



Published in final edited form as:

*Virology*. 2009 January 5; 383(1): 47–59. doi:10.1016/j.virol.2008.09.017.

## Evolution of proviral gp120 over the first year of HIV-1 subtype C infection

Vladimir Novitsky<sup>a,b</sup>, Stephen Lagakos<sup>c</sup>, Michaela Herzig<sup>a</sup>, Caitlin Bonney<sup>a</sup>, Lemme Kebaabetswe<sup>b</sup>, Raabya Rossenkan<sup>b</sup>, David Nkwe<sup>b</sup>, Lauren Margolin<sup>a</sup>, Rosemary Musonda<sup>b</sup>, Sikhulile Moyo<sup>b</sup>, Elias Woldegabriel<sup>b</sup>, Erik van Widenfelt<sup>b</sup>, Joseph Makhema<sup>b</sup>, and M. Essex<sup>a,b,\*</sup>

<sup>a</sup> Department of Immunology and Infectious Diseases, Harvard School of Public Health, FXB 402, 651 Huntington Avenue, Boston, MA, 02115, USA

<sup>b</sup> Botswana–Harvard School of Public Health AIDS Initiative Partnership, Private Bag BO 320, Bontleng, Gaborone, Botswana

<sup>c</sup> Department of Biostatistics, Harvard School of Public Health, 2-423, 665 Huntington Avenue, Boston, MA, 02115, USA

### Abstract

The evolution of proviral gp120 during the first year after seroconversion in HIV-1 subtype C infection was addressed in a case series of eight subjects. Multiple viral variants were found in two out of eight cases. Slow rate of viral RNA decline and high early viral RNA set point were associated with a higher level of proviral diversity from 0 to 200 days after seroconversion. Proviral divergence from MRCA over the same period also differed between subjects with slow and fast decline of viral RNA, suggesting that evolution of proviral gp120 early in infection may be linked to the level of viral RNA replication. Changes in the length of variable loops were minimal, and length reduction was more common than length increase. Potential N-linked glycosylation sites ranged  $\pm$ one site, showing common fluctuations in the V4 and V5 loops. These results highlight the role of proviral gp120 diversity and diversification in the pathogenesis of acute HIV-1 subtype C infection.

### Keywords

HIV-1 subtype C infection; Provirus; gp120; Diversity; Diversification; Evolution; Acute infection; Botswana

### Introduction

After decades of endemic evolution, HIV-1 has achieved enormous diversity in the worldwide HIV/AIDS epidemic, posing a significant challenge for an HIV vaccine design. A complex interplay between virus and host during HIV-1 infection results in the gradual accumulation of provirus variants. The constantly growing repertoire of archived proviral sequences represents a history of virus replication within the host, providing a continuous source for subsequent HIV replication and transmission. Despite the significant fraction of proviral DNA

\*Corresponding author. Harvard School of Public Health AIDS Initiative, FXB 402, 651 Huntington Avenue, Boston, MA, 02115, USA. Fax: +1 617 739 8348. E-mail addresses: messex@hsph.harvard.edu, lmelton@hsph.harvard.edu (M. Essex).

Nucleotide sequence accession numbers

The generated sequences were deposited in the GenBank except excluded hypermutated sequences. The accession numbers of the deposited 444 sequences are FJ378092 to FJ378535.

represented by replication-deficient sequences, the archived virus has an enormous potential for viral complementation and survival as a pool of viral quasispecies. It is likely that the amount of integrated provirus and extent of its diversity play important role in establishing and maintaining dual, triple and superinfections. Therefore, the contribution of accumulating provirus to the increase of viral diversity in the course of HIV-1 infection should not be underestimated.

