Intrapatient Diversity and Its Correlation with Viral Setpoint in Human Immunodeficiency Virus Type 1 CRF02_A/G-IbNG Infection

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:26881959

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Intrapatient Diversity and Its Correlation with Viral Setpoint in Human Immunodeficiency Virus Type 1 CRF02_A/G-IbNG Infection

Indu Mani,1 Peter Gilbert,2 Jean-Louis Sankale,1 Geoffrey Eisen,1 Souleymane Mboup,3 and Phyllis J. Kanki1*

Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts1; Statistical Center for HIV/AIDS Research and Prevention, Fred Hutchinson Cancer Research Center, Seattle, Washington2; and Laboratoire de Bactériologie-Virologie, Université Cheikh Anta Diop, Dakar, Senegal3

Received 9 April 2002/Accepted 22 July 2002

The human immunodeficiency virus type 1 (HIV-1) viral setpoint during the disease-free interval has been strongly associated with future risk of disease progression. An awareness of the correlation between viral setpoint and HIV-1 genetic evolution over time is important in the understanding of viral dynamics and infection. We examined genetic diversity in HIV-1 CRF02_A/G-IbNG-infected seroincident women in Dakar, Senegal; determined whether a viral setpoint kinetic pattern existed for CRF02_A/G-IbNG during the disease-free interval; and correlated viral load level and diversity. Samples were drawn during the disease-free interval from consenting CRF02_A/G-IbNG-infected, antiretroviral therapy-naïve female commercial sex workers in Dakar, Senegal. Based on sequential plasma RNA values, low and high viral setpoint groups were established. Intrapatient diversity and divergence over time was determined from earlier and later time point DNA samples from each person. Most individuals followed the viral setpoint paradigm. For each 1|−|log10 copy/ml of plasma increase in viral load, intrapatient diversity increased by 1.4% (P = 0.028). A greater diversification rate was observed in the high viral setpoint group than in the low viral setpoint group (P = 0.01). Greater nucleotide (P = 0.015) and amino acid (P = 0.048) divergences and a greater nucleotide divergence rate (P = 0.03) were found in the high viral setpoint group. There was no difference between the groups in the ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site. The greater intrapatient diversity, divergence, and diversification rates observed in the high viral setpoint group supports the notion that diversity is driven by cycles of viral replication resulting in accumulated mutations. Recognizing diversity potential based on viral load levels in individuals may inform the design of vaccines and therapies.
Maryland (22). The greatest diversity and divergence was noted in rapid progressors, followed by moderate progressors and finally by nonprogressors, and diversity had a negative correlation with CD4+ cell counts. This suggested that greater diversification was proportional to the rate of progression. McNearney et al. noted similar results for the env V3 region in a smaller study (25). These data supported the Nowak model, which assumes that viral replication, with the resulting accumulation of mutations, drives diversity and therefore antigenic variation, with the immune system controlling disease progression as long as it is able to recognize and control viral replication and emerging variants (30). This model suggests that host immune selection pressures are not driving diversity but rather that the diversity generated by viral replication is driving the disease. Many rounds of viral replication (and therefore high viral load levels) would eventually overwhelm the immune system with new viral variants and result in rapid progression to disease.

Wolinsky et al. studied the V3-V5 region of the env gene in men from the MACS cohort (47). They concluded that, for those in the rapid progressor group, there was little selective pressure for change and that therefore less diversity was evident in this group than in either the moderate or the slow progressor group. Quasispecies diversification and divergence were strongly correlated with a favorable clinical outcome, and this suggested that diversity driven by an effective host immune response resulted in significant pressure on the virus to evolve. Similar results were obtained for HIV-1-infected children in a study by Ganesan et al. (11).

Multiple intersubtype recombinant viruses have been identified in HIV-1-infected individuals, and several exist as major circulating recombinant forms (23). It is estimated that, worldwide, at least 9 million HIV-1 infections are due to CRF02_A/G-IbNG (23). CRF02_A/G-IbNG was first characterized from a sample collected in Ibadan, Nigeria (31). It is subtype A in the gag gene and the majority of the env gene, subtype G in the long terminal repeats, and a mosaic of A and G in the poly gene. It is epidemic in West and West Central Africa (3, 27) and is the predominant virus circulating in Senegal (38). The CRF02_A/G-IbNG-infected individuals included in our study were antiretroviral therapy-naive at the time of sampling. We sought to directly examine the genetic diversity of HIV-1 CRF02_A/G-IbNG in seroincident women during the disease-free interval in order to identify patterns of diversity at the viral setpoint, as opposed to during AIDS or primary infection. In this study, we examined intrapatient diversity and divergence in 12 HIV-1 CRF02_A/G-IbNG-infected seroincident women categorized into high and low viral load groups in Dakar, Senegal.

