Published online 26 September 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 18* ***5203–5216***
doi:10.1093/nar/gkl669

Sequence determinants of breakpoint location during HIV-1 intersubtype recombination

Heather A. Baird, Román Galetto[1], Yong Gao, Etienne Simon-Loriere[1], Measho Abreha, John Archer[2], Jun Fan[2], David L. Robertson[2], Eric J. Arts and Matteo Negroni[1,]*

Division of Infectious Diseases, Department of Medicine, Case Western Reserve University,
Cleveland, OH 44106, USA, [1]Unité de Régulation Enzymatique des Activités Cellulaires, CNRS URA 2185,
Institut Pasteur, 25 Rue du Dr Roux, Paris, Cedex 15, 75724 France and [2]Faculty of Life Sciences,
University of Manchester, Manchester M13 9PT, UK

Received July 20, 2006; Revised and Accepted August 30, 2006

ABSTRACT

Retroviral recombination results from strand switching, during reverse transcription, between the two copies of genomic RNA present in the virus. We analysed recombination in part of the envelope gene, between HIV-1 subtype A and D strains. After a single infection cycle, breakpoints clustered in regions corresponding to the constant portions of Env. With some exceptions, a similar distribution was observed after multiple infection cycles, and among recombinant sequences in the HIV Sequence Database. We compared the experimental data with computer simulations made using a program that only allows recombination to occur whenever an identical base is present in the aligned parental RNAs. Experimental recombination was more frequent than expected on the basis of simulated recombination when, in a region spanning 40 nt from the 5′ border of a breakpoint, no more than two discordant bases between the parental RNAs were present. When these requirements were not fulfilled, breakpoints were distributed randomly along the RNA, closer to the distribution predicted by computer simulation. A significant preference for recombination was also observed for regions containing homopolymeric stretches. These results define, for the first time, local sequence determinants for recombination between divergent HIV-1 isolates.

INTRODUCTION

The extensive sequence variability observed in the global HIV-1 epidemic originates from the rapid viral turnover (10^{10}–10^{12} viral particles per day) in an HIV-infected individual (1–3), combined with the high mutation rate during HIV-1 reverse transcription (10^{-4} per nucleotide) (4). Additionally, many studies have shown that recombination within HIV-1 intrapatient populations is pervasive and results in new viral variants, promoting the rapid selection of forms resistant to HIV-specific drug and immune pressure (5–8).

HIV-1 group M strains account for over 95% of infections worldwide, and can now be sub-divided into nine different subtypes or clades (A–D, F–H, J, K), which share 70–80% nt sequence identity in the envelope gene (*env*) (Los Alamos HIV Sequence Database, http://hiv-web.lanl.gov). Co-infection or superinfection with divergent strains can result in recombination between highly divergent viral genomes, e.g. generating intersubtype or intergroup recombinants (9,10). Some of these recombinant forms have major relevance in the pandemic, as indicated by the description, to date, of 34 circulating recombinant forms (CRFs) in the HIV Sequence Database. Partial and full genome sequencing of HIV-1 isolates from around the world indicates that, overall, at least 20% are chimeras of different HIV-1 subtypes (11–13). For instance, in East Africa, where the isolates that make the object of the present study were collected, intersubtype A/D, A/C, and D/C recombinant forms have rapidly emerged in the past 10 years due to the co-circulation of subtypes A, C and D (14).

Nearly all information on intersubtype HIV-1 recombinants is derived from sequences of isolates of established infections, which provides limited understanding on the forces involved in their genesis and replication success. Over the past decade, many studies have been directed at understanding the mechanisms of HIV-1 recombination either using cell-free reconstituted *in vitro* systems or using defective retroviral vectors in tissue culture based systems [see Refs (15) and (16) for reviews]. Recombinant DNA molecules have been proposed to arise during synthesis of either minus (−) (17–20) or plus (+) (21–23) DNA strand.

*To whom correspondence should be addressed. Tel: +33 01 45 68 85 05; Fax:+33 01 45 68 83 99; Email: matteo@pasteur.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

The contribution of the second type of mechanism to the overall frequency of recombination has been challenged by a mounting number of studies where both (−) and (+) strand recombination events could be separately screened, showing that template switching during (−) strand DNA synthesis is responsible for the majority of retroviral recombination events (17–20). Recombination occurring during (−) DNA strand synthesis is thought to follow a process known as copy choice (24), according to which the nascent DNA strand is transferred, from one to the other copy of genomic RNA present in the retroviral particle (defined as the donor and the acceptor RNAs, respectively). The degradation of the donor RNA by the reverse transcriptase (RT) encoded RNase H activity (25,26) is required to free the (−) strand DNA for strand invasion and transfer onto the acceptor RNA.

Structural features in the RNA (27–32) and stalling of DNA synthesis during reverse transcription (32–36) have both been shown to constitute ‘triggers’ for copy choice by following various proposed mechanisms [reviewed in Refs (16,37)]. These conclusions have been inferred, mostly in cell-free studies, by the correlation between the existence of preferential sites for strand transfer and the presence of RNA hairpins or positions of stalling of DNA synthesis in the vicinity. The physiological relevance of these observations remains to be assessed, particularly since the existence of preferential sites for copy choice *per se* remains to be proven when considering reverse transcription in infected cells. Indeed, recent studies on the reverse transcription products (RTPs) generated after a single infection cycle of cells in culture addressed the issue of whether recombination occurs randomly across the genome or if ‘hot spots’ for recombination exist at specific sites (20,38–40). Discordant conclusions have been reached using different pairs of subtype B HIV-1 isolates, with the identification of putative hot spots in one case (38) but not in another (40).

Previously we have shown that after a single infection cycle of cells in culture we can identify, in a short region of the *env* gene, the occurrence of preferential strand transfer between almost identical sequences derived from the LAI isolate from subtype B (39). In this case, the high level of switching was correlated to the presence of a stable RNA hairpin. Interestingly, the *env* C2 region was also identified as an intersubtype recombination hotspot using a dual infection/multiple cycle system (41). These studies, performed independently by the authors of the present study, were not directly comparable due to the use of different HIV-1 strains in the single and multiple cycle systems.

Most mechanistic studies on retroviral recombination have employed almost identical templates. However, recombinants involving genetically distant forms have been frequently identified and might also be those forms with a major impact on HIV evolution. In this case, the position where template switching occurs is the result of a complex overlay of factors, as the degree of sequence identity, RNA structures, and sequence features such as homopolymeric stretches (HPS). Here, we analyse the generation of A–D intersubtype recombinants along a segment spanning the regions from C1 to C4 of the gp120 coding sequence, and define the contribution of sequence motifs, identifiable by sequence comparison, that lead to template switching between genetically divergent natural isolates.

MATERIALS AND METHODS

Cells. 293T and U87.CD4.CXCR4 cells were grown in DMEM supplemented with 10% foetal calf serum (FCS), penicillin and streptomycin (from Invitrogen, Carlsbad, CA, USA) and maintained at 37°C with 10% CO_2. Selection for CD4 and CXCR4 expression in U87 cells was maintained with 300 μg/ml hygromycin/G418 and 1 μg/ml puromycin, respectively. MT4 cells were maintained in RPMI 1640 medium supplemented with 10% FCS and antibiotics at 37°C with 5% CO_2.

Viruses. Five syncytium-inducing (SI) HIV-1 strains were isolated from HIV-infected Ugandans as previously described (42). Two primary HIV-1 isolates were subtype A (115A and 120A), while the other three belonged to subtype D (89D, 122D and 126D), as deduced by phylogenetic analyses on the *env* sequences. Due to previous confusion in strain nomenclature, we have modified these virus names from those previously published (A14 is now 115A, A15 = 120A, D13 = 122D, D14 = 126D and D15 = 89D). All viral stocks were propagated and expanded in PHA-stimulated, IL-2 treated U87 cells as described (43). Tissue culture dose for 50% infectivity (TCID50) was determined for each isolate using the Reed and Muench method (44), and titers were expressed as infectious units per millilitre (IU/ml). All the *env* genes of these HIV-1 isolates have been previously sequenced (42).

Single cycle tissue culture assay. Single cycle recombination assays were completed using an assay system previously developed by our laboratory (Figure 1A–C) (39). HIV-1 envelope gene fragments from subtypes A and D (nt 6420–7520 of HXB2 isolate as reference) viral DNA were amplified by PCR from infected PBMC obtained from patients and cloned in pLac$^+$ and pLac$^-$ plasmids, which differ for the genetic marker present downstream (in the sense of reverse transcription) the sequence where recombination is studied. pLac$^+$ and pLac$^-$ plasmids carry, as genetic markers, either a functional *LacZ′* gene or a sequence complementary to a portion of the mRNA coding for *Escherichia coli malT* gene, respectively. All constructions were verified by sequencing. The transcomplementation plasmids, pCMVΔR8.2 (45) encoding HIV-1 gag, pol, and accessory proteins, and pHCMV-G (46) encoding the VSV envelope protein were co-transfected into 293T cells with the lac$^+$ and lac$^-$ genomic plasmids to produce defective retrovirus particles as described (39). Cells were previously plated at a density of 3.5×10^6 per 100 mm-diameter dish and transfected 16–20 h later. The medium was replaced 8 h after transfection, and the vector supernatants were recovered 36 h later. Non-internalised DNA was removed by treatment of the vector supernatants with DNaseI (1 μg/ml in the presence of 1 μM $MgCl_2$) for 30 min at 37°C. Viral titer was determined using the HIV-1 p24 enzyme-linked immunosorbent assay kit (PerkinElmer Life Sciences, Wellesley, MA, USA), and, when necessary, vector supernatants were concentrated by using Centricon® YM-50 centrifugal filter devices (Amicon-Millipore, Bedford, MA, USA) before transduction. MT4 cells were transduced with 200 ng of p24 antigen per 10^6 cells (an approximate multiplicity of infection of 20) in 35 mm dishes in a 500 μl volume. Two hours post-transduction, the cells were diluted up to a 4 ml volume with supplemented RPMI medium and maintained at 37°C in a 5% CO_2

