JVI Accepts, published online ahead of print on 22 January 2014
J. Virol. doi:10.1128/JVI.03136-13

Title:

15-20% of HIV substitution mutations are associated with recombination

Running title: HIV mutation and recombination

Timothy E. Schlub[1,2a], Andrew J. Grimm[2a], Redmond P. Smyth[3,4,5a], Deborah Cromer[6], Abha Chopra[7], Simon Mallal[7], Vanessa Venturi[6], Caryll Waugh[8,9], Johnson Mak[3,4,8,9*], Miles P. Davenport[2*]

Author Affiliations:

[1]School of Public Health, Sydney University, Sydney, New South Wales, Australia

[2]Complex Systems in Biology Group, Centre for Vascular Research, University of New South Wales, Sydney, NSW, Australia

[3]Centre for Virology, Burnet Institute, 85 Commercial Road, Melbourne, Victoria, 3004, Australia

[4]Department of Biochemistry and Molecular Biology, Department of Microbiology, Monash University, Clayton, Victoria, 3800, Australia

[5]Architecture et Réactivité de l'ARN, Université de Strasbourg, CNRS, IBMC, 15 rue René Descartes, 67084 Strasbourg, France

[6]Computational Biology Unit, Centre for Vascular Research, University of New South Wales, Sydney, Kensington, NSW, 2052, Australia

[7]Centre for Clinical Immunology and Biomedical Statistics, Royal Perth Hospital and Murdoch University, Perth, 6000, Australia

[8]School of Medicine, Faculty of Health, Deakin University, Geelong, Victoria, 3220, Australia

[9]Commonwealth Scientific and Industrial Research Organization, , Australian Animal Health Laboratory, Geelong, 3220, Australia.

[a] authors have contributed equally

[*]Corresponding authors

Corresponding Authors: MD: Phone: +61 2 9385 2762 Fax: +61 2 93851797 Email: m.davenport@unsw.edu.au; JM: Phone +61-3-5227-5062 Fax +61-3-5227-5555 Email: johnson.mak@deakin.edu.au

Abstract:

HIV undergoes a high rate of mutation and recombination during reverse transcription, but it is not known whether these events occur independently or are linked mechanistically. Here we use a system of silent marker mutations in HIV and a single round of infection in primary T-lymphocytes, combined with a high-throughput sequencing and mathematical modelling approach to directly estimate the viral recombination and mutation rates. From >7 million nt of sequences from HIV infection, we observe 4801 recombination events and 859 substitution mutations (≈1.51 and 0.12 events per 1000 nt respectively). We use experimental controls to account for PCR-induced and transfection-induced recombination and sequencing error. We find the single cycle virus-induced mutation rate is 4.6×10^{-5} mutations per nt after correction. By sorting our data into recombined and non-recombined sequences, we find a significantly higher mutation rate in recombined regions ($p=0.003$, Fisher's exact). We use a permutation approach to eliminate a number of potential confounding factors and confirm that mutation occurs around the site of recombination, and is not simply co-located in the genome. By comparing mutation rates in recombined and non-recombined regions we find that recombination-associated mutations account for 15-20% of all mutations occurring during reverse transcription.

Introduction:

The development of a genetically diverse quasispecies is one of the hallmarks of HIV infection. Genetic diversity underpins the ability of HIV to evolve resistance to antiretroviral therapy and to successfully evade the immune system. Viral diversity increases rapidly during initial infection (1) and is driven by mutation, recombination and population size. Mutations are primarily introduced during reverse transcription by the error prone reverse transcriptase (RT) enzyme (2) or by host cellular defense mechanisms (3, 4). These mutations can then be shuffled within the viral quasispecies by retroviral recombination, which relies on the co-packaging of two genetically distinct RNA genomes (for review see (5-9)). While the dimer initiation sequence (DIS) is not critical for HIV replication in primary cells (10-12), DIS is important to facilitate the recombination process, presumably by bringing genomic RNA sequences into close proximity during virion assembly and viral cDNA synthesis (13-17). Retroviral recombination occurs during reverse transcription when the RT enzyme switches between strands of the two co-packaged RNA genomes to produce a chimeric DNA molecule (5-7, 18).

Whether retroviral mutation is linked with recombination has been a subject of debate for decades, and numerous studies have measured HIV mutation and retroviral recombination rates *in vitro* and in cell culture (2, 19-31). Nevertheless, the question of 'whether mutation and recombination are associated' has not been resolved directly due to numerous inherent technical difficulties. To assess retroviral recombination, many studies utilize PCR to amplify and to clone out the engineered reporter genes within the retroviral vectors (or inter-subtype HIV constructs) from the infected cells,

76 which is followed by expression of cloned sequences in bacteria to indirectly assess
77 the recombination and mutation rate (32-34). Others have taken advantage of
78 retroviral vector or reporter systems, in which the expressions of reporter genes (such
79 as fluorescent proteins, cell surface markers, and/or antibiotic resistance genes) were
80 used to select cells containing recombined viral genomes for recombination and
81 mutation analyses (28-31, 35-39). As co-expression of reporter gene products in a
82 single cell is needed to identify recombination events; these systems generally use low
83 multiplicity of infections (MOIs) to avoid dual infection introducing experimental
84 bias (28-39). However, as these systems require expression of the gene products,
85 natural mutation-events or recombination-events that lead to non-productive
86 expression of reporter genes will be unaccounted for. It would have been
87 advantageous if recombination- and mutation-rates could be estimated directly from
88 the newly synthesized and/or integrated viral cDNA that is independent of MOI.
89 Furthermore, as these studies depend on PCR to amplify specific genomic sequences
90 for recombination and/or mutation analyses (28-30), the potential of PCR induced
91 recombination and mutation must be controlled for to limit over-estimation of these
92 two retroviral biological processes (40).
93
94 Using a transfection approach to generate heterozygous HIV to assess retroviral
95 recombination, we have previously shown that HIV recombination does not randomly
96 occur throughout the viral genome (20, 40). Based on the observed recombination
97 events in HIV *gag*, we have found that HIV-1 undergoes 1.35×10^{-3} recombination
98 events per nucleotide per replication cycle (20). It is estimated that HIV-1 has a
99 mutation rate that produces approximately 0.34 mutations per replication cycle (2, 19,

40, 41). Consequently, mutation and recombination associate relatively infrequently; and a large amount of sequencing data across the HIV genome is needed to determine whether recombination and mutations are linked. While reporter systems that can generate large quantities of data have been developed to measure either mutation rates or recombination rates on non-viral sequence (30, 35, 36, 38, 42-45), these systems, however, cannot simultaneously measure both recombination and mutation rates. Furthermore, these systems measure rates within foreign non-viral gene sequences. As *in vitro* studies have shown that template sequence and structure are important determinants of these processes (46, 47), the most accurate measurements will be derived from direct sequencing of the viral genome (20). However, before the advent of next-generation sequencing technology, generating the large amounts of sequencing data required to answer this question was not feasible.

We have previously described a method to quantify the rate of recombination by direct sequencing of the HIV-1 genome within infected primary T-lymphocytes (20, 40). This system addresses several issues that can otherwise bias the analysis of mutation and recombination rates using conventional approaches. Firstly, although recombination is most easily measured using highly divergent strains of HIV (as recombination can only be observed through the mixing of genetic marker points), the rate of recombination is greatly affected by the overall homology between genomes and processivity of RT (15, 16). Thus, recombination rate measurements using highly divergent HIV strains do not reflect the recombination occurring between highly related members of a viral quasispecies found within an infected individual (48). Our system addresses this issue by measuring recombination between two closely related

HIV-1 genomes that differ by silent mutations found in naturally occurring HIV sequences, and these HIV display identical replication kinetics in primary T-lymphocytes (20). Secondly, by sequencing viral cDNAs that are produced within 24 hours of infection (with fusion inhibitor T20 supplied 6 hours post-infection to block secondary infection), all HIV RT mutations and recombination that may lead to non-productive infection will also be accounted for. Our previous work also shows that low MOI is not vital for this type of recombination analysis and does not yield inter-virion recombination or homologous recombination of plasmid DNA from transfection that may bias data interpretation (20). Thirdly, we have developed experimental and analytical tools to account for artefacts that can compromise analysis. For example, the impact of multiple recombination events between two marker points (20), recombination during co-transfection of plasmids and PCR induced recombination (40) must be all minimised and controlled for in any analysis. Similarly, any analysis must take into account the possibility that coincident high or low rates of mutation and recombination (without the rate of mutation being specifically increased at sites of recombination) can lead to a spurious association.

In this study, we have utilized a system of marker sites introduced into HIV-1 to infect primary T-lymphocytes, followed by high-throughput sequencing. We have found that the previous estimate of nucleotide substitution rate is closely matched with our calculated 4.6×10^{-5} error rate using direct sequencing, and observe a significantly higher mutation rate in recombined regions. We eliminate a number of potential causes for this association to demonstrate a direct association between the

process of recombination and mutation. We show that 15-20% of the total mutations are associated with recombination.

