Published in final edited form as:
Infect Genet Evol. 2013 October ; 0: 361–368. doi:10.1016/j.meegid.2013.02.023.

Intra-Host Evolutionary Rates in HIV-1C *env* and *gag* during Primary Infection

Vlad Novitsky[a,b,*], **Rui Wang**[c], **Raabya Rossenkhan**[b], **Sikhulile Moyo**[b], and **M. Essex**[a,b]

[a]Harvard School of Public Health AIDS Initiative, Department of Immunology and Infectious Diseases, Harvard School of Public Health, 651 Huntington Avenue, Boston, Massachusetts 02115, United States of America

[b]Botswana–Harvard AIDS Institute, P/Bag BO 320, Gaborone, Botswana

[c]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, United States of America

Abstract

Background—HIV-1 nucleotide substitution rates are central for understanding the evolution of HIV-1. Their accurate estimation is critical for analysis of viral dynamics, identification of divergence time of HIV variants, inference of HIV transmission clusters, and modeling of viral evolution.

Methods—Intra-patient nucleotide substitution rates in HIV-1C *gag* and *env* gp120 V1C5 were analyzed in a longitudinal cohort of 32 individuals infected with a single viral variant. Viral quasispecies were derived by single genome amplification/sequencing from serially sampled blood specimens collected at median (IQR) of 5 (4–6) times per subject from enrollment (during Fiebig stages II to V) over a median (IQR) of 417 (351–471) days post-seroconversion (p/s). HIV-1C evolutionary rates were estimated by BEAST v.1.6.1 using a relaxed lognormal molecular clock model. The effect of antiretroviral therapy (ART) on substitution rates in *gag* and *env* was assessed in a subset of six individuals who started ARV therapy during the follow-up period.

Results—During primary HIV-1C infection, the intra-patient substitution rates were estimated at a median (IQR) of 5.22E-03 (3.28E-03–7.55E-03) substitutions per site per year of infection within *gag*, and 1.58E-02 (9.99E-03–2.04E-02) substitutions per site per year within *env* gp120 V1C5. The substitution rates in *env* gp120 V1C5 were higher than in *gag* ($p<0.001$, Wilcoxon signed rank test). The median (IQR) relative rates of evolution at codon positions 1, 2, and 3 were 0.73 (0.48–0.84), 0.67 (0.52–0.86), and 1.54 (1.21–1.71) in *gag*, and 1.01 (0.86–1.15), 1.05 (0.99–1.21), and 0.86 (0.67–0.94) in *env* gp120 V1C5, respectively. A first to the third position codon rate ratio > 1.0 within *env* was found in 25 (78.1%) cases, but only in 4 (12.5%) cases in *gag*, while a second to the third position codon rate ratio > 1.0 in *env* was observed in 26 (81.3%) cases, but in *gag* only in 2 (6.3%) cases ($p<0.001$ for both comparisons, Fisher's exact test). No ART effect on substitution rates in *gag* and *env* was found, at least within the first 3–4 months after ART initiation. Individuals with early viral set point ≥ 4.0 $\log_{10}$ copies/ml had higher substitution rates in *env* gp120 V1C5 (median (IQR) 1.88E-02 (1.54E-02–2.46E-02) vs. 1.04E

*Corresponding Author: Vlad Novitsky, Principal Research Scientist, Harvard School of Public Health AIDS Initiative, FXB 305A, 651 Huntington Avenue, Boston MA 02115, USA. Tel: +1-617-432-1225; fax: +1-617-739-8348; vnovi@hsph.harvard.edu.

(7.24E-03–1.55E-02) substitutions per site per year; p=0.017, Mann-Whitney sum rank test), while individuals with early viral set point ≥ 3.0 $\log_{10}$ copies/ml had higher substitution rates in *gag* (median (IQR) 5.66E-03 (3.45E-03–7.94E-03) vs. 1.78E-03 (4.57E-04–5.15E-03); p=0.028; Mann-Whitney sum rank test).

Conclusions—The results suggest that in primary HIV-1C infection, (1) intra-host evolutionary rates in *env* gp120 V1C5 are about 3-fold higher than in *gag*; (2) selection pressure in *env* is more frequent than in *gag*; (3) initiation of ART does not change substitution rates in HIV-1C *env* or *gag*, at least within the first 3–4 months after starting ART; and (4) intra-host evolutionary rates in *gag* and *env* gp120 V1C5 are higher in individuals with elevated levels of early viral set point.

Keywords

HIV-1; subtype C; evolutionary rates; substitution rates; primary infection

1. INTRODUCTION

HIV-1 is one of the fastest evolving RNA viruses. Multiple factors, such as the error-prone nature of viral reverse transcriptase, high replication rate, high population size within host, compartmentalization, frequent recombination, and superinfection, play important roles in HIV-1 evolutionary dynamics (Coffin et al., 1997; Perelson et al., 1996; Preston et al., 1988; Rambaut et al., 2004; Suzuki et al., 2000). Evolutionary rates, or rates of nucleotide substitution, represent the cumulative effects of virus mutation rate, generation time, effective population size and fitness, and are central for understanding the evolution of HIV-1 (Duffy et al., 2008; Jenkins et al., 2002). An accurate estimation of HIV-1 evolutionary rates is critical for analysis of viral dynamics, identification of divergence time of HIV variants, inference of HIV transmission clusters, and modeling of viral evolution. The HIV-1 evolutionary rates within the host are associated with disease progression (Edo-Matas et al., 2011; Lemey et al., 2007; Williamson, 2003; Wolinsky et al., 1996), which highlights the importance of a systematic assessment of intra-host evolutionary rates in HIV-1 infection.

The *nucleotide substitution rate* is defined as the number of nucleotide substitutions per site per year. Previous studies estimated the rate of nucleotide substitution in HIV-1 (inter-patient level) at about 1.0×10^{-3} per site per year (Duffy et al., 2008; Gojobori et al., 1990; Goudsmit and Lukashov, 1999; Korber et al., 2000; Korber et al., 1998; Leitner and Albert, 1999; Li et al., 1988; Salemi et al., 2001; Suzuki et al., 2000; Yusim et al., 2001), and pointed to different substitution rates among HIV-1 genes (Korber et al., 2000; Leitner and Albert, 1999; Li et al., 1988; Salemi et al., 2001). Using the maximum likelihood method, substitution rates in partial *gag* and *env* of HIV-1 were estimated at 2.5×10^{-3} per site per year (Jenkins et al., 2002). Applying a Bayesian framework and hierarchical models of phylogenetic analysis, intra-host substitution rates in HIV-1 *env* were estimated at 9.2×10^{-3} per site per year among disease progressors and 7.0×10^{-3} per site per year among long-term non-progressors (Edo-Matas et al., 2011). Analysis of synonymous and nonsynonymous rates using well-characterized datasets of prospectively followed individuals infected with HIV-1B (Shankarappa et al., 1999; Shriner et al., 2004) revealed intra-host evolutionary rates in *env* at 6.3×10^{-3} to 1.0×10^{-2} per site per year (Lemey et al., 2007; Lemey et al., 2006; Pybus and Rambaut, 2009). The intra-host evolutionary rates depend on the stage of infection and are lower as disease progresses (Lee et al., 2008; Pybus and Rambaut, 2009).