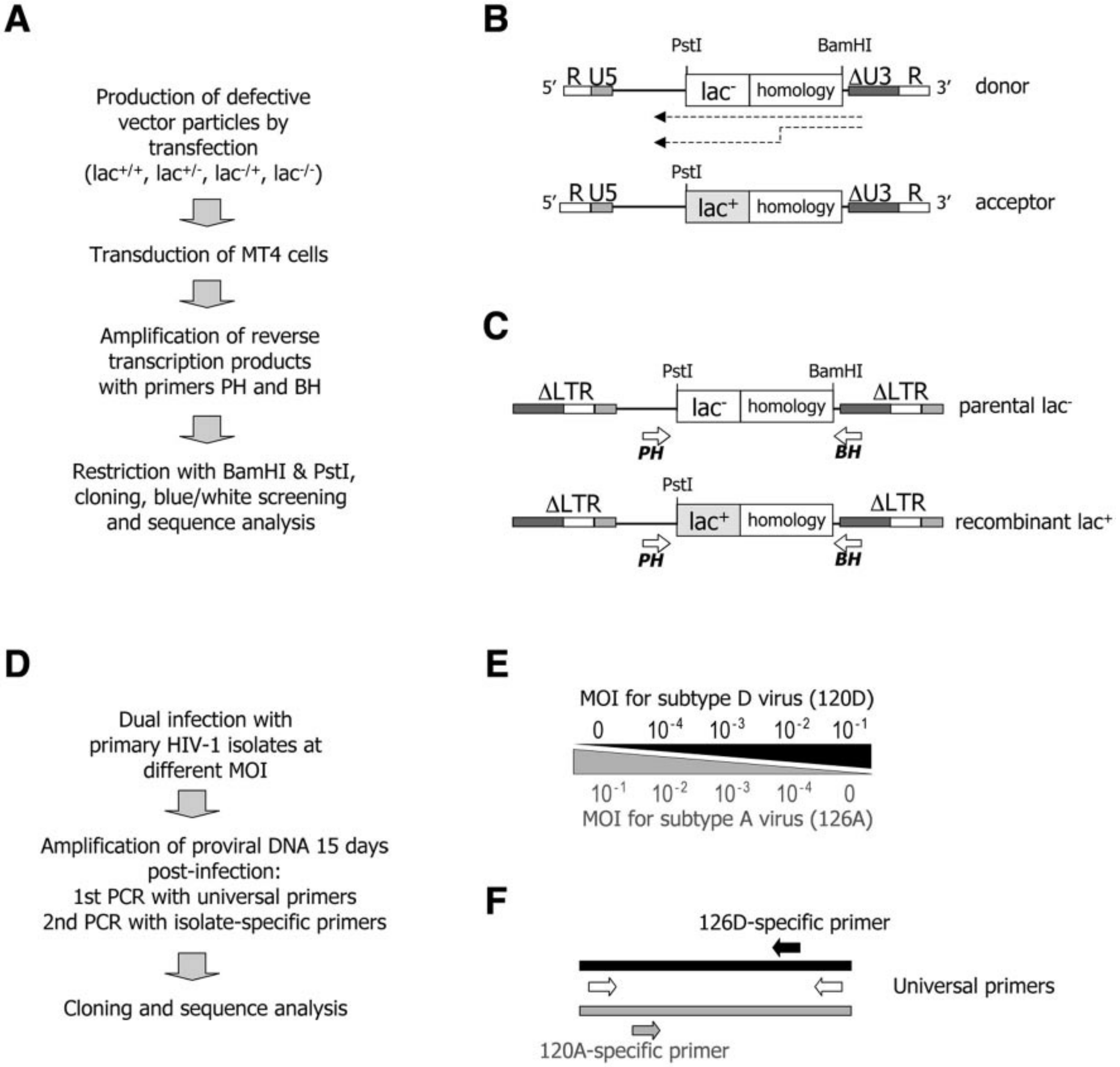
A
Production of defective vector particles by transfection (lac+/+, lac+/-, lac-/+, lac-/-)
Transduction of MT4 cells
Amplification of reverse transcription products with primers PH and BH
Restriction with BamHI & PstI, cloning, blue/white screening and sequence analysis
B
PstI
BamHI
5′ R U5
lac-
homology
ΔU3 R 3′
donor
PstI
5′ R U5
lac+
homology
ΔU3 R 3′
acceptor
C
PstI
BamHI
ΔLTR
lac-
homology
ΔLTR
parental lac-
PH
BH
PstI
ΔLTR
lac+
homology
ΔLTR
recombinant lac+
PH
BH
D
Dual infection with primary HIV-1 isolates at different MOI
Amplification of proviral DNA 15 days post-infection: 1st PCR with universal primers 2nd PCR with isolate-specific primers
Cloning and sequence analysis
E
MOI for subtype D virus (120D)
0 10-4 10-3 10-2 10-1
10-1 10-2 10-3 10-4 0
MOI for subtype A virus (126A)
F
126D-specific primer
Universal primers
120A-specific primer

Figure 1. Schematic representation of the systems. (**A**) outline of the single cycle infection tissue culture system. The structure of the genomic RNAs and reverse transcription products (RTP) are shown in (B) and (C), respectively. (**B**) lac$^-$ indicates a partially deleted, and thus non-functional, portion of the *malT* gene from *Escherichia coli*; lac$^+$ designates the *lacZ′* gene. The dotted lines in (B) represent reverse transcription leading to synthesis of parental (lac$^-$) or recombinant (lac$^+$) products identified in this assay. (**C**) the presence of partially deleted, and thereby non-functional, LTRs in the reverse transcription products is indicated as ΔLTR. The position where primers *PH* and *BH*, used for PCR amplification prior to cloning in *E.coli*, anneal on the RTP is shown (also see Materials and Methods, and Result sections for details). (**D**) outline of the multiple cycle tissue culture system. Dual infections of U87.CD4.CXCR4 cells with subtype A and D HIV-1 isolates were done with isolates 126D and 120A at different ratios of multiplicity of infection [MOI, given in (**E**)]. After the first round of replication, co-infected cells can produce both parental and heterodiploid viruses. Infection of new cells with heterodiploid virions can lead to intersubtype recombination. The scheme for PCR amplification of 126D/120A recombinants is outlined in (**F**). Black and grey bars indicate the portion of the proviral DNA coding for the part of the *env* gene studied. White arrows indicate schematically the position of annealing of the universal primers used for the first amplification reaction. The black and the grey arrow indicate the isolate specific primers used for nested amplification specific for the 126D/120A recombinants.

incubator for 40 h. The RTP were purified from the cytoplasmic fraction of transduced cells using the method described by Hirt (47), since most of the RTP will remain in a non-integrated form, particularly since our genomic vectors lack the FLAP sequence, shown to enhance nuclear import of RTP (48). Cells were lysed by incubation for 10 min at room temperature in a buffer containing 10 mM Tris–HCl pH 8.0, 10 mM EDTA and 0.6% SDS. High molecular weight DNA was removed by precipitation at a high salt concentration (NaCl 1 M) during 12–18 h on ice. The lysates were cleared by ultracentrifugation at 30 000 r.p.m. for 1 h and the supernatants treated with 100 µg/ml RNaseA for 1 h at 37°C and 100 µg/ml Proteinase K for 3 h at 50°C. After phenol/chloroform extraction, DNA was ethanol-precipitated and purified using the NucleoSpin® Extract clean-up kit (Macherey-Nagel, Düren, Germany). The purified double stranded DNA was digested with DpnI for 2 h at 37°C (in order to eliminate possible contaminating DNA of bacterial origin) prior to PCR amplification (20 cycles) with primers *BH* and *PH* (Figure 1C). The amplified product was purified after electrophoresis on agarose gel, digested with PstI and BamHI, ligated into an appropriate plasmid vector and transformed in *E.coli*. Plating on IPTG/X-Gal containing dishes allowed blue/white screening of recombinant and parental colonies, respectively (39). Recombination breakpoints were identified by full-length sequencing of the C1–C4 *env* region of the recombinant clones.

Estimation of the recombination rates in the single cycle system. The transfection of equal numbers of $pLac^+$ and $pLac^-$ plasmids leads to the production of similar quantities of each of the genomic RNAs as previously reported (39), since the same promoter (the viral U3 sequence) is present in both genomic plasmids. Given that the RNAs also share the same sequences for dimerization and encapsidation, these processes are expected to yield 50% of heterozygous, and 50% of homozygous particles, with an equal proportion (25%) of $lac^{+/+}$ and $lac^{-/-}$ vectors, following a Hardy–Weinberg distribution, as usually assumed (15). After transduction of MT4 cells, the RTP are amplified by PCR and cloned after digestion with BamHI and PstI in *E.coli*. This procedure will allow cloning $BamHI^+/lac^-$ and $BamHI^+/lac^+$ RTP, which can be generated by reverse transcription in homozygous $lac^{-/-}$ and in heterozygous particles (39). As a result, assuming that only one molecule of double stranded DNA is generated from each viral particle, one third of the total number of colonies will correspond to RTP issued from $lac^{-/-}$ vectors. The total number of colonies is therefore multiplied by 2/3 in order to consider only the RTP issued from heterozygous particles. The possibility of cloning products of cellular origin was also ruled out as described in (39). The number of white colonies (N) is corrected by a factor given by $n/48$, where n is the number of colonies that resulted from cloning of RTP after analysis of 48 white colonies. The global frequency of recombination (F) is therefore given by $F{=}b/\{2/3[N(n/48) + b]\}$, where b is the number of blue colonies. The recombination rate per nucleotide (f) within a given interval (i) is given by $f = F(x_i/X)/z_i$, where F is as above, x_i is the number of colonies where recombination was identified to have occurred within the interval considered, X is the total number of colonies on which mapping was performed, and z_i is the size in nucleotide of the interval.

HIV-1 dual infection assay. Different pairs of HIV-1 isolates were used to simultaneously infect U87 cells as described previously (43). We performed four separate dual infections of U87.CD4.CXCR4 with two HIV-1 isolates at different multiplicities of infection as indicated in (Figure 1D and E). One million U87.CD4.CXCR4 cells were incubated with these virus mixtures for 2 h at 37°C, 5% CO_2, washed and then resuspended in complete medium (1×10^6 cells/ml). Virus production during dual infection was monitored in the cell culture supernatants at days 3, 6, 9, 12 and 15 postinfection by the use of an endogenous RT assay (49). Cells were harvested at day 15, resuspended in DMSO/foetal bovine serum, and stored at −80°C for subsequent analysis.

PCR strategy to select HIV-1 recombinants. For all dual infection experiments, proviral DNA was extracted from lysed U87.CD4.CXCR4 cells using the QIAamp DNA Blood Kit (Qiagen). A ∼3 kb fragment of the HIV-1 *env* gene, encoding the gp120, was PCR amplified using the set of universal primers envB (AGAAAGAGCAGAAGACAGTGGCAATGA) and EnvN (CTGCCAATCAGGGAAGTAGCCTTGTGT). Isolate-specific primers internal to the previous *env* products were then used for a nested PCR amplification of subtype A/D recombinants using PCR primer ESA120 (AAGCATATGATGCAGAAGTAC) as sense primer specific for virus 120A, and EDS2 (TGTCAATTTCTCTTTCCCAC) as antisense primer specific for virus 126D. Both external and nested PCR were carried out in a 100 µl reaction mixture with defined cycling conditions (43). PCR-amplified products were separated on agarose gels and then purified using the QIAquick PCR Purification Kit (Qiagen). Control PCR amplifications were performed with subtype-specific DNA templates as previously described (41), to rule out the possibility of *Taq*-generated recombinants (50).

Nucleotide sequencing, phylogenetic, and recombination analysis. HIV-1 isolates were sequenced using primers previously described (43). The *env* sequences of the HIV-1 isolates used in this study and set of reference strains were aligned using the CLUSTAL X v.1.63b program (51). After sequencing of the recombinant clones, breakpoints were identified by visual inspection. Note, accuracy of breakpoint location is related to the length of the region between mismatches.

Simulation of recombinants. To generate a simulated recombinant sequence either the first or second parental sequence was randomly chosen to be used to make up the initial part of the artificial recombinant sequence. A random number between one and the length of aligned parental sequences was chosen to place the breakpoint. Breakpoint locations are then taken from the mismatches either side of this position. Breakpoints were not permitted to occur directly on mismatches.

