**In vivo** characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity

Gkikas Magiorkinis,1 Dimitrios Paraskevis,1 Anne-Mieke Vandamme,2 Emmanouil Magiorkinis,1 Vana Sypsa1 and Angelos Hatzakis1

**Correspondence**
A. Hatzakis
ahatzak@cc.uoa.gr

1National Retrovirus Reference Center, Department of Hygiene and Epidemiology, Athens University Medical School, Mikras Asias 75, 11527 Athens, Greece
2Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium

Recombination plays a pivotal role in the evolutionary process of many different virus species, including retroviruses. Analysis of all human immunodeficiency virus type 1 (HIV-1) intersubtype recombinants revealed that they are more complex than described initially. Recombination frequency is higher within certain genomic regions, such as partial reverse transcriptase (RT), *vif/vpr*, the first exons of *tat/rev*, *vpu* and *gp41*. A direct correlation was observed between recombination frequency and sequence similarity across the HIV-1 genome, indicating that sufficient sequence similarity is required upstream of the recombination breakpoint. This finding suggests that recombination *in vivo* may occur preferentially during reverse transcription through the strand displacement-assimilation model rather than the copy-choice model.

**INTRODUCTION**

Human immunodeficiency virus type 1 (HIV-1) belongs to the family *Retroviridae* and rates among the highest of variable human pathogens (Wain-Hobson, 1993). The extensive genetic diversity of the virus is caused by several factors, such as (1) the high misincorporation rate of reverse transcriptase (RT) in the absence of any proof-reading mechanisms (Dougherty & Temin, 1988; Preston et al., 1988; Roberts et al., 1988), (2) high virus production rate (Ho et al., 1995) and (3) homologous recombination (Robertson et al., 1995a, b).

Retrovirus recombination was first observed in avian tumour viruses (Vogt, 1971) and subsequently in other retroviruses (Clavel et al., 1989; Wong & McCarter, 1973). Recombination is thought to occur during the reverse transcription of the genomic RNA to the formation of double-stranded DNA (see Fig. 5A), providing the main strategy of retroviruses for exchanging genetic information between two heterogeneous RNAs and, thus, for tremendous genetic alterations within the viral genome. A prerequisite for recombination between two genetically diverse RNA molecules is their co-packaging in the same virion (Hu & Temin, 1990a, b).

Two different models have been proposed for the mechanism of retrovirus recombination (see Fig. 5B, C): the forced copy-choice model (Coffin, 1979; Vogt, 1971) and the strand displacement-assimilation model (Boone & Skalka, 1981a, b; Junghans et al., 1982). The forced copy-choice model proposes that template switching is driven mainly by breaks in the viral RNA, thus forcing the RT to switch to the other copy of the genomic RNA. According to this model, recombination occurs during the synthesis of minus-strand DNA. The various models proposed for strand transfer during minus-strand synthesis, and due not only to RNA breaks, have been known in general as ‘copy-choice’ (Vogt, 1971). In contrast, the strand displacement-assimilation model proposes that recombination occurs during the synthesis of plus-strand DNA, when, in one of the two ongoing second-strand cDNA synthesis complexes, an internally initiated DNA fragment is displaced by an upstream, growing DNA fragment, becoming free to hybridize and to continue second-strand synthesis from the other cDNA synthesis complex.

The impact of recombination on the evolution of HIV-1 was documented recently, showing that at least 10% of HIV-1 strains circulating comprise intersubtype recombinants (Kuiken et al., 2000; Robertson et al., 1995a, b). Furthermore, a significant proportion of the HIV-1 sequences that have been characterized initially by partial sequencing as non-recombinants were found to comprise intersubtype recombinants after analysis of their complete sequences (Kuiken et al., 2000). Three different classes of homologous recombination have been documented to occur in HIV-1: between strains of the same subtype...
(intrasubtype), between strains of different subtype (intersubtype) and between strains of different group (intergroup). In this study, 34 available full-length HIV-1 intersubtype mosaics were re-analysed to determine the in vivo properties of HIV-1 intersubtype recombination.

**METHODS**

**HIV-1 isolates.** A total of 34 full-length HIV-1 intersubtype recombinants, including all known HIV-1 'pure' subtypes (A–D, F–H, J and K) (Kuiken et al., 2000), available at the HIV database (http://hiv-web.lanl.hiv.gov) were re-analysed. Circulating recombinant form isolates were analysed only once. The accession numbers for the reference strains used were: M62320 (subtype A); M12508 (subtype B); U46016 (subtype C); K03454 (subtype D); AF005494 (subtype F1); U88826 (subtype G); AF005496 (subtype H); AF082395 (subtype J) and AJ249239 (subtype K).

