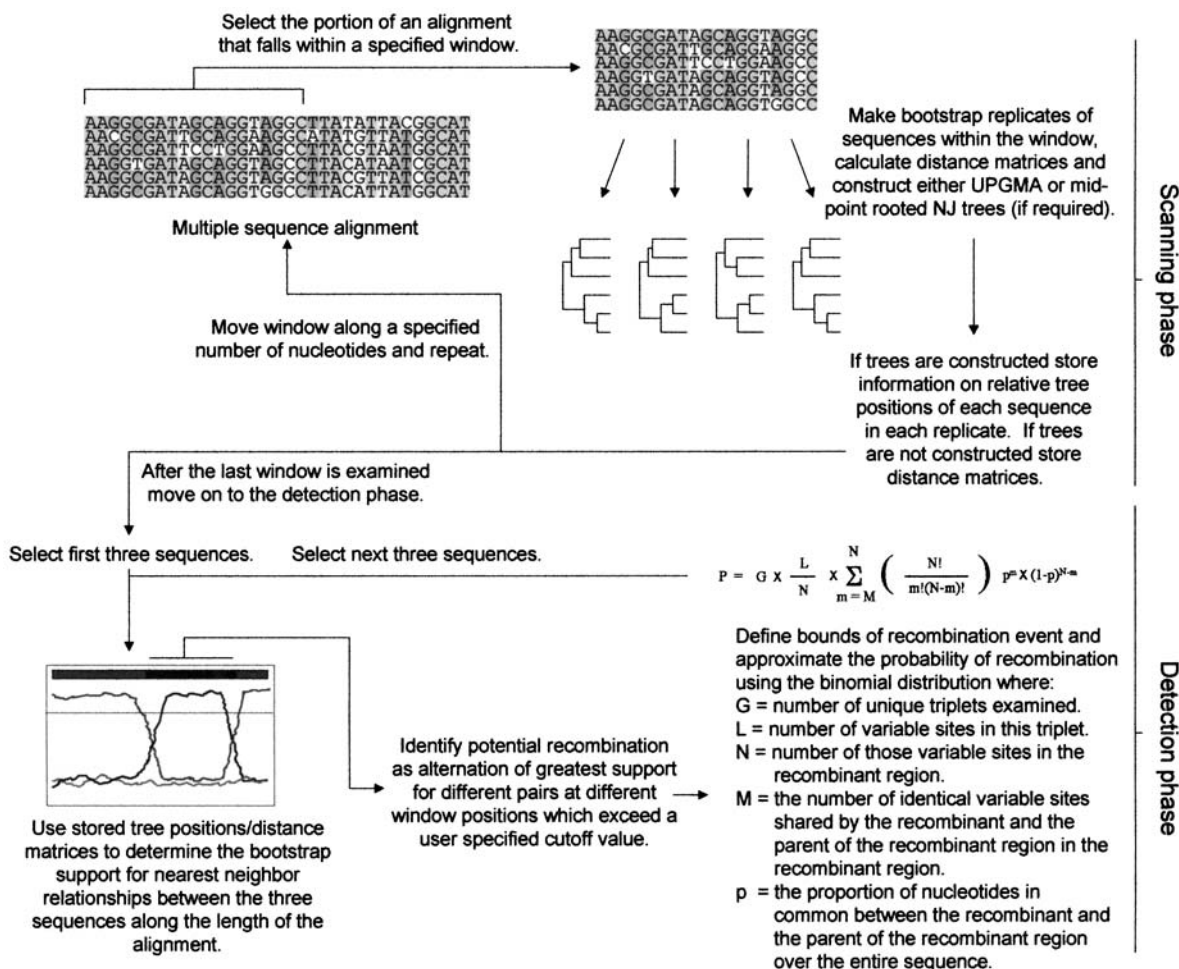


FIG. 1. The RECSCAN algorithm can be split into scanning and detection phases. During the scanning phase every tree or distance matrix for every bootstrap replicate at every window position is determined and stored for later analysis during the detection phase. Every combination of three sequences (or triplet) is individually examined during the detection phase for bootstrap evidence that one of the sequences may be alternatively more closely related to each of the other two sequences at different positions along its length. The probability that the pattern of sites within a potential recombinant region (the portion of the potentially recombinant sequence sharing high identity to the supposed parental sequence that would have contributed the smallest fraction of its sequence) could have occurred by a chance distribution of mutations (i.e., in the absence of recombination) is approximated using a Bonferroni corrected version of the binomial distribution as described by Martin and Rybicki.³

the tree positions of all other sequences in the alignment. An important point to note is that the trees that are generated are rooted—UPGMAs are by definition rooted but neighbor-joining (NJ) trees must be rooted at the midpoint of the longest path between two sequences in the tree. Rooting of trees is necessary to indicate which pair of sequences in a triplet shares a more recent ancestor. This solution increases the efficiency of the algorithm because only one full scan along the alignment is needed rather than one scan of the whole alignment for every triplet examined.