150 **Materials and Methods:**

151

152 **Molecular Clones**

153 The wild-type (WT) HIV-1 plasmid used was pDRNL(AD8) (49). The marker (MK)
154 plasmid (pDRNL(AD8)GagPolMarker plasmid) was created by introducing 15 and 35
155 genetic marker points by silent transition mutation in *gag* and *pol* of pDRNL(AD8),
156 respectively. pDNRL(AD8) is an R5-tropic strain of HIV, and thus infects the
157 activated/memory T-cell lymphocyte subset (50). This created a total of 46 intervals
158 spaced an average of 53.5 nucleotides apart (range 17-155, median 47 nucleotides)
159 where recombination and mutation could be simultaneously measured. The marker
160 sites were chosen on the basis that they were (i) silent mutations in the third position,
161 (ii) located in adjacent codons, (iii) A=>G mutations observed in the HIV sequence
162 database (HIV sequence compendium; www.hiv.lanl.gov). The only exceptions were
163 four markers in non-adjacent codons, and two markers with silent mutations in the
164 first position, and non-A=>G mutations [markers PR:$L_{23}L_{24}D_{25}$, and
165 RT:$R_{451}G_{452}R_{453}$), which introduced XhoI and NotI restriction sites into the genomes,
166 respectively. Two mutations in adjacent codons were used so that recombination
167 could be distinguished from mutations introduced during the PCR or sequencing
168 reactions. We reasoned that A to G mutations would have least impact on RNA
169 structure, as U can base pair with both A and G in RNA. Consequently, the
170 introduction of A to G marker points does not grossly disrupt base pairing within
171 RNA structure regions. Nevertheless, we did not modify regions of the HIV-1 genome
172 with known functional secondary structure and we preferentially modified regions of
173 the genome containing these mutations as natural polymorphisms (HIV sequence

compendium; www.hiv.lanl.gov). MK plasmid was constructed using standard molecular cloning techniques from two regions corresponding to *gag* and *pol* that were designed electronically and chemically synthesised. The 14-day replication kinetics of marker virus was indistinguishable from WT virus, and no protein or known RNA sequence elements were changed.

Viruses

Homozygous virus was produced by transfection of 293T cells with either WT or MK pDRNLAD8. In general, a total of 3 microgram of plasmid DNA containing proviral HIV genomes was used to transfect 293T cells for the production of HIV particles. Heterozygous virus was produced by co-transfection of equal masses of wild-type WT and MK pDRNLAD8 into 293T cells. When equal amounts of two HIV plasmids are co-transfected, co-packaging of RNA into virions is random (13). Consequently, we expect 50% heterozygous virions, 25% homozygous WT virions and 25% homozygous marker virions. Transfections were carried out with polyethylenimine (PEI; Polysciences), and transfection efficiencies were measured using a reverse transcriptase assay. Whilst under certain experimental conditions recombination has been observed during transfection (51-53), we have previously shown that these transfection conditions do not yield observable recombination (20). Furthermore, we have also converted the virion associated RNA to cDNA for direct sequencing, which shows that transfection induced recombination events have negligible influence on our results (see below). 36 hours post-transfection, virus-containing media was harvested, clarified by centrifugation at 1,462 x *g* for 30 minutes, and then passed through a 0.45µm filter to remove cellular debris. Purified virus was concentrated by

ultracentrifugation at 100,000 x *g* through a 20% sucrose cushion and stored at -80^{o}C. Virus was treated with 90units/mL benzonase (Sigma) for 15 minutes at 37^{o}C to remove contaminating plasmid DNA before use.

Infections

Stimulated peripheral blood lymphocytes (PBLs) from 3 separate blood donors (patients) were infected with equal mass of either homozygous or heterozygous virus (which contained a mix of homozygous and heterozygous virus), as determined by a HIV-1 antigen (p24 CA) micro ELISA assay (Vironostika). In this experimental setup, the ability to measure recombination is not affected by the MOI. Using viral cDNA synthesis as a surrogate marker for successful infection, retrospective analysis showed the average MOI to be 0.5 in this PBMCs. Efficient removal of plasmid DNA by benzonase treatment was confirmed for each sample using qPCR targeting the ampicillin gene. Only virus preparations where the infection has no or insignificant background PCR signals (<5% of the wild type infection) were used in subsequent infection and sequencing experiments (Table 1). We note that any potential carry over plasmid DNA would have the effect of reducing the observed HIV recombination and mutation. Six hours post-infection 10µg/mL T-20 (Roche) was added to the cells to prevent second round replication. 24 hours post-infection cells were pelleted and lysed in PCR lysis buffer containing 1× PCR buffer (Roche) with 0.5% vol/vol Triton-X100, 0.5% vol/vol NP-40 and 75 µg/ml proteinase K (Roche). 1x10^{6} cells were lysed per 100µl of PCR lysis buffer. Samples were incubated at 56°C for 1 h before proteinase K was inactivated at 95°C for 10 min, samples were then stored at –20°C. Lysates were diluted 1 in 10 before quantification by quantitative PCR.

Primers

PCR Primers were designed to span the 46 intervals as 14 overlapping amplicons of roughly 350 nucleotides. PCR primers were designed against regions that were identical between WT and MK plasmid according to the manufacturer's directions (http://www.finnzymes.com/tm_determination.html).

List of primers used are: Primer G1 sense [5'GGTGCGAGAGCGTCGGTATTAAG3']; Primer G1 antisense [5'CTGTGTCAGCTGCTGCTTGCTG3']; Primer G2 sense [5'TCCTCTATTGTGTGCATCAAAGGATAGATG3']; Primer G2 antisense [5'CCACTGTGTTTAGCATGGTATTTAAATCTTGTG3']; Primer G3 sense [5'CAAATGGTACATCAGGCCATATCACCTAG3']; Primer G3 antisense [5'CTGCATGCACTGGATGCAATCTATC3']; Primer G4 sense [5'GAAGGAGCCACCCCACAAGATTTA3']; Primer G4 antisense [5'GGTTCCTTTGGTCCTTGTCTTATGTCCAG3']; Primer G5 sense [5'GGAAGTGACATAGCAGGAACTACTAG3']; Primer G5 antisense [5'AGTCTTACAATCTGGGTTCGCATTTTGG3']; Primer G6 sense [5'AAACTCTAAGAGCCGAGCAAGCTTC3']; Primer G6 antisense [5'TGCCCTTCTTTGCCACAATTGAAACAC3']; Primer P1 sense [5'GCAGGAGCCGATAGACAAGGAACT3']; Primer P1 antisense [5'TAAAGTGCAGCCAATCTGAGTCAACAG3']; Primer P2 sense [5'AGAAATCTGCGGACATAAAGCTATAGG3']; Primer P2 antisense [5'GGAGTATTGTATGGATTTTCAGGCCCAA3']; Primer P3 sense [5'GTAAAATTAAAGCCAGGAATGGATGGC3']; Primer P3 antisense [5'GAAAAATATGCATCGCCCACATCCAG3']; Primer P4 sense [5'TGTGGGCGATGCATATTTTTCAGT3']; Primer P4 antisense [5'ATGGAGTTCATAACCCATCCAAAGGAATG3']; Primer P5 sense [5'CACCAGCAATATTCCAGTGTAGCATG3']; Primer P5 antisense [5'CTTTAATCCCTGCATAAATCTGACTTGCC3']; Primer P6 sense [5'GAACTCCATCCTGATAAATGGACAGTACAG3']; Primer P6 antisense [5'TTAAATGGCTCTTGATAAATTTGATATGTCCATTG3']; Primer P7 sense [5'CCACTAACAGAAGAAGCAGAGCTAGAACTG3']; Primer P7 antisense [5'CAGGTGGCTTGCCAATACTCTGTC3']; Primer P8 sense [5'AGGGTGCCCACACTAATGATGTGAAAC3']; Primer P8 antisense [5'AGTCTTCTGATTTGTTGTGTCCGTTAGG3']; Primer AMP sense [5'AACTCGCCTTGATCGTTGGG3']; Primer AMP antisense [5'TGTTGCCATTGCTACAGGCATC3']

Quantitative PCR

Quantitative PCR was performed on an MX3000 (Stratagene). Reverse transcription products were assessed using the HIV-1 specific primers M661/M667 (54) and background plasmid levels were assessed using primers directed against the ampicilin resistance gene, AMP (see primer list). Each PCR reaction contained 1× Brilliant II Master mix (Stratagene), 400 nM each primer and 5 µl cell lysate in a 15 µl reaction volume. For viral cDNA estimation, PCR conditions were an initial denaturation at 95°C for 15 min followed by 40 rounds of cycling at 95°C for 10 s, then 60°C for 30 s. Samples were compared to ACH2 cell standards.

PCR

With amplicon generation for next generation sequencing, PCR conditions were optimized to reduce the formation of artificial recombinants using the method outlined in Smyth et al. (40). PCR reactions were titrated to contain 2,500 copies of template DNA, 1x HF buffer (Finnzymes), 200 µM dNTP, 1 µM of each primer, 0.3 U of Phusion DNA polymerase (Finnzymes) in a 15µl total reaction volume. Plasmid DNA was titrated in cellular lysate from uninfected PBLs so that the DNA complexity of the control PCR reactions faithfully represented the experimental samples. PCR reactions were performed in quadruplicate and pooled to guard against PCR bias. PCR cycling conditions were 98 °C for 30 s followed 29 cycles of 98 °C for 10 s and 1 min for 72 °C.

