Network analysis of human and simian immunodeficiency virus sequence sets reveals massive recombination resulting in shorter pathways

Simon Wain-Hobson,1 Céline Renoux-Elbé,1 Jean-Pierre Vartanian1 and Andreas Meyerhans2

1Unité de Rétrovirologie Moléculaire, Institut Pasteur, F-75724 Paris cedex 15, France
2Department of Virology, University of the Saarland, D-66421 Homburg, Germany

The intrinsic recombination rate of human immunodeficiency virus (HIV) exceeds the point mutation rate by a factor of 10. As the majority of infected cells in vivo harbour multiple proviruses, the stage is set for rampant recombination. Therefore, it may be presumed that phylogenetic relationships and mutation frequencies will probably be affected by recombination. However, the proportion of homoplasies arising from recombination and mutation is not known. By studying the evolution of the hypervariable regions of the simian immunodeficiency virus envelope gene among four macaques, it is shown that homoplasies arise more from recombination than from point mutation. When recombination is accounted for, the minimum number of substitutions in a sequence set may be reduced by as much as 45%. In fact, the true number of point mutations in a set of HIV sequences tends to the number of discrete substitutions. Hence, lineages are younger than anticipated previously, although not in proportion to the ratio of the intrinsic recombination/point mutation rate. Recombination also inflates codon polymorphisms.

INTRODUCTION

Recombination is a common feature among viruses. Only for the negative-stranded RNA viruses is the phenomenon infrequent (Plyusnin et al., 2002; Sibold et al., 1999). For retroviruses, the initial description of recombination goes back more than 30 years (Kawai & Hanafusa, 1972; Vogt, 1971; Weiss et al., 1973). They are inherently recombinogenic. This follows from there being two copies of genomic RNA in a virion, while reverse transcriptase (RT) is of low processivity. As a consequence, the nascent strand may switch RNA templates during reverse transcription, giving rise to a mosaic provirus (Coffin, 1979).

The first estimation of a retrovirus recombination rate, i.e. the number of RNA crossovers in a single round of replication, was made for spleen necrosis virus (SNV), an oncogenic avian retrovirus. The rate was ~0.2 crossovers per genome per round (Hu & Temin, 1990; Zhang & Temin, 1993). In contrast, the recombination rate for the human immunodeficiency virus type 1 (HIV-1) lentivirus is of the order of about three per genome per round (Jetzt et al., 2000; Yu et al., 1998), more than 10-fold the rate for SNV. Importantly, for both viruses these recombination rates are 4- to 10-fold greater than the overall point mutation rates, which are ~0.05 per genome per round for SNV (Pathak & Temin, 1990) and ~0.25 per genome per round for HIV-1 (Mansky & Temin, 1995). In other words, by the time a mutation is made, template switching has occurred between 4 and 10 times.

Recombination is found at all levels of HIV genetics. Some HIV-1 M strains in global circulation are clearly composites of viruses from two to three clades (Carr et al., 1998; Hoelscher et al., 2001; McCutchan, 2000), while a few recombinants between HIV-1 M and O have been described (Peeters et al., 1999; Takehisa et al., 1999). Viral segments amplified from isolates or patient material have revealed numerous examples of recombination (Cheynier et al., 2001; Gratton et al., 2000; Vartanian et al., 1991). In an experimental setting, wild-type simian immunodeficiency virus (SIVmac) could be recovered from peripheral blood of macaques co-infected 15 days earlier by two viruses carrying deletions in either the vif or the nef genes (Wooley et al., 1997).

A recent report showed that the majority of HIV-1-infected splenocytes from two individuals harboured between one and eight proviruses, the mean being three to four per cell (Jung et al., 2002). Depending on the sequence divergence among proviruses within a single cell, the impact of recombination may be undetectable or easily identifiable.
For example, if the two genomic RNAs packaged in a retrovirus particle were identical, it would be impossible to identify a recombinant. However, if the two genomes were substantially different, then recombination could be readily discerned. In the above study on single splenocytes, up to 20–30% amino acid sequence variation was noted within the first two hypervariable regions of Env (Jung et al., 2002), including numerous recombinants. Multiple genetically divergent proviruses per cell means that the budding virions may contain RNA genomes derived from two different proviruses. Given m provirus copies per cell, there are $m(m+1)/2$ distinct ways to randomly assort m different genomic RNAs, among which will be $m(m-1)/2$ heterokaryons. For example, when $m=4$, there are 10 ways to re assort the genomes, of which six will be heterokaryons. Given a recombination rate of three crossovers per genome per cycle, virtually all heterokaryons will give rise to a unique mosaic structure.