Identification of breakpoints in sequences from the database. An alignment of envelope sequences was obtained from the Los Alamos HIV Sequence Database (http://hiv-web.lanl.gov/). For each intersubtype recombinant sequence identified in this alignment, a reduced alignment was generated including the recombinant itself, consensus sequences corresponding to the subtypes identified by the sequence database as being involved in the recombination event and a suitable outgroup (the consensus sequence from a different subtype). For each reduced alignment, modified versions of informative sites analysis and diversity plotting (52,53) were used to detect intersubtype recombination breakpoints. Breakpoints were tested by performing 1000 neighbor-joining bootstrap replicates, implemented with PAUP 4.0b10 (Swofford, 2002), on alignments from either side of putative breakpoints. Breakpoints were confirmed if the bootstrap replicates on both sides of the breakpoint were 75% or higher. In ambiguous cases, or where multiple breakpoints are in close proximity, the program SimPlot (54) was used to determine whether or not a breakpoint was false.

RESULTS

Recombination in HIV-1 after a single cycle of infection

Recombination within or between HIV-1 subtypes A and D was first examined using an experimental system in which HIV-1 copy choice recombination is detected after a single infection cycle of human cells in culture. Two isolates from subtype A (isolates 115 and 120) and three from subtype D (isolates 89, 122 and 126) were employed. Sequence similarity ranged from 67.6% to 73.4% between subtype A and D isolates, and 79.4% to 81.5% within the subtypes. Each virus isolate is designated by a number followed by the letter ‘A’ for subtype A or ‘D’ for subtype D. For each isolate the *env* gene was cloned into two types of HIV-1 genomic plasmids, previously described (39) (lac^{-} donor and lac^{+} acceptor) which were then used to produce heterozygous lac$^{+/-}$ and homozygous lac$^{-/-}$ and lac$^{+/+}$ VSV env-pseudotyped HIV-1 particles by transfection (Figure 1A). These defective virions were then used to transduce MT4 cells, and RTP were cloned as indicated in Figure 1A. The ratio of lac^{+} to total colonies provides an estimate of the recombination frequency in heterozygous particles (Figure 1A; see Materials and Methods). Intra and intersubtype recombination frequencies ranged from 4 to 10% in this study.

For each experiment, a control sample was run in which homozygous lac$^{+/+}$ and lac$^{-/-}$ vector viruses were produced separately, as previously described (39), and used to co-transduce MT4 cells, providing an estimate of the background of artifactual recombinant molecules generated during the experimental procedure. This frequency was at least 20 times lower than the recombination frequency obtained from heterozygous virions (data not shown), in accord with previous observations (39). It should be stressed that the sequence indicated as lac^{-} refers to a sequence unrelated to the lacZ sequence (a portion of the *MalT* gene, see Materials and Methods), making impossible the occurrence of reversion of a lac^{-} clone into a lac^{+} by mutation or recombination within the marker sequence.

Mapping recombination sites across the C1–C4 regions of gp120

In this study, the single cycle assay was used to detect recombination breakpoints over ~1100 nt encompassing conserved and variable regions of the envelope gene (C1–C4 *env* regions). A template switch during reverse transcription of the region of homology (Figure 1B) from a lac^{-} to a lac^{+} genomic RNA can be easily identified, after PCR amplification of RTP and cloning in bacteria, by following the blue phenotype of the colonies (Figure 1A; see Materials and Methods). Plasmids are then purified from the blue colonies and sequenced to identify recombination breakpoints. The identity of donor and acceptor isolates, for each experiment, is indicated by the name of the donor followed by the name of the acceptor and separated by a slash. For example, 115A/89D pair indicates that recombination resulted from template switching from the donor RNA template of subtype A, virus 115, onto the acceptor RNA template of the subtype D, virus 89. In each recombinant clone, the two closest residues specific to the donor and to the acceptor RNAs define the breakpoint where template switching had occurred. Figure 2A and B, illustrates the distribution of the breakpoints generated in the single cycle assay using four intersubtype and two intrasubtype HIV-1 pairs of donor and acceptor (Figure 2A and B, respectively). In one case (89D and 120A) both donor/acceptor possibilities were tested for a given pair of isolates (89D/120A and 120A/89D, Figure 2A). Breakpoints were mapped for 19–22 clones of the intrasubtype (115A/120A, 126D/122D) and intersubtype pairs (115A/126D, 89D/120A, 115A/89D). A more exhaustive mapping of breakpoints was performed with the 126D/120A pair by analysing 40 clones.

No mutations were introduced in the region of transfer in any of the 162 recombinants analysed, and no multiple recombinant breakpoints were detected. In our assay, only an odd number of template switching events could be detected on the basis of the recombinant phenotype of the bacterial colonies (BamHI^{+}/lac^{+}; Figure 1). The average recombination frequency found in these studies was 0.094. If multiple recombination events had a probability of occurrence independent from each other, the expected frequency of occurrence of triple recombinants would be given by $0.094^3 = 8.2 \times 10^{-4}$. This value corresponds to 0.87% of the recombinant population [$(8.2 \times 10^{-4})/(9.4 \times 10^{-2})$], making, in theory, the analysis of 115 recombinant clones (given by 1/0.0087) required to identify a triple recombinant in the population. Since here 162 clones were analysed, and no triple recombinants found, it can be concluded that the occurrence of a strand transfer event does not increase the probability of additional strand transfers, as would be expected if negative interference was occurring (40,55–57).

A non-homogeneous distribution of recombination breakpoints was observed for all pairs (Figure 2A, see *P*-values for Chi-square tests), with more recombination breakpoints mapping to C1, C2 and C3 regions. Frequent recombination was found in C1 and C2 with pairs 126D/120A, 115A/126D, and in both combinations with isolates 89D and 120A (in total, 81/101 breakpoints in C1 and C2, equal to 80.2%), while with 115A/89D pairs, most breakpoints clustered in C1 and C3 (14 out of 20, corresponding to 70.0%). In the two cases of intrasubtype recombination (Figure 2B), C2 and C3, but not C1, appeared to be preferential regions for copy choice recombination. In one case in particular (126D/122D, Figure 2B), C2 appeared to be strongly preferred as a region for recombination with 13 breakpoints out of 21 mapping in this region (61.9%).

Local analysis of recombination sites

A crucial issue in retroviral recombination studies is the identification of local regions where cross-overs occur at rates significantly higher than in others. The breakpoints mapped in (Figure 2A) suggest the existence of several regions of preferential recombination. Figure 3 shows the nucleotide alignment of the parental strains for the 126D/120A pair for the 5′ region (containing 33 of the 40 recombinant clones characterised for this pair of isolates). Filled triangles below the alignment give the location of the breakpoint with the white numbers indicating how many times that breakpoint was found. Apart from breakpoint O, all the individual breakpoints could be grouped in five areas spanning several contiguous (or almost-contiguous) breakpoints (A and B, C–F, G–J, K and L, and M and N in Figure 3). The clustering of breakpoints observed in these areas is significantly non-random, even when considering only the C1 and C2 regions (Chi-square test, $P < 0.001$), defining five areas prone to recombination. For instance, window K and L spans only 11 nt and harbours 4 breakpoints (recombination rate per nucleotide, 9.1×10^{-4}) whereas no recombinants

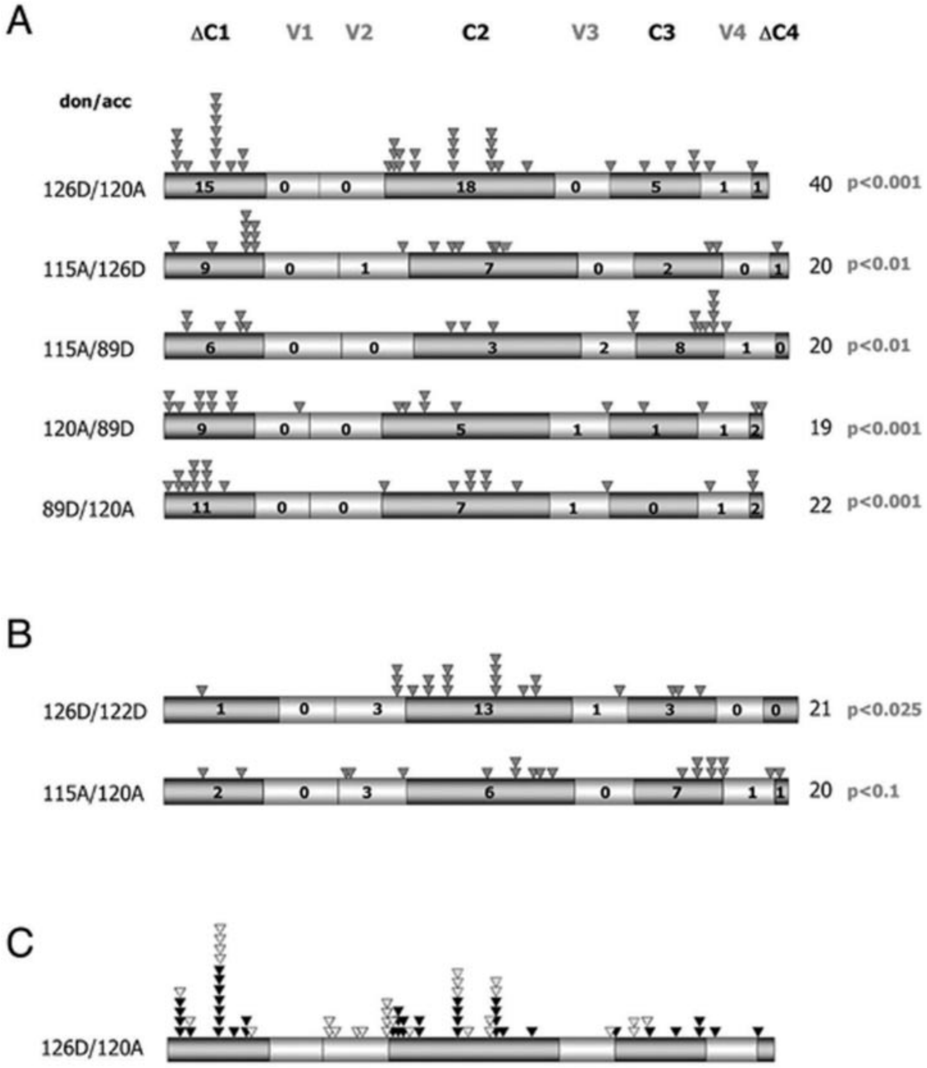

Figure 2. Distribution of breakpoints along the C1–C4 region of *env*. The region of the *gp120* gene studied for intersubtype (**A**) and intrasubtype (**B**) recombination is schematically shown, with the constant regions shaded. The different pairs used (as donor/acceptor) are shown on the left. The different sizes of the constant and variable regions for each pair reflect differences between sequence alignments, due to insertions and/or deletions present in one of the parental strains. Black numbers give the number of breakpoints identified in each region, and grey triangles their approximate position. The total amount of recombinants analysed for each pair is given on the right, together with the *P*-values (in grey) for Chi-square tests for a random distribution of the breakpoints across constant and variable regions. (**C**) Comparison of the distribution of recombination breakpoints in the 126D/120A pair after a single or multiple infection cycles (black and white triangles, respectively).

were found in the 59 and 44 nt regions that surround K and L (recombination rates per nucleotide, $\leqslant 4.0 \times 10^{-5}$ and $\leqslant 5.5 \times 10^{-5}$, respectively).