**Determination of recombination patterns.** DNA sequence alignments were performed using CLUSTAL_W, version 1.81 (Thompson et al., 1994). The recombination pattern of each sequence was resolved by bootscanning plots, as implemented in the SIMPLOT software, version 2.5 (Ray, 1998), and confirmed further by phylogenetic analysis using the neighbour-joining method (NJ) (Saitou & Nei, 1987) with Kimura’s two-parameter correction (Kimura, 1980), as implemented in PHYLIP, version 3.5c (Felsenstein, 1993) (see supplementary Fig. 1, available at http://vir.sgmjournals.org). Bootstrap analysis (100 replicates) was used to estimate the reliability of the constructed trees. Phylogenetic analysis was accomplished using the maximum-likelihood model (ML) with the Tamura–Nei evolutionary model (TrN), including $\gamma$-distributed rates heterogeneity among sites, as implemented in TREEPUZZLE, version 5.0.p6 (Schmidt et al., 2002). TrN was chosen as the best-fitting nucleotide substitution model in several pieces of the HIV-1 genome, according to the ML ratio test using the MODELTREE (Posada & Crandall, 1998) and PAUP*, version 4.0b10, programs (Swofford, 1998). To search for any potential relationships between the unclassified regions of the recombinant sequences and any HIV-1 sequences characterized previously, a BLAST search was performed using the default settings (http://www.ncbi.nlm.nih.gov/BLAST/). Phylogenetic analysis was then used to confirm the similarities obtained by BLAST.

**Frequency of recombination breakpoints.** According to the recombination pattern of each individual HIV-1 sequence, a database was created, including the co-ordinates of the breakpoints of all of the mosaic sequences. Based on this database, the frequency of recombination breakpoints adjusted per 1000 nucleotides was calculated along the HIV-1 genome. To determine thoroughly the recombination breakpoint frequency across the HIV-1 genome, a program termed BREAKSCAN was developed (available from the authors upon request). BREAKSCAN was developed to plot the frequency of recombination breakpoints in a sliding window along the genome. This was done by counting the total number of breakpoints within the window over all of the 34 recombinant strains and dividing by the total number of the recombinant sequences ($n=34$). Running BREAKSCAN, the frequency of recombination breakpoints was plotted in a sliding window of 1000 bp moving in steps of 50 bp (Fig. 1).

![Fig. 1. Frequency of recombination breakpoints (black) and extent of sequence similarity (grey) across the HIV-1 genome for a sliding window of 1000 bp.](image-url)
Independence of breakpoints. A small number of recombination breakpoints may have been inherited by the same progeny and, consequently, there might be a perturbation of their independence. To examine whether recombination events may share an immediate common ancestor, we identified all breakpoints localized within a similar region of the alignment (± 100 nt from the breakpoint position) that shared a common evolutionary history (HIV-1 subtype) in the adjacent regions. In the event that recombination events shared an immediate common ancestor, they should then cluster together in the region containing the breakpoint. Thus, to identify any potentially linked recombination events, we performed phylogenetic analysis using ML for all pairs of fragments (400 nt in length); these fragments included a breakpoint and shared a common HIV-1 subtype classification in the adjacent regions. Any recombination events found to share a common evolutionary history were excluded from all subsequent analyses.

Determination of sequence similarity across the HIV-1 genome. Sequence similarity across the HIV-1 genome was calculated by a newly developed program termed SIMSCAN (available from the authors upon request). SIMSCAN was designed to calculate the mean similarity (measured as the proportion of sites that are identical between two sequences) between different sequences in a sliding window moving along the genome. The mean intersubtype sequence similarity was determined by averaging all pairwise intersubtype similarities of nine sequences belonging to all HIV-1 'pure' subtypes characterized previously (subtypes A–D, F–H, J and K). Sequence similarity was plotted for a moving window of 1000 bp in steps of 50 bp using the SIMSCAN program and compared with the recombination breakpoint frequency in the same figure. In a similar way, we compared the recombination breakpoint frequency with the mean intersubtype evolutionary distances corrected by TrN with a χ²-distribution rates heterogeneity among sites, as estimated in TREEPUZZLE.