Little is known about intra-host evolutionary rates in HIV-1 non-B subtypes, particularly in subtype C. Evolutionary rates for HIV-1C were reported at 9.7×10^{-3} per site per year (Maljkovic Berry et al., 2007). Inter-patient evolutionary rates in HIV-1C were estimated at $0.05–2.95\times10^{-3}$ per site per year for *gag* and at $3.1–4.8\times10^{-3}$ per site per year for *env*

(Walker et al., 2005). Abecasis et al. obtained similar estimates of substitution rates for HIV-1C *env*, and about 3 times lower rates for HIV-1C *pol* (Abecasis et al., 2009).

In this study we assessed the *in vivo* intra-host substitution rates in HIV-1 subtype C *gag* and the V1C5 region of *env* gp120 during primary infection. Our sample set of prospectively collected HIV-1C quasispecies from 32 subjects can be compared to a comprehensive set of HIV-1B *env* sequences described by Shankarappa et al. (Shankarappa et al., 1999). While the follow-up period in our study was shorter (~400 days p/s vs. about 8–10 years), the sample size was larger (n=32 vs. n=8). We assessed levels and distribution of intra-host substitution rates in primary HIV-1C infection, compared rates between HIV-1C *gag* and *env*, evaluated the relative rates of HIV-1C evolution across codon positions, estimated the effect of antiretroviral therapy (ART) on viral substitution rates, and addressed potential associations between viral evolutionary rates and HIV-1 RNA load in the early phase of HIV-1C infection.

2. MATERIAL AND METHODS

2.1

Study subjects were enrolled in the primary HIV-1 subtype C infection cohort in Botswana (Novitsky et al., 2009b; Novitsky et al., 2009c; Novitsky et al., 2008) from April 2004 to April 2008. The study was approved by Institutional Review Boards in Botswana and the US. Written informed consent was obtained from each participant. The entire cohort was comprised of 42 individuals with estimated time of seroconversion (Novitsky et al., 2009a; Novitsky et al., 2011b; Novitsky et al., 2009b; Novitsky et al., 2010b; Novitsky et al., 2009c). However, 10 of those individuals were infected with multiple viral variants (Novitsky et al., 2011c). Therefore, only a subset of 32 individuals who were infected with unrelated viral variants with limited diversity of viral quasispecies at the time point of sampling closest to the estimated time of seroconversion (HIV infection with a single viral variant) were analyzed in this study, including 6 subjects identified with acute HIV-1 infection (Fiebig stage II) and 26 subjects identified soon after seroconversion within Fiebig stage IV or V (Table 1). Time “zero” corresponded to seroconversion, as a more accurate, measurable, and less subjective time point than estimation of the time of HIV infection (Novitsky et al., 2010b). There were 5 male (1 acute) and 27 female (5 acute) participants. The median age at enrollment was 28 years (IQR 25–34, range 20–56). All subjects were Botswana nationals, and all infections were HIV-1 subtype C (Novitsky et al., 2009a; Novitsky et al., 2009c). Serial measurements of HIV-1 RNA and CD4 T cell counts were performed over time (Novitsky et al., 2011a; Novitsky et al., 2009c). The early viral set point was estimated as a mean value of HIV-1 RNA measurements in plasma during the period 100 to 300 days post-seroconversion (p/s) (Novitsky et al., 2011a; Novitsky et al., 2010a). The median (IQR) follow-up period was 417 (351–471) days post-seroconversion in *env*, and 418 (350–471) days p/s in *gag*. Six of 32 (19%) subjects initiated ART within the observed period of time due to a drop in CD4+ T cells.

2.2. HIV-1C sequences

The methodology of generating *gag* and *env* quasispecies was presented in detail elsewhere (Novitsky et al., 2009a; Novitsky et al., 2011b; Novitsky et al., 2009b; Novitsky et al., 2010b; Novitsky et al., 2011c). Briefly, for both cDNA and DNA templates, a strategy of single-genome amplification by limiting dilutions (Liu et al., 1996; Palmer et al., 2005) was employed with some modifications. The analyzed region of *gag* corresponded to nucleotide positions 841 to 2,217 of HXB2, and the analyzed region in *env* corresponded to nucleotide positions 6,615 to 7,757 of HXB2. The FastStart High Fidelity Enzyme Blend (Roche Diagnostics, Indianapolis, IN) was used for all PCR amplifications. Amplicons were

purified by Exo-SAP (Dugan et al., 2002), and were sequenced directly at both strands on the ABI 3730 DNA Analyzer using BigDye technology. Sequence contigs were assembled by SeqScape v.2.6. Multiple sequence alignments were performed by using MUSCLE v.3.5 (Edgar, 2004) followed by minor manual adjustments in BioEdit (Hall, 1999). Analysis was performed using separate alignments for each subject. Only intra-host insertions and deletions were present in alignments. In contrast to the alignment of viral sequences that originate from different hosts, the number of indels in the intra-host alignment was relatively small, and therefore, the individual alignments were not gap-stripped. The obtained sequences were tested by HYPERMUT v.2.0 (Rose and Korber, 2000) and hypermutated sequences were excluded from analysis.

2.3. Estimating viral substitution rates

To estimate evolutionary rates in HIV-1C *gag* and *env* gp120 we used a comprehensive BEAST package v.1.7 (Drummond and Rambaut, 2007) that incorporates phylogenetic uncertainty into estimates of evolutionary parameters (Gray et al., 2011) as described previously (Stadler et al.). Briefly, the optimal evolutionary model estimated by jModelTest (Posada, 2008) was used for inferring the maximum likelihood (ML) tree by PhyML (Guindon et al., 2009; Guindon et al., 2010). The ML tree was used as a starting tree in Bayesian analysis. The independent HKY+G model was used for each codon, which allowed us to estimate relative rates of evolution at each codon position and codon rate ratio values of the 1^{st} to 3^{rd} codon position, and 2^{nd} to 3^{rd} codon position. We employed a relaxed molecular clock approach that explicitly models and estimates the level of rate variation among lineages (Drummond et al., 2006). The posterior distribution for the coefficient of variation that does not include zero indicates that the relaxed clock model provides a better fit to the data than the strict clock model (Drummond et al., 2006; Gray et al., 2011). The prior on the HKY transition/transversion parameter (κ) was Gamma distribution with a shape of 0.05 and a scale of 20. The prior on the alpha shape parameter (α) was an exponential distribution with a mean of 1. The ucld.stdev (S) parameter of the lognormal relaxed clock had an exponential prior with mean of 0.333. Viral evolutionary rates before and after the ART were estimated in a subset of six subjects who started ART during the follow-up period (Table 1) by assigning temporal partitions corresponding to the sampling intervals before and after initiation of ART.

