Consistent risk group-associated differences in human immunodeficiency virus type 1 \(vpr\), \(vpu\) and V3 sequences despite independent evolution

Carla L. Kuiken,1,* Marion T. E. Cornelissen,1 Fokla Zorgdrager,1 Susan Hartman,1 Adrian J. Gibbs2 and Jaap Goudsmit1

1 Human Retrovirus Laboratory, Department of Human Retrovirology, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands and 2 Research School of Biological Sciences, Australian National University, Canberra, Australia

Human immunodeficiency virus type 1 \(vpr\), \(vpu\) and V3 sequences from 15 homosexual men and 19 intravenous drug users in the Amsterdam Cohort studies were analysed. Previously, we reported that V3 domains of viruses from drug users are distinguishable from those of homosexual men on the basis of two silent mutations. Phylogenetic analysis of \(vpr\), \(vpu\) and V3 shows that differences in all three regions correlate with risk group. Two positions in both \(vpr\) and \(vpu\) were found to differ significantly between the risk groups. The distinguishing positions were confirmed for sequences from 11 Scottish and four German samples. The three regions show relatively independent evolution patterns; they resulted in different phylogenies, the only stable clustering being based on the risk group distinction. Pairwise differences between sequences of the genes were moderately correlated (around 0.30). Surprisingly, when only silent changes were counted, the correlations dropped almost to zero, indicating that the evolution towards independence was more advanced in the silent than in the non-silent positions. This suggests that selection at the amino acid level is not the primary driving force for the independent evolutionary behaviour of the genes. Recombination, combined with restrictions on certain amino acids because of epistatic interactions between the genes, could be an alternative explanation of this phenomenon.

Introduction

\(vpr\) and \(vpu\) are two of the accessory genes of human immunodeficiency virus type 1 (HIV-1). They are located just 5' of the envelope gene, with \(vpu\) partially overlapping it. The proteins are dispensable \textit{in vitro}, since laboratory strains can be defective for either gene (Dedera et al., 1989; Cohen et al., 1988). However, \textit{in vivo} both are usually intact, and probably have a role in the virus life cycle (Westervelt et al., 1992; Klimkait et al., 1990).

The \(vpr\) gene encodes a 96 amino acid protein. It is also found in HIV-2 and some of the simian immunodeficiency virus (SIV) isolates known. Moreover, similarities between the \(vpr\) and \(vpx\) genes in the HIV–SIV group suggest that \(vpx\) may actually have arisen from a duplication of the \(vpr\) gene. Large quantities of Vpr protein are packaged in the virion, suggesting that it may have a role in the early stages of infection, before integration of the DNA (Yuan et al., 1990). Several studies have shown that the protein is involved in virus production in macrophages, but not (or less pronounced) in T cells (Balliet et al., 1994; Ogawa et al., 1989; Cohen et al., 1990; Yuan et al., 1990; Connor et al., 1995; Balotta et al., 1994). Vpr has the same function in HIV-2 (Hattori et al., 1990). Recent evidence indicates that it has a role in the regulation of latent infection (Levy et al., 1994) through the control of activation of both CD4 T cells and macrophages (Levy et al., 1993).

The \(vpu\) gene has no known homologue in HIV-2 and SIV. Vpu protein is synthesized in infected cells, but not packaged in the virion (Strebel et al., 1989). Available evidence indicates that Vpu is mainly associated with particle maturation and budding (Strebel et al., 1989; Klimkait et al., 1990; Terwilliger et al., 1989). It has also been shown that independent of this mechanism, Vpu is involved in CD4 degradation (Willey et al., 1994; Chen et al., 1993; Raja et al., 1994; Buonocore et al., 1994), which in turn affects HIV-1 particle production because...
CD4 molecules capture gp160 envelope precursors on their way to the cell surface.

