AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 11, Number 11, 1995
Mary Ann Liebert, Inc.

Sequence Note

Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning

MIKA O. SALMINEN,[1] JEAN K. CARR,[1] DONALD S. BURKE,[2] and FRANCINE E. McCUTCHAN[1]

The Human Immunodeficiency Virus Type I (HIV-1) generates genetic diversity both by the accumulation of point mutations and by recombination.[1–3] Evidence indicates that both of these processes are actively contributing to the diversity of viral forms comprising the AIDS pandemic. Among internationally collected HIV-1 isolates, at least eight distinct genotypes in the main (M) or prevalent group and a variety of outlier (O) forms are now recognized.[4] Intergenic recombinants, with interspersed segments of genetic material from two or more parental genotypes, have also been described.[5–7] In phylogenetic analyses, recombinant forms often go unrecognized; indeed, published trees of HIV-1 *gag* or *env* genes have often identified them as distant members of established genotypes or as "new" genotypes.[8,9] Here we describe "bootscanning" for the recognition and analysis of recombinant genomes.

Recombinant HIV-1 isolates were initially identified when phylogenetic trees were constructed from sequential segments of available HIV-1 sequences.[6,10] It became apparent that about 10% of the sequences of sufficient length for analysis contained interspersed segments of genetic material from two different genotypes. There were also sequence segments that could not be assigned to any known genotype, and these were interspersed with those that joined established genotypes with high bootstrap values.[11] Figure 1 depicts the positions of these recombinant viruses in phylogenetic trees of HIV-1 *gag* and *env* genes. The behavior of the chimeric forms was unpredictable; some joined established genotypes (usually accompanied by bootstrap values lower than 100%), while others formed branches nearer the main trunk. Still others gave the appearance of "new" genotypes. Thus the traditional methods of phylogenetic analysis were not always sufficient to distinguish mosaic from nonmosaic genomes.

We have developed a new procedure for the identification of HIV-1 recombinants and for mapping of their recombination breakpoints. We employed reference sequences of established HIV-1 genotypes, as compiled in the HIV Database Compendium[4] and methods of phylogenetic analysis as implemented in the PHYLIP package, version 5.3c.[12] Analysis was on a SUN workstation using Genetics Data Environment (GDE) version 2.2, and some functions were automated using shell scripts. Beginning with a query sequence 1.5KB or greater in length, a multiple sequence alignment was built using five sequences of each of the established genotypes (or fewer if five were not available) plus an outgroup; the alignment was edited to minimize the number of insertion/deletion events. Gaps were excluded from the alignment using masks. Alignments were then divided into sequential, 50% overlapping segments of 200–500 bases. Bootstrapped phylogenetic analysis (100–1000 replicates) was subsequently applied to each segment using maximum parsimony, maximum likelihood, or distance matrix methods. Nonrecombinant genomes joined an established genotype with bootstrap values greater than 70% for all of the overlapping segments; recombinant genomes clustered with two parental genotypes alternately.

Recombinant genomes were further analyzed by reducing the number of sequences included in the analysis. The parental genotypes, either represented as consensus sequences or as a single reference sequence, plus an outgroup, were again aligned with the query sequence and the analysis repeated as described above. The bootstrap value with which the query sequence clustered with the three know sequences was then plotted. Most of the overlapping segments joined one of the parental genotypes with high bootstrap values; those flanking a transition from one genotype to another showed intermediate values (Fig. 1); we assigned the recombination breakpoints to the midpoint of the transition.

Some fragments exhibited anomalous behavior. They either showed equivalent, low bootstrap values with both parental genotypes and with the outgroup, or they showed elevated bootstrap values with the outgroup. We were able to resolve these fragments either by increasing or by decreasing the length of the segment examined, respectively. In Fig. 1, segments re-

[1] Henry M. Jackson Foundation for the Advancement of Military Medicine, Rockville, Maryland 20850.
[2] Division of Retrovirology, Walter Reed Army Institute of Research, Rockville, Maryland 20850.