Statistics. The association between sequence similarity and recombination breakpoint frequency was modelled using a linear regression model and the coefficient of determination ($r^2$) was estimated. To detect a possible phase difference between the two series, this procedure was repeated by moving the similarity plot either upstream or downstream in steps of 50 bp over a range of 1000 bp relative to the recombination frequency plot and a maximum $r^2$ value was obtained. To examine whether the recombination breakpoints are more clustered than would be expected by chance, we compared the observed versus the expected number of breakpoints in the major gene products, where the expected number of breakpoints per gene region was estimated by averaging the total number of breakpoints per sequence length in the alignment using the chi-squared test.

Simulations. To test for the statistical significance of the association between the sequence similarity and the recombination breakpoint frequency of the maximum $r^2$ value and to validate the statistical significance of the difference between the observed and the randomly derived frequency of recombination breakpoints, recombination was simulated using a uniform distribution of breakpoints and the distributions of the maximum $r^2$ value. The chi-squared values obtained from the simulated datasets were used as the null. More specifically, the positions of the recombination breakpoints were simulated using a Monte Carlo-based algorithm according to the following assumption: the positions of the breakpoints were distributed uniformly along the genome and, more specifically, within positions 400–7800 of the alignment. Recombination breakpoints were simulated, for simplicity, at positions multiple of 50 nt, since the accuracy for detection of the breakpoints is below 50 nt.

RESULTS

Our re-analysis revealed that approximately half of the sequences are more complex recombinants than described initially. The overall number of breakpoints was found to be 247 with a median (range) of 7 (2–15). An intersubtype recombination breakpoint was defined as a region of approximately 100–200 nt in length in which the HIV-1 subtype classification was significantly different (supported by > 75% NJ bootstrapping and > 70% ML puzzling steps) in the adjacent pieces. Moreover, we identified 14 breakpoints within regions that shared an immediate common ancestor, thus corresponding to non-independent recombination events excluded from all subsequent analyses. The results of this re-analysis provided an updated dataset of the 233 recombination breakpoint positions based on all intersubtype HIV-1 recombinants available to date.

First, the frequency of recombination breakpoints, adjusted per 1000 nt, was plotted along the HIV-1 genome (Fig. 2). According to Fig. 2, there is only a twofold fluctuation of the breakpoint frequency throughout the HIV-1 genome, e.g. within the pol gene, the frequency of breakpoints was
twofold higher than that in gag or env. In contrast, where a more detailed approach was used allowing the mapping of the recombination breakpoints within a sliding window of 1000 bp (BREAKSCAN), the frequency of intersubtype recombination was found to differ considerably (fivefold) within certain genomic regions (Jetzt et al., 2000). Recombination frequency peaks were found in RT, in the accessory genes vif and vpr, in the first exons of tat/rev, in vpu and in gp41, indicating that they provide intersubtype recombination hot spots. Conversely, in gp120 and in p17/p24 (gag), the frequency of recombination was the lowest throughout the complete HIV-1 genome. To examine whether the recombination breakpoints are more clustered than expected by chance, we compared the observed versus the expected number of breakpoints in the major gene products. The result of this analysis revealed that the distribution of breakpoints observed was significantly different from the expected one ($P < 0.002$), suggesting that intersubtype recombination does not occur randomly across the HIV-1 genome.

The shape of the recombination breakpoint frequency plot prompted us to examine for any potential association with sequence similarity within the HIV-1 genome. Comparison of results obtained from the BREAKSCAN and SIMSCAN programs in a single plot (Fig. 1) indicated that there is a correlation between the outlines of the two curves, suggesting that recombination occurs more frequently adjacent to genomic regions with higher conservation. Furthermore, Fig. 1 suggests that there is a ‘phase difference’ between the two plots and, more specifically, the similarity peaks appear upstream of the recombination hot spots. To resolve the phase difference between sequence similarity and recombination breakpoint frequency, the coefficient of determination ($r^2$) between the two was determined by moving the similarity plot upstream or downstream in steps of 50 bp over a range of 1000 bp relative to the recombination frequency plot (Fig. 3A). The coefficient of determination was highest ($r_{max}^2 = 0.69$) for a phase difference of 700 bp upstream to the recombination frequency plot (Fig. 3A). The coefficient of determination was similar ($r_{max}^2 = 0.67$)