2.4. Statistical analysis

Data are summarized with medians (inter quartile range for 25% and 75%). Comparisons of continuous outcomes between two groups were based on the Mann-Whitney Rank Sum test/ Wilcoxon test. Analysis of relative evolutionary rates at different codon positions was performed using One Way Analysis of Variance. Associations between substitution rates and HIV-1 RNA in plasma were assessed using linear regression analysis. All reported p-values are 2-sided and not adjusted for multiple analyses.

2.5. Accession numbers

A total of 1,963 HIV-1 subtype C sequences used in this study were deposited to GenBank previously with the following accession numbers: GQ275453–GQ275667, GQ275750–GQ276051, GQ276137–GQ276247, GQ870874–GQ870904, GQ276248–GQ276284, GQ276418–GQ276698, GQ276766–GQ276795, GQ276797–GQ277080, GQ277179–GQ277222, GQ277224–GQ277293, GQ277295–GQ277374, GQ277404–GQ277443, GQ277446–GQ277569, GQ375107–GQ375128, GQ870874–GQ870939, GQ870957–GQ871036, and GQ871050–GQ871183. Additionally, 2,219 HIV-1C sequences were deposited to GenBank within this study: accession numbers KC628761–KC630979. The supplementary Table S1 includes information for all viral sequences used in this study including sequence ID, former sequence ID (if any), GenBank accession numbers, patient

code, source of amplification, date of sampling, and the number of days from estimated seroconversion.

3. RESULTS

3.1. Intra-host evolutionary rates in HIV-1C *gag* and *env* gp120 V1C5

The *gag*- and *env*-based maximum likelihood trees with all subjects, 5 sequences per subject randomly selected from different time points of sampling, are shown in Supplementary Figures S1A and S1B for *gag* and *env*, respectively. The individual phylogenetic trees for each study subject for HIV-1C *gag* are presented in Supplementary Figures S02 to S33, and for the HIV-1C V1C5 region of *env* gp120 are shown in Supplementary Figures S34 to S65.

The individual intra-patient nucleotide substitution rates within HIV-1C *gag* and *env* gp120 V1C5 among study subjects are presented in Figure 1A, while a summary is shown in Figure 1B. The median (IQR) of the nucleotide substitution rate within HIV-1C *gag* was 5.22×10^{-3} (3.28×10^{-3} to 7.55×10^{-3}) substitutions per site per year ranging from 3.39×10^{-4} to 2.25×10^{-2}. The median (IQR) of the substitution rate in *env* gp120 V1C5 was 1.58×10^{-2} (9.99×10^{-3} – 2.04×10^{-2}) substitutions per site per year of infection, ranging from 1.88×10^{-4} to 5.85×10^{-2}. The observed evolutionary rates were significantly higher in *env* gp120 V1C5 (~ 3-fold) than in *gag* ($p<0.001$, Wilcoxon test). The distribution of nucleotide substitution rates in *env* gp120 V1C5 and *gag* was skewed toward smaller values, suggesting that the majority of HIV-1C-infected individuals exhibit relatively low evolutionary rates, while a small proportion of subjects maintain elevated levels of viral evolutionary rates during primary HIV-1C infection. The linear regression analysis demonstrated a statistically significant ($p=0.015$) but weak (adjusted $R^2 = 0.155$) direct correlation between substitution rates in *env* gp120 V1C5 and *gag* (Fig. 1C).

3.2. Relative evolutionary rates at codon positions

Within HIV-1C *gag*, the median (IQR) relative rates of evolution were 0.73 (0.48 to 0.84), 0.67 (0.52 to 0.86), and 1.54 (1.21 to 1.71) at codon positions 1, 2, and 3, respectively. Thus, in *gag*, the relative evolutionary rates at the 1st and 2nd codon positions were significantly lower than at the 3rd codon position ($p<0.001$, One Way Analysis of Variance), which is expected for a continuous protein-coding region. In contrast, within the analyzed region of HIV-1C *env* gp120 V1C5, the median (IQR) relative rates of evolution at codon positions 1, 2, and 3 were 1.01 (0.86 to 1.15), 1.05 (0.99 to 1.21), and 0.86 (0.67 to 0.94), respectively. Therefore in *env* gp120 V1C5, the relative evolutionary rates at the 1st and 2nd codon positions were significantly higher than at the 3rd codon position ($p<0.001$, One Way Analysis of Variance), suggesting a substantial selection pressure in the analyzed region of HIV-1C *env*.

The codon rate ratio values can be used to estimate the ratio of the evolution rates at different codon positions. The codon rate ratio values of the 1st to 3rd codon positions, and the 2nd to 3rd codon positions, were found to be lower in HIV-1C *gag* than in *env* gp120 V1C5 ($p<0.001$ for both comparisons, Wilcoxon test; Fig. 2). The first to the third position codon rate ratio > 1.0 was found in only 4 (12.5%) *gag* cases but in 25 (78.1%) *env* cases, while the second to the third position codon rate ratio > 1.0 was observed in only 2 (6.3%) *gag* cases but in 26 (81.3%) *env* cases ($p<0.001$ for both comparisons, Fisher's exact test).

3.3. ART and viral evolutionary rates

To address whether initiation of ART alters the levels of substitution rates in *gag* and/or in *env* we compared viral evolutionary rates before and after the ART in a subset of six subjects who dropped CD4 levels and started ART during the follow up period. The median

(IQR) time of ART initiation was 234 (162 to 339) days p/s. The median (IQR) time of follow-up after ART initiation was 114 (43 to 216) days. Viral substitution rates in *gag* decreased in 3 of 6 subjects, and increased in the other 3 subjects (Fig. 3A left panel). In *env* gp120 V1C5, substitution rates increased in one subject, and decreased in the other 5 subjects (Fig. 3A, right panel). Overall, the substitution rates in *gag* and *env* gp120 V1C5 did not differ before and after initiation of ART ($p>0.1$ for both comparisons), at least within the first 3–4 months after starting ART. Analysis of relative evolutionary rates at codon positions demonstrated a similar non-significant difference between viral substitution rates before and after subjects were placed on ART. The changes in codon rate ratio values of the 1st to 3rd codon position (Fig. 3B), and 2nd to 3rd codon position (Fig. 3C), did not reach statistical significance in either *gag* or *env* gp120 V1C5 ($p>0.1$ for all comparisons).