The possibility that the breakpoints identified more frequently reflected a bias in the PCR amplification and cloning procedure, rather than the relative abundance of these recombinants among the RTP, was investigated. Four plasmids were used, each containing the sequence of one different recombinant clone from the 126D/120A pair. Each recombinant was chosen as to represent a different breakpoint, defined in Figure 3 by letters D, H, J and L. These plasmids, we refer to as recD, recH, recJ and recL, were mixed in the following proportions, respectively: 5, 22.5, 50 and 22.5%. The mixture was amplified by PCR under the same conditions as for the RTP isolated from transduced cells, and the amplified material cloned as described above. If amplification and cloning did not bias the relative amounts of each recombinant DNA, the same proportions present before amplification are expected to be found in the cloned DNA. As shown in Table 1, the proportions observed by sequencing 48 clones were consistent with those of the starting DNAs, supporting the interpretation that the number of times each type of recombinant was cloned in our recombination assays reflects its relative abundance among the RTPs.

Mapping breakpoints in replication-competent intersubtype HIV-1 *env* recombinants from dual infections

We then wanted to detect and characterise intersubtype recombination breakpoints in the HIV-1 *env* gene following multiple rounds of HIV-1 replication and to compare these

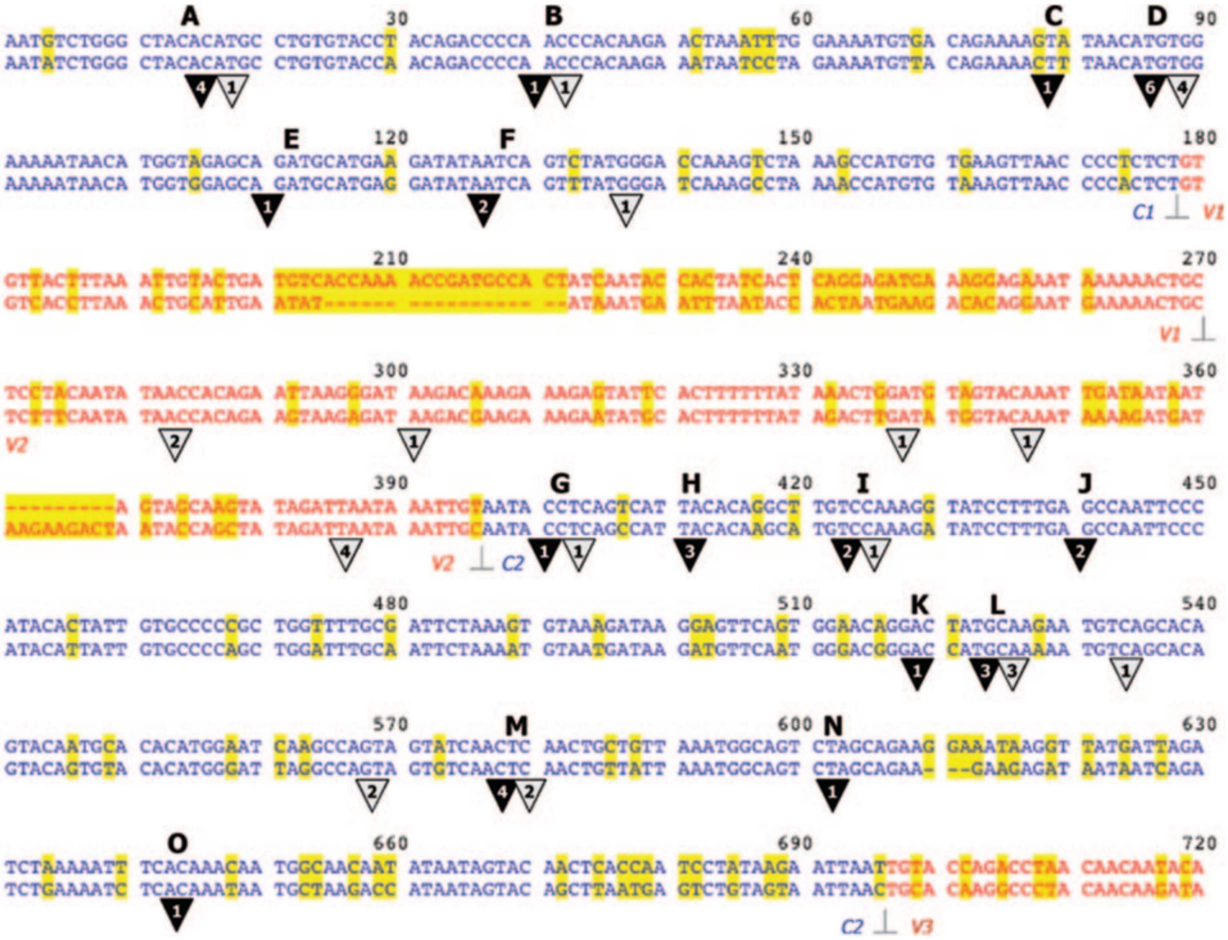
A 30 B 60 C D 90
AATGTCTGGG CTACACATGC CTGTGTACCT ACAGACCCCA ACCCACAAGA ACTAAATTTG GAAAATGTGA CAGAAAAGTA TAACATGTGG
AATATCTGGG CTACACATGC CTGTGTACCA ACAGACCCCA ACCCACAAGA AATAATCCTA GAAAATGTTA CAGAAAACTT TAACATGTGG
4 1 1 1 1 6 4
E 120 F 150 180
AAAAATAACA TGGTAGAGCA GATGCATGAA GATATAATCA GTCTATGGGA CCAAAGTCTA AAGCCATGTG TGAAGTTAAC CCCTCTCTGT
AAAAATAACA TGGTGGAGCA GATGCATGAG GATATAATCA GTTTATGGGA TCAAAGCCTA AAACCATGTG TAAAGTTAAC CCCACTCTGT
1 2 1
C1 ⊥ V1
210 240 270
GTTACTTTAA ATTGTACTGA TGTCACCAAA ACCGATGCCA CTATCAATAC CACTATCACT CAGGAGATGA AAGGAGAAAT AAAAAACTGC
GTCACCTTAA ACTGCATTGA ATAT------ ---------- --ATAAATGA ATTTAATACC ACTAATGAAG ACACAGGAAT GAAAAACTGC
V1 ⊥
300 330 360
TCCTACAATA TAACCACAGA ATTAAGGGAT AAGACAAAGA AAGAGTATTC ACTTTTTTAT AAACTGGATG TAGTACAAAT TGATAATAAT
TCTTTCAATA TAACCACAGA AGTAAGAGAT AAGACGAAGA AAGAATATGC ACTTTTTTAT AGACTTGATA TGGTACAAAT AAAAGATGAT
V2
2 1 1 1
390 G H 420 I J 450
---------A GTAGCAAGTA TAGATTAATA AATTGTAATA CCTCAGTCAT TACACAGGCT TGTCCAAAGG TATCCTTTGA GCCAATTCCC
AAGAAGACTA ATACCAGCTA TAGATTAATA AATTGCAATA CCTCAGCCAT TACACAAGCA TGTCCAAAGA TATCCTTTGA GCCAATTCCC
4 V2 ⊥ C2 1 1 3 2 1 2
480 510 K L 540
ATACACTATT GTGCCCCCGC TGGTTTTGCG ATTCTAAAGT GTAAAGATAA GGAGTTCAGT GGAACAGGAC TATGCAAGAA TGTCAGCACA
ATACATTATT GTGCCCCAGC TGGATTTGCA ATTCTAAAAT GTAATGATAA GATGTTCAAT GGGACGGGAC CATGCAAAAA TGTCAGCACA
1 3 3 1
570 M 600 N 630
GTACAATGCA CACATGGAAT CAAGCCAGTA GTATCAACTC AACTGCTGTT AAATGGCAGT CTAGCAGAAG GAAATAAGGT TATGATTAGA
GTACAGTGTA CACATGGGAT TAGGCCAGTA GTGTCAACTC AACTGTTATT AAATGGCAGT CTAGCAGAA- --GAAGAGAT AATAATCAGA
2 4 2 1
O 660 690 720
TCTAAAAATT TCACAAACAA TGGCAACAAT ATAATAGTAC AACTCACCAA TCCTATAAGA ATTAATTGTA CCAGACCTAA CAACAATACA
TCTGAAAATC TCACAAATAA TGCTAAGACC ATAATAGTAC AGCTTAATGA GTCTGTAGTA ATTAACTGCA CAAGGCCCTA CAACAAGATA
1
C2 ⊥ V3

Figure 3. Mapping breakpoints in the C1–C2 region with 126D/120A pair. The 720 nt at the 5′ end of the *env* region studied are shown. Top row, isolate 120A, bottom, 126D (proviral sense DNA sequences are given). Blue and red residues belong to constant and variable regions, respectively, as deduced by amino acid sequence comparison with the HXB2 reference sequence (68). The border between constant and variable regions is indicated. The residues in a yellow background highlight base differences between the two parental isolates. Hyphenations indicate deleted residues. White numbers in black triangles indicate the number of individual breakpoints mapped, in the interval above, in the single cycle system; while black numbers in grey triangles correspond to breakpoints identified after multiple replicative cycles. The breakpoint is defined by the sequence bordered by the two closest base substitutions in the two parental isolates. Black letters refer to recombination breakpoints identified using the single cycle system, as referred in the text.

Table 1. Test for the existence of bias in the amplification and cloning procedure of recombinants

Recombinant	Proportion of total starting DNA (y)	Expected number of clones (n) in total of 48	Observed clones in a total of 48
recD	0.050	2.4 ± 1.5	2
recH	0.225	10.8 ± 3.3	9
recJ	0.500	24 ± 4.9	25
recL	0.225	10.8 ± 3.3	12

The number of clones (n) expected to be found among 48 clones analysed (third column), was calculated as 48 y, where y is the fraction of total input DNA for each specific recombinant designated in the second column. Error bars were calculated as $n^{1/2}$.

breakpoints to those generated with the single cycle cell-based assay. Selection of *env* recombinants in the single-cycle assay is not dependent on HIV-1 gp120 function or the replication efficiency of the resulting intra- or intersubtype HIV-1 recombinants. We have previously described a method for infecting peripheral blood mononuclear cells with two or more primary HIV-1 isolates of different subtypes (43). Using a heteroduplex tracking assay (HTA), we were able to measure dual virus production, derive relative fitness values of each isolate, and to detect HIV-1 recombination in *env* (43) after heterodiploid virus generation.