A diverse distribution of HIV-1 subtypes within the HIV/AIDS epidemic has resulted in worldwide dominance by HIV-1 subtype C (Hemelaar et al., 2006; McCutchan, 2006; Osmanov et al., 2002). It is believed that the highest incidence and prevalence of HIV-1 infection in the world is caused by subtype C in southern African countries. Significant progress has been made recently in studying HIV-1 subtype C (Bredell et al., 2007; Coetzer et al., 2007; Ndung'u, Renjifo, and Essex, 2001; Ndung'u et al., 2000, 2006; Novitsky et al., 2002a, 2001a, 2003a, 2001b, 2002b, 2007a; Rademeyer et al., 2004; Rong et al., 2007a; van Harmelen et al., 1998, 2001, 1997, 1999; Williamson et al., 1995, 2003; Williamson, 2003; Zhang et al., 2005, 2006), providing strong evidence for subtype-specific differences on genetic and functional levels (Gnanakaran et al., 2007; Li et al., 2006a; Rademeyer et al., 2007; Rong et al., 2007b). However only few recent studies addressed viral quasispecies in HIV-1 subtype C infection (Rong et al., 2007a; Rousseau et al., 2008; Salazar-Gonzalez et al., 2008). Lingering evolution of subtype C quasispecies is understudied, at least as compared with a series of comprehensive studies on evolution of subtype B viruses (Delwart et al., 2002; Freel et al., 2003; Frost et al., 2005; Gottlieb et al., 2008; Herbeck et al., 2006; Keele et al., 2008; Learn et al., 2002; Leitner, Kumar, and Albert, 1997; Ritola et al., 2004; Shankarappa et al., 1999). A thorough analysis of the sequence of transmitted viruses and the structures of their envelopes may yield clues to help guide vaccine design (Johnston and Fauci, 2007).

It is believed that the population of viral quasispecies in acute HIV-1 infection is relatively uniform; however both homogeneous and heterogeneous virus populations have been reported at or soon after virus transmission (Delwart et al., 1994; Gottlieb et al., 2007a, 2008; Keele et al., 2008; Learn et al., 2002; Long et al., 2000; Markham et al., 1998; Poss et al., 1995; Salazar-Gonzalez et al., 2008; Shankarappa et al., 1999; Wolfs et al., 1992; Zhang et al., 1993; Zhu et al., 1993). A linear increase of viral diversity was estimated at the level of about 1% per year in the C2–V5 region of HIV-1 subtype B *env* (Herbeck et al., 2006; Shankarappa et al., 1999).

Selection pressure is distributed unevenly across HIV-1 Env (Choisy et al., 2004; Gaschen et al., 2002) with clear subtype-specific differences (Gnanakaran et al., 2007). In HIV-1 subtype C, the C3 region of gp120 includes the highly variable  $\alpha$ 2-helix and, together with the V4 region, was shown to be the main target for early neutralizing antibodies (NeuAb) (Moore et al., 2008). Shorter lengths of V1–V5 loops and fewer glycosylation sites, particularly in the V1–V2 region, were associated with development of early strain-specific NeuAb (Gray et al., 2007). The number of potential N-linked glycosylation sites (PNGS) and the sequence length of the C2–V5 region of Env were shown to be relatively stable in infants infected with HIV-1 subtype C (Zhang et al., 2005).

The goals of this case series study were to characterize the evolution of gp120 in acute subtype C infection, to test whether viral diversity and diversification differ between viral RNA load phenotypes of slow and fast decline, and to evaluate changes in the length of variable loops and the number of PNGS in acute HIV-1 subtype C infection in adults.

## Results

Two distinct viral RNA phenotypes were evident at the very early stage of HIV-1 subtype C infection (Fig. 1A, individual curves of viral RNA load, and 1B, early viral RNA set point). The line of evidence for viral RNA phenotypes included the following: 1) differences in decline of viral RNA from peak between groups was confirmed by levels of viral RNA and slopes of viral RNA decline within first 2 months post-seroconversion; 2) differences in the level of an early viral RNA set point from 50 to 200 days post-seroconversion; 3) differences in the levels of CD4+ T cell counts; and 4) difference in decline of CD4+ T cells below 200 within 1 year after seroconversion. The observed phenotypes of viral RNA, namely *slow* and *fast decline* of viral RNA, provided a rationale for looking for potential differences in early viral evolution among subjects experiencing either slow or fast decline of viral RNA at early stage of HIV-1 infection. We hypothesized that viral diversity and diversification are higher in subjects with slow decline of viral RNA and high early viral RNA set point. Conversely, subjects with fast decline of viral RNA and a low level of early viral RNA set point have lower viral diversity and diversification.