Transfection-induced recombination

283 To assess the rate of transfection-induced recombination RNA was extracted from
284 heterozygous virus using phenol chloroform based TRI reagent (Sigma Aldrich)
285 according to the manufacturer's recommendations and reverse transcribed into cDNA
286 using SuperScript^{TM}III (SSIII) (Invitrogen Life Technologies) and gene specific
287 primer GAG4(4195)R: 5'ACATTTCCAACAGCCCTTTTTCCTAG 3'. To control
288 for *in vitro* cell-free reverse transcription and PCR induced recombination, control
289 samples consisting of homozygous WT virus and homozygous MK virus were mixed
290 in equal quantities (based on p24 values) prior to RNA extraction, and were reverse
291 transcribed in parallel with RNA extracted from heterozygous virus. Reverse
292 transcription was performed in the presence and absence of SSIII, the latter condition
293 providing a control for any plasmid contamination carried over from transfection.
294 Real time PCR was used to estimate viral cDNA copy number against a standard
295 curve based on plasmid pDRNL(AD8) using primers GAG1(2945)F: 5'
296 GAGATGGGTGCGAGAGCGTC 3' and GAG1 (3314)R: 5'
297 TGTGTCAGCTGCTGCTTGCTG 3'. Twenty replicate wells containing 2,500
298 copies of template viral cDNA were amplified using optimized PCR conditions to
299 reduce the formation of artificial recombinants as outlined in Smyth et al. (40).
300
301 **454 Sequencing**
302 PCR amplicons spanning *gag* and *pol* were pooled for each blood donor. Unique
303 6 nucleotide identifiers (barcodes) were then individually attached by parallel tagged
304 sequencing to allow multiplexing of different blood donors on the same sequencing
305 run (55, 56). Single-stranded sequencing libraries were constructed from 1 µg of
306 initial starting material using the 454-library preparation kit (Roche) according to the

manufacturer's directions. Libraries were quantified and stored at − 20 °C until further use. Emulsion PCR and sequencing were performed according to standard GS FLX titanium procedures. In order to avoid excessive resampling of the same DNA strand, the 454-library was constructed from PCR products deriving from at least 10 times more viral cDNA molecules than the maximum number of sequencing reads allow in a reaction. For each full 454-sequencing run, the 454-library consists of PCR products from 4,000 PCR reactions, with each reaction containing 2,500 templates for amplification. This results in a 454-library consisting PCR products derived from ten million original templates, with a maximum of one million sequencing reads per 454-sequencing reaction. For PCR products derived from every 2500 copies of original template, only 10% (250) of its next generation sequencing read was used for analysis to minimize re-sampling of the same original sequence that could bias the estimation of the recombination and the mutation rates.

Sequence alignment

Sequencing data was processed to remove the 6-nucleotides barcode and assigned to a sample only upon a perfect barcode match. To reduce the missense error rate of 454, sequences were also removed if they were of poor quality (such as containing ambiguous nucleotides ('Ns') or not full length. All sequencing analysis was performed using software custom written in BioRuby (www.bioruby.org). In order to count mutations and other events, the sequences were aligned against the consensus sequence of the wild type and marker type version of the amplicon (plus a margin of 100 nucleotides on both ends of the amplicon) using *needle*. The parameters for needle were a gap opening penalty of 3.0, and a gap extension penalty of 0.5.

Segments which contained an ambiguous marker (a mutation within the marker) were discarded from the analysis.

Measuring recombination rate

The conversion of raw rates of chimera formation into recombination events per nucleotide (REPN) and statistical comparisons between recombination rates were performed using the methods in Schlub et al. (20). Briefly, recombination is detected by monitoring the linking of marker points from wild type and marker type genomes into a single genome. However, simply measuring the crude number of recombination events divided by the number of nucleotides can be misleading. Our approach is described in detail elsewhere (20), and takes into account a number of factors inherent in the marker system:

1) Multiple recombination events can occur between two marker points. Consequently, all odd numbers of recombination events between two markers (1,3,5,7,9,….etc.) will be counted as a single recombination event, while even numbers of recombination events between two markers (2,4,6,8,10,….etc.) will go undetected.
2) We infect cells using a mix of heterozygous and homozygous virions (and in the latter, recombination will go undetected). Thus, we directly estimate the proportion of heterozygous virions for each sample.
3) The widths of our intervals vary substantially, and therefore we must take into account different likelihoods of multiple recombination events.

353 Our mathematical procedure accounts for these factors to estimate the
354 recombination rate within the experimental system. This estimated recombination
355 rate minimizes the chi square value

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

356

357 Where k is the number of intervals between markers and Ei and Oi and the expected
358 and observed number of detected recombinations for interval i. The expected number
359 of detected recombination events in any given interval Ei is calculated by

$$E = s \sum_{j=1}^{\left[\frac{L_i+1}{2}\right]} C(L_i, 2j-1) r^{2j-1} (1-r)^{L_i-2j+1}$$

360

361 Where: s is the number of heterozygous sequences; L_i is the nucleotide length of the
362 interval; $[(L_i+1)/2]$ is the integer part of $(L_i+1)/2$; and $C(L_i,k)$ is the binomial
363 coefficient for picking k unordered outcomes from L_i possibilities.

364

365 **Estimating mutation rate per recombination: Method 1**

366 To estimate the background rate of mutation and the rate of mutation induced by
367 recombination, we consider the number of point mutations in intervals with and
368 without recombination. Intervals derived from homozygous virions do not display
369 recombination regardless of RT template switching activity. Therefore we initially
370 limit ourselves to sequences that are known to be derived from heterozygous
371 sequences (sequences which contain at least one observable recombination). The
372 mutation rate in heterozygous intervals without recombination, m_b, provides an
373 estimate of the background mutation rate alone. The mutation rate per nucleotide in

intervals with recombination, m_r, then represents the cumulative effect of background mutation and mutation induced from the recombination event. Thus,

$$m_r = m_b + p/L$$

where p is the mutation rate per recombination, and L is the length of the interval. As m_r, m_b is measurable from the data, and L is known, p can be calculated. Despite the large dataset, as mutation rates and recombination rates are relatively low, mutation rates for individual intervals may have large amounts of variation by random chance alone. Thus mutation rate calculations will not be informative if calculated on a per interval basis. To overcome this, we calculate m_r and p from the cumulative mutation rate over all heterozygous intervals with and without recombination, and in this case substitute L with the average length of intervals displaying recombination.

Estimating mutation rate per recombination: Method 2

Method 1 estimates the background mutation rate per nucleotide and an additional mutation rate per recombination event. However, this method only uses a fraction of the available data (sequences known to be heterozygous). Additionally, method 1 does not compensate for recombination events occurring at a greater frequency than one per interval (although this is estimated to be minimal). To adjust for these factors, we develop a second method to measure mutation rates as described below.

To estimate the background rate of mutation and the rate of mutation induced by recombination, we consider the number of point mutations in intervals with recombination (N_I), heterozygous intervals without recombination (contains a

recombination somewhere on the sequence) (N_2), and intervals of unknown ancestry (no recombination on the remainder of the sequence) without recombination (N_3). Thus for each interval, over all sequences

$$N_1 = mI_1 + p\frac{x}{R}s_1$$

$$N_2 = mI_2 + p\frac{y}{1-R}s_2$$

$$N_3 = mI_3 + \frac{y}{1-R}s_3 + (x+y)s_4$$

where: m is the background rate of mutation; I_1, I_2 and I_3 is the number of informative nucleotides in intervals with recombination, heterozygous intervals without recombination and intervals of unknown ancestry without recombination respectively; p, is the rate of mutation per recombination; R is the probability of observing a recombination over that interval; s_1 and s_2 are the number of intervals with recombination and heterozygous intervals without recombination respectively: s_3 and s_4 are the estimated number of intervals derived from heterozygous virions that have no recombination along the entire sequence (and are thus of unknown heterozygous/homozygous ancestry), and the estimated number of homozygous intervals respectively; and where

$$x = P_1 + 3P_3 + 5P_5 + \ldots$$

$$y = 2P_2\ 4P_4 + 6P_6 + \ldots$$

with P_i the probability of i recombinations occurring in a single interval. Thus the terms x/R and $y/(1-R)$ represent the expected number of recombinations per interval in heterozygous intervals with and without recombination respectively. The term $(x+y)$ represents the expected number of recombinations to occur in a homozygous interval (where recombination is unobservable).

To calculate a single mutation rate per nucleotide and mutation rate per recombination over all intervals, we optimize the expected values of N_1, N_2 and N_3 to minimise the summed square error to the observed counts over all intervals. As N_3 represents the majority of our dataset, weighting of N_1, N_2 and N_3 errors may be employed to amplify their effect on the mutation rate estimations. However, such weighting did not substantially change mutation rate calculations in this dataset, and do not affect the conclusions drawn. Optimisations were performed in *Matlab* (The Mathworks Inc. v7.1.0.124 (R14)) with function *fmincon*.

Estimating mutation rate per recombination: Subtracting controls

As the mechanism for individual mutations and recombinations cannot be determined, we cannot remove experimentally-induced mutation and/or recombination prior to the analysis and comparison of rates. Rather we estimate rates of mutation, recombination and mutation per recombination in the biological sample, and then separately estimate these in the controls. We then show that the contribution of experimental mutation/recombination to the mutation rate per recombination is minimal. Specifically, by estimating the proportion of recombination events attributable to experimental factors, and their corresponding mutations, we estimate that only 9-15%

441 of the recombination associated mutation rate observed following infection could be
442 due to experimental factors. This was subtracted from the biological sample rate to
443 estimate the recombination associated mutation rate attributable to viral infection
444 processes only (RT, RNA polymerase II, and potentially host nucleic acid editing
445 enzymes such as APOBEC3G). This rate is then used to calculate the mutation rate
446 per recombination for viral infection only.
447
448 **Statistical Analysis**
449 Statistical and mathematical analysis was carried out in *Graphpad Prism* and *Matlab*
450 respectively. Fisher's exact tests were used to compare overall rates of mutation. To
451 eliminate the effect of 'co-incident hotspots', we developed a permutation approach to
452 investigate the relationship between mutation and recombination (Figure 3A, 3B). In
453 this approach, each sequence interval (the region between two marker sites) was
454 classified as recombined (R) or non-recombined (NR). The classification of R or NR
455 was then randomly permuted (reshuffled) 10,000 times to generate the expected
456 distribution of mutations within R and NR intervals if mutation and recombination
457 were not associated. To avoid confounding variables, we only reshuffled within the
458 same interval, same amplicon, same direction of sequencing, and same patient. By
459 comparing the observed difference in R and NR mutation rates with those obtained
460 from random reshuffling, we eliminated the confounding effects of coincident
461 hotspots, and Simpson's Paradox.