**MATERIALS AND METHODS**

**Study population.** Since 1970, the government of Senegal has required that commercial sex workers register and receive periodic testing and treatment for sexually transmitted diseases. A cohort of consenting commercial sex workers attending the Institut Hygiene Social clinic in Dakar, Senegal, has been followed and studied since 1985, as described previously (38). For the purposes of this study, a subset from this cohort consisting of seroincident CRF02_A/G-IbNG HIV-1-infected individuals was identified. Nucleotide sequencing and subtype characterization were performed as previously described (38). Individuals were included in the study if (i) at least two plasma RNA measurements with a total observation time of at least 6 months were available, (ii) if the individuals were not dually infected with HIV-1 and HIV-2, and (iii) if the samples were not collected after the individual had been diagnosed with AIDS. To ensure that samples were collected after the acute stage of infection had been reached, all initial RNA viral load samples were required to be drawn at least 2 months after the estimated seroconversion date, calculated as the midpoint between the date when the last negative blood sample was collected and the date when the first positive blood sample was collected. The mean of all plasma viral load determinations was calculated for each person. A mean viral load less than 4,000 RNA copies/ml of plasma was considered low, and one of more than 30,000 RNA copies/ml of plasma was considered high. Individuals were thereby stratified into two groups, the high viral load group (n = 7) and the low viral load group (n = 5), and of course none of the individuals had progressed to AIDS at the time of sampling. Two DNA samples from each individual in the high and low viral load groups were selected, one sample collected at an earlier time and one collected at a later time, with a required observation time of more than 6 months, the purpose being to maximize the observation time (Table 1). There were no significant differences in observation period between the high viral load (median observation period, 2.4 years) and the low viral load (median observation period, 3.2 years) groups (P = 0.64 [Wilcoxon rank sum test]).

**Viral load assays.** The HIV-1 plasma RNA viral load was quantified by two branched DNA (bDNA) assays (Quantiplex version 2.0 and Quantiplex version 3.0; Bayer Diagnostics, Tarrytown, N.Y.). The viral RNA loads of 10 subjects were quantified by the Quantiplex version 2.0 assay (lower limit of detection, 50 RNA copies/ml) and those of the two remaining subjects were quantified by the Quantiplex version 3.0 assay (lower limit of detection, 50 RNA copies/ml). EDTA-preserved plasma samples were stored at −70°C prior to quantification. Studies have shown that both bDNA assay versions are strongly and significantly correlated with one another and demonstrate excellent linearity, precision, and concordance (6, 21) in studies with duplicate plasma samples. The Quantiplex version 3.0 assay appears to be more accurate that is approximately twofold higher than that provided by the Quantiplex version 2.0 assay (6, 21), though this difference is greater at lower viral load levels; that is, the differences between the assays seem to become less pronounced with increasing viral load levels. This suggests that the Quantiplex version 3.0 assay might overestimate or that the Quantiplex version 2.0 assay might underestimate at lower viral load levels (21). The two samples quantified by the Quantiplex version 3.0 assay in our study were from two individuals in the high viral load group, and the difference in the accuracy of quantification between the two versions would be minimal at that level.

**DNA amplification and sequencing by PCR.** Proviral DNA was extracted from peripheral blood mononuclear cells with the Qiagen QIAmp Blood Midi kit (Qiagen, Inc., Chatsworth, Calif.). The DNA was resuspended in Tris-EDTA buffer, and its concentration was determined by measurement of the optical density at 260 nm. One microgram of DNA was used for each PCR. For each