A study of V3 sequences from a group of HIV-1 infected people in the Amsterdam Cohorts of homosexuals and intravenous drug users has revealed that the virus variants that were found in the homosexual men and haemophiliacs could be distinguished by two mutations from variants found in drug users (Kuiken et al., 1993). These mutations were both silent, and one of them was found with very high consistency. Subsequently, in a collection of over 100 homosexual seroconversion sequences the ‘drug user variant’ was found only once (Kuiken et al., 1996). In a second study, using only published V3 sequences from homosexuals and drug users (and their sexual partners) from other countries in Europe, and judging by the most conserved distinctive mutation drug users and homosexuals from Germany (Hamburg) and the United Kingdom (Scotland: Edinburgh and surrounding regions) could also be reliably distinguished on the basis of the V3 sequences. Sequences obtained from samples from Italian drug users indicated that the epidemic there consists of a mixture of the two virus variants (Kuiken & Goudsmit, 1994).

As the most consistent mutations were silent, we assumed that the distinction was epidemiological in origin, and represented a founder effect: the drug user population appears to have become infected by one accidental variant, which differs from the other variants by a number of mutations. To further investigate this hypothesis, we have sequenced other regions of the HIV-1 genome.

In this paper we report the results of an analysis of the vpr and vpu sequences from 34 HIV-1 infected participants in the Amsterdam Cohorts of homosexuals and drug users, supplemented with those from 11 Scottish and four German samples. The data show that the distinction between the risk groups is present in both vpr and vpu.

Methods

Origins of sera. For sequencing of vpr, vpu and V3, serum samples were obtained from 15 homosexual men and 19 intravenous drug users, all participants in the Amsterdam Cohorts (Van den Hoek et al., 1988; De Wolf et al., 1987). All drug user samples were obtained 3 months to 1 year after seroconversion, and form a random selection from all participants who seroconverted between 1986 and 1991. Samples from homosexuals were either seroconversion samples (seroconversion dates between 1985 and 1988) or the first seropositive sample for those who were already HIV infected when they entered the Cohort study (sampling date 1984 or 1985).

In addition, four German samples [two from homosexuals, two from drug users; a subset of those previously discussed in Kuiken & Goudsmit (1994)] and 11 Scottish samples [five from homosexuals and six drug users and sexual contacts of these; unrelated to the samples discussed in Kuiken & Goudsmit (1994)] were used.

RNA isolation and sequencing. Genomic RNA was isolated from 100 µl of serum or plasma according to Boom et al. (1990). The RNA was converted to cDNA, which was subjected to a nested PCR. For V3, the method was as described previously (De Wolf et al., 1994). Primers used for vpr were 5'VPR-1 (5' GATCTCTCACATCTGGCCT 3') (I = inosine) and 3'VPR-4 (5' CTCTTTCTGATCATGAGAT-GCC 3'), final MgCl2 concentration 2.4 mM, for the first PCR, and 5'VPR-2-Sp6 (5' ATTTAGGTGACACTATGCGCTTGCA- GTTTAIGAA 3') and 3'VPR-3-T7 (5' TAATACGACTCACTATAGT- AAGGAAAGCAACACTTTTACATAA 3'), both at a final MgCl2 concentration of 2.4 mM, for the second PCR. For vpu the primer sets were 5'VPU-1 (5' GCATCTCCTAATGGCAAGGAAG 3') and 3'VPU-4 (5' ATATGCTTTAGCATCTGACCAAATA 3') for the first PCR (final MgCl2 concentration 3.0 mM) and 5'VPU-2-Sp6 (5' ATTTAGGTGACACTATGCGCTTGCA- GTTTAIGAA 3') and 3'VPU-3-T7 (5' TAATACGACTCACTATAGT- CATTTATCAACGTAGCCACAGAC 3') (final MgCl2 concentration 3.5 mM) for the second PCR. In all cases, sequences were rejected if they contained more than 1% heterogeneous or illegible positions after two sequencing attempts. A minimum of six clones was then made by cloning the PCR product in an AT cloning system. In three Scottish samples, two subpopulations were clearly present; in these three cases, one subpopulation was randomly selected after it was established that the relevant nucleotide positions were the same in both populations. Sequencing was done using an automated sequencer (Applied Biosystems).

Sequence analysis. Alignment was done by eye, keeping codons intact. Phylogenetic analysis was done using the program MEGA (Kumar et al., 1994). The neighbour-joining method (Saitou & Nei, 1987) was used to construct phylogenetic trees based on distances calculated using the Kimura 2-parameter estimation method (Kimura, 1980). The variability within sets of sequences was calculated using Hamming distances between all pairs of sequences (Hamming, 1986), expressed as percentages of the total sequence length (including gaps). These were averaged over all sequences in a set to calculate the variation. Pearson correlation coefficients between the Hamming distances on two genes/regions were used as a measure of co-variation. Positions containing gaps or ambiguities were excluded from all analyses.