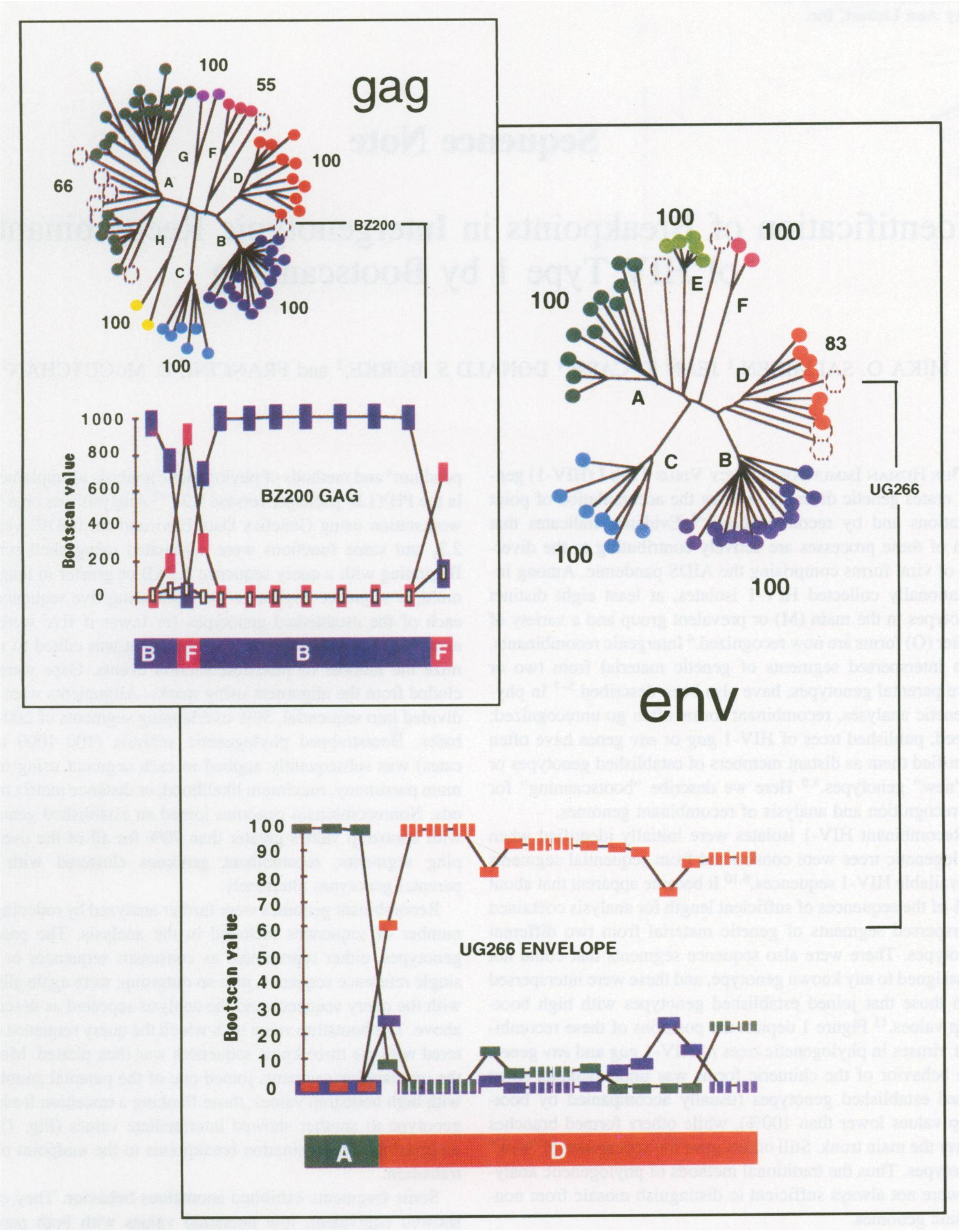

FIG. 1. Resolution of the chimeric structure of HIV-1 genes by bootscanning. Neighbor-joining trees of complete *gag* and *env* genes of HIV-1 isolates are shown, together with bootstrap values on the major branches identifying genotypes A through H. Solid circles represent isolates not known to be recombinant; open circles indicate positions of identified recombinants. Isolates BZ200 from Brazil and UG266 from Uganda were analyzed for recombination breakpoints in *gag* and *env* genes, respectively, by bootscanning. For BZ200, consensus sequences of genotypes B and F were used and the outgroup was the genotype H isolate VI557[8]; 1000 bootstrap replicates of maximum likelihood were used on 300-bp segments overlapping by 150 bp. The *env* gene of UG266 was analyzed similarly, with consensus sequences from genotypes A and D and genotype B isolate MN as outgroup, using 100 bootstrap replicates. Some areas of the UG266 *env* gene had to be resolved by analysis of larger segments (hatched symbols). Likewise, the small segment of genotype F in isolate BZ200 was resolved only by analysis of successively smaller segments in the region. Additional, unresolved breakpoints may remain in the examples shown.

quiring this adjustment of the segment length are indicated with patterned symbols.