**TABLE 1. Profiles of two groups in study population**

<table>
<thead>
<tr>
<th>Group and patient identification no.</th>
<th>Mean raw viral load (RNA copies/ml)</th>
<th>Time interval between DNA samples (y)</th>
<th>Seroconversion date (mo/yr)</th>
<th>No. of viral load determinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low viral load</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2723</td>
<td>616</td>
<td>3.3</td>
<td>1/96</td>
<td>7</td>
</tr>
<tr>
<td>608</td>
<td>733</td>
<td>2.6</td>
<td>12/93</td>
<td>3</td>
</tr>
<tr>
<td>225</td>
<td>1,521</td>
<td>6.5</td>
<td>6/91</td>
<td>4</td>
</tr>
<tr>
<td>3065</td>
<td>2,845</td>
<td>2.3</td>
<td>3/97</td>
<td>3</td>
</tr>
<tr>
<td>1067</td>
<td>3,394</td>
<td>3.2</td>
<td>10/94</td>
<td>2</td>
</tr>
<tr>
<td>High viral load</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1196</td>
<td>35,692</td>
<td>5.4</td>
<td>4/93</td>
<td>4</td>
</tr>
<tr>
<td>915</td>
<td>43,028</td>
<td>7.1</td>
<td>1/92</td>
<td>7</td>
</tr>
<tr>
<td>394</td>
<td>68,440</td>
<td>0.75</td>
<td>4/96</td>
<td>2</td>
</tr>
<tr>
<td>2587</td>
<td>92,730</td>
<td>4.2</td>
<td>1/95</td>
<td>4</td>
</tr>
<tr>
<td>2909</td>
<td>126,423</td>
<td>2.4</td>
<td>4/98</td>
<td>3</td>
</tr>
<tr>
<td>2997</td>
<td>31,738b</td>
<td>1.4</td>
<td>9/96</td>
<td>4</td>
</tr>
<tr>
<td>2425</td>
<td>216,422b</td>
<td>1.3</td>
<td>3/96</td>
<td>4</td>
</tr>
</tbody>
</table>

* Unless otherwise indicated, viral load was determined by Quantiplex version 2.0 bDNA assay.

b Viral load determined by Quantiplex version 3.0 assay.
individual, three independent PCRs were performed under identical conditions for each sample at both time points to minimize resampling bias (42). To amplify the HIV-1 C2V3 env region, a nested PCR was performed with two sets of primers, WT1-WT2 for the first round and KK30-KK40 for the second round, as previously described (38, 48). Additionally, a single gag fragment was sequenced from each individual to ensure subtype concordance between the gag and the env genes, further verifying the classification of CRF02_AG-IbNG. A nested PCR with two primer sets (p168-p190 for the first round and p91-p92 for the second round) was performed to amplify the HIV-1 gag 3'-p24/5'-p7 region, as previously described (35). Negative controls were included in each amplification reaction. The PCR products were purified by agarose gel electrophoresis and by using purification columns (gel extraction kit; Qiagen, Inc.). All purified products were subsequently cloned into the pCR2.1 vector (Topo-TA Cloning; Invitrogen, Carlsbad, Calif.). Several clones were processed for each PCR. The plasmid preparation for the double-stranded DNA sequencing was performed by alkaline lysis with silica columns (S.N.A.P. Mini-Prep kit; Invitrogen). Samples were sequenced by dye terminator cycle sequencing with Taq polymerase (PerkinElmer Applied Biosystem Division, Foster City, Calif.) and an automatic sequencer (ABI 373; Perkin-Elmer Applied Biosystem Division). Each earlier and later time point sample was processed separately, from DNA extraction to sequence analysis.

**Sequence analysis and statistical methods.** For all generated C2V3 viral sequences, a multiple alignment was performed with the Clustal package of multiple alignment programs (Clustal X 1.64), with minor manual adjustments when necessary and complete removal of positions which contained gaps (45). Reference sequences from the HIV sequence database of the Los Alamos National Laboratory were included in certain analyses. Phylogenetic analyses were performed by the neighbor-joining method for both env and gag regions. The alignments were corrected for multiple substitutions, and reliability was estimated by 1,000 bootstrap resamplings (37). Consensus sequences were generated for each individual's earlier and later sequences by using the SeqPublish program (HIV sequence database; Los Alamos National Laboratory). DNA and protein distance analyses were executed with the DNAdist and Prodist programs by using the Phylip 3.572 package (9). Statistical analyses were performed with S-Plus version 6.0 software (Insightful Corporation, Seattle, Wash.) and another software program (version 4.0: Stata Corporation, College Station, Tex.). Viral load values for each individual were log_{10} transformed for the purpose of statistical analysis.