Recombination can generate homoplasies that are the same character state in different genomes. However, homoplasies can also arise by independent RT misincorporation, which will be referred to as point mutation. Given that the HIV-1 recombination rate is 10-fold greater than the rate of point mutation, when in doubt it may be reasonable to rule in favour of recombination underlying any homoplasy as opposed to point mutation. Yet the overall rate of retrovirus point mutation is of course an average; the $K_m$ and $V_{max}$ of some substitutions, particularly transitions, will be greater than others and context dependent (Ricchetti & Buc, 1990). Thus, it might be argued that a few sites, constituting hot or warm spots, might rival recombination as an explanation for the existence of homoplasies. The question requiring an empirical answer is, do homoplasies in lentiviral genomes arise mainly from recombination or mutation?

If rampant recombination generates large numbers of homoplasies then what are the effects on branch lengths? Certainly, standard phylogenic methods ignore recombination. Computer simulations in which sequences were recombined show that branch lengths were overestimated (Schierup & Hein, 2000). Network analyses such as the SplitsTree program are more appropriate to describing homoplasies, recombination and sequence space (Bandelt & Dress, 1992; Eigen & Nieselt-Struwe, 1990; Huson, 1998). Yet sequence space is enormous. For example, 100 variable sites – typical of different molecular clones from the same isolate – means connecting $4^{100}$ (~10$^{60}$) points in sequence space. Inevitably, constraints have been introduced into the SplitsTree program. For example, no hypercube presentation of the sequence data is possible. Nonetheless, useful information can still be recovered (Cheynier et al., 2001; Kils-Hütten et al., 2001; Plikat et al., 1997). Here, it is shown that for HIV and SIV, homoplasies are mainly the result of recombination. The implications for the analysis of virus evolution are discussed.

## METHODS

The HIV-1 and SIV sequence data sets have been published previously (Jung et al., 2002; Pelletier et al., 1995; Plikat et al., 1997; Vartanian et al., 1997). Nucleic acid sequences were aligned using the multiple sequence alignment algorithm, as implemented in CLUSTAL W (Thompson et al., 1994). Gap penalty parameters were set to 3-0 for opening a new gap and 0-05 for extension of an existing gap, the output format was set to multisequence format. Sequence alignments were translated to NEXUS format using a modified version of READSEQ and used as input for SplitsTree, version 2.4 (Bandelt & Dress, 1992; Huson, 1998). When indels were present, sequences were compiled and coded using INDELCODE (Cheynier et al., 2001). For each data set, the most parsimonious path was mapped out on the phylogram.

Consider a set of four closely related sequences (Fig. 1A). Sequences I and N differ by single substitutions (A2G and T6C) from the parental sequence H, while D encodes the same two substitutions. These sequences can be placed at the vertices of a square of unit length of one mutation (Fig. 1B). D can be derived from either I or N by an independent transition, either A2G or T6C (Fig. 1C). In this representation, the same substitution at the same site, yet in a different genome, i.e. homoplasy, shows up as parallel lines. The alternative solution is that sequence D arose by recombination of sequences I and N, somewhere between bases 2 and 6 (Fig. 1C). It is clear that the results of the two different molecular events, recombination and point mutation, are indistinguishable by a posteriori sequence analyses.

Importantly, for a set of sequences, SplitsTrees discard substitutions...
at sites that violate the constraints of the program rather than discard complete sequences (Bandelt & Dress, 1992; Huson, 1998). The proportion of sites removed is indicated in what is called the fit. When all substitutions at all variable sites are retained, the fit is 100. When sites are stripped, the fit falls approximately in proportion to the number of sites stripped. The topology of the sequences is correct for the subset of variable sites retained but information has been removed.