For this study we have performed dual infections of U87.CD4.CXCR4 cells with isolates 120A and 126D at different multiplicity of infection (Figure 1D). Dual infection and subsequent recombination was monitored 15 days post-infection using quantitative PCR and HTA. Intersubtype *env* recombinants from dual infections were amplified with subtype-specific primers and cloned into the pCR®2.1-Topo vector (e.g. pCR-A/D env). The primers used for amplification were chosen in such a way as to follow the occurrence of recombination where 126D constituted the donor isolate and 120A the acceptor (analogous to the single cycle assay). To rule out the possibility of *Taq*-generated recombinants, proviral DNA from single infections were mixed and used as template for PCR amplification. The fraction of recombinants to total PCR products resulted to be <0.1%/Knt in *env*, 20-fold less than the intersubtype recombination frequencies observed in the dual infections and 100-fold less than single cycle recombination frequencies (H. A. Baird, M. Negroni and E. J. Arts, unpublished data). Thus, these artificial recombinants did not substantially contribute to recombination breakpoints in our subsequent analyses.

Thirty recombinant clones were sequenced, from the condition of double infection with an equal amount of each parental virus, and breakpoints identified as described in Materials and Methods. Most sites for recombination could be found in the C1 and C2 regions (17 of 30 clones, Figure 2C), although this preference was less pronounced than in the single cycle system, mostly due to the appearance of breakpoints in V2 (9 of 30), with 4 of these 9 cross-overs mapping to the V2–C2 border. Besides these additional breakpoints in the V2 region, there was striking similarity in the 126D/120A recombination hotspots identified in the single cycle and multiple cycle assays. The breakpoints displayed the same non-random distribution across the C1–C4 region with the most intense recombination mapping in C1 and C2. For example, four breakpoints out of 30 were identified in the D window in (Figure 3), compared to the 6/40 in this window derived from the single cycle. Minor differences in the precise localisation of the breakpoints were observed occasionally, as in the K and L region or around breakpoint M (Figure 3). These differences could reflect statistical fluctuations due to limited sampling, or could be indicative of a selection for replication competent forms slightly different from those generated in the absence of selection.

Concurrent to the study on the position of the sites of recombination breakpoints, we investigated the frequency of occurrence of recombination in this portion of the *env* gene (H. A. Baird, M. Negroni and E. J. Arts, unpublished data). Recombinants could not be detected prior to 5 days or when the MOI was reduced to 0.001 or less, observations that can be accounted for by the need for the occurrence of dual infection in order to produce heterozygous virus, a process that requires 2–3 days and a sufficiently high viral titer. At an MOI of 0.01, the production of recombinants (120A/126D and 126D/120A) was only 2%/Knt of the parental viruses 120A and 126D viruses. This low frequency of recombination (~5-fold less than in the single cycle system) could reflect a low frequency of co-infected cells, a reduced fitness of most recombinants, or both.

Recombination and overall sequence similarity

A tendency for clustering of breakpoints in the constant regions underscores the role of the degree of sequence identity between parental isolates in the recombination process, confirming previous observations (58,59). To evaluate whether this parameter alone is sufficient to account for this distribution of breakpoints, we plotted the recombination rate per nt for constant and variable regions for each isolate pair, as a function of the degree of sequence identity between the respective parental strains in each of these regions (Figure 4). The distribution obtained clearly indicates that, as the degree of sequence identity increases, so does the probability of observing high recombination rates. However, no significant correlation exists (Pearson's product moment correlation $r^2 = 0.208$), as would be expected if the degree of identity were the unique determinant of recombination rate. Rather, in several cases, despite high sequence conservation between donor and acceptor RNAs, the recombination rate was low (bottom right corner of Figure 4). Although a certain degree of sequence identity is a prerequisite for a region to harbour efficient recombination, average sequence identity seem not sufficient to determine whether a region will be a site of intense recombination or not.

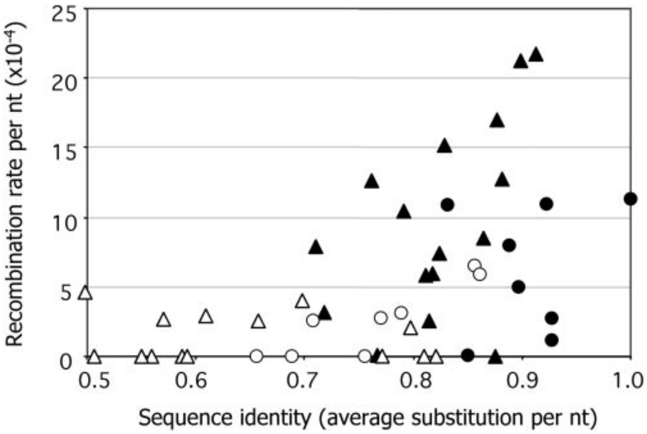

Figure 4. Effect of the sequence identity of C and V regions on recombination. For each pair of isolates studied, a degree of sequence identity was calculated for each individual constant and variable region. The recombination rate per nucleotide in each of these regions was then computed and plotted as function of sequence identity. Circles and triangles represent data from intra- and inter-subtype recombinants, respectively. Constant and variable regions are given by filled and empty symbols, respectively. Only regions presenting >0.5 of similarity are shown, no recombinants having been found in regions with lower similarity.

Analysis of the region of recombination by comparison with simulated recombinant data

Starting from the analysis of the sequence of the recombinants generated from these natural isolates we considered sequence features that have been suggested to affect copy choice in retroviruses and RNA viruses starting from such sequence analyses: presence of HPS, level of sequence identity, and base composition (41,59–63).

To assess the importance of these parameters, we compared the breakpoints generated experimentally with breakpoints generated randomly, on the same sequences, using a computer program, as described in Materials and Methods. The rationale is that no biological property of the nucleic acid sequences (such as primary or secondary structure, or ability to induce pausing of reverse transcription) will affect the distribution of the simulated breakpoints and, if a given parameter plays a role in the generation of the experimental breakpoints, the two datasets will differ when compared on the basis of that parameter.

The program defines the position of breakpoints by choosing random numbers that are reported on the alignment of two parental isolates (see Materials and Methods). This analysis is performed by generating, for each pair of parental strains, the same number of recombinant samples as for the experimental dataset (i.e. 40 samples for 126D/120A, 20 for 115A/126D, and so on, see Figure 2), and repeated—1000 times simulations for each pair. Breakpoint positions were rejected if the same residue was not present in the aligned donor and acceptor RNAs. In the experimental dataset it is impossible to define exactly on which nucleotide of a breakpoint strand transfer occurred, while in the computer simulations this information is available. To compare the results obtained in the two cases, also in the computer simulations the whole breakpoint that included the selected nucleotide was considered as potential region of transfer

Sequence features of the breakpoints

A parameter that has been suggested to be important is the presence of HPS, due to their role in inducing stalling of DNA synthesis. We therefore computed how frequently the breakpoints identified included a HPS, considering for this analysis all sequences of at least three identical residues. For each breakpoint, in the experimental dataset as well as for the simulated breakpoint, the number of HPS were counted. HPS were considered as such when at least three identical and consecutive nucleotides were found. If more than one HPS was present in the breakpoint, each HPS was computed for calculation. The values in the experimental dataset were significantly higher than those for the simulated data. For HPS of at least 3 nt, 114 were found within the experimental breakpoints data set, a value significantly higher ($P < 0.05$) than those found in 1000 simulations, in which 114 HPS (of 3 nt or more) were identified in only 14 cases. The discrepancy between the two datasets was even more marked when only HPS spanning at least four residues were considered. In this case 51 HPS were found in the experimental dataset, while in all 1000 simulations values <51 were always observed ($P < 0.001$).

Concerning the base composition, the presence of regions rich in A and/or U content has been suggested to favour template switching in several RNA viruses, by facilitating the melting of the nascent DNA from the donor RNA. When we analysed this parameter, the composition of experimental and simulated breakpoints in 1000 independent randomisations resulted similar, with A/U contents of 62.3% (±8.4%) and 63.4%, respectively, suggesting that this parameter does not influence the probability of occurrence of recombination. No significant differences were observed between experimental and the simulated breakpoints even when considering other nucleotide compositions than the A/U content (data not shown).

Length of the breakpoints

The degree of sequence similarity is expected to be particularly crucial near the 3′ end of the nascent DNA, since this region guides the correct positioning of the acceptor RNA in proximity of the elongating HIV-1 DNA during the transfer process. We analysed how the context of sequence similarity in the proximity of the 3′-OH of the nascent DNA influences template switching. To address this issue, we first plotted the number of times breakpoints were identified as a function of their length in nucleotides (Figure 5). As a control, 1000 independent randomisations were run. Significant negative correlations (Pearson's product moment correlation) between the length and frequency of the selected breakpoint was found in all simulations (the average was −0.6551 ± 0.0009). The typical result of one randomisation is shown in Figure 5. This trend is due to the greater number of shorter potential breakpoint regions than longer regions. In the experimental set of data, instead, a scattered distribution was observed (P-value −0.2111), mostly due to a lower incidence of breakpoints spanning 5 or less nt, and a preference for the use of breakpoints in the 15–25 nt range. Thus, the experimental never falls into the range of the simulated data ($P < 0.001$). This indicates a strong tendency for recombination breakpoints not to be located in regions close to mismatches.

Features of the breakpoint and the preceding region

The analysis above is informative only of features of the breakpoint itself. The region preceding the breakpoint, in the sense of reverse transcription, should also influence the likelihood of template switching by stabilising the annealing

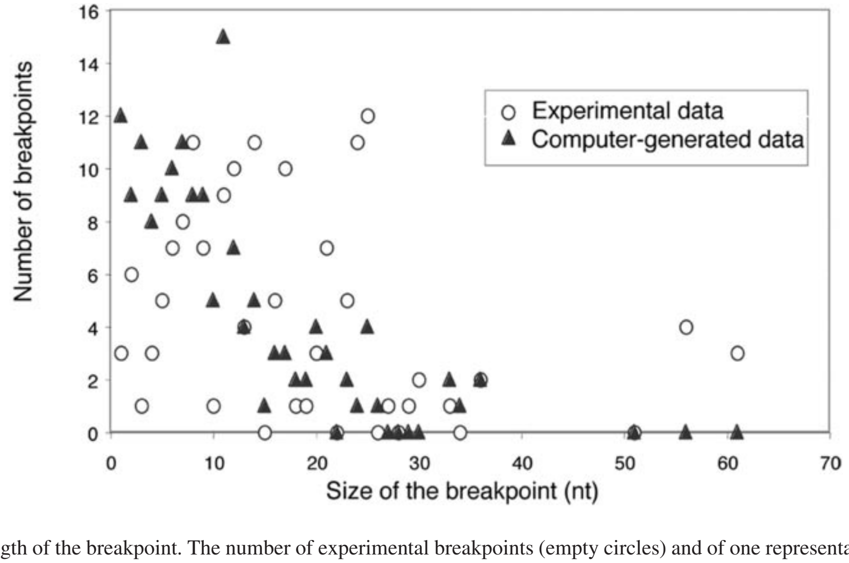

Figure 5. Copy choice and length of the breakpoint. The number of experimental breakpoints (empty circles) and of one representative set of computer-generated breakpoints (filled triangles) is plotted as function of the length of the breakpoint, in nucleotides. Results from all combinations of pairs of parental isolates were pooled.