**Nucleotide sequence accession numbers.** All nucleotide sequences were submitted to GenBank and provided with accession numbers (AF526650 to AF526676).

**RESULTS**

**HIV-1 viral load assays in RNA plasma.** All twelve women evaluated in this study had viruses that clustered with CRF02_AG-IbNG in both the env and the gag fragments. For the low viral load group, the mean viral load ± the standard deviation was 1.822 ± 1.250 RNA copies/ml of plasma (3.02 ± 0.283 log_{10} RNA copies/ml), and for the high viral load group, the mean viral load ± the standard deviation was 87.782 ± 66.203 RNA copies/ml of plasma (4.78 ± 0.304 log_{10} RNA copies/ml). Interestingly, 11 of the 12 individuals had a mean fluctuation in viral load measurement of <0.5 log_{10} RNA copies/ml, which is indicative of a setpoint pattern in viral kinetics (5, 16). A single individual had a viral load measurement that fluctuated by 0.85 log_{10} RNA copies/ml. These results suggest that our study subjects could be characterized by a stable viral setpoint kinetic pattern, as described earlier.

**Earlier and later intrapatient diversity.** We generated and sequenced multiple C2V3 env clones from each time point. Panels a and b of Fig. 1 show two neighbor-joining phylogenetic trees including all earlier and later time point clones for each sample in the low and high viral load groups, respectively. There was a distinct difference between the high and low viral load groups in the clustering patterns of clones. Two of five low viral load individuals (subjects 2723 and 3065) showed evidence of separate clustering between the times the samples were collected, while the remainder revealed mixed earlier and later clones on branches. Four of five low viral load individuals (subjects 2723, 668, 225, and 1067) displayed horizontal branch lengths for the majority of their later clones that were similar to or shorter than those displayed for earlier clones (Fig. 1a), suggesting that later clones have fewer nucleotide changes (less divergence) than clones do. In the high viral load group, five of seven individuals (subjects 1196, 915, 2587, 2909, and 2997) revealed distinct clustering between all earlier and later clones, and there was evidence of a similar pattern emerging in a sixth individual (subject 2425) (Fig. 1b). Additionally, four of seven individuals (subjects 1196, 915, 2587, and 2997) displayed later clones with horizontal branch lengths that were greater than those of earlier clones, while the remaining individuals (subjects 394, 2909, and 2425) showed evidence of a similar trend in later clones. The most dramatic clustering and the longest clone branch lengths in later samples occurred in the two individuals with the greatest observation times, subjects 1196 and 915. Later clones from members of the high viral load group displayed longer branch lengths and therefore greater divergence those from members of the low viral load group. This phylogenetic analysis suggests that the high viral load group exhibited a higher magnitude of nucleotide divergence over time, a divergence not evident in analyses of earlier samples.

Intrapatient diversity was defined as the mean pairwise nucleotide distance, calculated by measuring the distances between all sequences from a single individual at a single time point, and was calculated at both time points for each individual (Fig. 2a). There was no significant difference in intrapatient diversity between the high (median diversity, 0.018) and low (median diversity, 0.012) viral load groups for the earlier time point (P = 0.88 [Wilcoxon rank sum test]). Therefore, both groups exhibited a similar level of intrapatient diversity at the earlier time point. The difference in intrapatient diversity at the later time point between the high (median diversity, 0.042) and low (median diversity, 0.018) viral load groups was not significant (P = 0.073 [Wilcoxon rank sum test]), although the trend was toward a greater intrapatient diversity in the high viral load group at the later time points, suggesting that high viral load subjects have a greater tendency towards development of a more diverse quasispecies population over time. There was a significant association between the log_{10}-transformed viral load and later intrapatient diversity, with the linear regression model predicting that for each 1|log_{10} copy/ml increase in mean viral load, the diversity increased by 0.014 U (95% confidence interval [CI], 0.0017 to 0.026 U; P = 0.028). Additionally, there was no significant association between the mean log_{10}-transformed viral load and intrapatient diversity at earlier time points (P = 0.54). These models suggest that the intrapatient diversity is positively correlated over the course of infection as a function of the viral load setpoint.