### Phylogenetic relationships

All eight subjects in this study were infected by HIV-1 subtype C, which was evident by clustering with four HIV-1 subtype C references (Fig. 2). Clustering with subtype C references was supported in the NJ analysis by bootstrap value of 100 (data not shown). The evolutionary history of proviral sequences was inferred by the maximum likelihood method. As expected, viral quasispecies belonging to different subjects formed distinct clusters (Fig. 2). Subjects with slow decline of viral RNA and a high early viral RNA set point (shown in red) generally had longer and more diversified branches while subjects with fast decline and low early viral RNA set point (shown in green) demonstrated more compact branches in the phylogenetic tree. Two subjects from the group with slow decline of viral RNA, 1811 and 5018 demonstrated distinct intra-patient clusters of quasispecies suggesting a potential infection with more than a single viral variant. The topology of viral sequences along the branches connecting two intra-patient clusters suggested a potential recombination between viral quasispecies belonging to different clusters.

To test the possibility of dual infection, genetic pairwise distances between subjects' MRCA were compared with intra-patient distances including total and intra-cluster distances for cases 1811 and 5018 (Fig. 3). Distances between subjects' MRCA had a mean of  $25.2\% \pm 3.2\%$  ranging from 18.3% to 30.5%. The total intra-patient distances in 1811 and 5018 were  $3.3\% \pm 2.3\%$  and  $1.7\% \pm 1.7\%$ , respectively, while intra-cluster distances were  $0.8\% \pm 0.5\%$  and  $0.7\% \pm 0.5\%$  in subject 1811, and  $0.4\% \pm 0.5\%$  and  $1.1\% \pm 0.3\%$  in subject 5018. Not surprisingly, the pairwise intra-patient distances between clusters accounted for the upmost level of viral intra-patient diversity and were  $5.9\% \pm 0.5\%$  and  $4.1\% \pm 0.4\%$  in 1811 and 5018, respectively. All intra-patient distances in cases 1811 and 5018 were significantly lower than inter-patient distances ( $p < 0.001$  for all comparisons). Therefore, the comparison of inter- and intra-patients distances, and the branch topology in the phylogenetic tree provided evidence that subjects 1811 and 5018 are unlikely to be dually infected with distinct unrelated viral isolates. In contrast, both cases 1811 and 5018 are more likely to represent HIV-1 infection with more than a single viral variant that apparently originated from the pool of diversified viruses at the time of transmission or multiple transmission events from the same (per subject) donor.

We tested 1811 or 5018 sequences for potential recombinants. The presence of breakpoint in bootstrap sliding window analysis was used as criteria for recombination. The branch topology allowed us to identify potential candidates located in-between clusters that are likely to represent recombinants between clusters. In case 1811, ten sequences originating from 91 to 350 days after seroconversion were identified as potential recombinants (Fig. 4A). In nine cases

break-points were identified and these sequences were classified as recombinants. No evidence for recombination was found in a single sequence 350.08. The location of breakpoints was not uniform suggesting a number of independent recombination events. The breakpoints at the end of V3 loop and on the edge of C3 region and V4 loop were observed in multiple viral variants. A total of eight sequences were tested for potential recombination in subject 5018 (Fig. 4B). Only two sequences sampled at day 97 after seroconversion demonstrated breakpoints. None of the remaining sequences tested provided evidence for recombination.

The overall topology of gp120 sequences in subject 1811 suggested a congruent evolution of viral quasispecies within two clusters accompanied by recombination of viral variants between clusters. Interestingly, the cluster 2 was represented by a minor viral variant at 16 days but by a major variant at 97 days, implying a dynamic interplay between apparent immune pressure and virus over the first 3 months after seroconversion. In subject 5018, viral variants representing the seroconversion sequences (shown by a diamond) were evident in cluster 1, but were not seen in cluster 2. It is possible that viral variants representing cluster 2 did exist at the time of seroconversion but were not amplified apparently due to a low copy number.