Additional multivariate analysis of the sequences was done using the program PCOORD (Higgins, 1992). Numbers of silent and non-silent changes were calculated according to Nei & Gojobori (1986) using the program PCoord (Higgins, 1992). For these calculations, the vpr and vpu sequences were truncated to span only the open reading frames of the genes.

Assessing the strength of risk group association. To evaluate the strength of the association between the positions and the risk groups, the following method was used. For each position in the sequences, a pairwise comparison matrix was constructed consisting of zeros and ones, indicating for each pair of sequences whether they contained identical or different nucleotides at that position (position matrix). A second matrix contained a similar same/different index indicating whether each pair of sequences came from the same or a different risk group (risk group matrix). A Pearson correlation coefficient was then calculated between corresponding cells of the position matrix and the risk group matrix, yielding a position-wise indication of the association between sequence differences and group membership.

The correlation coefficient, in the present case, was calculated on the pairwise differences, which are not statistically independent. Although this does not alter its interpretation, the associated significance test loses its validity. Instead, a simulation method was employed to assess
the correlation that one can expect to find accidentally, when no real
association is present. The sequences were randomly divided into two
groups 100 times, and for all positions in all three regions the
correlation was calculated; these never rose above 0.33. Therefore
actual correlations larger than 0.33 were considered significant. The
descriptive level for the resulting criterion is difficult to calculate due
to dependencies in the data, but the criterion is considerably stricter than
1%. Because the criterion is so conservative, difficulties due to multiple
testing are avoided.

Results
Independent variation of the genes
In both \textit{vpr} and \textit{vpu}, variation was lower than in V3. In \textit{vpr}, the variation in the drug user group (mean 3.5\%,
SD 1.3) was slightly lower than that in the homosexual
group (mean 4.3\%, SD 1.3). In \textit{vpu}, no difference between
the groups was seen (drug user: mean 6.4\%, SD 1.7;
homosexuals: mean 6.0\%, SD 1.5). For the drug user
group the V3 variation was 8.6\% (SD 2.3) versus 6.5\%
(SD 1.9) in the homosexual group. In the drug user group,
there were two samples that contained very similar virus
variants (13072 and 15038: < 1\% difference in all genes).
An epidemiological link between these persons could not
be established, but the samples may have been a few
transmissions apart. Ratios of silent to non-silent
changes (calculated per number of silent/nonsilent sites)
were 1.12 for V3, 3.3 for \textit{vpr} and 1.8 for \textit{vpu}. None of the
sequences contained any mutations that generated stop
codons.

To investigate the evolutionary differences between
the genes, we looked at the extent to which differences
between viruses in one gene or region were associated

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig1a.png}
\caption{Fig. 1(a). For legend see page 787.}
\end{figure}
Fig. 1(b). For legend see facing page.

with differences in another. This co-variation was calculated by correlating the numbers of pairwise differences between the sequences of one gene with those between the corresponding sequences of the other. The genes do not co-vary strongly: the Pearson correlation coefficient between the pairwise distances in vpr and vpu was 0·36; between vpu and V3, 0·31; and between vpr and V3, 0·32. Correlations between the numbers of silent changes were 0·05 (vpr and vpu), 0·20 (vpr and V3) and 0·07 (vpu and V3). These numbers are much lower, which is surprising because the numbers of silent changes presumably reflect the evolutionary distances more accurately, being less subject to various pressures. [However, silent changes are not always selectively neutral (e.g. Coffin, 1992).] Absolute numbers of silent changes were large enough for meaningful comparison (vpr: median 4, range 0–11; vpu: median 3, range 0–9; V3: median 4·5, range 0–13·5). The significance of the
Fig. 1. Phylogenetic trees based on the \( vpr \) (a), \( vpu \) (b) and V3 (c) regions. Double lines indicate branches leading to HIV-1 sequences from intravenous drug users, single lines those leading to HIV-1 sequences from homosexual men.