**Diversification and diversification rate.** In this study, diversification is defined as the change in intrapatient diversity, or the difference between earlier and later time point intrapatient diversity. The diversification rate is defined as the diversification divided by the observation time (the length of time between collection of earlier and later samples). Examination of the diversification rate revealed a significant difference, almost
FIG. 1. Low (a) and high (b) viral setpoint group neighbor-joining phylogenetic trees of multiple clones of HIV-1 env C2V3 sequences. The designations of earlier clones appear in roman type, while those of later clones appear in italics. In the clone designations, the letters E and L respectively indicate early and late time point clones, the letters A, B, and C indicate different PCRs, and the final digit indicates the clone number sequenced. The tree was constructed using 1,000 bootstrap resamplings. The number of resamplings carried out is indicated at each major branch point. Bar, 1% nucleotide divergence.
FIG. 1—Continued.
FIG. 2. Box-and-whisker plots of intrapatient diversity at earlier and later time points (a) and diversification rates (b), stratified by viral load in each case. The line in the middle of the box represents the median and the box extends from the 25th percentile ($x_{25}$) to the 75th percentile ($x_{75}$), which is defined as the interquartile range (IQR). The whiskers extend to the upper and lower adjacent values. The upper adjacent value is less than or equal to $x_{75} + 1.5\times$IQR and the lower adjacent value is greater than or equal to $x_{25} - 1.5\times$IQR. Any observed points more extreme than the adjacent values are referred to as outside values or outliers and are plotted individually.
200-fold, between the high (median diversification rate, 0.0054 [0.54%] per year) and low (median diversification rate, 0.000030 [0.0003%] per year) viral load groups ($P = 0.010$ [Wilcoxon rank sum test]). The test was also significant (120-fold difference) when time was excluded (median diversification, 1.2 and 0.01% for high and low groups, respectively; $P = 0.018$ [Wilcoxon rank sum test]). Thus, the high viral load group displayed a diversification and a diversification rate that were significantly higher than those displayed by the low viral load group (Fig. 2b). However, linear regression models did not demonstrate statistically significant associations ($P$ was 0.075 and 0.12 for the high and low viral load groups, respectively). These differences in the diversification and the diversification rate suggest the possibility that diversity is influenced by the number of cycles of viral replication and the proportional mutation rate.

**Nucleotide divergence between all earlier and later viral sequences.** Divergence can be defined as the distance from a founder sequence to a later sequence. The average of nucleotide distance between all earlier and later sequences was calculated for each individual, and this single value approximated the nucleotide divergence from earlier to later virus populations in a single individual. This divergence value was significantly different for the high (median distance, 0.054) and the low (median distance, 0.029) load subjects ($P = 0.015$ [Wilcoxon rank sum test]), with the high viral load group subjects displaying a greater level of divergence from earlier to later time points (data not shown). Linear regression models identified a positive linear correlation between the log$_{10}$-transformed viral load level and the nucleotide distance measurement, with a time-adjusted slope of 0.30 (95% CI, 0.083 to 0.52; $P = 0.012$) and nonadjusted slope of 0.28 (95% CI, 0.058 to 0.50; $P = 0.018$). A positive correlation between viral load level and the magnitude of nucleotide divergence was identified, with high viral load subjects displaying greater divergence over time.

**Amino acid divergence between all earlier and later viral sequences.** The understanding of amino acid diversity and divergence is important inasmuch as the majority of immune responses to infection occur at the protein level. The average amino acid distance between all earlier and later sequences was calculated for each individual, and this single value approximated the amino acid divergence from earlier and later virus populations in a single individual. The amino acid divergence was significantly different for the high and low viral load groups (median divergence, 0.26 versus 0.11; $P = 0.048$ [Wilcoxon rank sum test]) with the high viral load subjects exhibiting a significantly higher amino acid divergence from earlier to later time points (data not shown). There was a positive but nonsignificant linear correlation between the protein distance and the log$_{10}$-transformed viral load measurement ($P = 0.094$). This suggested that the greater nucleotide divergence noted in the high viral load group resulted in amino acid changes and therefore a greater amino acid divergence associated with high viral load levels. The increase in potential nonsynonymous changes in this population may provide evidence of additional selective pressures in individuals carrying high viral loads.