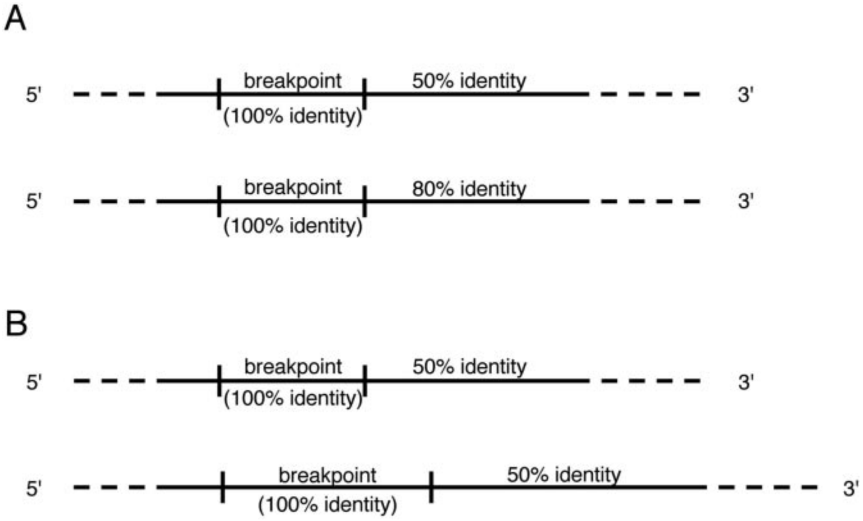

Figure 6. Rationale for the analysis of the impact of local sequence identity on recombination. (**A**) breakpoints of the same size, but presenting a region at their 3′ differing in degree of similarity. (**B**) breakpoints differing in size but presenting the same degree of similarity at their 3′.

of the nascent DNA onto the acceptor RNA. Two breakpoints of the same size, for instance, are not expected to harbour recombination at similar frequencies if the region on their 3′ presents different degrees of identity between the parental sequences (Figure 6A). On the other hand, the size of the breakpoint should also be taken into account, since two breakpoints presenting the same degree of identity at their 3′ end, but differing in size (Figure 6B) should differ in recombination frequency, being higher for the longer breakpoint region. Both, the size of the breakpoint and the degree of identity at its 3′ end must therefore be taken into account. We thereby considered for this analysis a window that starts from the 5′ end of the breakpoint. To do so, we considered a 30 nt window, this size being >90% of the potential breakpoint regions in our data (Figure 5). Simulated breakpoints were also used as control for this analysis. Since the degree of sequence divergence does not affect the distribution of the simulated data, while it is expected to do so for the experimental data, differences between the two sets of data should allow the inference of how identity influences the generation of the experimentally observed breakpoints. To perform this analysis, the numbers of observed and random breakpoints were plotted as a function of the number of discordant nt between the parental sequences in the 30 nt window (Figure 7A). A higher frequency was observed in the experimental data for zero or one discordant residue since, in a total of 162 recombinants, respectively, 14 and 48 recombinants were found in the experimental dataset, in contrast to the 2.6 and 8.6 recombinants found, on average, in simulated data. A sharp decrease was instead observed for the presence of the second discordant residue, with the values for the two sets of data that became comparable. For the presence of more discordant residues, the two curves followed a similar trend.

Recombination appeared therefore to be skewed toward the preference for breakpoints with only an extremely narrow window of discordant nucleotides (0 or 1) at their 3′ end. Recombinants occurring preferentially in these regions corresponded to 31.5% of the total as calculated by the formula (14 + 48 − 2.6 − 8.6)/162, where 14 and 48 correspond to the number of recombinants in the categories presenting no or one discordant residue, respectively, in the experimental dataset; 2.6 and 8.6 indicates those that fell in these categories in the simulated dataset, and 162 is the total number of recombinants analysed.

The window was then shifted stepwise in the 3′ direction. By doing so, we expected the differences between observed and random data to disappear when the shift was such so as to make us consider a distance from the breakpoint at which sequence identity is no longer influential. As shown in Figure 7, a peak was still observed in the experimental data, for a 5 nt shift (Figure 7B) and, even if still present, it decreased for a 10 nt shift (Figure 7C). In addition, in the latter case, the peak was shifted slightly on the right of the graph indicating than for the region from 11 to 40 nt away from the 5′ end of the breakpoint the presence of up to two discordant residues does not have such dramatic effects as it was the case for the 1–30 nt window (Figure 7A). Also in these cases, the curves for the experimental data followed a trend similar to that of the simulated data (Figure 7B and C). For further shifts, the peak of the experimental data faded progressively (Figure 7D) and, eventually, disappeared (Figure 7E and F).

DISCUSSION

We have characterised recombinants between HIV-1 isolates belonging to subtypes A and D, such recombinants being frequently encountered in natural infections, particularly in East Africa. Recombination in the C1–C4 region of the *gp120* gene was studied, using two tissue culture systems, one where infection is limited to a single cycle and the other where multiple cycles follow an initial dual infection.

The frequency of recombination in the single cycle assay varied from 4 to 10%, for a sequence spanning ∼1100 nt. The extrapolation of these frequencies to the whole genome suggests the occurrence of approximately one template

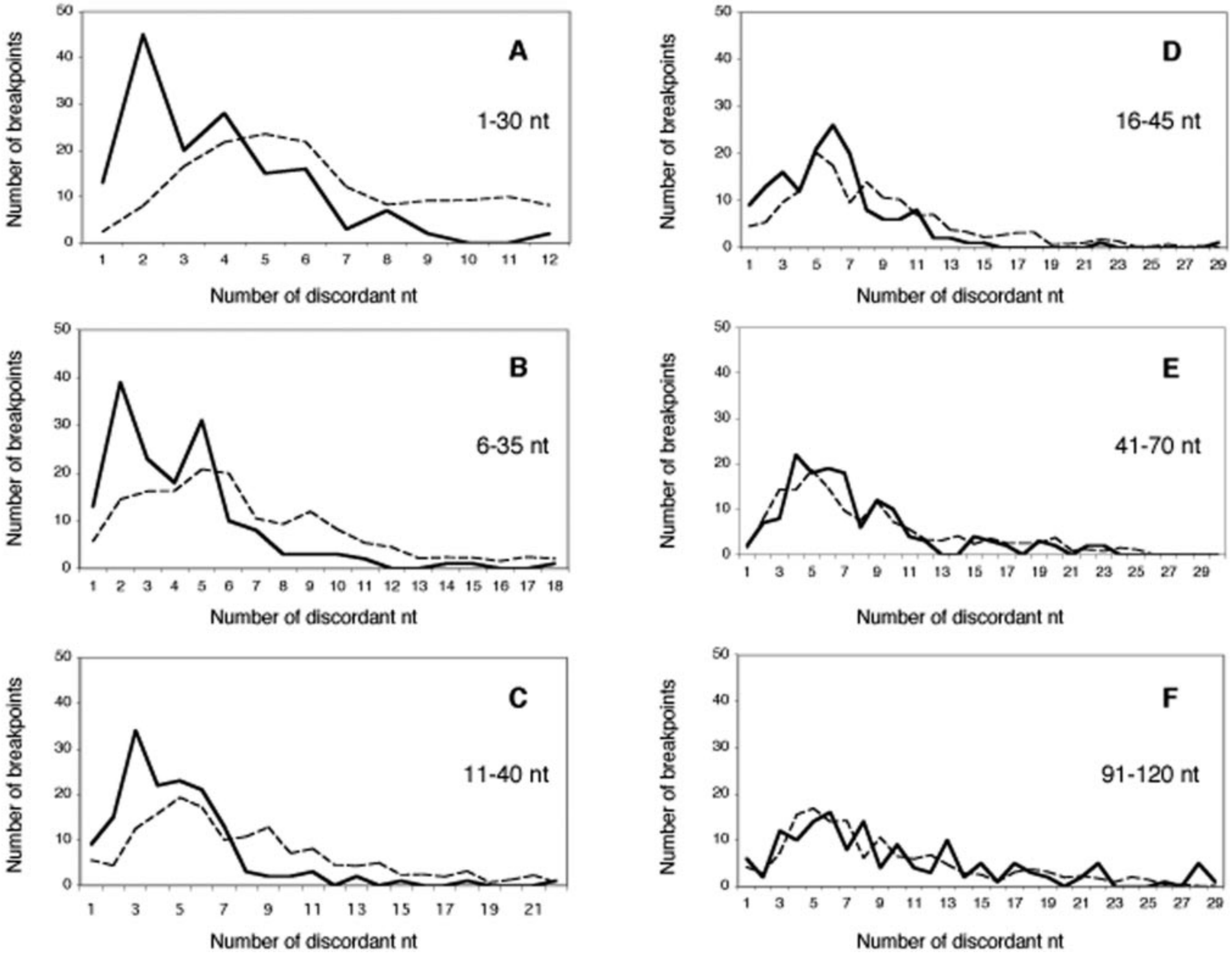

Figure 7. Sequence identity in a 30 nt window on the 3′ of the 5′ border of breakpoints. The frequency of experimentally observed (solid line) or computer-generated breakpoints (dotted line) is plotted as a function of the number of nucleotides that, in a 30 nt window on the 3′ of the 5′ border of the breakpoint, are different between the parental sequences (**A**). (B–F) same analysis as in (A), but shifting the window in the 3′ direction by 5 (**B**), 10 (**C**), 25 (**D**), 40 (**E**), or 90 (**F**) nt.

switching per infection cycle, in a heterozygous virion, a rate of recombination lower than those reported on the whole genome between highly related subtype B strains (20,39,40). This could be due to the use of more divergent isolates in the present study, and to features of the portion of genome studied that, containing stretches of highly divergent sequences, decreases the probability of recombination. In fact, breakpoints clustered in regions coding for the conserved domains of the *gp120* gene (Figure 2).

After multiple cycles of infection with the 126D/120A pair, the C1 and C2 recombination hotspots found after a single cycle of infection with these strains were still present, but the emergence of V2 recombinants represents a significant enrichment with respect to those generated by the copy-choice mechanism in a single round. Replication competent recombinants with breakpoints in V2 may be selected due to the plasticity of V1/V2 loops in the encoded gp120 and the ability of Env to accommodate significant genetic change and still retain host cell entry functions. Regions C1 and C2 code for portions of the protein that undergo important conformational changes upon binding to the CD4 receptor (64) and may therefore be more structurally constrained than the hypervariable domains. It could be possible that the subtle shifts in the positions of the breakpoints in these regions observed from the single to the multiple cycle system (Figure 3) reflect a 'fine-tuning' for the generation of replication competent, recombinant gp120 molecules. An analysis of recombinant isolates in the database indicates that the proportion of breakpoints found after a single infection cycle approximates that of the circulating recombinants, apart for V1/V2 and C2 regions, for which a different proportion of breakpoints was found in the database (Figure 8). This observation is potentially indicative of selection having an influence on the outcome of recombination in these two regions.