Difference between the correlations of the numbers of silent changes could not be directly tested, however, because again they are based on pairwise comparisons. A very conservative estimate (using Fisher's \( r \)-to-\( Z \) transformation with \( N \) set equal to the number of sequences) indicates that the 95\% confidence interval around the 0.20 correlation has a width of 0.24, making the difference between 0.20 and 0.05 non-significant.

This phenomenon of relatively independent variation is also reflected in the phylogenetic analyses (Fig. 1). Analyses based on the three genes yielded different phylogenies. Only the I3072/I5038 pair consistently clustered together. Several small clusters with moderate to high bootstrap values (> 0.60) formed in phylogenetic analyses based on all three genes [e.g. H0008/H0617 (0.66) in \( vpu \), and H0569/H0617 (0.92) and I0090/I5032 (0.96) in \( vpr \)]. Some of these clustered pairs appeared in more than one of the trees, but none was present in all (except the related viruses mentioned above). Bootstrap values in the analyses based on individual genes were generally low. This instability was reflected in an analysis based on all three genes together, which yielded a tree with three stable groupings (not shown): one with the related drug user sequences (100\%), one with the two drug user sequences with homosexual characteristics (see below, 100\%), and the separation between the two risk groups (86\%). Two drug users who had sequences with 'homosexual' characteristics formed a separate sub-cluster within the homosexual cluster in all trees except the one based on \( vpu \) sequences.
Fig. 2. vpr, vpu and V3 regions (based on the consensus sequence) depicting the mutations characteristic of homosexual and drug user virus populations. Letters underneath the alignments indicate amino acid differences that result from the mutations: in X/Y, X indicates the amino acid associated with the homosexual variant, Y the one in the drug user group. If no translation is given, the change is silent. In the vpu sequence, translations based on both Epu and env reading frames are shown. The numbering of the regions is relative to the starting point of the sequences used in this study. * Amino acid changes in the Env reading frame.

**Risk-group related differences between the sequences**

Phylogenetic analysis clearly separated the two risk groups in trees based on each of three regions (Fig. 1). The drug user group contained two samples that yielded virus which did not conform to the risk group classification (I3057 and I0070). These two samples mostly clustered together and were grouped with the homosexual samples. The two groups could also be distinguished on the basis of each region using principal coordinate analysis (Higgins, 1992), again suggesting that some positions differed consistently between the two groups.

To evaluate the strength of the association for each position, Pearson correlations were used (see Methods). As noted previously for V3, associations varied from almost perfect to very weak. In Table 1, all correlations over 0.2 are included. In the Dutch group, correlations that are significant according to the criterion defined above are marked with an asterisk. In both vpr, vpu and V3, two positions were found to be significantly associated with risk group. In V3, both changes were silent. In vpr, one of the two changes was non-silent (Q/R). In vpu, both significant changes were silent in the vpu reading frame (one was located in the stop codon). They are both located in the section that also encodes...
Risk group differences in HIV-1 vpr and vpu

Table 1. Pearson correlations between nucleotides present at each position and risk group membership for selected positions in vpr, vpu and V3

<table>
<thead>
<tr>
<th>Position</th>
<th>Dutch samples (n = 34)</th>
<th>Foreign samples (n = 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V3-15</td>
<td>0.36*</td>
<td>0.36</td>
</tr>
<tr>
<td>V3-41</td>
<td>0.33</td>
<td>0.30</td>
</tr>
<tr>
<td>V3-60</td>
<td>0.33</td>
<td>0.22</td>
</tr>
<tr>
<td>V3-129</td>
<td>0.73*</td>
<td>0.74</td>
</tr>
<tr>
<td>V3-208</td>
<td>0.33</td>
<td>0.15</td>
</tr>
<tr>
<td>V3-230</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>vpr-48</td>
<td>0.77*</td>
<td>0.69</td>
</tr>
<tr>
<td>vpr-70</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td>vpr-245</td>
<td>0.33</td>
<td>0.35</td>
</tr>
<tr>
<td>vpr-272</td>
<td>0.77*</td>
<td>0.76</td>
</tr>
<tr>
<td>vpu-189</td>
<td>0.30</td>
<td>0.37</td>
</tr>
<tr>
<td>vpu-222</td>
<td>0.71*</td>
<td>0.48</td>
</tr>
<tr>
<td>vpu-250</td>
<td>0.33</td>
<td>0.17</td>
</tr>
<tr>
<td>vpu-261</td>
<td>0.48*</td>
<td>0.39</td>
</tr>
<tr>
<td>vpu-274</td>
<td>0.30</td>
<td>0.35</td>
</tr>
</tbody>
</table>

* Significant correlation (see Results).