**Consensus sequence analysis of nucleotide and amino acid divergence.** Consensus nucleotide and amino acid sequences were generated for all earlier and later time point samples from each subject. These alternate divergence measures were evaluated to confirm the previous findings. The nucleotide divergence between earlier and later consensus sequences was significantly greater in the high viral load group (median divergence, 0.027 [high viral load group] versus 0.0092 [low viral load group]; $P = 0.023$ [Wilcoxon rank sum test] [data not shown]). Additionally, the amino acid divergence between earlier and later consensus sequences was significantly greater in the high viral load group (median divergence, 0.13 [high viral load group] versus 0.044 [low viral load group]; $P = 0.023$ [Wilcoxon rank sum test] [data not shown]).

**$dn/ds$ ratios as a measure of positive selection pressure.** Characterization of nucleotide substitutions in a gene or genome involves several main forces, including mutation generation, random genetic drift, purifying or negative selection, and occasionally positive selection (12). The ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (dn/ds ratio) is a measure of positive selection and is frequently used in the HIV-1 system. A $dn/ds$ ratio of greater than 1 may be considered indicative of positive selective pressure. We calculated the $dn/ds$ ratios for our groups of high and low viral load individuals through the SNAP program (Los Alamos Database), based on the method employed by Nei and Gojobori (28). The calculated $dn/ds$ ratios were not normally distributed; therefore, analysis by the nonparametric Wilcoxon rank sum test was conducted on both direct and loge-transformed values. There was no significant difference between the high and low viral load groups with regard to $dn/ds$ ratios (median $dn/ds$ ratio, 1.20 [high viral load group] versus 0.435 [low viral load group]; median loge, $dn/ds$ values, 0.19 [high viral load group] versus 0.83 [low viral load group]; $P = 0.168$ [Wilcoxon rank sum test]). This argues against strong positive selection on one of the two populations but indicates that there may have been a greater number of nonsynonymous substitutions in the high viral load group than in the low viral load group (median $dn$ value, 0.0567 [high viral load group] versus 0.0232 [low viral load group]; median loge, $dn$ value, −2.87 [high viral load group] versus −3.76 [low viral load group]; $P = 0.028$ [Wilcoxon rank sum test]). This argues against strong positive selection on one of the two populations but indicates that there may have been a greater number of nonsynonymous substitutions in the high viral load group than in the low viral load group (median $dn$ value, 0.0567 [high viral load group] versus 0.0232 [low viral load group]; median loge, $dn$ value, −2.87 [high viral load group] versus −3.76 [low viral load group]; $P = 0.028$ [Wilcoxon rank sum test]). This argues against strong positive selection on one of the two populations but indicates that there may have been a greater number of nonsynonymous substitutions in the high viral load group than in the low viral load group (median $dn$ value, 0.0567 [high viral load group] versus 0.0232 [low viral load group]; median loge, $dn$ value, −2.87 [high viral load group] versus −3.76 [low viral load group]; $P = 0.028$ [Wilcoxon rank sum test]).

**Interaction between the effects of viral load and time and its**
effect on nucleotide and amino acid divergence. The rate of viral divergence over time appeared to differ by viral load level. This interaction was noted for viral load measured as a binary variable (Fig. 3a) and as a continuous quantitative variable (Fig. 3b.) Fig. 3a shows a plot of the nucleotide and amino acid divergence (mean nucleotide distance and mean protein distance between all earlier and later clones) versus the observation time separately by viral load group. In each plot, the best-fit regression lines for the low and high viral load groups are substantially nonparallel, illustrating the interaction. Statistical tests for interaction in the linear regression model gave $P$ values of 0.148 and 0.089 for mean nucleotide and mean protein distances, respectively. Statistical tests for interaction in additional linear regression models for consensus nucleotide distances ($P = 0.033$) and consensus protein distances ($P = 0.052$) were performed (data not shown). The results were sensitive to the individual with low viral load in the bottom right corner of each plot; this and the small sample size imply that formal inferences should be made cautiously. Figure 3b illustrates the interaction when viral load is considered quantitatively for nucleotide and amino acid divergence. In each plot, the black line represents the estimated increase in mean divergence (100%) per $1 - [\log_{10}\text{increase in viral load}]$ as a function of time and the gray lines are 95% pointwise confidence intervals. The steepness of the (positive) slope indicates the strength of interaction, with a zero slope indicating no interaction. Statistical tests for interaction in the linear regression model gave $P$ values of 0.175 and 0.102 for mean nucleotide and mean protein distances, respectively. Statistical tests for interaction in additional linear regression models for consensus nucleotide distances ($P = 0.069$) and consensus protein distances ($P = 0.097$) were performed (data not shown). In sum, the data suggest that the rate of viral divergence over time was modified by viral load, with a higher viral load leading to a considerably higher divergence rate. The relationship between divergence and time was different between high and low viral load levels, with evidence of a strong interaction between observation time and viral load level.