This is illustrated in Fig. 2. The 16 taxa set describes 11 site polymorphisms in six codons of the HIV-1 nef gene (Plikat et al., 1997) that highlight the problem (Fig. 2A). The taxa have N465 as their origin, although this information is not necessary for generating a SplitsTree. When all 16 taxa were included, a dendrogram with eight mutations was generated (Fig. 2B). Yet simple inspection of the sequences suffices to show that sequences 465 and 482 are not

---

**Fig. 2.** Effects of excluding polymorphic sites and sequences in SplitsTree analyses. A collection of 16 nef sequences (Plikat et al., 1997) have been pared down to six polymorphic codons selected to show the effects of site or sequence stripping. (A) Sequence N465 is the origin while the numbering of polymorphic nucleotides in the codons is with respect to the first base of the nef ATG codon. Only differences with respect to N465 are shown. (B) SplitsTree with all 16 sequences or taxa. The network is star-like as the fit is $\approx 100\%$. Only eight substitutions were incorporated, sites 43, 404 and 563 having been stripped. Because of this, a number of sequences are apparently identical when in fact no two sequences are, see (A). (C) As an alternative, removal of just one sequence, N504, gave rise to a complex network including all 11 polymorphic sites. (D) N504 differs from N500 by a single A563G transition, which was already present in the network and is highlighted in bold (see Fig. 3C). This substitution can be added back manually, so generating three sides of a square with a common vertex, something the SplitsTree program cannot do. In the present example, excluding a single sequence generates a fuller tree in which all informative sites were represented.

---

http://vir.sgmjournals.org 887

Network analysis of HIV/SIV sequence sets
Animal 20-188, All taxa in
Fit=100.0 ntax=40 nchar=323 const=294

Animal 20-526, 6 hypermutants excluded
Fit=100.0 taxa=35 nchar=323 const=296

Animal 20-402, 3 hypermutants and
sequence 213 excluded
Fit=100.0 ntax=38 nchar=323 const=285

Animal 193, 9 sequences excluded
Fit=100.0 ntax=33 nchar=326 const=272
identical. Indeed no two sequences are the same. The fit of ~59 indicates that some sites have been stripped, notably sites 43, 404 and 563. As a general rule, the lower the fit the more a collection of sequences resembles a star phylogeny, Fig. 2(B) being typical.

An alternative to removing sites is to remove the minimum number of sequences while retaining a fit of 100. In the present example, removing just sequence N504 resulted in a 15 taxa network with a fit of 100 (Fig. 2C). Furthermore, all 11 polymorphic sites were incorporated. Inspection of N504 shows that it differs from N500 by a single A636G transition (Fig. 2A). Yet this transition was already extensively represented in the network (Fig. 2C, highlighted in bold). The A636G transition coupling N504 to N500 can be added back manually to the figure, so generating a structure with three faces of a cube with a common vertex (Fig. 2D), something that the SplitsTree program cannot construct. Not surprisingly, therefore, a cube cannot be constructed either, even though it merely represents all transitions from, for example, AAA to GGG. However, if two sequences at adjacent or opposing vertices of the cube are removed, SplitsTrees can describe six of the eight sequences making up a cube as adjacent squares or a regular hexagon, respectively.

In contrast to site stripping, which may lead to erroneous topologies (Alber et al., 2001; Barbork et al., 1998; Holmes et al., 1999a; b; Smith et al., 1999) (Fig. 2B), eliminating the minimum number of sequences has the advantage of generating a correct topology for those sequences retained. This trial and error solution has been used in a few studies to analyse mutation rates and matrices among sets of closely related HIV or SIV sequences (Cheynier et al., 2001; Kils-Hütten et al., 2001; Plikat et al., 1997).

RESULTS

From a posteriori sequence analyses, it is not possible per se to distinguish between a homoplasy arising from recombination or point mutation. Experimental data are necessary. One way of distinguishing between the two is to compare the evolution of the same sequence in different macaques. If there are homoplases due to point mutation, perhaps resulting from hot or warm spots intrinsic to a particular molecular clone or virus strain, then they should show up in different animals. However, genomes from different macaques kept in isolation simply cannot recombine with each other. Hence, the comparison of sequences from different animals should reveal homoplases resulting uniquely from point mutations.