In the examples shown, and in a wider analysis including several full-length recombinant HIV-1 genomes as well as full length *gag* and *env* genes, we were able to assign all of the segments to one of the two parental genotypes using a criterion bootstrap value of 70% or greater, and to map recombination breakpoints within 100 bases (data not shown). The breakpoints identified here correlated well with published reports using informative site distribution analysis.[6,10] However, some additional breakpoints were found; the upstream segment of genotype F in the *gag* gene of isolate BZ200 is one example (Fig. 1). The bootscanning procedure may provide additional power to resolve short, interspersed segments from two genotypes.

The subdivision of available sequence into segments of equal length is intrinsically problematic, since the information density, related to the fraction of nonconserved nucleotide positions, varies widely in different segments. We have successfully used segments of 200 bases in the *env* gene, and increased to 300 and 500 bases in the more conserved *gag* and *pol* genes, respectively, in our initial analyses, and have noted for each of these genes that smaller segments are usually uninformative. The precision with which recombination breakpoints can be identified may vary across the genome.

Available reference isolates of HIV-1 may not represent the full spectrum of genotypes comprising the global epidemic. This may result in difficulty at the first stage of analysis, when parental genotypes of a recombinant form are identified. A related problem comes into play at the second stage, when available reference isolates of a single genotype may be diverse in the region examined; neither a consensus sequence nor a selected reference isolate may prove satisfactory for resolving the parentage of particular segments. Together, these problems have not significantly compromised our ability to resolve chimeric structures, but the procedure may gain power as the database of HIV-1 sequence information becomes more complete.

We are unaware of any intrinsic limitation on the number of cross-over points within a recombinant HIV-1 genome. Indeed, our unpublished data, including several full-length and additional partial HIV-1 chimeras, indicate that most recombinant genomes harbor several breakpoints unevenly distributed across the genome. There is a significant possibility that segments containing closely interspersed segments of different genotypes remain unresolved in our analyses.

Bootscanning would appear to represent a powerful approach to the identification and mapping of recombinant genomes, with wide application. Computer software to automate the bootscanning procedure, adjusting both segment size and overlap to gain maximal resolution, is under development. Such software will enable a thorough and systematic evaluation of available sequence data, with the goals of identifying reference isolates of HIV-1 that show no evidence of recombination and, possibly, additional HIV-1 recombinants. When completed, this software will be made available.

REFERENCES

1. Li WH, Tanimura M, and Sharp PM: Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 1988;5:313–330.
2. Hu WS and Temin HM: Retroviral recombination and reverse transcription. Science 1990;250:1227–1233.
3. Coffin JM: Retroviridae and their replication. In: *Virology* (Fields BN and Knipe DM, eds.). Raven Press, New York, 1990, pp. 1437–1500.
4. Myers G, Korber B, Wain-Hobson S, Jeang KT, Henderson LE, and Pavlakis GN (eds.): *Human Retroviruses and AIDS: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences.* Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 1994.
5. Sabino EC, Shaper EG, Morgado MG, Korber BTM, Diaz RS, Bongertz V, Cavalcante S, Galvao-Castro B, Mullins JI, and Mayer A: Identification of human immunodeficiency virus type-1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. J Virol 1994;68:6340–6346.
6. Sharp P, Robertson D, McCutchan F, and Hahn B: Recombination in HIV-1. Nature (London) 1995;374:124–126.
7. Leitner T, Escanilla D, Marquina S, Wahlberg J, Brostrom C, Hansson HB, Uhlen M, and Albert J: Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. Virology 1995;209:136–146.
8. Louwagie J, McCutchan F, Peeters M, Brennan TP, Sanders-Buell E, Eddy G, van der Groen G, Fransen K, Gershey-Damet GM, Deleys R, and Burke DS: Phylogenetic analysis of *gag* genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS 1993;7:769–780.
9. Louwagie J, Janssens W, Mascola J, Heyndricks L, Hegerich P, van der Groen G, McCutchan FE, and Burke DS: Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type-1 (HIV-1) isolates of African origin. J Virol 1995;69:263–271.
10. Hahn BH, Robertson DL, McCutchan FE, and Sharp PM: *Recombination and Diversity of HIV: Implications for Vaccine Development.* Neuvième Colloque des Cent Gardes Paris, 1994.
11. Felsenstein J: Confidence limits on phylogenies: An approach using the bootstrap. Evolution 1985;39:783–791.
12. Felsenstein J: PHYLIP-phylogenetic inference package (version 3.2). Cladistics 1989;5:164–166.

Address reprint requests to:
Mika O. Salminen
Henry M. Jackson Foundation
1600 East Gude Drive
Rockville, Maryland 20850