3.4. Viral substitution rates and HIV-1 RNA load in plasma

We assessed potential associations between the levels of HIV-1 RNA in plasma and intra-host substitution rates in HIV-1C *gag* and *env* gp120 V1C5 in the linear regression model, and using two thresholds of HIV-1 RNA in plasma, 3.0 $\log_{10}$ and 4.0 $\log_{10}$ copies/ml. There was weak (adjusted $R^2=0.165$) but statistically significant ($p=0.012$) association between substitution rates in *gag* and HIV-1 RNA load. No statistically significant association was found between substitution rates in *env* and HIV-1 RNA load. When analyzed within groups with the defined threshold of HIV-1 RNA load, the substitution rates in HIV-1C *gag* were higher in subjects with early viral set point above 3.0 $\log_{10}$ copies/ml than in subjects with set point below 3.0 $\log_{10}$ (median (IQR) 5.66×10^{-3} (3.45×10^{-3} to 7.94×10^{-3}) vs. 1.78×10^{-3} (4.57×10^{-4} to 5.15×10^{-3}); $p=0.028$; Mann-Whitney sum rank test). However, no difference in the *gag* substitution rates was observed between subjects with early viral set point above and below 4.0 $\log_{10}$ copies/ml ($p=0.33$). In contrast, substitution rates in *env* gp120 V1C5 did not differ between subjects with set point below and above 3.0 $\log_{10}$ copies/ml ($p=0.20$), but reached statistical significance between subjects with early viral set point greater than and lower than 4.0 $\log_{10}$ copies/ml (median (IQR) 1.88×10^{-2} (1.54×10^{-2} to 2.46×10^{-2}) vs. 1.04×10^{-2} (7.24×10^{-3} to 1.55×10^{-2}) substitutions per site per year; $p=0.017$; Mann-Whitney rank sum test).

4. DISCUSSION

Using a Bayesian phylogenetic framework we estimated the intra-host evolutionary rates in two structural HIV-1C genes, *gag* and *env*, during primary infection. We demonstrated substantial heterogeneity in the substitution rates between HIV-1C-infected subjects during the early stage of infection, provided evidence of higher rates in *env* than in *gag*, demonstrated differential patterns in the relative evolutionary rates at different codon positions in *gag* and *env*, assessed rates before and after initiation of ART, and evaluated associations between nucleotide substitution rates and levels of HIV-1 RNA load.

The HIV-1C nucleotide substitution rates were estimated at 5.22×10^{-3} (3.28×10^{-3} to 7.55×10^{-3}) substitutions per site per year in *gag*, and at 1.58×10^{-2} (9.99×10^{-3} to 2.04×10^{-2}) substitutions per site per year in *env* gp120. As expected, our estimates of intra-host evolutionary rates in HIV-1C infection are higher than estimates of the inter-host rates, and are slightly higher than previous estimates of intra-host rates in HIV-1C by Walker et al. (Walker et al., 2005) and Abecasis et al. (Abecasis et al., 2009). One of the potential reasons (which is also a study limitation) is related to the nature of frequent serial sampling and including transient viral polymorphisms that might not be fixed yet. As a result, our estimates of the HIV-1C evolutionary rates could be higher than estimates made over a longer period of time with less frequent sampling. It is known that frequent sampling over a short time period may result in estimates that approach the viral mutation rate (Duffy et al., 2008).

Reasons underlying evolutionary rate variation between subjects are likely to include viral replication and population size, virus–host interactions and viral adaptation. While natural selection is the dominant force determining intra-host evolutionary dynamics of HIV-1 (Pybus and Rambaut, 2009), it is not surprising that the extent of selection pressure during the course of HIV infection differs between subjects and can cause variation in nucleotide substitution rates between individuals. In addition, upon virus transmission to a new host, reverse mutations toward subtype consensus (Friedrich et al., 2004; Leslie et al., 2004; Li et al., 2007; Novitsky et al., 2009b) are likely to affect evolutionary rates due to differences in the host genetic makeup between the previous and new hosts. Viral transient polymorphisms including deleterious mutations can also contribute to variability in HIV substitution rate between subjects (Edwards et al., 2006; Lemey et al., 2006).

As expected, the intra-host evolutionary rates in HIV-1C primary infection were higher in *env* than in *gag*, which is in line with previous studies (Abecasis et al., 2009; Walker et al., 2005). Three codon positions in the protein-coding genes have very different substitution rates and evolutionary dynamics (Yang and Yoder, 2003). The analysis of relative evolutionary rates across codon positions suggested a substantial difference between *env* and *gag* in primary HIV-1C infection in the context of selection pressure. The relative nucleotide substitution rates in the V1C5 region of gp120 were slightly above 1.0 at codon positions 1 and 2, while only 0.86 at codon position 3. This observation provides evidence for a strong selective pressure on the V1C5 region of gp120 during primary HIV-1C infection. In contrast, the relative evolutionary rates in *gag* were higher at the third codon position, highlighting differences between the two structural proteins during primary HIV-1C infection. The proportions of individuals with the codon rate ratio above and below 1.0 in *gag* and *env* provided additional evidence for difference between these genes.

Initiation of ART dramatically reduces the level of viral replication. Thus, it was important to address whether intra-host evolutionary rates differ before and after ART administration. In the limited subset of subjects who started ART during the follow-up period, and the reduced number of analyzed time points due to “before” and “after” subdivision, we found no evidence for statistically significant changes in the viral substitution rates, at least over the first 3–4 months after initiation of ART. In subjects with the highest evolutionary rates in *env,* OC and OE, the mean HIV-1 RNA load over the early stage of infection was also high, 5.11 and 5.40 $\log_{10}$ copies/ml, respectively. However in the next three subjects with the highest evolutionary rates in *env,* OX, OZA, and PA, the early stage HIV-1 RNA load was relatively low or modest, 1.78, 3.56, and 4.13 $\log_{10}$ copies/ml, respectively. Interestingly, none of the subjects with high evolutionary rates dropped their CD4 count and started ART during the follow-up period. We cannot exclude the possibility that nucleotide substitution rates might change at a later stage of ART.