462 **Results:**

463 Measurement of recombination and mutation

464 We used a system of markers introduced into the HIV *gag* and *pol* to simultaneously
465 measure the rate of recombination and mutation during a single round of replication
466 (Figure 1A). A total of 46 intervals spaced an average of 53.5 nucleotides apart (range
467 17-155, median 47 nucleotides) were used. The positions of these mutations were
468 chosen due to their natural polymorphism in the HIV database, and this marker HIV
469 also has identical replication kinetics to our parental wild type control. Heterozygous
470 virus was produced by co-transfection of 293T cells with an equal mass of wild-type
471 (WT) and marker (MK) plasmid, leading to a mixture of virions containing WT
472 homozygous (25%), MK homozygous (25%), and WT-MK heterozygous (50%) co-
473 packaged RNA genomes in a manner as previously described (13, 14, 20). The ratio
474 of homozygous and heterozygous virion in our HIV infection stock was also
475 internally monitored for each infection based on the frequency of non-recombined
476 WT and MK sequences (20). This mixture of heterozygous and homozygous virus
477 was used to separately infect freshly stimulated T-lymphocytes from three separate
478 blood donors. Following a single round of infection, viral DNA was extracted, PCR
479 amplified and sequenced (Figure 1B). As previously reported, we have chosen to use
480 a high fidelity and high processivity DNA polymerase (40), Phusion, that has a
481 significant higher fidelity rate than other polymerase, with an estimated error rate of
482 4.5×10^{-7} per base pair (57)

483

484 In addition to our experimental sample, we included a number of controls that account
485 for sequencing error and PCR-induced recombination and mutation (Figure 1D, 1E

and Table 2, 'DNA control' and 'Homozygous', respectively). The DNA control consisted of PCR amplification and subsequent sequencing of plasmid DNA (Figure 1D, Table 2, 'DNA control'). This sample allowed us to control for the rate of recombination and sequencing error during PCR amplification and 454 sequencing. Plasmid DNA was titrated in cellular lysate from uninfected PBLs so that the DNA complexity of the control PCR reactions faithfully represented the experimental samples. The second control involved infecting cells with a mixture of homozygous (containing identical co-packaged RNA genomes) WT or MK virus, before PCR amplification and sequencing (Figure 1E, Table 2, 'Homozygous control'). In this case, recombination during HIV reverse transcription is expected to occur at a normal rate, but effectively be 'silent', as the homozygous virus simply recombines onto an identical RNA strand. However, recombinant chimeras between WT and MK DNA strands may still occur during subsequent PCR amplification after cells are lysed. Thus, the level of PCR-induced recombination in this sample should occur at exactly the same rate as during heterozygous virus infection. This second recombination control also directly monitors the levels of inter-virion recombination events, demonstrating the levels of MOI used in these infections do not bias the observed recombination rate (Table 2). Finally, this control incorporates the possibility of multiple rounds of infection, which are not expected to occur due to addition of fusion inhibitors 6 hours after infection and the termination of infection 24 hours post infection.

In response to reviewer concerns, a further round of controls was performed to assess whether transfection-induced recombination (TIR) might be occurring during the

initial stage of transfecting the plasmids. TIR might occur when the WT and MK plasmids are co-transfected, and undergo recombination *in vitro*. To exclude this possibility, we sequenced the virus produced by transfected cells (Figure 1C, Table 3). Because this involved a reverse-transcription step performed on viral RNA, it is possible that any recombination events observed in the viral sequences may have arisen from the reverse transcription rather than TIR. Therefore we included controls where WT and MK plasmids were transfected separately into cells, and the resultant homozygous viruses were only combined for the stage of reverse transcription. These assays showed that transfection of the two different plasmids did not result in more recombination than seen with reverse transcription of homozygous virions, indicating very little or no TIR (approximately 0.006×10^{-3} REPN, less than 0.5% of the total recombination rate; Table 3). Moreover, the total level of recombination seen was so low that even if all of these recombination events were caused by TIR (and not from the reverse-transcription step), they would not account for any significant proportion of recombination events seen in the data.

For each experimental condition, six pairs of overlapping *gag* PCR amplicons and eight pairs of overlapping *pol* PCR amplicons were used to amplify the corresponding viral sequences. These PCR amplicons cover all the marker points in *gag* and *pol*, and have an approximate length of 350 nucleotides that are optimal for 454 pyro-sequencing. Each PCR amplicon contains between 5 to 7 marker points, hence, 4 to 6 intervals each for our analysis.

Following sequencing and alignment, we obtained on average 3364 sequences per interval (range 1290-6889) for the experimental sample. Each marker position was then classified as WT, MK, or ambiguous (if it was not identical to either a WT or MK sequence). Each interval between markers was then classified as recombined (R) if adjacent markers were WT and MK, non-recombined (NR) if adjacent markers were identical, or ambiguous if either adjacent marker could not be classified. We calculated the recombination rate using a method that takes into account the proportion of homozygous sequences (where viral-induced recombination cannot be detected), as well as the possibility of multiple undetectable recombination events between marker sets (described in detail in reference (20)). From this we estimated the overall recombination rate in our experimental samples (1.51 per 1000 nt), and in our controls (Figure 1F and Table 2, 'Infection – total').

We next studied the rate of mutation in these samples (Figure 1G and Table1 'Infection – total'). To avoid re-sampling of the viral cDNA that could lead to an inaccurate estimation of mutation rate in HIV genome, 10 fold excess of viral cDNA were used as templates for 454 sequencing, ie, more than 10 million distinct viral cDNA templates were used for a complete run (1 million reads) of 454 sequencing. In the experimental sample, from a total of 7,180,712 informative nucleotides we observed 859 mutations, giving an overall mutation rate in of 0.120 per 1000 nt. This mutation rate represents the cumulative effect of experimentally induced "background" PCR and sequencing error, and the viral RT induced mutation. We estimated the background rate by amplifying and sequencing a plasmid DNA with a 50:50 mix of the WT and MK sequences in the presence of cell lysates from

uninfected PBL. Here our sample consists of 4,931,016 nt, from which we observed 366 mutations, giving a mutation rate of 0.0742 per 1000 nt (Table 2, 'DNA'). We note that our background rate of substitution mutations from PCR amplification and sequencing is significantly lower than some other estimates, and this difference likely reflects the high fidelity nature of Phusion DNA polymerase in our analyses. In addition, removal of low quality sequencing reads (ie: containing ambiguous nucleotides ('n') or that were not full length amplicons) also kept the rate low. The mutation rate attributable to viral factors (including RT, RNA polymerase II, and, potentially host nucleic acid-editing enzymes (e.g. APOBEC3G) is the difference between the error rate in the biological sample and plasmid DNA control. This rate is $0.120 - 0.0742 = 0.0458$ per 1000 nt. This is similar to the rate of 1.4×10^{-5} to 4×10^{-5} previously estimated for HIV-1 (2, 19, 41).

Higher mutation rate in recombined intervals

In order to test whether recombination was associated with mutation, we compared the mutation rate in intervals where recombination is and is not observed. We found a significantly higher mutation rate in recombined intervals than in non-recombined intervals (0.181 / 1000 nt vs. 0.117 / 1000 nt, $p = 0.003$, Fishers exact). However, since our procedure involved PCR amplification, and we have previously demonstrated that PCR induced recombination occurs at a low but significant rate (20), it is possible that PCR-induced recombination during sample preparation was responsible for the observed association between recombination and mutation. To exclude this possibility, we analysed the control infections where we infected cells

581 with a 50:50 mix of homozygous WT and MK virus. Because of the low rate of PCR-
582 induced recombination, we sequenced twice as many samples, leading to a total of
583 15,407,974 informative nucleotides from control samples. As previously described
584 (40), the rate of PCR induced recombination in this sample (0.050 per 1000 nt) was
585 much lower than the cumulative effect of viral RT induced + PCR-induced
586 recombination measured with heterozygous infection (1.51 per 1000 nt). PCR induced
587 recombination was associated with a higher mutation rate (mutation rate of 0.537 per
588 1000 nt in R sequences versus 0.118 per 1000 nt in NR, $p < 0.0001$ Fishers exact).
589 Although mutation was increased with PCR-induced recombination, the overall rate
590 of PCR induced recombination was much too low to account for our observed
591 difference in the heterozygous HIV infection samples. That is, correcting for the
592 sample size, we expect that around 3% of recombination events are due to PCR error,
593 and that in our total sample these PCR-recombined regions would contain
594 approximately 5 mutations (including both 'background' and recombination-
595 associated mutations). Thus, less than 10% (5/54) of total mutations in the
596 recombined regions could be attributed to PCR-induced recombination. Similarly,
597 transfection-induced recombination accounted for <0.5% when we sequenced the
598 virion genomes directly (Table 3) and is too low to contribute to the observed rates.
599
600 Although we observed an overall higher rate of mutation in recombined regions, this
601 could have occurred for a number of potential reasons. Firstly, the apparent
602 association between mutation and recombination could have been artificially created
603 by pooling data over the various intervals and sequences from various amplicons.
604 This could occur in a scenario where 'coincident recombination and mutation hotspots'

led to higher incidence of both mutation and recombination at particular segments, even though recombined intervals had the same mutation rate as non-recombined intervals within the segment (Figure 2A, pale orange box). This would be an example of the so-called 'Simpson's Paradox' (58). A second potential confounder is the effect of ‘error-prone cells’ or 'error-prone strands’ driving the observed association. That is, if some cells, RT enzymes or viral RNA strands produce a particularly high rate of both mutation and recombination in any particular set of amplicon, then mutation and recombination will appear associated, even if it is only because mutation is higher all along the ‘error-prone strand’ and not higher in all recombined intervals on the strand in general (Figure 2B, pale pink boxes).

Mutation is associated with recombination

To eliminate the effect of ‘co-incident hotspots’, we developed a permutation test (Figure 3A, 3B, Materials and Methods). In this approach, recombination classification (R or NR) for each sequence interval (the region between two marker was randomly permuted to generate the expected distribution of mutations within R and NR intervals if mutation and recombination were not associated. By comparing the empirical difference in R and NR mutation rates with those obtained from permutation, the probability of randomly observing an association between recombination and mutation assuming no underlying association can be eliminated with the confounding effects of coincident hotspots, and Simpson's Paradox eliminated. From 10,000 reshuffling runs we observed only 51 runs where the difference in mutation rates was as great as that observed experimentally. Thus, the

association between mutation and recombination is significant (p=0.0102, two-tailed test), and not affected by the confounding effects of coincident hotspots.