The second solution is a more extreme modification of the BOOTSCAN method, both in concept and in speed. As with the first solution, only one full scan along the alignment is made but, rather than taking note of relative positions of sequences within trees, only the raw pairwise distances between sequences in the bootstrap replicates are considered. This eliminates the need to estimate trees. In considering only distances this solution is somewhat of a BOOTSCAN-RIP hybrid, but is also equivalent to individually scanning triplets using bootstrapped UPGMAs rather than NJ trees. Omitting the tree inference component of the analysis makes the scanning phase approximately 60-fold faster—not quite as quick as RIP, but considerably faster than BOOTSCAN.

A second feature of the RIP algorithm included in RECSCAN is a statistical test for recombination. As with BOOTSCAN, the first stage of identifying recombination with RECSCAN is the detection of bootstrap support in excess of a user specified cutoff (usually ~70%) that groups one sequence alternatively with each of the other two sequences in a triplet. Approximate recombination breakpoint positions are assumed to be the midpoints of transitions between high bootstrap support grouping potential recombinant sequences with different parental sequences. Once the boundaries of a potential recombinant region have been determined, the probability that the pattern of variable nucleotide positions within the recombinant region could have occurred in the absence of recombination (i.e., by chance) is approximated using a modified Bonferroni corrected version of the binomial distribution as previously described by Martin and Rybicki³ (Fig. 1).

We examined simulated datasets with different variations of RECSCAN to both test their recombination detection power and ensure that our modifications to the algorithm had no undesirable effects. The simulated datasets used were previously generated for a study comparing the recombination detection power of 14 recombination detection methods.⁵ Briefly, 20 groups of 100 10-sequence genealogies were simulated using the coalescent with recombination. Each group of genealogies was simulated with one of five different degrees of recombination (recombination parameter $\rho = 4Nr = 0, 1, 4, 16, \text{ or } 64$ recombination events in the whole population from which the sample comes from, per site per generation) and one of four different degrees of genetic diversity ($\theta = 10, 50, 100, \text{ or } 200$ substitutions in the population per site per generation). A recombination parameter $\rho = 0$ allows the determination of a false-positive rate under different levels of nucleotide diversity. Ten sequences 1000 nucleotides in length were evolved on the simulated genealogies using the Hasegawa-Kishino-Yano⁶ nucleotide substitution model with a gamma distribution shape parameter ($\alpha = \infty, 2, 0.5, \text{ or } 0.05$).⁷ These simulated parameters span the range of recombination rates,

genetic diversity, and rate heterogeneity typically observed in HIV sequence data from single individuals.^{5,7}

Our analyses of the simulated datasets indicated that the NJ tree and distance scanning variants of our algorithm (both with window size = 100, step size = 20, bootstrap replicates = 100, bootstrap cutoff = 70, and the Jukes-Cantor, 1969 [JC] substitution model) had nearly identical power when using a Bonferroni corrected binomial p value cutoff of 0.05 (Fig. 2). Neither variant reported false positives in more than 7% of datasets with no recombination (the expected error rate with a 0.05 p value cutoff is approximately 5%). Detection power for both variants remained nearly identical for different window sizes (300, 200, 100, and 50; data not shown), step sizes (10, 20, and 50; data not shown), numbers of bootstrap replicates (50, 100, and 1000; data not shown), and substitution models (JC, Kimura two-parameter;⁸ data not shown).