In a previous study of the evolution of the first two hypervariable regions (V1V2) of the SIV envelope protein, data from four animals at two time-points were reported (Pelletier et al., 1995). The sequences were under very weak selection pressure, a finding reiterated for an analysis of the equivalent regions in the HIV-1 Env protein (Kils-Hütten et al., 2001). The interesting feature of the SIV data set was the fact that the infections were initiated by inoculation with DNA of an infectious molecular clone. Accordingly, all the sequences were derived from the same founder sequence and not a collection of variants derived during preparation of stock virus that would otherwise be used for inoculation.

For three animals, 20-188, 20-526 and 20-402, virtually all sequences could be incorporated individually in SplitsTrees with a fit of 100 (Fig. 3A–C). There was evidence of some parallelograms but they were generally few. In contrast, for animal i-963, there was far greater evidence of network formation (Fig. 3D). Homoplases, i.e. the same substitution at the same site, among sequences from the four animals are highlighted in colour. There were 15 homoplases among a total of 125 discrete substitutions, 14 shared between two animals and one (A68G) between three animals. However, this A68G transition was present only in sequences taken at 19 months post-infection, indicating that it is hardly a hot spot of mutation. Equally, a substitution connecting a taxon to the parental SIVmac239 in one tree was rarely connected to SIVmac239 in another tree, again indicating that they cannot be described as hot spots. Overall, the fraction of true homoplases due to point mutations between animals represent a minority of total number of substitutions (excess substitutions = 14 × 1 + 1 × 2 = 16; excess/total discrete substitutions = 16/125, ~13 %), indicating that this is probably also true of substitutions within an animal. Homoplases between animals could be broken down into 10 non-synonymous substitutions, four synonymous substitutions and one transversion generating an in-phase stop codon. Such a distribution is close to that expected from weak purifying selection, as noted before (Pelletier et al., 1995).

Fig. 4(A) describes the minimum path length connecting the 33 taxa in the i-963 network in the absence of recombination. A total of 64 substitutions was necessary, including 11 homoplases highlighted in red. As only three of them were noted in the three other monkeys, the majority of homoplases in this network were unique to the animal. As there is little evidence for hot spots of substitution, the extensive networking for monkey i-963 must reflect mainly the effects of recombination.
As diversity is needed to identify recombination, it is not surprising that monkey i-963, which showed the greatest degree of variation, also showed the greatest extent of networking. These observations concord well with the 10-fold higher rate of HIV-1 recombination over mutation. As soon as a mutation is made during provirus formation, it will be spread to other genomes via reassortment and crossing over in the next round. When counting the number of mutations in a data set, recombination will inflate the number of discrete substitutions. A trivial and hypothetical example was illustrated in Fig. 1. Without recombination, three substitutions were necessary to connect all sequences in the set. With recombination, only two substitutions were necessary. What is the situation for a real data set? Taking the i-963 network as an example, the minimum path length connecting the 33 SIV V1V2 env sequences in the network was 64 substitutions long (Fig. 4A), of which 11 were homoplasies (shown in red). If recombination is allowed, then a number of sequences are clearly recombinants, for example, 110, 119, 202 and 208 to cite a few (Fig. 4B). In this representation, putative recombination pathways are shown by converging orthogonal arrows (shown in red). The minimum path length over the same network is reduced to 53 substitutions and ≥12 putative crossovers (red) when recombination is factored in; this can be viewed as an inflation of 45 % when ignoring recombination. The most striking examples of recombination are exemplified by B10-4 and B7-4. Both could be explained by a single recombination between sequences en route for B9-4* and B7-2*. Indeed B7-4 is a ‘pure’ recombinant without further additional point mutations with respect to putative ancestors, B9-4* and S (Fig. 6). B10-4 is a recombinant that could have had missing intermediates Q and R as ancestors (Fig. 6).

For branches not in a network, it may not be concluded that they represent a series of consecutive point mutations. Firstly, the recombination rate is 10-fold greater than the point mutation rate, meaning that in all probability, recombination must be suspected. Secondly, some intermediates may either have escaped sampling or have become extinct. Referring back to Fig. 5(A, B), the unique branch lengths connecting sequences B9-4* (Fig. 5A, encircled) and B7-4 (Fig. 5B, encircled), respectively, to their networks
cannot be prima-facie evidence of a series of point mutations but artificial lineages arising from insufficient sampling. In contrast, the branch of five substitutions leading to B7-2* (Fig. 5B) remained a branch in Fig. 6. It may be supposed that with greater sampling this too might prove to be part of a network. Given that only a trivial fraction of HIV genomes in an individual are sampled, lone branches would seem inevitable.