In the linear regression analysis we demonstrated that higher nucleotide substitution rates in *gag* but not in *env* are associated with higher HIV-1 RNA levels in plasma during primary HIV-1C infection. In addition rates in both *gag* and *env* were higher among individuals with higher levels of HIV-1 RNA, although the threshold for early viral set point differed for *gag* and *env*. The observed associations suggest that higher levels of HIV replication result in higher intra-host nucleotide substitution rates, at least during primary infection.

In summary, the study utilized a unique prospective set of HIV-1C *gag* and *env* quasispecies to estimate intra-host evolutionary rates during primary HIV-1C infection, to assess the effect of ART initiation on rates, and to evaluate associations between evolutionary rates and early HIV-1 RNA. Better understanding of the ranges and dynamics of nucleotide substitution rates could help to improve accuracy in estimating divergence time of circulating HIV variants and reconstruction of HIV transmission histories.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to all participants of the *Tshedimoso* study in Botswana. We thank the study clinical team for their dedication and outstanding work in the clinic and outreach. We thank the Botswana Ministry of Health, Gaborone City Council clinics, and the Gaborone VCT *Tebelopele* for their ongoing support and collaboration. Finally, we thank Lendsey Melton for excellent editorial assistance. The primary HIV-1 subtype C infection study in Botswana, the *Tshedimoso* study, was supported and funded by NIH grant R01 AI057027. This work was supported in part by NIH grants D43 TW000004 (RR) and R37 AI24643 (RW), and also through the AAMC FIC/Ellison Overseas Fellowships in Global Health and Clinical Research (RR). The results of the study were presented partially at the 17th International Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology in Belgrade, Serbia, August 27–31, 2012.

References

Abecasis AB, Vandamme A-M, Lemey P. Quantifying Differences in the Tempo of Human Immunodeficiency Virus Type 1 Subtype Evolution. J Virol. 2009; 83:12917–12924. [PubMed: 19793809]

Coffin, JM.; Hughes, SH.; Varmus, HE. Retroviruses. Cold Spring Laboratory Press; Cold Spring Harbor, N.Y: 1997.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006; 4:e88. [PubMed: 16683862]

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007; 7:214. [PubMed: 17996036]

Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008; 9:267–276. [PubMed: 18319742]

Dugan KA, Lawrence HS, Hares DR, Fisher CL, Budowle B. An improved method for post-PCR purification for mtDNA sequence analysis. J Forensic Sci. 2002; 47:811–818. [PubMed: 12136989]

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

Edo-Matas D, Lemey P, Tom JA, Serna-Bolea C, van den Blink AE, van 't Wout AB, Schuitemaker H, Suchard MA. Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models. Mol Biol Evol. 2011; 28:1605–1616. [PubMed: 21135151]

Edwards CTT, Holmes EC, Pybus OG, Wilson DJ, Viscidi RP, Abrams EJ, Phillips RE, Drummond AJ. Evolution of the Human Immunodeficiency Virus Envelope Gene Is Dominated by Purifying Selection. Genetics. 2006; 174:1441–1453. [PubMed: 16951087]

Friedrich TC, Dodds EJ, Yant LJ, Vojnov L, Rudersdorf R, Cullen C, Evans DT, Desrosiers RC, Mothe BR, Sidney J, Sette A, Kunstman K, Wolinsky S, Piatak M, Lifson J, Hughes AL, Wilson N, O'Connor DH, Watkins DI. Reversion of CTL escape-variant immunodeficiency viruses in vivo. Nat Med. 2004; 10:275–281. [PubMed: 14966520]

Gojobori T, Moriyama EN, Kimura M. Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci U S A. 1990; 87:10015–10018. [PubMed: 2263602]

Goudsmit J, Lukashov VV. Dating the origin of HIV-1 subtypes. Nature. 1999; 400:325–326. [PubMed: 10432109]

Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol Biol. 2011; 11:131. [PubMed: 21595904]

Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 2009; 537:113–137. [PubMed: 19378142]

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59:307–321. [PubMed: 20525638]

Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 1999; 41:95–98.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol. 2002; 54:156–165. [PubMed: 11821909]

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. Timing the ancestor of the HIV-1 pandemic strains. Science. 2000; 288:1789–1796. [PubMed: 10846155]

Korber B, Theiler J, Wolinsky S. Limitations of a molecular clock applied to considerations of the origin of HIV-1. Science. 1998; 280:1868–1871. [PubMed: 9669945]

Lee HY, Perelson AS, Park SC, Leitner T. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. PLoS Comput Biol. 2008; 4:e1000240. [PubMed: 19079613]

Leitner T, Albert J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci U S A. 1999; 96:10752–10757. [PubMed: 10485898]

Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, Barroso H, Taveira N, Rambaut A. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput Biol. 2007; 3:e29. [PubMed: 17305421]

Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. AIDS Rev. 2006; 8:125–140. [PubMed: 17078483]

Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, Tang Y, Holmes EC, Allen T, Prado JG, Altfeld M, Brander C, Dixon C, Ramduth D, Jeena P, Thomas SA, St John A, Roach TA, Kupfer B, Luzzi G, Edwards A, Taylor G, Lyall H, Tudor-Williams G, Novelli V, Martinez-Picado J, Kiepiela P, Walker BD, Goulder PJ. HIV evolution: CTL escape mutation and reversion after transmission. Nat Med. 2004; 10:282–289. [PubMed: 14770175]

Li B, Gladden AD, Altfeld M, Kaldor JM, Cooper DA, Kelleher AD, Allen TM. Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. J Virol. 2007; 81:193–201. [PubMed: 17065207]

Li WH, Tanimura M, Sharp PM. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol. 1988; 5:313–330. [PubMed: 3405075]

Liu SL, Rodrigo AG, Shankarappa R, Learn GH, Hsu L, Davidov O, Zhao LP, Mullins; JI, Haynes BF, Pantaleo G, Fauci AS. HIV Quasispecies and Resampling. Science. 1996; 273:413c–417.