Although intensive efforts have been made to understand the triggers for recombination using closely related sequences, no systematic search for distinctive sequence features that govern HIV recombination has been led to date using natural isolates. This is an important issue, though, since recombinants involving distantly related strains not

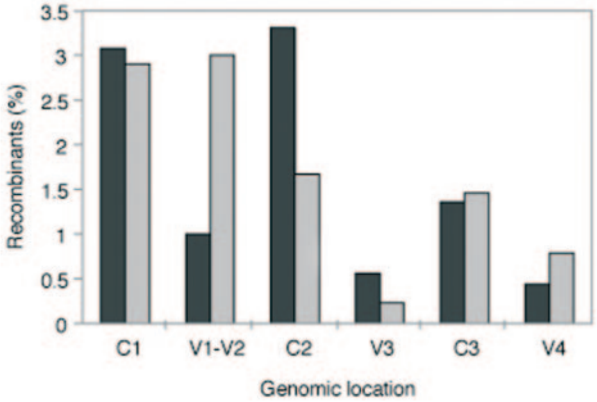

Figure 8. Comparison of genomic location of breakpoints in the experimental dataset and in recombinant forms available from the HIV sequence database. The percentages of recombinants in the various constant and variable regions found in the present work (data cumulated for all pairs) and in all recombinants in the database are plotted as dark and light grey bars, respectively. There were 116 intersubtype recombinants in the envelope region in the database, including 16 CRFs. Note, each CRF was counted only once. Of these there were 57 recombinants, with a total of 90 non-identical breakpoints in the region corresponding to the experimentally generated recombinants. As only a small portion of the C4 region (~20% of its total length) was included in the experimental data, comparison with the database recombinants was not carried out.

only are frequently isolated in patients but also very probably have a major impact on HIV evolution. By comparing experimental to simulated breakpoints we analysed three main parameters that, based on sequence analysis, permitted insights into the probability that a genomic region will undergo recombination during reverse transcription. A major difficulty in such an analysis is constituted by the intrinsic impossibility in determining the exact position of transfer within each breakpoint. Here we try to overcome this problem by considering two parameters in parallel. First we analyse the breakpoint itself, and then we consider a portion that, apart from the breakpoint, also includes part of the region preceding it, in the sense of reverse transcription.

The comparison between experimental and simulated data permits insights into the influence of sequence divergence on recombination, paradoxically underlining and minimising its importance. We observed that breakpoints presenting zero or one discordant residue, between donor and acceptor RNA in a region spanning 30 nt from the 5′ of the breakpoint, are significantly more abundant in the experimental dataset than in the simulated data. An abrupt decrease in the frequency of recombinants is then observed for the presence of more than one discordant residue, with the probability of recombination dropping to that expected by a random distribution, as that of the simulated dataset. By using a sliding window approach, we could define that discrimination for highly similar sequences exists for regions up to 40 nt upstream (in the sense of reverse transcription) the 5′ border of the breakpoint (Figure 7C), with up to two discordant residues tolerated in the 40 nt region. Altogether, these observations underscore the importance of sequence identity, and define strict requirements for the preferential occurrence of copy choice in regions of the genome of natural isolates.

How can these results be accounted at the mechanistic level? It can be conceived that, when a large region of high similarity is available, the nascent DNA initially anneals to the acceptor RNA from the 5′ side of the DNA to progressively propagate then towards the 3′ of the nascent DNA through a mechanism sensitive to the presence of even only a few discordant residues. Sequence identity would only be a prerequisite for this mechanism, which would be modulated, in each specific case, by other parameters. The structure of the genomic RNA could be such a parameter, as previously suggested. Indeed, by using identical donor and acceptor RNAs, with the exception of a few base substitutions introduced for mapping purposes, we identified a recombination hot spot in infected cells in a 18 nt long region of the C2 portion of the *gp120* gene of the LAI isolate. The presence of a RNA hairpin in that region was required for the existence of the hot spot, and template switching was proposed to occur following a branch migration mechanism involving a double stranded region on the acceptor RNA. Interestingly, the introduction of a single discordant nucleotide was sufficient to abolish the occurrence of preferential recombination in that region (65), and discordant residues are known to inhibit dramatically branch migration (66).

At the same time, a surprising observation, downplaying the importance of sequence identity, is that these sequence identity requirements are fulfilled for the generation of only a minority of the recombinant molecules, no deviation from the distribution of the simulated data being observed for ~70% of the recombinants (as deduced in the section ‘Features of the breakpoint and the preceding region’). Noteworthy, the only constraint applied to the selection of breakpoints in the computer simulation was to exclude transfer of the 3′ end of the nascent DNA on a mismatched nucleotide between nascent DNA and acceptor RNA. It appears therefore that, if the extent of sequence similarity in the region preceding the breakpoint is not extremely high, template switching is closer to random, where the crucial parameter becomes the probability that the nascent DNA, guided by its overall complementarity with acceptor RNA, transfers its 3′ end on a position where the utmost 3′ residue is complementary to the corresponding one on the acceptor RNA. Reverse transcription would therefore follow a mode where the 3′ part of the nascent DNA is frequently transferred on the acceptor RNA and, if a matching residue is encountered, synthesis continues on this molecule.

Another parameter generally assumed to be important for recombination is the presence of HPS in the region of interest. These sequence motifs have been shown to be associated to pausing of DNA synthesis, a parameter suggested to be important for promoting template switching, based on observations in reconstituted *in vitro* assays. We establish here, for the first time, a direct correlation between frequency of recombination and the presence of HPS within the breakpoint in a cell culture-based experimental system. The correlation is stronger when considering stretches spanning at least four residues. Whether the correlation between recombination and HPS implies that pausing is the mechanisms promoting template switching in these sequences remains to be determined. For instance, runs of G or C residues were found to be more effective in inducing stalling of synthesis *in vitro*, than A or U runs (67), while here only a mild bias for the presence of polyC stretches in recombination regions was observed (data not shown). Along these lines, in a previous work in which we dissected a hot spot in a hairpin structure (65), the hot region contained a HPS constituted by four U residues, and mutating this motif, did not significantly affect the pausing pattern in the region, while it decreased the frequency of recombination by more than 2-fold. It is therefore possible that other parameters than pausing, as the efficiency of annealing of complementary nucleic acids, contribute to preferential recombination in HPS-containing regions.

In conclusion, this study defines the general rules that determine the probability of copy choice to occur between divergent natural strains of HIV-1, opening the way to mechanistic studies on this aspect of the mating of genetically divergent strains.

ACKNOWLEDGEMENTS

Research for this study was performed at Institut Pasteur, Paris, France and at Case Western Reserve University, Cleveland, Ohio. M.N. was supported by grant 02172 from the Agence Nationale de Recherches sur le SIDA (ANRS), and grant 51005-02-00/AO16-2 from SIDACTION. E.J.A. was supported by NIAID, NIH grants AI49170, AI57005, and AI43645-02. R.G. was supported by a fellowship from SIDACTION. Single cycle infection assays were carried out in biosafety level 2 facilities at Institut Pasteur. Work with replication-competent viruses was performed in the biosafety level 2 and 3 facilities of the CWRU Center for AIDS Research (AI25879). J.A. is supported by a BBSRC studentship and J.F. by a University of Manchester OSS award. Funding to pay the Open Access publication charges for this article was provided by Sidaction, grant 51005-02-00/AO16-2 to M.N.

Conflict of interest statement. None declared.

REFERENCES

1. Wei,X., Ghosh,S.K., Taylor,M.E., Johnson,V.A., Emini,E.A., Deutsch,P., Lifson,J.D., Bonhoeffer,S., Nowak,M.A., Hahn,B.H. *et al.* (1995) Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, **373**, 117–122.
2. Ho,D.D., Neumann,A.U., Perelson,A.S., Chen,W., Leonard,J.M. and Markowitz,M. (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, **373**, 123–126.
3. Perelson,A.S., Neumann,A.U., Markowitz,M., Leonard,J.M. and Ho,D.D. (1996) HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. *Science*, **271**, 1582–1586.
4. Preston,B.D., Poiesz,B.J. and Loeb,L.A. (1988) Fidelity of HIV-1 reverse transcriptase. *Science*, **242**, 1168–1171.
5. Leitner,T., Escanilla,D., Marquina,S., Wahlberg,J., Brostrom,C., Hansson,H.B., 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.
6. Sabino,E.C., Shpaer,E.G., Morgado,M.G., Korber,B.T., Diaz,R.S., Bongertz,V., Cavalcante,S., Galvao-Castro,B., Mullins,J.I. and Mayer,A. (1994) Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. *J. Virol.*, **68**, 6340–6346.
7. Shriner,D., Rodrigo,A.G., Nickle,D.C. and Mullins,J.I. (2004) Pervasive genomic recombination of HIV-1 *in vivo*. *Genetics*, **167**, 1573–1583.
8. Charpentier,C., Nora,T., Tenaillon,O., Clavel,F. and Hance,A.J. (2006) Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J. Virol.*, **80**, 2472–2482.
9. Robertson,D.L., Sharp,P.M., McCutchan,F.E. and Hahn,B.H. (1995) Recombination in HIV-1. *Nature*, **374**, 124–126.
10. Peeters,M., Liegeois,F., Torimiro,N., Bourgeois,A., Mpoudi,E., Vergne,L., Saman,E., Delaporte,E. and Saragosti,S. (1999) Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. *J. Virol.*, **73**, 7368–7375.
11. Peeters,M. and Sharp,P.M. (2000) Genetic diversity of HIV-1: the moving target. *Aids*, **14**, S129–S140.
12. Osmanov,S., Pattou,C., Walker,N., Schwardlander,B. and Esparza,J. (2002) Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. *J. Acquir. Immune Defic. Syndr.*, **29**, 184–190.
13. Peeters,M., Toure-Kane,C. and Nkengasong,J.N. (2003) Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. *Aids*, **17**, 2547–2560.
14. Harris,M.E., Serwadda,D., Sewankambo,N., Kim,B., Kigozi,G., Kiwanuka,N., Phillips,J.B., Wabwire,F., Meehen,M., Lutalo,T. *et al.* (2002) Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate. *AIDS Res. Hum. Retroviruses*, **18**, 1281–1290.
15. An,W. and Telesnitsky,A. (2002) HIV-1 genetic recombination: experimental approaches and observations. *AIDS Rev.*, **4**, 195–212.
16. Negroni,M. and Buc,H. (2001) Mechanisms of retroviral recombination. *Annu Rev. Genet.*, **35**, 275–302.
17. Anderson,J.A., Bowman,E.H. and Hu,W.S. (1998) Retroviral recombination rates do not increase linearly with marker distance and are limited by the size of the recombining subpopulation. *J. Virol.*, **72**, 1195–1202.
18. Yu,H., Jetzt,A.E., Ron,Y., Preston,B.D. and Dougherty,J.P. (1998) The nature of human immunodeficiency virus type 1 strand transfers. *J. Biol. Chem.*, **273**, 28384–28391.
19. Zhang,J., Tang,L.Y., Li,T., Ma,Y. and Sapp,C.M. (2000) Most retroviral recombinations occur during minus-strand DNA synthesis. *J. Virol*, **74**, 2313–2322.
20. Jetzt,A.E., Yu,H., Klarmann,G.J., Ron,Y., Preston,B.D. and Dougherty,J.P. (2000) High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J. Virol*, **74**, 1234–1240.
21. Boone,L.R. and Skalka,A.M. (1993) In Skalka,A.M. and Goff,S.P. (eds), *Reverse Transcriptase*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 119–133.
22. Hu,W.S., Pathak,V.K. and Temin,H.M. (1993) In Skalka,A.M. and Goff,S.P. (eds), *Reverse Transcriptase*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 251–274.
23. Magiorkinis,G., Paraskevis,D., Vandamme,A.M., Magiorkinis,E., Sypsa,V. and Hatzakis,A. (2003) *In vivo* characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity. *J. Gen. Virol.*, **84**, 2715–2722.
24. Vogt,P.K. (1973) In Silvestri,L. (ed.), *The Meeting 'Possible Episomes in Eukaryotes'*. North Holland Publishing Company, Amsterdam, Holland, pp. 35–41.
25. Oyama,F., Kikuchi,R., Crouch,R.J. and Uchida,T. (1989) Intrinsic properties of reverse transcriptase in reverse transcription. Associated RNase H is essentially regarded as an endonuclease. *J. Biol. Chem.*, **264**, 18808–18817.
26. Krug,M.S. and Berger,S.L. (1989) Ribonuclease H activities associated with viral reverse transcriptases are endonucleases. *Proc. Natl Acad. Sci. USA*, **86**, 3539–3543.
27. Kim,J.K., Palaniappan,C., Wu,W., Fay,P.J. and Bambara,R.A. (1997) Evidence for a unique mechanism of strand transfer from the transactivation response region of HIV-1. *J. Biol. Chem.*, **272**, 16769–16777.
28. Negroni,M. and Buc,H. (2000) Copy-choice recombination by reverse transcriptases: reshuffling of genetic markers mediated by RNA chaperones. *Proc. Natl Acad. Sci. USA*, **97**, 6385–6390.
29. Moumen,A., Polomack,L., Roques,B., Buc,H. and Negroni,M. (2001) The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucleic Acids Res.*, **29**, 3814–3821.