In the case of 'error-prone strands', the confounding effect of a high mutation and high recombination rate of a subset of sequences would lead to a spurious association between recombination and mutation. In this scenario, we expect that sequences with a high recombination rate would also have a high mutation rate (Figure 2B). Therefore the mutation rate should be on average higher in recombined strands (in both recombined and non-recombined intervals) than it is in non-recombined strands. To investigate this, we classified all non-recombined (NR) intervals into those on a strand with recombination elsewhere (NR_R), and those on a strand with no recombination (NR_{NR}). If the association between mutation and recombination is due to error prone strands we expect that the mutation rate of $NR_R > NR_{NR}$. Alternatively, if the association between recombination and mutation is not due to error-prone strands we expect that the mutation rate of $NR_R \leq NR_{NR}$. The case of mutation rate of $NR_R < NR_{NR}$ can occur because NR_R intervals are truly non-recombined (they are known to be derived from a heterozygous infection as they have an observed recombination elsewhere), whereas, NR_{NR} will be a mix of heterozygous intervals without recombination, homozygous intervals without recombination and homozygous intervals with undetectable recombination. The homozygous intervals with undetectable recombination will increase the mutation rate in these intervals so that $NR_R < NR_{NR}$. Upon analysis, we found that the mutation rate in non-recombined intervals in strands with recombination was lower than the rate on strands without recombination ($NR_R < NR_{NR}$, $p < 0.004$ using reshuffling approach). This indicates

652 that the association between recombination and mutation was not a product of
653 confounding effects such as error prone strands.
654
655 This analysis of mutation rates on NR_R and NR_{NR} strands provides a useful
656 independent verification that neither PCR-induced recombination nor transfection-
657 induced recombination was a factor driving the association between mutation and
658 recombination. That is, both PCR-induced and transfection-induced recombinations
659 occur without consideration for whether the virions were homozygous or
660 heterozygous. Hence, if the association were due to either of these mechanisms we
661 would expect the rate of mutation in NR_R and NR_{NR} to be the same.
662
663 Pattern of mutation in recombined and non-recombined sequences
664 Given the increased mutation in recombined intervals, we sought to investigate
665 whether the pattern of mutation in R intervals was significantly different from that in
666 NR intervals. Because many NR intervals have come from homozygous virions, and
667 thus have undetected recombination, we compared mutations in non-recombined
668 intervals from recombined sequences (NR_R; where any odd number of recombination
669 events should have been detectable). We found no substantial differences in the
670 pattern of mutation observed in NR_R and recombined sequences (Table 4).
671
672 Recombination and frameshift mutations
673 The above results have analysed the rate of substitution mutations. However, it has
674 also been suggested that recombination may be associated with insertions and
675 deletions (indels) (59). To investigate this, we performed the same analysis as

676 described above to investigate whether the indel rate was higher in R or NR intervals
677 (Table 2). Although we observed a trend towards higher rate of indels in recombined
678 sequences (6.989 vs. 6.572 indels per 1000 nt), this was not significant using a two-
679 tailed permutation test (p=0.31). One reason for the inability to demonstrate a
680 consistent significant difference in indel rates is the high background indel rate due to
681 the difficulties of identifying homopolymer length during 454 sequencing (60).
682 Indeed, the rate of indels in the DNA control is higher than that of the experimental
683 sequences (6.694 vs. 6.589 indels per 1000 nt, p=0.028 (Fischer's exact)). The 454
684 sequencing indels in the DNA sample are preferentially clustered around longer
685 homopolymer tracts. Thus, we filtered the indel detection algorithm to exclude indels
686 arising from homopolymer tracts of ≥3 nt. This reduced the indel rate in both R and
687 NR sequences, although the difference was still not significant (p=0.21, two-tailed
688 permutation test).
689
690 One issue with comparing all intervals with observed recombination with all intervals
691 with no observed recombination is that the latter include homozygous sequences in
692 which silent recombination (and the associated rate of mutation) occurred on a
693 homozygous virion, and so was not observable. Thus, including these 'silently
694 recombined' sequences in the non-recombined group will push up the observed
695 mutation rate. In order to try to exclude this, we can limit our analysis to non-
696 recombined regions that were observed on recombined sequences (NRR). That is,
697 because recombination was observed elsewhere on a sequence, we know that the
698 sequence came from a heterozygous virion, and thus silent recombination was not
699 possible. The indel rate in R was significantly higher than the indel rate in NR_R

(6.989 vs 6.265 indels per 1000 nt, p = 0.0464, two-tailed permutation test). Additionally, when the data were filtered to eliminate indels from homopolymers (a limitation of 454 pyro-sequencing) of ≥3 n.t., the recombination rate in R was significantly higher than in NR_R (2.02 vs. 1.55, p = 0.003, two-tailed permutation test). While the rate of indels is significantly higher in R compared with NR_R sequences, the high background rate of indels complicates the analysis.

Discussion:

Genetic diversity in HIV primarily arises during reverse transcription via the introduction of mutations that are then shuffled between viral genomes by recombination. It is generally thought that recombination and mutation are not associated, although previous studies addressing this question have led to conflicting results (28-30). In this study, we have employed a novel HIV marker system and high-throughput pyro-sequencing system to simultaneously measure mutation and recombination rates directly on an authentic HIV-1 genome. Our comprehensive analysis demonstrates that recombination is associated with point mutation in HIV infection of primary T cells.

Three previous studies have attempted to evaluate the potential linkages between recombination and mutation in retroviruses (28-30). In the first two studies, 65,000 and 4,100 base pairs were sequenced in spleen necrosis virus (SNV) vector and HIV genomes (with anticipated 2 and 0.1 mutation events), yet zero and two mutations were detected, respectively, leading the authors to conclude that any association was absent or uncertain (28, 29). In a third study, a larger analysis was carried out in which five recombined sequences were observed to contain substitution mutations, leading the authors to conclude that around 6% of recombination resulted in mutation (30). However, the authors have pointed out that the background mutation rate of this study was four times higher than expected. Moreover, the estimated rate of recombination-associated-mutation would lead to a higher than theoretically possible mutation rate in HIV, even if it was the only source of mutation and excluding other mechanisms, such as APOBEC-mediated mutations; That is, if there is a 6% chance

731 of mutation per recombination, and HIV undergoes 5-15 recombination events per
732 genome (20, 35, 45), then this would lead to 0.3-0.9 mutation-associated
733 recombination per genome. However, only ~0.3-0.4 mutations are usually observed
734 (2, 19), suggesting this rate is higher than compatible with the observed mutation
735 rates. The authors speculate that this high level of mutation might result from the two
736 rounds of reverse transcriptions and subsequent PCR (30). With this higher than
737 expected mutation rates and the potential confounding recombination events (30), it is
738 difficult to utilize the dataset to determine the potential linkage between retroviral
739 mutation and recombination.
740
741 In the current study, we analysed over twenty seven million nt of sequencing data
742 (including seven million nt of viral genome from heterozygous HIV infection),
743 observing 859 mutations and 4801 recombination events during HIV infection. Using
744 this system, we find an overall mutation rate of 0.0458 mutations per 1000 nt, per
745 replication cycle and a recombination rate of 1.51 events per 1000 nt, per replication
746 cycle, which are similar to the rates previously reported for HIV-1. Furthermore,
747 using appropriate controls and statistical analyses, we demonstrated that this
748 association between mutation and recombination represents a biological association
749 rather than experimental or statistical artefacts.
750
751 The association between recombination and mutation can be explained in three ways:
752 mutation increases the chance of recombination; recombination increases the chance
753 of mutation; or the chance of individual mutation and recombination events can be
754 influenced simultaneously by some other factor. Our analysis cannot identify the

direction of causality. However, if the direction is such that mutation increases the chance of recombination, the effect of each mutation on the overall recombination rate will be minimal as the mutation rate is approximately 100 fold smaller than the recombination rate. Conversely, as recombination occurs much more frequently than mutation, if recombination influences mutation, this may have a significant impact on the overall mutation rate and mutation hotspots.

Assuming the scenario where recombination increases the probability of mutation, we calculated what combination of background mutation rate per nucleotide and additional mutation rate per recombination event best fits the experimental data, thus estimating the proportion of mutations that are attributable to recombination. Using the overall rate of recombination and the rates of mutation in R, NR_{NR} and NR_{R} sequences we use two mathematical methods to calculate the recombination induced mutation rate. After subtracting the recombination associated mutation rate from sequencing and PCR, we find that each viral recombination has 0.5-0.6 % chance of inducing a mutation in the same interval. This corresponds to viral recombination inducing approximately 0.07-0.09 mutations per genome (of 9600 base pair length), representing between 15% and 20% of all viral associated mutations.