Env; in the env reading frame, both changes were non-silent (vpu-222, R/K; vpu-261, S/N).

Other northern European samples

Previously, we reported that samples from Germany and Scotland showed the same risk group-related distinction as the Dutch ones. vpr and vpu sequences from 11 Scottish samples (five homosexuals and six drug users) and four German samples (two homosexuals, two drug users) were analysed to assess whether the same pattern was also found in these genes. One sample from a Scottish homosexual was negative for vpr despite repeated isolation attempts, so for vpr only four Scottish homosexual samples were available.

Two Scottish samples from heterosexual women were assigned to the drug user group, because the Scottish epidemic is most severe among drug users (Brettle et al., 1987; Peutherer et al., 1985), and because in northern Europe the largest risk factor for women who do not use intravenous drugs is sexual contact with a drug user. Both samples yielded a V3 sequence that was unmistakably of the drug user type. Judging by the two significant positions in the V3 region, one Scottish drug user sample (IUK60), from a male drug user, yielded a sequence with clear homosexual characteristics. All other samples fell into the correct group based on V3 position 129 and vpr position 272. For vpr-48, one homosexual sample deviated from the expected pattern; for V3-15, two drug user samples had a different nucleotide than expected. Consistency in vpu was lower: for both positions (222 and 261), three of seven homosexual samples adhered to the Dutch pattern. All drug user samples except IUK60 showed the expected pattern in vpu-222, and five of these did so in vpu-261.

Discussion

The finding that the intravenous drug user population in The Netherlands appeared to be infected with an HIV variant that is distinct from the one found in homosexuals and haemophilia patients presumably reflects an epidemiological founder effect: the ancestral virus infecting the first drug user victim of the HIV epidemic had several unique mutations which, being selectively neutral or nearly so, remained present in the quasispecies descending from it. The two mutations found in V3 were both silent, and therefore presumably have little or no effect on the fitness of the virus. Furthermore, haemophilia patients without exception carried the ‘homosexual’ variant, which, for the Netherlands (where blood donations are given voluntarily and without financial compensation) is compatible with this epidemiological model. Later on, the same distinction was found to be present in other European countries as well, namely Germany and Scotland (with high consistency) and Italy (where four out of nine drug user samples contained the ‘homosexual’ variant) (Kuiken & Goudsmit, 1994). It is possible that all these drug user infections are epidemiologically related.

The V3-129 mutation is found in only four out of 47 sequences listed in the Los Alamos compendium alignment (Myers et al., 1994), all of these from the United States: HIV-1 BAL, a macrophage-tropic isolate from a New York patient, and three from intravenous drug users from Baltimore (HIVUS712, -715 and -716). These four exclusively also have vpu-261, and three have V3-15. The mutation in vpu-222, however, is found in 13 of the sequences listed in the alignment. vpr sequences are not available for these isolates. In a set of 30 V3 sequences from the United States (NIAID), three out of four drug user sequences had the V3-129 mutation, as well as two heterosexual transmission cases from Puerto Rico. Other anecdotal evidence suggests that the Dutch drug user variant is also associated with drug use in the United States, though less consistently than in mid-Europe: based on V3-129, it is present in three of 29 sequences from control subjects in the Florida dentist case, one of these from a drug user (Ciesielsky et al., 1992); it is found in the only drug user in a set of three index persons from New York City and their partners (Zhu et al., 1993), and again in a drug user in a set of four United States patients studied by Haigwood et al. (1993). In a set of 11 samples
from Baltimore drug users, five adhered to this pattern (V. V. Lukashov and others, unpublished).

The present study was undertaken to investigate whether a risk group-related distinction found in V3 could also be found in vpr and vpu. The results show that this distinction is clearly present in both these genes. This is consistent with a founder effect: accidental differences present in the variant that was the founder of the epidemic in drug users would be scattered throughout the genome. Most of the positions found to discriminate between the risk groups in the Dutch dataset could be confirmed independently using sequences from a Scottish and a German sample set.