For the low diversity alignments ($\theta = 10$), using a distance scan with a 99.9% bootstrap cutoff, 1000 bootstrap replicates, and with no binomial p value cutoff proved more powerful than both the distance and NJ scans with a 0.05 binomial p value cutoff (Fig. 2). However, for higher diversity alignments ($\theta > 10$), scans using the 0.05 binomial p value cutoffs were more powerful. Importantly, decreasing the bootstrap cutoff to 99.5% with 1000 bootstrap replicates resulted in an excessive rate of false positives (Fig. 2). This may seem surprising when one considers that 70% bootstrap support in an NJ tree is widely considered to be significant. It is, however, an example of how severe multiple comparison correction problems can be when using the BOOTSCAN method to explore for evidence of recombination (as opposed to simply using it to describe recombination). We should note that during the analysis of a 10-sequence alignment, 1000 nucleotides in length using a 100 nucleotide scanning window, 10 independent windows are examined for each of the 120 triplets examined, i.e., a total of 1200 bootstrapped rooted three sequence NJ trees need to be constructed. If a bootstrap cutoff of 70% were equivalent to a probability cutoff of 0.05, one would expect to encounter approximately one false positive result for every 20 trees examined or, put another way, 120 false positives per full alignment analyzed.

RECSCAN performs quite well when compared with other recombination detection methods. The MAXIMUM χ^2 method⁹ is one of the most powerful nonparametric recombination detection methods yet published⁵ and is only substantially more powerful than RECSCAN when analyzing datasets with low diversity ($\theta = 10$; Fig. 2). Whereas RECSCAN is a phylogenetic method that relies on identification of recombination through detecting alterations in tree topologies, MAXIMUM χ^2 is a substitution distribution method that identifies recombination through detection of deviations from an expected statistical distribution of substitutions.⁵ The MAXIMUM χ^2 and other substitution distribution methods do not require alterations in tree topologies to identify recombination and therefore they should be, and generally are, more powerful than phylogenetic methods.^{5,10,11} With respect to recombination detection power, RECSCAN clearly outperformed other phylogenetic methods such as RECPARS¹² and PLATO¹³ (Fig. 2).

We used RECSCAN to screen a set of 24 real nucleotide sequence alignments that had been previously assembled and used to compare the “consensus” detection power and “consensus” false-positive rate of a range of other nonparametric recombination detec-