Fig. 5. Minimum path lengths for HIV-1 env V1V2 sequences derived from laser microdissected single nuclei. (A) SplitsTree network (fit = 100) for nine sequences from cell B9. Bold type identifies the minimum path length without and with recombination. (B) A comparable network for nine sequences from cell B7 (fit = 100). Ignoring recombination inflated the path lengths by 5/31 (16%) and 3/31 (9%) for these two cases, respectively. Two lone branches leading to sequences B9-4* and B7-4 are encircled as they are, in fact, part of a fuller network (see Fig. 6).

Fig. 6. Increasing the number of taxa in a network reveals further recombination. By combining sequences from different cells from the same patient, 13 taxa encoding 45 mutations could be built into a broad network (fit = 100). Without recombination, the network included 14 homoplasies (top). Allowing as few as five crossovers, the minimum path length was 31 mutations (bottom). Some hypothetical intermediates that could have recombined are noted as Q, R and S. For this network, ignoring recombination inflated the minimum path length by 14/31 or 45%. Sequences B9-4* and B7-4 are encircled (see Fig. 5).
When may mutation generate extensive networking? Mono-
tonous G→A hypermutation of retroviral genomes occurs
during negative-strand DNA synthesis under conditions of
an unbalanced dTTP and dCTP concentration (Martinez
et al., 1994; Vartanian et al., 1997). Overall, G→A sub-
stitution frequencies of up to 0.3 per G residue are possible
(Vartanian et al., 2002). As an example, a small segment of
a set of hypermutated HIV sequences generated in vitro is
shown in Fig. 7(A). SplitsTree analyses of the sequences
generated extensive networks (Fig. 7B). Thus, when the
point mutation rate is very high, ~10^4-fold greater than
the normal rate of ~0.25 x 10^-5 per base per cycle, as is the
case here, extensive networking due to point mutation is
possible. In turn, this indicates that at normal mutation
rates, recombination is the major mechanism underlying
network formation.

DISCUSSION
Rampant recombination among HIV genomes provides
a challenge to genetic analysis if only by its tempo and
magnitude in vivo. Once a little diversity has been generated,
virtually every provirus will be genetically distinct. Because
the m.o.i. is generally high in culture, as it is in vivo,
recombination is also an issue when culturing HIV ex vivo.
Only with a short-term culture derived from an infec-
tious molecular clone will the effects of recombination be
minimized. Purifying selection must be weeding out the
effete recombinant genomes, just as it does deleterious point
mutations. However, even this situation is not simple.
Firstly, a cell harbouring two defective, yet transcriptionally
active, proviruses can produce heterokaryons, which, in
the next round, can recombine to yield viable progeny.
Secondly, given that the average provirus copy number per
cell is three to four in vivo (Jung et al., 2002), even if a
recombinant is functionally viable, in the next round it may
recombine with a different genome to yield yet another
recombinant of unpredictable fitness.

As can be seen when scoring the number of mutations,
ignoring recombination inflates the minimum path length
connecting sequences in any data set (Figs 1, 4–6). Among
the examples shown there was an overestimation of between
10 and 45%. This conclusion is derived empirically. It is
intriguing that the upper value is close to the 50% predicted
from the simplest conceptual example shown in Fig. 1(C).
Hence, from the temporal standpoint, intrapatient HIV
lineages may be generally younger than might be otherwise
anticipated.