A distinction between viruses obtained from drug users and homosexuals has been found in other studies as well. A recent study by Albert et al. (1994) showed a clear and consistent distinction in a set of reverse transcriptase sequences. Our analysis of these sequences showed that three positions in these reverse transcriptases are consistently different between the risk groups (not shown). Analysis of a small set of 10 Swedish V3 sequences suggests that, although again the sequences differ between risk groups, this distinction is not the same one that is found in the Dutch, Scottish and German viruses (not shown). A Madrid study showed a distinction to be present in gag as well as env (fusion domain) sequences (Rojas et al., 1994), but again, based on V3 sequences, this distinction seems to be different from the Dutch/German/Scottish distinction (C. López-Galindez, personal communication).

Combining all these results, we can conclude that it is likely that several different founder effects are present in the European HIV epidemic. The fact that the ‘Dutch distinction’ is also present in part of the United States drug user population suggests that this particular variant may have originated from the United States, and then having crossed the Atlantic subsequently spread among Dutch, German and Scottish, and some Italian drug users. A sizeable proportion of Amsterdam drug users are German or Italian; a study among Edinburgh drug users showed that several of them had shared needles in Amsterdam (Brette et al., 1987). The source of the homosexual infection could be a different one. At present, data are too scarce to draw more inferences about the origin of the other founder effects, or about the ways in which the virus has spread through Europe. The hypothesis that the northern European drug user variant is derived from a United States variant is being investigated at present.

The three regions investigated here appear to have evolved more or less independently. The correlations between pairwise differences were small, and the only clustering that was reinforced when the sequences were analysed together was the homosexual/drug user split. This instability may be caused by the fact that the sequences are rather short. Also, phylogenetic trees in populations of closely related sequences typically have a star shape, with short internal branches and long external ones (Holmes et al., 1995); this results in very unstable trees, where removal of one or two sequences often results in a different structure.

Although the data have to be interpreted with care due to the above considerations, alternative explanations can be suggested. One is that the genes are evolving in different directions because of different selection pressures. However, this model is hard to reconcile with the fact that the correlations between the pairwise distances drop almost to zero when they are based on only silent mutations. Although silent positions may also be under some selective pressure in coding regions (Coffin, 1992), it is hard to explain why they would be more diverged than non-silent ones. This finding may reflect some form of epistatic interaction between the genes, whereby certain amino acids encoded by one gene necessitate or facilitate the presence of specific amino acids encoded by the other.

Independent evolution in different genes has been reported before in sequences from one patient (Pedroza Martins et al., 1991; Delassus et al., 1991; Vartanian et al., 1991; Howell et al., 1991), and frequent recombination events have been suggested as an explanation. Increasingly, reports suggest that recombination occurs frequently in HIV-1 (Monken et al., 1995; Diaz et al., 1995; Zhu et al., 1995; Robertson et al., 1995). It is possible that the independent evolution found in the present study is a reflection of recombination, where the risk group distinction is preserved because the recombining viruses generally are from the same risk group and share the same genotype. In this light, one might expect that correlations between distances would be inversely proportional to the distance between the genes (and thus, to the probability of recombination). This does not appear to be the case in the present dataset, but more data are needed to draw this conclusion more firmly.

We would like to express our gratitude to Dr Michael Schreiber of the Bernhard Nocht Institut für Tropenmedizin in Hamburg, Germany, and to Dr Andrew Leigh Brown of the University of Edinburgh, Scotland, for generously donating serum samples that were used for this study; to Dr Edward Holmes of the University of Oxford, for thoughtful comments; to Drs Jan Albert, Thomas Leitner and Cecilio López-Galindez, for sharing unpublished results; and to Drs Tuofu Zhu (Aaron Diamond Institute, New York) and Carol Ciesielsky (Centers for Disease Control, Atlanta) for unearthing the risk groups associated with published sequences.

References

endoplasmic reticulum: requirement of both the anchor and cytoplasmic domain of CD4. Virology 204, 357–366.


(Received 14 September 1995; Accepted 14 December 1995)