**Fig. 3.** (A) Phase difference between recombination breakpoint frequency and sequence similarity. The coefficient of determination ($r^2$) between the recombination breakpoint frequency and sequence similarity is plotted against the phase difference between the two, calculated in steps of 50 bp with a maximum phase difference of 1000 bp. The window sizes for either the recombination frequency or the similarity plot were 1000 bp. (B) Coefficient of determination ($r^2$) between the recombination breakpoint frequency and the similarity plot when the window size varied from 800 to 2000 bp in steps of 100 bp for the recombination frequency. (C) Coefficient of determination ($r^2$) between the recombination breakpoint frequency and the similarity plot when the window size varied from 400 to 1200 bp in steps of 100 bp for the similarity plot.
when the evolutionary distances using TrN with a \( \gamma \)-distribution rates heterogeneity among sites was used instead of the uncorrected sequence similarity.

To examine whether this phase difference was affected by the window size of the similarity or the recombination frequency plots, the correlation coefficient between them was determined for different window sizes (Fig. 3B, C). In particular, similar results for the phase difference were obtained when the window size varied from 800 to 2000 and from 400 to 1200 bp in steps of 100 bp for recombination frequency and similarity plots, respectively, in all possible combinations (Fig. 3). The minimum window size of the recombination frequency plot used in this analysis was set to 800 bp, estimated according to Sturges' rule (Sturges, 1926). Furthermore, the minimum window size for the similarity plot was set to 400 bp, considering that such a fragment contains enough phylogenetic signal throughout the HIV-1 genome. According to this analysis, the correlation between the recombination frequency and similarity plots was maximized only for a phase difference of approximately 650 bp upstream to the recombination frequency and it was not affected by the window size.

To test for the statistical significance of the estimated \( r_{\text{max}}^2 \) value (\( r_{\text{max}}^2 = 0.69 \)), we estimated the distribution of the coefficient of determination (\( r^2 \)) between the similarity and the frequency of recombination under the assumption that the positions of the breakpoints are distributed uniformly across the genome. More specifically, the location of the breakpoints was simulated in 1000 replicates of 34 sequences and then the correlation between recombination frequency and similarity was calculated for each dataset using a sliding window of 1000 nt, in the same way as in the original data. The value obtained for the real dataset (\( r^2 = 0.69 \)) was found to lie at the extreme 5% of the distribution of the 1000 maximized coefficients of determination, indicating that there is a statistically significant association between the frequency of recombination breakpoints and sequence similarity (Fig. 4).

**DISCUSSION**

Our re-analysis revealed that intersubtype recombination in HIV-1 is not a random process but occurs more or less frequently in certain genomic regions. The identification of intersubtype recombination hot spots was based on the analysis of all full-length HIV-1 intersubtype recombinants using a sensitive method, such as the sliding window approach, which allows monitoring for any potential differences in the recombination frequency in every single partial genomic region of the HIV-1 genome.

The identification of recombination hot spots may be important for the development of retrovirus constructs used as reagents for HIV-1 vaccine development (Kuiken et al., 2000). Additionally, hot spots render them the most appropriate regions for detecting HIV-1 intersubtype recombinants and thus facilitating the HIV-1 molecular epidemiological studies based on the analysis of partial

**Fig. 4.** Distribution of 1000 simulated coefficients of determination between the frequency of recombination breakpoints and sequence similarity obtained under the assumption that breakpoints are distributed uniformly along the genome. Black bars indicate the extreme 5% of the distribution.
genomic regions. Furthermore, in genomic regions where recombination occurs less frequently, the assumption that virus divergence is dominated by point mutations is not violated, thus rendering them most appropriate for phylogenetic analyses and molecular clock calculations. The existence of certain hot spots within different genomic regions could be explained possibly by any structural features of the RT, which causes template jumping, or any potential functional constraints between different genomic regions belonging to diverse HIV-1 subtypes (Gao et al., 1996; Paraskevis et al., 2000; Worobey & Holmes, 1999).