An alternate method to discern the aforementioned interaction was to substitute a divergence rate (divergence divided by observation time) for the divergence measures. There was a significant positive correlation between the nucleotide and amino acid divergence rates and the $\log_{10}$-transformed viral load measurement (by both the Wilcoxon rank sum test and Spearman’s correlation coefficient) in most of the divergence measures. Nucleotide and amino acid divergence rates are plotted against viral load in Fig. 3c, and the positive correlation between viral load level and divergence is evident.

**DISCUSSION**

Biologically relevant changes in viral replication occur at $>0.5\ log_{10}$ RNA copies/ml, and it has therefore been suggested that changes of $<0.5\ log_{10}$ RNA copies/ml over time are indicative of the viral setpoint (5, 16). The data for the majority of individuals evaluated in this cohort appear to corroborate the viral setpoint paradigm. The viral load measurements of 11 of 12 of our subjects fluctuated by $<0.5\ log_{10}$ RNA copies/ml, while a single individual had a variability of $0.85\ log_{10}$ RNA copies/ml. All RNA measurements were taken during the disease-free interval. Our subjects provided a unique opportunity to evaluate viral loads during the disease-free interval for a non-B subtype. While the majority of setpoint studies have described subtype B-infected individuals undergoing antiretroviral therapy at some point during sampling, our samples were collected from antiretroviral therapy-naïve individuals, providing an opportunity to elucidate the natural history of viral kinetics in the absence of additional selective pressures.

Many studies have evaluated HIV-1 subtype B diversity at single and sequential time points and its correlation with disease progression (1, 7, 11, 15, 18, 20, 22, 24, 25, 40, 41, 43, 47). These studies included individuals treated with antiretroviral drugs and sampled both during the disease-free interval and once clinical AIDS was established. Some studies evaluated both HIV-1 seroincident and seroconverters. A few studies did not directly examine sequence when calculating diversity, while others did not control for resampling bias. To our knowledge, this is the first study to examine the relationship between viral load level and genetic diversity in a non-B subtype during the disease-free interval. We attempted to characterize the relationship between genetic diversity and the HIV-1 RNA viral load and also sought to determine whether the relationship was linear and positively or negatively correlated. Additionally, this design provided an opportunity to examine diversity during the disease-free interval, unaltered by the viral dynamics of clinical AIDS or acute infection. Not surprisingly, the magnitude of diversity was positively correlated to viral load level, such that individuals carrying high viral loads demonstrated a greater intrapatient diversity, diversification, diversification rate, and divergence over time in their nucleotide and protein sequences than did individuals carrying low viral loads. Further, the greater the observation time, the greater the generated divergence between the earlier and the later samples in the high viral load group. Accordingly, intrapatient diversity was similar between individuals carrying high and low viral loads earlier in infection, with a trend towards a greater intrapatient diversity in the high viral load group later in infection. Much of this may be attributable to mutations generated by increasing rounds of viral replication in the individuals carrying high viral loads as well as to rapid virus turnover. Reverse transcriptase has a high error rate and an inadequate proofreading and repair mechanism. Hence, the mutation rate of HIV-1 has a profound effect on genetic variability (4), and if diversity generated by mutation rate is directly proportional to rounds of replication, this diversity will be positively compounded by time. Not surprisingly, the high levels of nucleotide diversity and divergence found in the high viral load group was magnified in individuals with the greatest observation time, a finding similar to what has been observed for subtype B infections (17). It therefore seems quite likely that viral replication rate and the resultant level of circulating virus is the primary driving force in the generation of nucleotide diversity in this population of CRF02_A/G-IbNG-infected individuals.