One plausible explanation for this observation is that recombination is mutagenic, and thus mutations are introduced into the viral genome at the site of recombination (24, 26, 27, 61). Indeed, early *in vitro* investigations showed that HIV-1 RT frequently adds non-templated nucleotides at the 3' ends of nascent DNA during reverse transcription and that these non-templated nucleotides are misincorporated upon

779 strand transfer (25, 27). In support of this mechanism, the mutation rate for MLV, a
780 related retrovirus, is reported to be 1000 fold higher at the site of first strand transfer
781 than other regions of the genome (62). However, *in vitro* recombination and first
782 strand transfer occur at template ends and it is not known whether this corresponds to
783 recombination at internal positions within the genome analyzed in this study. It is
784 reported that non-templated addition is highly specific for purines [A>G>>T>>C], yet
785 there was no difference in the mutation spectrum observed between recombined and
786 non-recombined intervals. This suggests that this mechanism of mutagenic
787 recombination does not take place, or that the mutation spectrum during a natural
788 infection cycle is different to that observed *in vitro*. Another mechanism of mutation
789 at recombination sites is referred to as 'slippage synthesis' (24). This occurs due to
790 the misalignment of the 3' end of the nascent DNA onto the acceptor RNA template.
791 This type of recombination-induced mutagenesis is expected to be highly dependent
792 on local sequence characteristics of the template. Indeed, one *in vitro* study
793 demonstrated that whilst one recombination location was associated with a 30%
794 mutation rate, another recombination location was not associated with mutations (21).
795 Differences in template sequence may explain why some studies support the notion
796 that recombination causes mutation (21, 24, 25) whereas other studies do not (63).
797 One important advantage of our system is that mutation and recombination rates are
798 measured directly on the HIV-1 genome, meaning that these results are not biased by
799 foreign gene sequences.
800
801 Another explanation for the association between mutation and recombination is that
802 mutation increases the likelihood of template switching, rather than template

803 switching being mutagenic (22, 23, 64). HIV-1 RT is capable of extending
804 mismatched template primers, but this extension is associated with pausing of DNA
805 synthesis (22). The most widely accepted model for retroviral recombination, the
806 dynamic copy choice model, suggests that the rate of recombination is determined by
807 the dynamic steady state between DNA polymerase and RNase-H activities (65). This
808 model predicts that increasing the RNase-H to polymerase ratio by stalling DNA
809 synthesis will increase the local rate of recombination. In agreement with this model,
810 synthesis from a mismatched primer increased strand transfer by 50% compared to a
811 complementary primer, and this was associated with a significant pause in synthesis
812 (22, 23). Furthermore, in an *in vitro* template-switching assay, a lower frequency of
813 mutations was observed on the donor template when in the presence of an acceptor
814 template, implying that mutation induces template switching onto the acceptor (64).
815
816 In this study, the direction of causality is unclear. Indeed, it is plausible that
817 recombination induced mutation and mutation induced recombination are both
818 responsible for the observed association. Regardless of the causality, our study
819 demonstrates a direct linkage between recombination and mutation in driving overall
820 viral evolution. In the event that the association is only due to recombination-induced
821 mutation, our calculations show that up to 20% of mutations result from
822 recombination. It is unclear whether the observed association between mutation and
823 recombination provides an evolutionary advantage to the virus, or is simply a result of
824 the mechanisms of transcription of the error prone RT. Given the importance of
825 recombination and mutation to the evolution of drug resistance and immune escape,
826 dissecting parameters and molecular determinants that regulate these events will be

vital to define the process of HIV evolution. Such information is likely to be invaluable for understanding the emergence of immune escape and anti-viral drug resistant HIV in the course of HIV infection and AIDS pathogenesis.

References

1. Maldarelli F, Kearney M, Palmer S, Stephens R, Mican J, Polis MA, Davey RT, Kovacs J, Shao W, Rock-Kress D, Metcalf JA, Rehm C, Greer SE, Lucey DL, Danley K, Alter H, Mellors JW, Coffin JM. 2013. HIV Populations Are Large and Accumulate High Genetic Diversity in a Nonlinear Fashion. J Virol 87:10313-10323.
2. Mansky LM, Temin HM. 1995. Lower In Vivo Mutation Rate of Human Immunodeficiency Virus Type 1 than That Predicted from the Fidelity of Purified Reverse Transcriptase. J. Virol. 69:5087-5094.
3. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. 2003. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature. 424:99-103.
4. Zhang H, Yang B, Pomerantz RJ, Zhang C, Arunachalam SC, Gao L. 2003. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature. 424:94-98.
5. Johnson SF, Telesnitsky A. 2010. Retroviral RNA dimerization and packaging: the what, how, when, where, and why. PLoS Pathog 6.
6. Moore MD, Hu WS. 2009. HIV-1 RNA dimerization: It takes two to tango. AIDS Rev 11:91-102.
7. Negroni M, Buc H. 2001. Mechanisms of retroviral recombination. Annual Review Of Genetics. 35:275-302.
8. D'Souza V, Summers MF. 2005. How retroviruses select their genomes. Nat Rev Microbiol 3:643-655.
9. Paillart JC, Shehu-Xhilaga M, Marquet R, Mak J. 2004. Dimerization of retroviral RNA genomes: an inseparable pair. Nat Rev Microbiol 2:461-472.
10. Hill MK, Shehu-Xhilaga M, Campbell SM, Poumbourios P, Crowe SM, Mak J. 2003. The Dimer Initiation Sequence Stem-Loop of Human Immunodeficiency Virus Type 1 Is Dispensable for Viral Replication in Peripheral Blood Mononuclear Cells. J Virol 77:8329-8335.
11. Jones KL, Sonza S, Mak J. 2008. Primary T-lymphocytes rescue the replication of HIV-1 DIS RNA mutants in part by facilitating reverse transcription. Nucleic Acids Res 36:1578-1588.
12. Huthoff H, Das AT, Vink M, Klaver B, Zorgdrager F, Cornelissen M, Berkhout B. 2004. A human immunodeficiency virus type 1-infected individual with low viral load harbors a virus variant that exhibits an in vitro RNA dimerization defect. J Virol 78:4907-4913.
13. Chen J, Nikolaitchik O, Singh J, Wright A, Bencsics CE, Coffin JM, Ni N, Lockett S, Pathak VK, Hu WS. 2009. High efficiency of HIV-1 genomic RNA packaging and heterozygote formation revealed by single virion analysis. Proceedings Of The National Academy Of Sciences Of The United States Of America. 106:13535-13540.

14. Moore MD, Nikolaitchik OA, Chen J, Hammarskjold ML, Rekosh D, Hu WS. 2009. Probing the HIV-1 genomic RNA trafficking pathway and dimerization by genetic recombination and single virion analyses. PLoS pathogens 5:e1000627.
15. Balakrishnan M, Fay PJ, Bambara RA. 2001. The kissing hairpin sequence promotes recombination within the HIV-I 5' leader region. J Biol Chem 276:36482-36492.
16. Balakrishnan M, Roques BP, Fay PJ, Bambara RA. 2003. Template dimerization promotes an acceptor invasion-induced transfer mechanism during human immunodeficiency virus type 1 minus-strand synthesis. J Virol 77:4710-4721.
17. Dykes C, Balakrishnan M, Planelles V, Zhu Y, Bambara RA, Demeter LM. 2004. Identification of a preferred region for recombination and mutation in HIV-1 gag. Virology 326:262-279.
18. Basu VP, Song M, Gao L, Rigby ST, Hanson MN, Bambara RA. 2008. Strand transfer events during HIV-1 reverse transcription. Virus Res 134:19-38.
19. Mansky LM. 1996. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. AIDS Res Hum Retroviruses 12:307-314.
20. Schlub TE, Smyth RP, Grimm AJ, Mak J, Davenport MP. 2010. Accurately measuring recombination between closely related HIV-1 genomes. PLoS Comput Biol 6:e1000766.
21. Wu W, Palaniappan C, Bambara RA, Fay PJ. 1996. Differences in Mutagenesis during Minus Strand, Plus Strand and Strand Transfer(Recombination) Synthesis of the HIV-1 *Nef* Gene *In Vitro*. Nucleic Acids Res. 24:1710-1718.
22. Palaniappan C, Wisniewski M, Wu W, Fay PJ, Bambara RA. 1996. Misincorporation by HIV-1 Reverse Transcriptase Promotes Recombination via Strand Transfer Synthesis. J. Biol. Chem. 271:22331-22338.
23. Diaz L, DeStefano JJ. 1996. Strand transfer is enhanced by mismatched nucleotides at the 3' primer terminus: a possible link between HIV reverse transcriptase fidelity and recombination. Nucleic Acids Res 24:3086-3092.
24. Wu W, Blumberg BM, Fay PJ, Bambara RA. 1995. Strand Transfer Mediated by Human Immunodeficiency Virus Reverse Transcriptase in Vitro Is Promoted by Pausing and Results in Miscorporation. J. Biol. Chem. 270:325-332.
25. Peliska JA, Benkovic SJ. 1994. Fidelity of in vitro DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase. Biochemistry 33:3890-3895.
26. Peliska JA, Benkovic SJ. 1992. Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase. Science. 258:1112-1118.

27. Patel PH, Preston BD. 1994. Marked Infidelity of Human Immunodeficiency Virus Type 1 Reverse Transcriptase at RNA and DNA Template Ends. Proc. Natl. Acad. Sci. USA 91:549-553.
28. Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, Preston BD, Dougherty JP. 2002. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J Virol 76:11273-11282.
29. Bircher LA, Rigano JC, Ponferrada VG, Wooley DP. 2002. High fidelity of homologous retroviral recombination in cell culture. Arch Virol 147:1665-1683.
30. Chin MP, Lee SK, Chen J, Nikolaitchik OA, Powell DA, Fivash MJ, Jr., Hu WS. 2008. Long-range recombination gradient between HIV-1 subtypes B and C variants caused by sequence differences in the dimerization initiation signal region. J Mol Biol 377:1324-1333.
31. Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol 74:1234-1240.
32. Simon-Loriere E, Galetto R, Hamoudi M, Archer J, Lefeuvre P, Martin DP, Robertson DL, Negroni M. 2009. Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus. PLoS Pathog 5:e1000418.
33. Galetto R, Moumen A, Giacomoni V, Veron M, Charneau P, 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.
34. Moumen A, Polomack L, Unge T, Veron M, Buc H, 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.
35. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM. 2004. From The Cover: Dynamics of HIV-1 recombination in its natural target cells. Proceedings Of The National Academy Of Sciences Of The United States Of America. 101:4204-4209.
36. Chin MP, Chen J, Nikolaitchik OA, Hu WS. 2007. Molecular determinants of HIV-1 intersubtype recombination potential. Virology 363:437-446.
37. Motomura K, Chen J, Hu WS. 2008. Genetic recombination between human immunodeficiency virus type 1 (HIV-1) and HIV-2, two distinct human lentiviruses. J Virol 82:1923-1933.
38. Chen J, Rhodes TD, Hu WS. 2005. Comparison of the genetic recombination rates of human immunodeficiency virus type 1 in macrophages and T cells. J Virol 79:9337-9340.
39. Chen J, Powell D, Hu WS. 2006. High frequency of genetic recombination is a common feature of primate lentivirus replication. J Virol 80:9651-9658.