Maljkovic Berry I, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY, Bruno W, Leitner T. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J Virol. 2007; 81:10625–10635. [PubMed: 17634235]

Novitsky V, Lagakos S, Herzig M, Bonney C, Kebaabetswe L, Rossenkhan R, Nkwe D, Margolin L, Musonda R, Moyo S, Woldegabriel E, van Widenfelt E, Makhema J, Essex M. Evolution of proviral gp120 over the first year of HIV-1 subtype C infection. NIHMSID # 79286. Virology. 2009a; 383:47–59. [PubMed: 18973914]

Novitsky V, Ndung'u T, Wang R, Bussmann H, Chonco F, Makhema J, De Gruttola V, Walker BD, Essex M. Extended high viremics: a substantial fraction of individuals maintain high plasma viral RNA levels after acute HIV-1 subtype C infection. AIDS. 2011a; 25:1515–1522. [PubMed: 21505307]

Novitsky V, Wang R, Baca J, Margolin L, McLane MF, Moyo S, van Widenfelt E, Makhema J, Essex M. Evolutionary gamut of in vivo Gag substitutions during early HIV-1 subtype C infection. Virology. 2011b; 421:119–128. [PubMed: 22014506]

Novitsky V, Wang R, Bussmann H, Lockman S, Baum M, Shapiro R, Thior I, Wester C, Wester CW, Ogwu A, Asmelash A, Musonda R, Campa A, Moyo S, van Widenfelt E, Mine M, Moffat C, Mmalane M, Makhema J, Marlink R, Gilbert P, Seage GR 3rd, DeGruttola V, Essex M. HIV-1 subtype C-infected individuals maintaining high viral load as potential targets for the "test-and-treat" approach to reduce HIV transmission. PLoS One. 2010a; 5:e10148. [PubMed: 20405044]

Novitsky V, Wang R, Margolin L, Baca J, Kebaabetswe L, Rossenkhan R, Bonney C, Herzig M, Nkwe D, Moyo S, Musonda R, Woldegabriel E, van Widenfelt E, Makhema J, Lagakos S, Essex M. Timing constraints of *in vivo* gag mutations during primary HIV-1 subtype C infection. PLoS One. 2009b; 4:e7727. [PubMed: 19890401]

Novitsky V, Wang R, Margolin L, Baca J, Moyo S, Musonda R, Essex M. Dynamics and timing of in vivo mutations at Gag residue 242 during primary HIV-1 subtype C infection. Virology. 2010b; 403:37–46. [PubMed: 20444482]

Novitsky V, Wang R, Margolin L, Baca J, Rossenkhan R, Moyo S, van Widenfelt E, Essex M. Transmission of Single and Multiple Viral Variants in Primary HIV-1 Subtype C Infection. PLoS One. 2011c; 6:e16714. [PubMed: 21415914]

Novitsky V, Woldegabriel E, Kebaabetswe L, Rossenkhan R, Mlotshwa B, Bonney C, Finucane M, Musonda R, Moyo S, Wester C, van Widenfelt E, Makhema J, Lagakos S, Essex M. Viral Load and CD4+ T Cell Dynamics in Primary HIV-1 Subtype C Infection. J Acquir Immune Defic Syndr. 2009c; 50:65–76. [PubMed: 19295336]

Novitsky V, Woldegabriel E, Wester C, McDonald E, Rossenkhan R, Ketunuti M, Makhema J, Seage GR 3rd, Essex M. Identification of primary HIV-1C infection in Botswana. NIHMSID # 79283. AIDS Care. 2008; 20:806–811. [PubMed: 18608056]

Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, Rock D, Falloon J, Davey RT Jr, Dewar RL, Metcalf JA, Hammer S, Mellors JW, Coffin JM. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. J Clin Microbiol. 2005; 43:406–413. [PubMed: 15635002]

Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in Vivo: virion clearance rate, infected cell life-span, and viral generation time. Science. 1996; 271:1582–1587. [PubMed: 8599114]

Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008; 25:1253–1256. [PubMed: 18397919]

Preston BD, Poiesz BJ, Loeb LA. Fidelity of HIV-1 reverse transciptase. Science. 1988; 242:1168–1171. [PubMed: 2460924]

Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009; 10:540–550. [PubMed: 19564871]

Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nat Rev Genet. 2004; 5:52–61. [PubMed: 14708016]

Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. Bioinformatics. 2000; 16:400–401. [PubMed: 10869039]

Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M, Vandamme AM. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB journal: official publication of the Federation of American Societies for Experimental Biology. 2001; 15:276–278. [PubMed: 11156935]

Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol. 1999; 73:10489–10502. [PubMed: 10559367]

Shriner D, Shankarappa R, Jensen MA, Nickle DC, Mittler JE, Margolick JB, Mullins JI. Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. Genetics. 2004; 166:1155–1164. [PubMed: 15082537]

Stadler T, Kouyos R, von Wyl V, Yerly S, Boni J, Burgisser P, Klimkait T, Joos B, Rieder P, Xie D, Gunthard HF, Drummond AJ, Bonhoeffer S. Estimating the basic reproductive number from viral sequence data. Mol Biol Evol. 2012; 29:347–357. [PubMed: 21890480]

Suzuki Y, Yamaguchi-Kabata Y, Gojobori T. Nucleotide Substitution Rates of HIV-1. AIDS Rev. 2000; 2:39–47.

Walker PR, Pybus OG, Rambaut A, Holmes EC. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. Infect Genet Evol. 2005; 5:199–208. [PubMed: 15737910]

Williamson S. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol. 2003; 20:1318–1325. [PubMed: 12777505]

Wolinsky SM, Korber BT, Neumann AU, Daniels M, Kunstman KJ, Whetsell AJ, Furtado MR, Cao Y, Ho DD, Safrit JT. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection [see comments]. Science. 1996; 272:537–542. [PubMed: 8614801]

Yang Z, Yoder AD. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol. 2003; 52:705–716. [PubMed: 14530137]

Yusim K, Peeters M, Pybus OG, Bhattacharya T, Delaporte E, Mulanga C, Muldoon M, Theiler J, Korber B. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2001; 356:855–866.

Highlights

- HIV-1C quasispecies from 32 subjects analyzed prospectively
- We assessed levels and distribution of intra-host substitution rates in primary HIV-1C infection
- We compared rates between HIV-1C *gag* and *env*
- We evaluated the relative rates of HIV-1C evolution across codon positions
- We estimated the effect of ART on viral substitution rates
- We addressed associations between viral evolutionary rates and HIV-1 RNA load