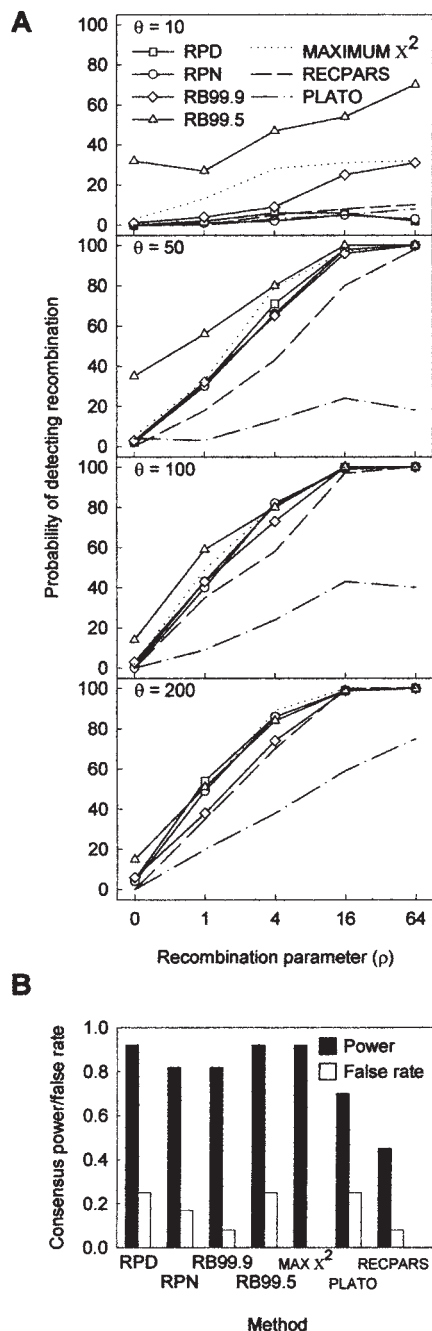


FIG. 2. The recombination detection power of different versions of RECSCAN compared to that of the MAXIMUM χ^2 ,⁹ RECPARS,¹² and PLATO¹³ methods as determined in Posada and Crandall⁵ and Posada.¹¹ The MAXIMUM χ^2 method is one of the most powerful nonparametric recombination detection methods yet published, whereas RECPARS and PLATO are phylogenetic methods that, like the RECSCAN method, identify recombination by detecting discrepancies between phylogenetic signals in different parts of an alignment.⁵ RPD and RPN = RECSCANS with Bonferroni corrected binomial P value calculation using distance and neighbor-joining scans, respectively (with other settings as mentioned in the text). RB99.9 and RB99.5 = RECSCANS with inference of recombination using 1000 bootstrap replicates with 99.9% and 99.5% bootstrap cutoffs, respectively (with other settings as mentioned in the text). (A) Power and false-positive rates determined using simulated datasets. Each panel represents the analysis of 500 simulated alignments 1000 nucleotides in length evolved under the Hasegawa–Kishino–Yano model of evolution with α (shape of the gamma distribution of rate variation among sites) = ∞ and five different degrees of recombination (ρ ; 100 alignments per value of ρ). Alignments represented in the different panels were evolved with different degrees of nucleotide diversity (θ). Whereas $\rho = 0, 1, 4, 16,$ and $64,$ respectively, indicates an average of 0, 3, 12, 48, and 192 recombination events in the evolutionary history of each of the alignments examined, two sequences chosen at random from alignments with $\theta = 10, 50, 100,$ and 200 would be expected to differ at an average of approximately 1%, 5%, 9%, and 17% of their sites, respectively. (B) Consensus power and false-positive rates determined using 24 real datasets as described in Posada.¹¹

tion methods.¹¹ In terms of detection power, all versions of the algorithm (both the distance and NJ tree-scanning versions using binomial P value calculation and distance scanning versions using a bootstrap support cutoff) outperformed the RECPARS and PLATO methods (Fig 2B). The distance scanning versions using either a 0.05 binomial P value cutoff or 99.5% bootstrap cutoff were as powerful as the MAXIMUM χ^2 method. However, all versions of the algorithm except that using a 99.9% bootstrap cutoff had a consensus false-positive rate in excess of 15%. While this might seem cause for concern it should be noted that this is a consensus false-positive rate¹¹ and not a false-positive rate in the same vein as that determined using the simulations. Whereas the high

consensus power rating simply indicates that RECSCAN detects recombination in all but one or two of the 12 alignments in which most (>50%) other methods also detect recombination, the false-positive rate similarly indicates that RECSCAN also detects recombination in between one and three of the 12 alignments in which most other methods do not detect recombination.

RECSCAN is a powerful exploratory recombination detection method that couples the benefits of BOOTSCAN's phylogenetic sensitivity with the speed and statistical rigor of RIP. An implementation of RECSCAN is available within our recombination detection and analysis package, RDP2, which can be downloaded free of charge from <http://darwin.uvigo.es/rdp/rdp.html>.

ACKNOWLEDGMENTS

We would like to thank Mika Salminen for informative discussions on the BOOTSCAN method, the National Research Foundation of South Africa (D.P.M.), U.S. National Institutes of Health (R01-GM55276) (K.A.C., D.P., D.P.M.), and the “Ramón y Cajal” programme of the Spanish government (D.P.) for partially funding the development and distribution of RDP2 and our work in recombination.

REFERENCES

1. Salminen MO, Carr JK, Burke DS, and McCutchan FE: Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses* 1995;11:1423–1425.
2. Siepel AC, Halperen AL, Macken C, and Korber B: A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res Hum Retroviruses* 1995;11:1413–1416.
3. Martin D and Rybicki E: RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* 2000;16:562–563.
4. Holmes EC, Worobey M, and Rambaut A: Phylogenetic evidence for recombination in Dengue virus. *Mol Biol Evol* 1999;16:405–409.
5. Posada D and Crandall KA: Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc Natl Acad Sci USA* 2001;98:13757–13762.
6. Hasegawa M, Kishino K, and Yano T: Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 1985;22:160–174.
7. Posada D and Crandall KA: Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* 2001;18:897–906.
8. Felsenstein J: PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* 1989;5:164–166.
9. Maynard Smith J: Analyzing the mosaic structure of genes. *J Mol Evol* 1992;34:126–129.
10. Drouin G, Prat F, Ell M, and Clarke GDP: Detecting and characterizing gene conversions between multigene family members. *Mol Biol Evol* 1999;16:1369–1390.
11. Posada D: Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Mol Biol Evol* 2002;19:708–717.
12. Hein J: Reconstructing evolution of sequences subject to recombination using parsimony. *Math Biosci* 1990;98:185–200.
13. Grassly NC and Holmes EC: A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol* 1997;14:239–247.

Address reprint requests to:
Darren P. Martin
University of Cape Town
Observatory
7925 South Africa

E-mail: Darren@science.uct.ac.za