40. Smyth RP, Schlub TE, Grimm A, Venturi V, Chopra A, Mallal S, Davenport MP, Mak J. 2010. Reducing chimera formation during PCR amplification to ensure accurate genotyping. Gene 469:45-51.
41. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. Journal of virology 84:9864-9878.
42. Chen J, Dang Q, Unutmaz D, Pathak VK, Maldarelli F, Powell D, Hu WS. 2005. Mechanisms of nonrandom human immunodeficiency virus type 1 infection and double infection: preference in virus entry is important but is not the sole factor. J Virol 79:4140-4149.
43. Chin MP, Rhodes TD, Chen J, Fu W, Hu WS. 2005. Identification of a major restriction in HIV-1 intersubtype recombination. Proceedings Of The National Academy Of Sciences Of The United States Of America. 102:9002-9007.
44. Rhodes T, Wargo H, Hu WS. 2003. High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication. J Virol 77:11193-11200.
45. Rhodes TD, Nikolaitchik O, Chen J, Powell D, Hu WS. 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.
46. Moumen A, Polomack L, Roques B, Buc H, Negroni M. 2001. The HIV-1 repeated sequence R as a robust hot-spot for copy choice recombination. Nucleic Acids Res. 29:3814-3821.
47. Harrison GP, Mayo MS, Hunter E, Lever AM. 1998. Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5' and 3' of the catalytic site. Nucleic Acids Res 26:3433-3442.
48. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping LH, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH, Shaw GM. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proceedings of the National Academy of Sciences of the United States of America 105:7552-7557.
49. Englund G, Theodore TS, Freed EO, Engelman A, Martin MA. 1995. Integration Is Required for Productive Infection of Monocyte-Derived Macrophages by Human Immunodeficiency Virus Type 1. J. Virol. 69:3216-3219.
50. Bleul CC, Wu L, Hoxie JA, Springer TA, Mackay CR. 1997. The HIV Coreceptors CXCR4 and CCR5 Are Differentially Expressed and

Regulated on Human T Lymphocytes. Proc. Natl. Acad. Sci. USA 94:1925-1930.

51. Rauth S, Song KY, Ayares D, Wallace L, Moore PD, Kucherlapati R. 1986. Transfection and homologous recombination involving single-stranded DNA substrates in mammalian cells and nuclear extracts. Proceedings Of The National Academy Of Sciences Of The United States Of America. 83:5587-5591.

52. Sprengel R, Varmus HE, Ganem D. 1987. Homologous recombination between hepadnaviral genomes following in vivo DNA transfection: implications for studies of viral infectivity. Virology 159:454-456.

53. Wake CT, Vernaleone F, Wilson JH. 1985. Topological requirements for homologous recombination among DNA molecules transfected into mammalian cells. Mol Cell Biol 5:2080-2089.

54. Zack JA, Arrigo SJ, Weitsman SR, Go AS, Haislip A, Chen IS. 1990. HIV-1 entry into quiescent primary lymphocytes: molecular analysis reveals a labile, latent viral structure. Cell 61:213-222.

55. Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 35:e97.

56. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD. 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic Acids Res 35:e91.

57. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America 108:9530-9535.

58. Simpson EH. 1951. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society B 13:238-241.

59. Baird HA, Galetto R, Gao Y, Simon-Loriere E, Abreha M, Archer J, Fan J, Robertson DL, Arts EJ, Negroni M. 2006. Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. Nucleic Acids Res 34:5203-5216.

60. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437:376-380.

61. DeStefano JJ, Raja A, Cristofaro JV. 2000. In vitro strand transfer from broken RNAs results in mismatch but not frameshift mutations. Virology 276:7-15.

62. Kulpa D, Topping R, Telesnitsky A. 1997. Determination of the site of first strand transfer during Moloney murine leukemia virus reverse transcription and identification of strand transfer-associated reverse transcriptase errors. Embo J 16:856-865.
63. DeStefano J, Ghosh J, Prasad B, Raja A. 1998. High fidelity of internal strand transfer catalyzed by human immunodeficiency virus reverse transcriptase. J Biol Chem 273:1483-1489.
64. Diaz L, Cristofaro JV, DeStefano JJ. 2000. Human immunodeficiency virus reverse transcriptase base misincorporations can promote strand transfer. Arch Virol 145:1117-1131.
65. Nikolenko GN, Svarovskaia ES, Delviks KA, Pathak VK. 2004. Antiretroviral drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase increase template-switching frequency. J Virol 78:8761-8770.

Figures Legends:

Figure 1: Measuring recombination and mutation rates

(A) A schematic of the marker system. Marker plasmid (MK) was generated through the introduction of silent markers into *gag* and *pol* of wild-type (WT) HIV (shown as green dots). The position of the amplicons are marked as red horizontal bars. (B) The experimental infection. Marker (MK) plasmid and wild-type (WT) plasmid DNA was co-transfected into 293T cells to produce a mixture of heterozygous and homozygous virus. Virus was then used to perform a single round of infection in primary T-cells. DNA was subsequently extracted for PCR and high throughput sequencing. (C,D,E) Controls for experimentally induced recombination and mutations. (C) The first control consisted of amplifying and sequencing heterozygous virus reverse transcribed *in vitro*. This control assesses the rate of recombination occuring during co-transfection of viral plasmid. (D) This control involved PCR amplification of a mixture of MK and WT plasmid DNA to assess the rate of mutation and recombination during amplification and sequencing. (E) Two separate transfections of either WT plasmid or MK were used to produce homozygous WT and MK virus. These homozygous virus preparations were combined 50:50 and used to infect cells. Thus, any recombination during reverse transcription occurs onto an identical strand (since the virus is homozygous), and is undetectable. The rate of mutation is the same as the experimental sample, and only experimentally (PCR-) induced recombination is measured. (F,G) Recombination and mutation rates in each region of the experimental sample. The average rates that would produce this distribution are show as dashed lines.

Figure 2: Dissecting the association between recombination and mutation

An experimental association between mutation and recombination may be observed due to either "coincident hotspots" (A) or "error-prone strands" (B), even though there is no real mechanistic association. (A) A coincident hotspots scenario can occur if there is a region of the genome with both high mutation and high recombination rate (yellow box). In this scenario, overall we see that 2 of the 5 recombined segments (R) (40%) have mutations, and only 4 of the 13 (31%) non-recombined segments (NR) have a mutation. However, this apparent association is simply due to the high mutation and recombination rate in the yellow region. There is no higher mutation rate in R segments when individual regions are analysed separately. That is, there are no mutations in the recombined regions outside the 'hotspot', and an equal rate of recombination in R and NR within the 'hotspot' (ie: 2/3 segments have mutations for both R and NR). (B) The presence of 'error prone strands' (orange) with high mutation and recombination can also lead to an erroneous association. Here, we would observe 3/7 (43%) recombined regions contain a mutation, and 5/17 (29%) of non-recombined regions contain a mutation. However, there are no mutations in recombined regions outside the error prone strands, and in the error prone strands there are a total of 3 / 4 regions mutated in both the recombined and non-recombined regions.

Figure 3: Permutation of recombination and mutation sites

An excess of mutations in recombined regions can be produced simply by having regions of the genome with high mutation and recombination rates ("coincident

hotspots, Figure 2A"). To overcome this, we perform a permutation test as follows. (A) We classify each interval on each sequence to be recombined (R) and/or mutated (▲). (B) With each reshuffle, the recombination status of intervals is randomly permuted. To eliminate confounding factors (such as co-incident hotspots), recombination status is only permuted within the same interval, amplicon, direction of sequencing and patient. Reshuffling 10000 times generates the distribution of mutation rates that would occur if recombination and mutation were not mechanistically associated. Statistics are calculated by comparing the original un-permuted data with the generated null distribution.

Table 1: Removal of contaminating plasmid DNA by benzonase treatement.

Efficient removal of contaminating plasmid DNA by benzonase treatment was confirmed by qPCR quantification of HIV and amplicillin gene sequences using specific primers and the same standard based on serial dilutions of the NL43 plasmid. HIV specific primers were 5'TTAAATGGCTCTTGATAAATTTGATATGTCCATTG3' P7 sense and 5'CCACTAACAGAAGAAGCAGAGCTAGAACTG3' P7 antisense. Ampicillin specific primers were 5' AACTCGCCTTGATCGTTGGG 3' Amp sense and 5' TGTTGCCATTGCTACAGGCATC 3' Amp antisense.