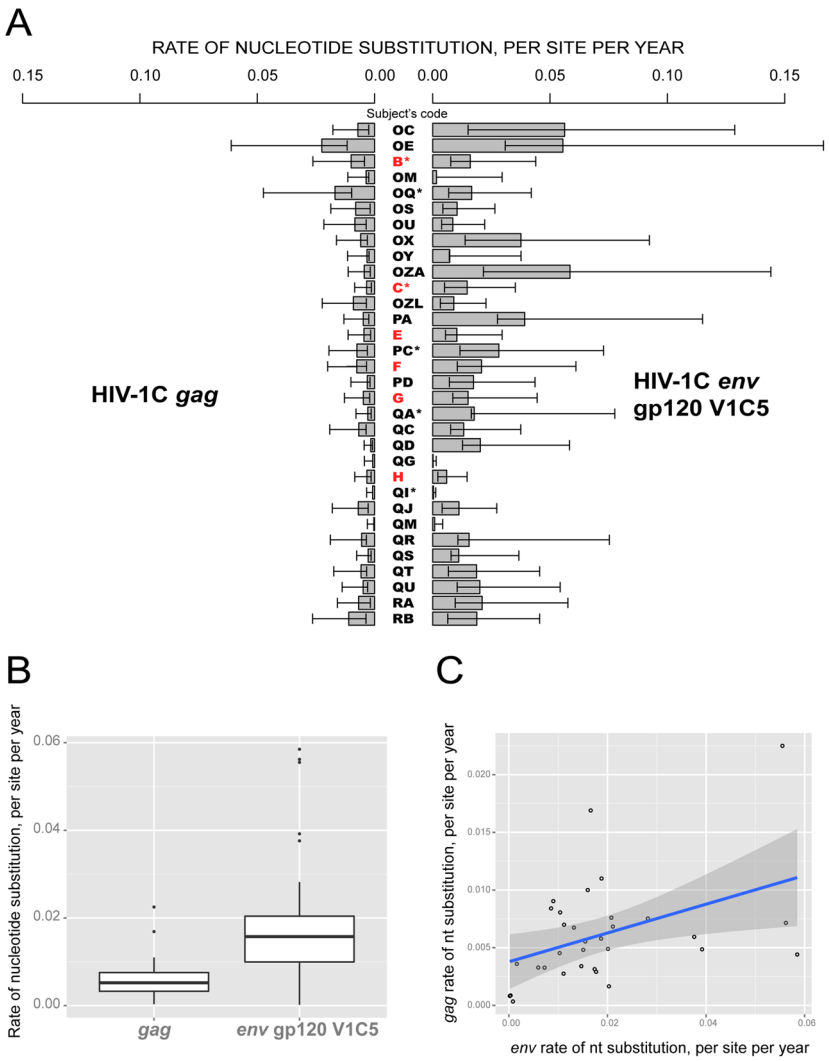

Figure 1.

Nucleotide substitution rates within HIV-1C *gag* and *env* gp120 V1C5 among 32 study subjects infected with a single viral variant. **A:** Individual intra-host rates of nucleotide substitution, shown as substitutions per site per year with lower HPD within bars and upper HPD outside bars. Subjects identified within acute HIV-1C infection are highlighted in red. Asterisks highlight subjects who initiated ART during the follow-up period. **B:** Summary of nucleotide substitution rates within HIV-1C *gag* and *env* gp120 V1C5. The boxplots highlight the median and the 1st and the 3rd quartiles of nucleotide distribution rates in *gag* and *env* gp120 V1C5. Whiskers correspond to the "1.5 rule": less than Q1 – 1.5*IQR and greater than Q3 + 1.5*IQR. Open circles indicate outliers. **C:** Analysis using linear regression model to assess potential association between substitution rates in *env* gp120 V1C5 and *gag*.

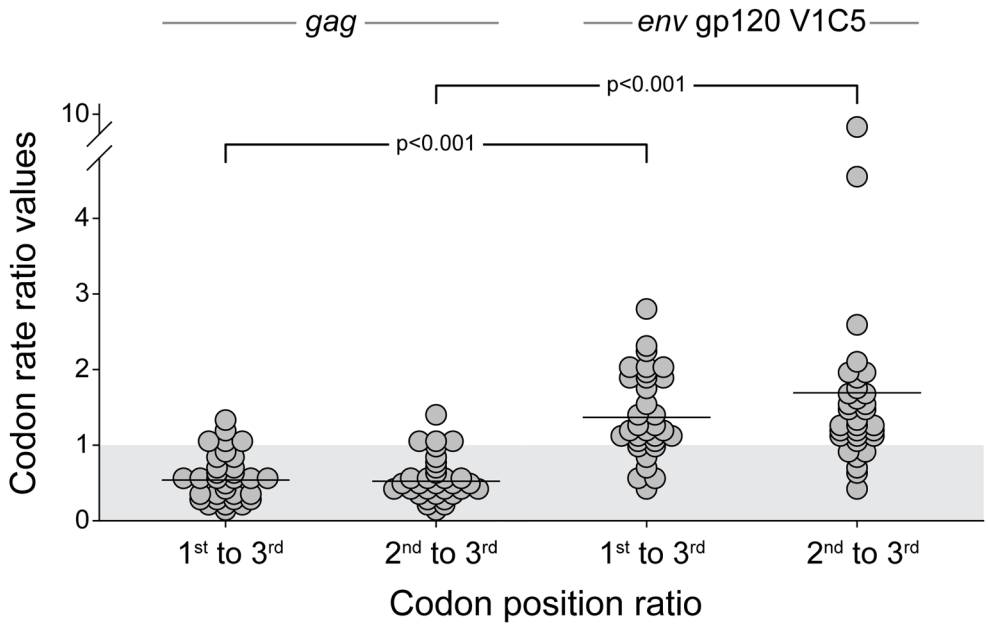

Figure 2.

The codon rate ratio values of the 1st to 3rd codon position, and 2nd to 3rd codon position in HIV-1C *gag* and *env* gp120 V1C5. The gray shadow at the bottom highlights the area below 1.0 and denotes the proportion of subjects below and above 1.0 for each codon rate ratio.