On the other hand, based on our analysis, sequence similarity is of vital importance for intersubtype recombination events, as has been considered previously (Worobey & Holmes, 1999). It is of particular interest that in C2–V3 region of env, which is the most variable region of the HIV-1 genome (Kuiken et al., 2000), no recombination breakpoints were observed. Furthermore, the importance of sequence similarity in the efficiency of recombination has been demonstrated previously in vitro, where the rate of recombination between two non-homologous RNAs was 100 to 1000 times less than that between two entirely homologous sequences (Zhang & Temin, 1993). Thus, in accordance with previous findings about the correlation between sequence similarity and recombination, we provide direct evidence based on data observed in vivo: first that sequence similarity drives template jumping in retrovirus homologous recombination and second that sequence dissimilarity may be the predominant constraint in recombination events between quite divergent strains. According to our analysis, there might be a similarity threshold for template switching and recombination. In the case of HIV-1 intersubtype recombination, where there is considerable divergence between different lineages, recombination occurs more frequently in the most well-conserved regions, such as in pol. These findings are in accordance with recent findings from in vitro studies (Iglesias-Sanchez & Lopez-Galindez, 2002).

Although different parameters such as selection, protein functionality and epidemiological reasons may have played a role in the successful spread of the HIV-1 intersubtype recombinants detected until now, we provide evidence that sequence similarity provides a major constraint in the process of the generation of HIV-1 recombinants. This means that this particular association between sequence similarity and recombination may affect the formation of mosaic strains before other factors such as selection or virulence play their role in the spread of newly generated recombinants.

On the other hand, the expected outcome of recombination frequency should be quite different if the evolutionary selection was the main driving force: recombination should be more frequent adjacent to divergent regions, such as env, and less frequent at conserved regions, such as the accessory genes and pol/RT (see supplementary Fig. 2, available online at http://vir.sgmjournals.org), resulting in negative correlation between sequence similarity and recombination frequency. While point mutations do happen randomly because of the fortuity of RT errors, they are strongly selected by evolutionary forces, resulting in more conserved regions, such as pol, and less conserved regions, such as env. Surprisingly, recombination does not seem to be affected by evolution in the same way and the observed pattern must be affected by another, less obvious, underlying mechanism.

Moreover, our analysis revealed a newly identified phase difference of approximately 650 nt between sequence similarity and recombination frequency. According to the models proposed for retrovirus recombination, hybridization between the nascent DNA and the acceptor molecule is required for priming an efficient continuation of DNA synthesis (Hu & Temin, 1990b; Negroni & Buc, 2001). Thus, if sequence similarity is important in this process, it will appear as a phase shift with respect to the recombination breakpoint: downstream if recombination occurs during first-strand cDNA synthesis and upstream when during second (or plus)-strand synthesis (Fig. 5). Our observations are thus more supportive of the strand displacement-assimilation model to occur in vivo.

Previous in vitro experiments documented that retrovirus recombination may occur during both plus- and minus-strand DNA synthesis (Hu & Temin, 1990a, b; Junghans et al., 1982), while it is strongly believed to occur mainly during minus-strand DNA synthesis through the copy-choice model (Jetzt et al., 2000; Negroni & Buc, 2001). It is interesting that since the first observation of the DNA–H structures through electron microscopy and the phrasing of the strand displacement-assimilation model, no other evidence has ever supported this model as predominant in any other retrovirus (Negroni & Buc, 2001). On the contrary, our analysis provides evidence that strand displacement-assimilation might play a significant role in intersubtype recombination, although further study is required to assess the potential of the strand displacement-assimilation model in the generation of HIV-1 intersubtype recombinants.

Recombination is influential for retrovirus evolution and it has been estimated that for HIV-1, recombination occurs, on average, more frequently than the mutation accumulation rate (Jetzt et al., 2000; Negroni & Buc, 2001). The evolutionary advantages of homologous recombination are well established (Burke, 1997). In depth analysis reveals that recombination rises as a primitive form of sexual reproduction, a type of amphigony (Temin, 1991). Therefore, it may serve as an emergency exit from the accumulation of deleterious mutations known as ‘Muller’s ratchet’ (Chao, 1990). Furthermore, it gives the potentiality of ‘evolutionary broad jumping’ between peaks of fitness; it allows more efficient exploration of the sequence space and increases genetic diversity.

In brief, our findings indicated that there are certain intersubtype recombination hot spots across the HIV-1 genome and that sequence similarity plays a pivotal role in
recombination events between different subtypes, indicating that sufficient sequence similarity is required upstream of the recombination breakpoint and suggesting that the strand displacement-assimilation model might provide the dominant model for HIV-1 inter_subtype recombination.

REFERENCES


...
G. Magiorkinis and others


Ray, S. C. (1998). SIMPLOT. Division of Infectious Diseases, Johns Hopkins University School of Medicine, Baltimore, USA.