Table 2: Summary of mutation and recombination rates

For each sample, the number of informative nucleotides from recombined or non-recombined intervals is indicated, as well as the number of recombination events, point mutations and indels (frameshift errors). We included controls for measuring the

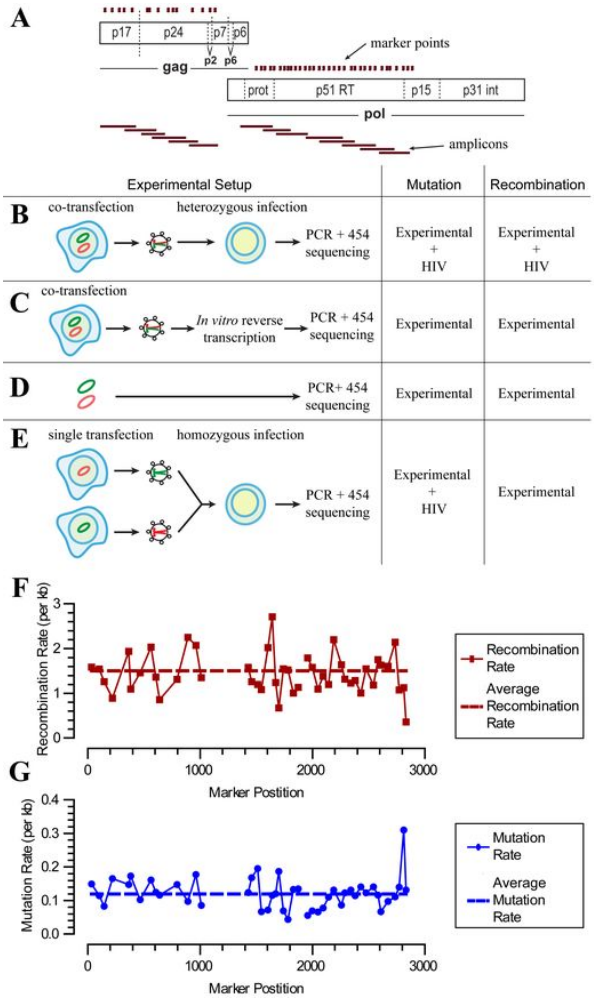
magnitude of recombination and mutation resulting from PCR and 454 sequencing: DNA control – where plasmid DNA was amplified and spiked with cellular lysates, to control for any impact of these on amplification efficiency/fidelity, measuring PCR recombination and mutation from PCR and 454 sequencing; and homozygous control – where two homozygous virus infection (one MK, one WT) were combined after lysis and before PCR amplication, so that recombination during reverse transcription is silent and recombination during PCR is observable, measuring PCR recombination and mutation of viral cDNA and 454 sequencing. The experimental infection samples are further broken up into recombined and non-recombined sequences. The latter are also broken up into non-recombined sequences on recombined strands and on non-recombined strands.

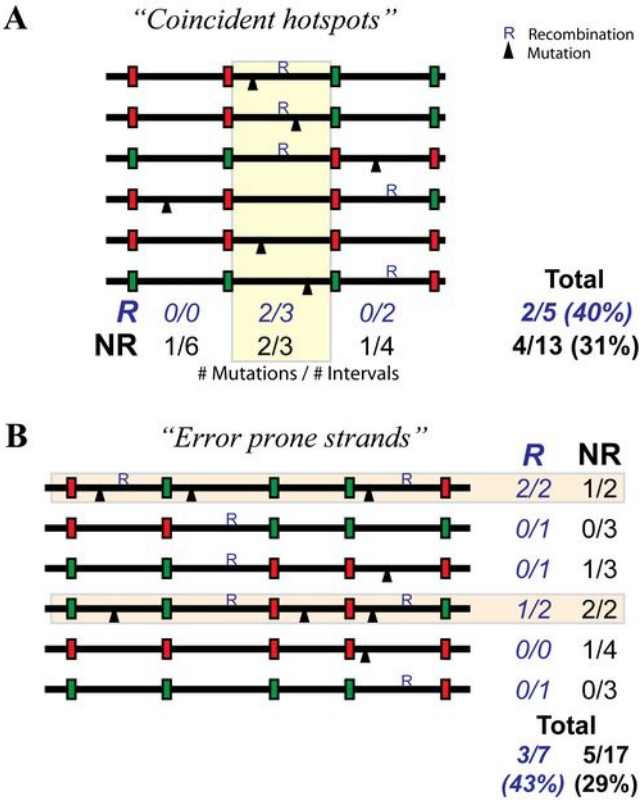
Table 3: Summary of mutation and recombination for transfection control

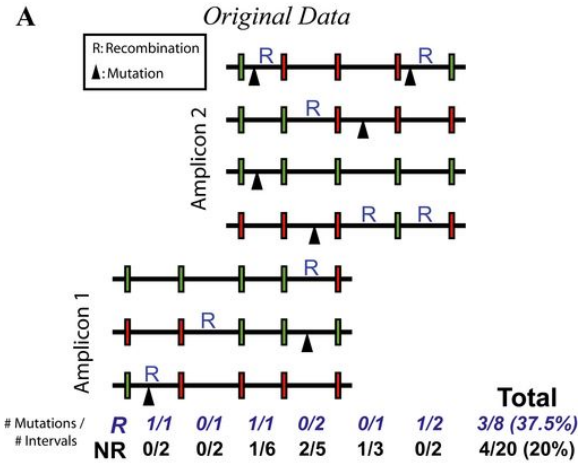
To exclude the possibility that recombination during transfection of plasmids is biasing our results, we sequence virus produced by transfected cells (row 1). This rate will measure the cumulative effect of transfection induced recombination and the reverse transcription step using Superscript III (see Materials and Methods). To determine the level of reverse transcription recombination using Superscript III, we sequence virus produced by cells separately transfected with either MK plasmids or WT plasmids only that are mixed before reverse transcription (row 2). We also repeat the PCR control (DNA control) for this assay and the experimental sample.

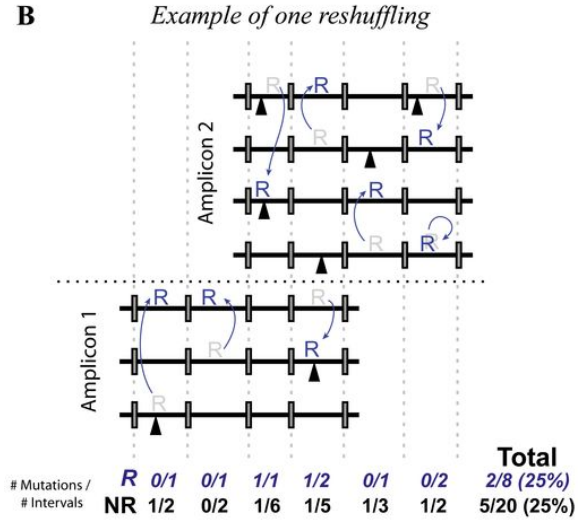
Table 4: Mutation pattern in recombined and non-recombined intervals

1168 The frequency of substitution mutations in A, C, G and T, as well as the nucleotide to
1169 which the nucleotide is mutated is indicated for recombined intervals (top) and for
1170 non-recombined intervals from recombined sequences (NR_{NR}) (middle). These
1171 mutation patterns are the cumulative product of experimental factors (such as
1172 sequencing error) and viral factors. To give some indication of the contribution of
1173 sequencing error to the pattern seen, the frequency of substitution mutations in the
1174 plasmid DNA control is shown in the bottom panel.
1175









	Donor	Sample	HIV copies	Ampicillin copies	% Background
1	JL10	Intervirion	13250	163.3	1.23
2	JL10	Wild-Type	9559	121.8	1.27
3	RS13	Intervirion	11700	127.6	1.09
4	RS13	Wild-Type	1746	29.56	1.69
5	WJ2	Intervirion	5822	44.36	0.76
6	WJ2	Wild-Type	4148	51.93	1.25

Sample	Total nucleotides	Recombination Events	Recombination rate per 1000nt	Point mutations	Mutation rate per 1000nt	Indels	Indel rate per 1000nt
DNA control	4931016	37	0.024	366	0.074	33008	6.70
Homozygous control	15407974	338	0.050	1830	0.119	98281	6.38
Infection – total	7180712	4801	1.51	859	0.120	47316	6.59
Recombined sequences (R) - total	297908	4801	NA	54	0.181	2082	6.99
Non-recombined sequences (NR) - total	6882804	0	NA	805	0.117	45234	6.57
NR sequences on NR strands	6312179	0	NA	762	0.121	41659	6.60
NR sequences on R strands	570625	0	NA	43	0.075	3575	6.27

	Total nucleotides	Recombination Events	Recombination rate per 1000 nt (REPN)
Transfection induced recombination control	22,600,749	58	0.006
Superscript III control	3,261,995	4	0.005
DNA control	24,893,615	31	0.004
Heterozygous Infection	6,558,479	4243	1.59

Recombined interval

Original nucleotides	Nucleotide count	Proportion mutated $\times 10^{-5}$ (count)	Nucleotide mutation as a % of total mutations (count)			
			A	C	G	T
A	111789	8.1 (9)	0	3.7 (2)	13.0 (7)	0.0 (0)
C	56088	12.5 (7)	3.7 (2)	0	1.9 (1)	7.4 (4)
G	66473	45.1 (30)	46.3 (25)	3.7 (2)	0	5.6 (3)
T	63558	12.6 (8)	5.6 (3)	5.6 (3)	3.7 (2)	0

Non recombined interval from recombined sequence

Original nucleotides	Nucleotide count	Proportion mutated $\times 10^{-5}$ (count)	Nucleotide mutated as a % of total mutations (count)			
			A	C	G	T
A	220054	5.0 (11)	0	4.7 (2)	20.9 (9)	0.0 (0)
C	101042	4.9 (5)	0.0 (0)	0	2.3 (1)	9.3 (4)
G	124497	16.9 (21)	41.9 (18)	0.0 (0)	0	7.0 (3)
T	125032	4.8 (6)	0.0 (0)	11.6 (5)	2.3 (1)	0

Plasmid DNA control

Original nucleotides	Nucleotide count	Proportion mutated $\times 10^{-5}$ (count)	Nucleotide mutated as a % of total mutations (count)			
			A	C	G	T
A	1892627	4.9 (99)	0	1.1 (4)	20.8 (76)	4.4 (16)
C	883997	12.4 (117)	5.7 (21)	0	2.2 (8)	20.8 (76)
G	1074185	11.4 (131)	25.4 (93)	0.5 (2)	0	6 (22)
T	1080207	4.5 (52)	1.9 (7)	10.7 (39)	0.5 (2)	0