30. Balakrishnan,M., Fay,P.J. and Bambara,R.A. (2001) The kissing hairpin sequence promotes recombination within the HIV-I 5′ leader region. *J. Biol. Chem.*, **276**, 36482–36492.
31. Moumen,A., Polomack,L., Unge,T., Veron,M., Buc,H. and Negroni,M. (2003) Evidence for a mechanism of recombination during reverse transcription dependent on the structure of the acceptor RNA. *J. Biol. Chem.*, **278**, 15973–15982.
32. Derebail,S.S. and DeStefano,J.J. (2004) Mechanistic analysis of pause site-dependent and -independent recombinogenic strand transfer from structurally diverse regions of the HIV genome. *J. Biol. Chem.*, **279**, 47446–47454.
33. DeStefano,J.J., Mallaber,L.M., Rodriguez-Rodriguez,L., Fay,P.J. and Bambara,R.A. (1992) Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase. *J. Virol.*, **66**, 6370–6378.
34. Wu,W., Blumberg,B.M., Fay,P.J. and Bambara,R.A. (1995) Strand transfer mediated by human immunodeficiency virus reverse transcriptase *in vitro* is promoted by pausing and results in misincorporation. *J. Biol. Chem.*, **270**, 325–332.
35. Roda,R.H., Balakrishnan,M., Kim,J.K., Roques,B.P., Fay,P.J. and Bambara,R.A. (2002) Strand transfer occurs in retroviruses by a pause-initiated two-step mechanism. *J. Biol. Chem.*, **277**, 46900–46911.
36. Lanciault,C. and Champoux,J.J. (2006) Pausing during reverse transcription increases the rate of retroviral recombination. *J. Virol.*, **80**, 2483–2494.
37. Galetto,R. and Negroni,M. (2005) Mechanistic features of recombination in HIV. *AIDS Rev.*, **7**, 92–102.
38. Zhuang,J., Jetzt,A.E., Sun,G., Yu,H., Klarmann,G., Ron,Y., Preston,B.D. and Dougherty,J.P. (2002) Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J. Virol.*, **76**, 11273–11282.
39. Galetto,R., Moumen,A., Giacomoni,V., Veron,M., Charneau,P. and Negroni,M. (2004) The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot *in vivo*. *J. Biol. Chem.*, **279**, 36625–36632.
40. Levy,D.N., Aldrovandi,G.M., Kutsch,O. and Shaw,G.M. (2004) Dynamics of HIV-1 recombination in its natural target cells. *Proc. Natl Acad. Sci. USA*, **101**, 4204–4209.
41. Quinones-Mateu,M.E., Gao,Y., Ball,S.C., Marozsan,A.J., Abraha,A. and Arts,E.J. (2002) *In vitro* intersubtype recombinants of human immunodeficiency virus type 1: comparison to recent and circulating *in vivo* recombinant forms. *J. Virol.*, **76**, 9600–9613.
42. Gao,Y., Paxinos,E., Galovich,J., Troyer,R., Baird,H., Abreha,M., Kityo,C., Mugyenyi,P., Petropoulos,C. and Arts,E.J. (2004) Characterization of a subtype D human immunodeficiency virus type 1 isolate that was obtained from an untreated individual and that is highly resistant to non-nucleoside reverse transcriptase inhibitors. *J. Virol.*, **78**, 5390–5401.
43. Quinones-Mateu,M.E., Ball,S.C., Marozsan,A.J., Torre,V.S., Albright,J.L., Vanham,G., van Der Groen,G., Colebunders,R.L. and Arts,E.J. (2000) A dual infection/competition assay shows a correlation between *ex vivo* human immunodeficiency virus type 1 fitness and disease progression. *J. Virol.*, **74**, 9222–9233.
44. Reed,L.J. and Muench,H. (1938) A simple method of estimating fifty percent endpoints. *Am. J. Hyg.*, **27**, 493–497.
45. Naldini,L., Blomer,U., Gallay,P., Ory,D., Mulligan,R., Gage,F.H., Verma,I.M. and Trono,D. (1996) *In vivo* gene delivery and stable transduction of non-dividing cells by a lentiviral vector. *Science*, **272**, 263–267.
46. Yee,J.K., Miyanohara,A., LaPorte,P., Bouic,K., Burns,J.C. and Friedmann,T. (1994) A general method for the generation of high-titer, pantropic retroviral vectors: highly efficient infection of primary hepatocytes. *Proc. Natl Acad. Sci. USA*, **91**, 9564–9568.
47. Hirt,B. (1967) Selective extraction of polyoma DNA from infected mouse cell cultures. *J. Mol. Biol*, **26**, 365–369.
48. Zennou,V., Petit,C., Guetard,D., Nerhbass,U., Montagnier,L. and Charneau,P. (2000) HIV-1 genome nuclear import is mediated by a central DNA flap. *Cell*, **101**, 173–185.
49. Torre,V.S., Marozsan,A.J., Albright,J.L., Collins,K.R., Hartley,O., Offord,R.E., Quinones-Mateu,M.E. and Arts,E.J. (2000) Variable sensitivity of CCR5-tropic human immunodeficiency virus type 1 isolates to inhibition by RANTES analogs. *J. Virol.*, **74**, 4868–4876.
50. Bradley,R.D. and Hillis,D.M. (1997) Recombinant DNA sequences generated by PCR amplification. *Mol. Biol. Evol.*, **14**, 592–593.
51. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
52. Robertson,D.L., Hahn,B.H. and Sharp,P.M. (1995) Recombination in AIDS viruses. *J. Mol. Evol.*, **40**, 249–259.
53. van Cuyck,H., Fan,J., Robertson,D.L. and Roques,P. (2005) Evidence of recombination between divergent hepatitis E viruses. *J. Virol.*, **79**, 9306–9314.
54. Lole,K.S., Bollinger,R.C., Paranjape,R.S., Gadkari,D., Kulkarni,S.S., Novak,N.G., Ingersoll,R., Sheppard,H.W. and Ray,S.C. (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.*, **73**, 152–160.
55. White,R.L. and Fox,M.S. (1974) On the molecular basis of high negative interference. *Proc. Natl Acad. Sci. U S A*, **71**, 1544–1548.
56. Hu,W.S., Bowman,E.H., Delviks,K.A. and Pathak,V.K. (1997) Homologous recombination occurs in a distinct retroviral subpopulation and exhibits high negative interference. *J. Virol.*, **71**, 6028–6036.
57. Rhodes,T.D., Nikolaitchik,O., Chen,J., Powell,D. and Hu,W.S. (2005) Genetic recombination of human immunodeficiency virus type 1 in one round of viral replication: effects of genetic distance, target cells, accessory genes, and lack of high negative interference in crossover events. *J. Virol.*, **79**, 1666–1677.
58. Zhang,J. and Temin,H.M. (1994) Retrovirus recombination depends on the length of sequence identity and is not error prone. *J. Virol.*, **68**, 2409–2414.
59. An,W. and Telesnitsky,A. (2002) Effects of varying sequence similarity on the frequency of repeat deletion during reverse transcription of a human immunodeficiency virus type 1 vector. *J. Virol.*, **76**, 7897–7902.
60. Nagy,P.D. and Bujarski,J.J. (1996) Homologous RNA recombination in brome mosaic virus: AU-rich sequences decrease the accuracy of crossovers. *J. Virol.*, **70**, 415–426.
61. Nagy,P.D. and Bujarski,J.J. (1997) Engineering of homologous recombination hotspots with AU-rich sequences in brome mosaic virus. *J. Virol.*, **71**, 3799–3810.
62. Nagy,P.D. and Bujarski,J.J. (1998) Silencing homologous RNA recombination hot spots with GC-rich sequences in brome mosaic virus. *J. Virol.*, **72**, 1122–1130.
63. Nagy,P.D., Ogiela,C. and Bujarski,J.J. (1999) Mapping sequences active in homologous RNA recombination in brome mosaic virus: prediction of recombination hot spots. *Virology*, **254**, 92–104.
64. Chen,B., Vogan,E.M., Gong,H., Skehel,J.J., Wiley,D.C. and Harrison,S.C. (2005) Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature*, **433**, 834–841.
65. Galetto,R., Giacomoni,V., Veron,M. and Negroni,M. (2006) Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. *J. Biol. Chem.*, **281**, 2711–2720.
66. Panyutin,I.G. and Hsieh,P. (1994) The kinetics of spontaneous DNA branch migration. *Proc. Natl Acad. Sci. USA*, **91**, 2021–2025.
67. Klarmann,G.J., Schauber,C.A. and Preston,B.D. (1993) Template-directed pausing of DNA synthesis by HIV-1 reverse transcriptase during polymerization of HIV-1 sequences *in vitro* [published erratum appears in *J Biol Chem* 1993 Jun 25;**268**(18):13764]. *J. Biol. Chem.*, **268**, 9793–9802.
68. Korber,B., Foley,B.T., Kuiken,C., Pillai,S.K. and Sodroski,J.G. (1998) In Korber,B., C.L.,K., Foley,B., Hahn,B., McCutchan,F., Mellors,J.W. and Sodroski,J. (eds), *Human Retroviruses and AIDS 1998*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, pp. III-102–111.