When factoring in recombination, the minimum number
of point mutations is simply the number of discrete sub-
stitutions across all variable sites in the data set. This
obviously reduces statistical power when analysing muta-
tions in terms of non-synonymous and synonymous
substitutions (Zanotto et al., 1999). Furthermore, by
ignoring both phylogeny and recombination, two-by-two
sequence comparisons used to compute ratios of non-
synonymous to synonymous substitutions and thence
infer positive selection are equally erroneous (Evans et al.,
1999; Price et al., 1997). The statistical analysis of codon
polymorphisms has received considerable attention of late
(Nielsen & Yang, 1998; Yamaguchi-Kabata & Gojobori,
2000; Yang et al., 2000). In fact, sequence data in Fig. 2
corresponded to nef codons that were identified in a PAML

![Fig. 7. Extensive network formation when the mutation rate is very high. (A) A set of G→A hypermutated HIV-1 U5
sequences generated in an in vitro reaction with a highly biased [dTTP]:[dCTP] ratio (Vartanian et al., 1997). The overall
frequency of mutation was ~0.1 per G residue, which is in the order of 10^4-fold greater than the overall HIV-1 mutation rate
(Mansky & Temin, 1995). As G→A transitions are produced co-incident with DNA synthesis, even if recombination occurred,
it would not produce networks for the RNA templates that were error free. (B) Network formation can be attributed solely
to multiple point mutations. Even in this situation, sequences A07, A09, A12, A14, A15 and A18 were removed to obtain a fit of
100. For the full set of 18 taxa the fit was ~9.]
analyses as being under statistically significant positive selection (Nielsen & Yang, 1998; Yang et al., 2000; data not shown). The minimum path length connecting all 16 taxa in Fig. 2(D) requires 17 substitutions, among which there are seven homoplasies (Fig. 8A). Allowing for recombination, the minimum of substitutions necessary is 11 (Fig. 8B), the number of discrete substitutions in the 16 taxa data set (Fig. 2A). In short, each substitution could have occurred just once and spread by subsequent recombination. It would seem that inferring positive selection from the statistical analysis of polymorphic codons may be invalid if recombination is taken into account.

With so much recombination going on, why doesn’t multidrug-resistant HIV come up earlier? There are three restrictions limiting the emergence of multidrug resistance. The first is a truism: less replication means less recombination and less mutation. Secondly, HIV replication is highly localized (Blancou et al., 2001; Cheynier et al., 1994, 1998), occurring in discrete sites throughout the body. The massive destruction of infected cells (Pelletier et al., 1995; Wain-Hobson, 1993) and virus (Ho et al., 1995; Wei et al., 1995) means that mutants do not mix freely. Genomes encoding drug resistance mutations can exist in separate sites but, until they find themselves in the same cell, recombination cannot accelerate the process. In this context, it would be interesting to determine the mean provirus copy number per cell in individuals under HAART (highly active anti-retrovirus therapy). The final factor may well be the least important. In principle, it should be easier to generate a gp41 fusion inhibitor/protease inhibitor double mutant where the crucial bases conferring resistance are ~5 kb apart, rather than a protease/RT double mutant where the sites are separated by <1 kb. However, as the recombination rate is so high, this is unlikely to be a major restriction.

This situation where the recombination rate exceeds the mutation rate is not without precedent. For numerous bacteria, including Neisseria species, Streptococcus pneumoniae and Staphylococcus aureus, evolutionary change at neutral loci is more likely to occur by recombination than by point mutation (Feil et al., 1999, 2000, 2001). Splits decomposition analyses have revealed extensive networks (Alber et al., 2001; Holmes et al., 1999a; Smith et al., 1999). Estimates of the recombination to substitution rates may be as high as 100:1 (Feil & Spratt, 2001), although precise recombination rates are lacking. Yet, as the point mutation rates for RNA viruses and retroviruses are far higher than those of bacteria, the sheer profusion and tempo by which retrovirus recombinants arise distinguish them from their bacterial counterparts.

In conclusion, HIV and SIV sequence polymorphisms in a data set are strongly influenced by recombination. When recombination is catered for by the SplitsTrees program, the

http://vir.sgmjournals.org
minimum number of substitutions in a data set necessary to explain sequence complexity is the number of unique mutations. Accordingly, a sequence set is younger than would have been thought previously.

**ACKNOWLEDGEMENTS**

We wish to thank Drs D. Haydon for PAML analyses, N. Müller-Lantzsch for continued support and J. Diez for helpful discussions. This work was supported by grants from the Institut Pasteur, the ANRS and the Deutsche Forschungsgemeinschaft.

**REFERENCES**


Network analysis of HIV/SIV sequence sets