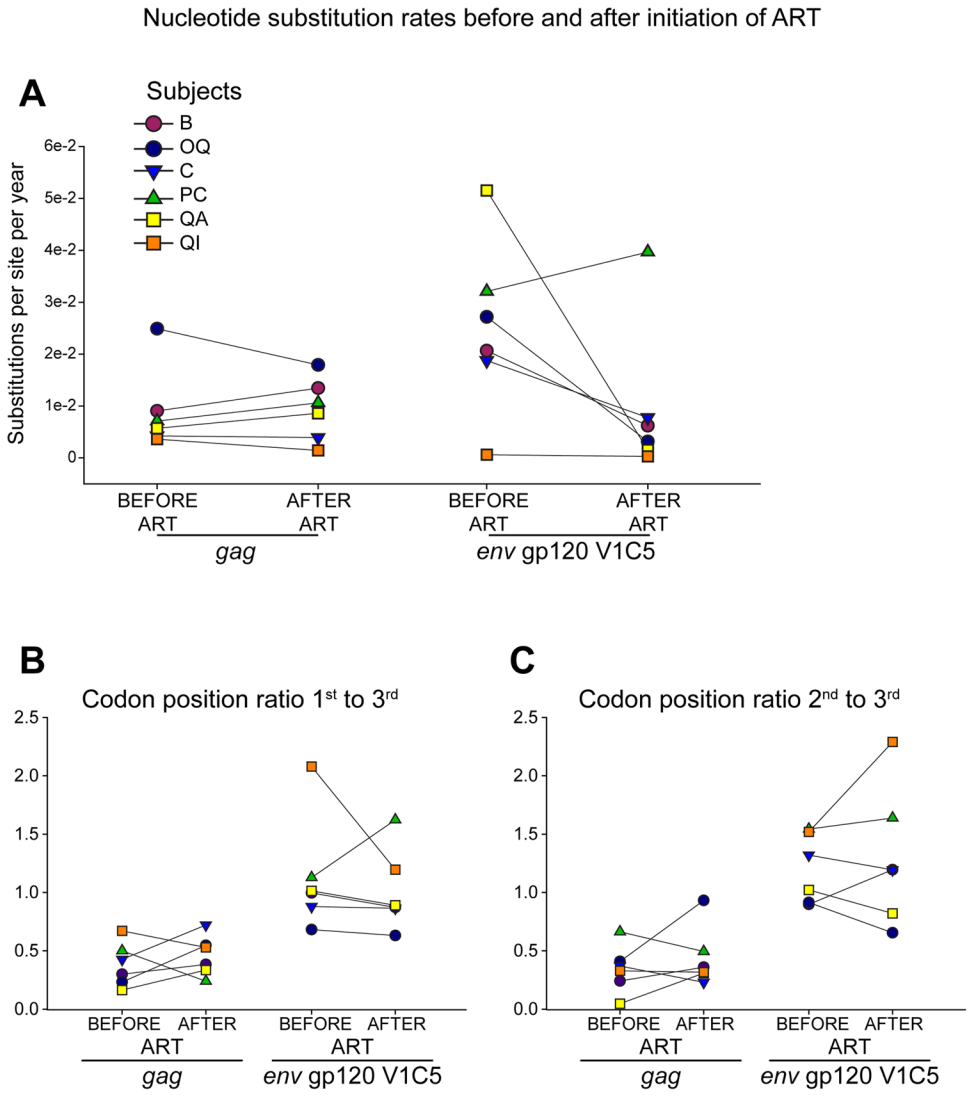

Figure 3.

Nucleotide substitution rates before and after initiation of ART in six subjects who initiated ART during the follow-up period. **A:** Nucleotide substitutions per site per year in HIV-1C *gag* and *env* gp120 V1C5. **B:** Codon position ratio 1st to 3rd. **C:** Codon position ratio 2nd to 3rd.

Infect Genet Evol. Author manuscript; available in PMC 2014 October 01.

Table 1

Study subjects. Age is shown at enrollment. HIV-1 RNA load is shown as mean for measurements over the period from 100 to 300 days post-seroconversion (p/s). Time of ART initiation is presented only for subjects who started ART within the follow-up period (subjects who started ART at later time points are not shown).

Patient code	Gender	Age	HIV-1 RNA load, mean 100–300 days p/s	*gag* quasispecies				*env* quasispecies				Time of ART initiation (if any) during follow-up, days p/s
				Earliest time point, days p/s	Latest time point, days p/s	N of time points	N of *gag* sequences	Earliest time point, days p/s	Latest time point, days p/s	N of time points	N of *env* sequences	
OC-2381	M	25	5.11	27	1,446	10	159	86	1,496	12	149	
OE-2530	M	26	5.40	52	236	5	68	52	236	4	66	
B-2865	F	20	5.12	10	223	6	72	10	223	6	78	199
OM-2869	F	23	4.11	86	178	2	38	86	178	2	25	
OQ-2990	F	20	5.61	52	415	5	44	52	415	5	75	270
OS-3079	M	42	3.74	62	1,213	10	150	62	583	7	82	
OU-3091	F	27	3.71	13	471	7	116	13	471	8	78	
OX-3251	F	46	1.78	58	555	7	57	58	464	6	66	
OY-3244	F	37	3.36	51	514	8	48	51	514	7	33	
OZA-3285	F	31	3.56	78	540	7	73	78	540	7	60	
C-3312	F	32	5.44	4	445	7	143	4	445	6	95	362
OZL-3354	F	30	3.47	6	1,259	8	97	6	1,259	6	62	
PA-3390	F	24	4.13	65	227	4	67	65	227	4	44	
E-3430	F	35	3.51	30	404	8	99	30	404	7	89	
PC-3481	F	32	5.32	54	418	6	80	54	418	6	51	388
F-3505	F	53	4.28	7	447	7	45	7	447	7	58	

Patient code	Gender	Age	HIV-1 RNA load, mean 100–300 days p/s	*gag* quasispecies				*env* quasispecies				Time of ART initiation (if any) during follow-up, days p/s
				Earliest time point, days p/s	Latest time point, days p/s	N of time points	N of *gag* sequences	Earliest time point, days p/s	Latest time point, days p/s	N of time points	N of *env* sequences	
PD-3508	F	25	4.22	59	428	5	44	59	428	5	44	
G-3603	M	34	3.93	4	438	6	85	4	345	5	69	
QA-5099	M	28	5.09	67	389	4	64	6	389	4	42	150
QC-5340	F	22	4.01	80	415	4	47	80	415	4	56	
QD-5355	F	27	4.30	79	420	4	59	79	420	4	48	
QG-5511	F	37	1.96	48	415	5	11	48	415	4	14	
H-5582	F	26	3.90	16	347	6	88	16	378	7	99	
QI-5715	F	27	5.34	20	476	5	44	20	476	4	48	112
QJ-5768	F	29	3.59	8	471	7	77	8	471	6	68	
QM-5849	F	37	2.49	48	415	4	21	48	357	3	27	
QR-5943	F	26	3.37	7	485	5	58	7	485	4	51	
QS-6020	F	26	2.22	44	351	5	31	44	351	4	34	
QT-6024	F	24	3.59	44	349	5	64	44	349	5	77	
QU-6029	F	25	5.06	84	266	4	43	84	266	4	55	
RA-6343	F	56	4.25	54	422	6	84	54	422	4	54	
RB-6380	F	28	5.25	52	332	4	40	52	332	4	69	
Median		28		50	419	6	64	50	417	5	59	
IQR1		25		15	351	5	44	12	351	4	47	

Patient code	Gender	Age	HIV-1 RNA load, mean 100–300 days p/s	*gag* quasispecies				*env* quasispecies				Time of ART initiation (if any) during follow-up, days p/s
				Earliest time point, days p/s	Latest time point, days p/s	N of time points	N of *gag* sequences	Earliest time point, days p/s	Latest time point, days p/s	N of time points	N of *env* sequences	
IQR3		34		60	472	7	84	60	471	6	76	