All individuals generated diversity and divergence proportional to their viral load level: individuals with high viral load were characterized by high diversity and divergence and individuals with low viral load were characterized by lower diversity and divergence, as has been described previously in studies of HIV-1 subtype B (22, 25). Increased nucleotide diversity
FIG. 3. (a) Nucleotide and amino acid divergence (y axis) plotted against observation time in years (x axis) with regression lines. Squares, high viral load subjects; circles, low load subjects. (b) Slope of nucleotide and amino acid divergence (y axis, represented by mean distances) plotted against observation time (x axis) for all subjects. Solid line, estimated (percent) increase in mean divergence per $10^{-\log_{10}}$ increase in viral load as a function of time interval; broken line, 95% pointwise confidence interval; dashed horizontal line, zero line. (c) Number (log_{10} transformed) of viral RNA copies/ml of plasma (y axis) for each subject plotted against nucleotide and amino acid divergence rates (x axis, mean distance divided by observation time.).
and divergence which resulted in concordant amino acid divergence was noted in high viral setpoint individuals. Accordingly, individuals carrying high viral loads possessed a greater number of nonsynonymous nucleotide changes. This might be consistent with initial strong immune pressures on the virus population until rapidly replicating variants predominated to push the individual to AIDS (30) or on individual virulent variants resulting due to the accumulation of nonsynonymous changes during the progression to AIDS (40). There was no significant difference in the dn/ds ratio between high and low viral load individuals. However, one cannot rule out the possibility of positive selective forces acting on the high viral load population, or of the absence of positive selection on the low viral load population. Even the presence of small selective forces, whether positive or negative in nature, could influence variation in the population (4). The magnitude of sequence divergence noted in the high viral setpoint individuals in this study is most likely a direct function of a higher viral replication rate and duration of infection (time) to compound resulting error-associated mutations. Alternatively, one cannot rule out new, unique, and virulent viral variants that emerge as a result of high replication rate in the individuals with high viral setpoints and continue to reestablish new populations as infection time progresses.

Unlike many of the published studies of diversity, which evaluated subtype B-infected homosexual men, this study was an evaluation of a cohort composed solely of women infected with non-subtype B virus through a heterosexual mode of transmission. The role of gender-specific differences in the biology of HIV disease pathogenesis and transmission remains to be fully characterized, although the data suggest that such differences may exist in viral load (8, 10, 36, 42) and pathogenesis (8, 10). There is also evidence that HIV-1-infected women from Mombasa, Kenya, harbor viruses of greater heterogeneity than do HIV-1-infected men of the same population (19). Such gender-specific differences in biology and transmission may affect the interpretation of diversity in this population as well as the general applicability of any findings to other HIV-1-infected populations.

It is also possible that diversity seen in the env C2V3 region may not be fully representative of diversity across the entire genome and that selective forces towards positive selection in this region may not be applicable to forces across the entire genome. However, the V3 region is an important functional region for virus tropism and pathogenicity, cytopotoxic T-lymphocyte responses, humoral antibody responses, and coreceptor usage and is also a region of frequent mutations (22). Therefore, examination of mutations and diversity in this region may represent genetic diversity in functionally important domains more susceptible to selection. Future studies to further characterize the relationship between genetic diversity and viral load are clearly needed.

It is also necessary to address and better clarify the role of immunologic correlates and their associations with viral load. We have recognized an interaction between genetic diversity, viral setpoint, and duration of infection. Therefore, elucidation of the etiology and directionality of these variables and their influences on one another may provide additional prognostic markers for disease progression in regions where non-B subtypes predominate. In particular, accounting for such diversity during vaccine design might allow for selection of more broadly neutralizing epitopes for use in negative partners in discordant couples, babies of HIV-1-infected mothers, or in other individuals at risk for transmission by viruses that establish high viral setpoints. Elucidation of the interactions between diversity and markers of disease progression will be pivotal for vaccine design, therapy development, and disease pathogenesis studies.

ACKNOWLEDGMENTS

We thank Mary Fran McLane, Christopher Mullins, Donald Hamel, Abdoulaye Dieng-Sarr, and Aissatou Gueye-Ndiaye for consistent support and assistance and Wei-Kung Wang for critical review and suggestions.

This work was supported in part by U.S. Army grant DAMD 17-95-C-5005 as well as by NIH grants AI 43879 and AI 467274.

REFERENCES