### Diversity and diversification

Viral population heterogeneity of the proviral gp120 was studied prospectively over the first year after seroconversion (Fig. 5A). The population of proviral quasispecies was relatively homogeneous at seroconversion evident from a mean value of  $0.21\% \pm 0.15\%$  and a tight range from 0.015 to 0.36% in most subjects except the case 1811 infected with more than a single viral variant. At day 16, the overall diversity in subject 1811 was  $1.45\% \pm 2.24\%$  but only  $0.35\% \pm 0.23\%$  within cluster 1 (no data for cluster 2 because a single sequence available at day 16). At seroconversion, the extent of proviral diversity did not differ between subjects with slow and fast decline of viral RNA ( $p=0.18$  and  $p=0.09$  for uncorrected and cluster-corrected comparisons). However after seroconversion, proviral diversity in gp120 varied between subjects and the difference reached statistical significance between groups with different viral RNA phenotype. Fig. 5B demonstrates that proviral diversity in subjects with slow decline of viral RNA is higher than in subjects with fast decline of viral RNA load. During the first 200 days after seroconversion, the mean diversity was  $0.24\% \pm 0.18\%$  in the group with fast decline of viral RNA, while it was  $1.40\% \pm 0.65\%$  and  $0.73\% \pm 0.22\%$  for uncorrected and cluster-corrected distances in the group with slow decline of viral RNA (Fig. 5C;  $p=0.029$  and  $p=0.006$  for uncorrected and cluster-corrected comparison, respectively). However, the significance between groups was lost at later time points, in part, probably due to increased fluctuations apparent from high standard deviations in groups (Fig. 5C; 201+ days comparisons). A gradual increase of proviral diversity was evident by positive slopes in all subjects by 200 days after seroconversion, ranging from  $7.6E-06$  in subject 3603 to  $2.0E-04$  in subject 1811. Slopes of proviral diversity differed between groups of viral RNA phenotypes and were higher in subjects with slow decline of viral RNA for the period 0 to 200 days after seroconversion ( $p=0.031$ ). By the end of the first year of infection the mean diversity of proviral quasispecies in gp120 increased to  $1.72\% \pm 1.04\%$  for uncorrected and  $1.32\% \pm 0.76\%$  for the cluster-corrected distances. The diversity was higher in the group with slow decline of viral RNA at the time period 301+ days in the uncorrected analysis ( $p=0.015$ ; 95% CI for difference of means 0.54 to 2.87) but was not seen in the cluster-corrected analysis ( $p>0.05$ ).

The shape of viral diversification resembled graphs of viral diversity (Fig. 6A). The mean diversification of viral quasispecies from MRCA for the period from 0 to 200 days after seroconversion was  $1.02\% \pm 1.04\%$  and  $0.89\% \pm 0.86\%$  for uncorrected and cluster-corrected distances, respectively. For the later time points, the mean diversity from MRCA increased to  $1.25\% \pm 0.79\%$  and  $1.57\% \pm 1.17\%$  for uncorrected and cluster-corrected distances, respectively. This was close but a little higher than the previously reported 1%-per-year increase of *env*

C2V5 diversity in HIV-1 subtype B infection (Herbeck et al., 2006; Shankarappa et al., 1999). Similarly to proviral diversity, viral diversification from MRCA differed between the viral RNA phenotypes. During 0 to 200 days after seroconversion diversity from MRCA reached  $1.74\% \pm 1.04\%$  and  $2.01\% \pm 1.40\%$  for uncorrected and corrected distances in the group with slow decline of viremia, while remaining at mean of  $0.30\% \pm 0.22\%$  in subjects with fast decline of viral RNA load (Fig. 6B;  $p=0.029$  and  $p=0.044$  for uncorrected and cluster-corrected comparisons). At later time points, only a trend to a higher viral diversification from MRCA was observed in the group with slow decline of viral load ( $p=0.095$  and  $p=0.079$  for uncorrected and cluster-corrected comparisons), which can be attributable to higher inter-patient variation and relatively small sample size.

### Length of variable loops and number of PNGS

To address changes in length of variable loops in proviral gp120, we quantified the amino acid length of five variable loops and C3 region over the first year of infection and computed the difference in amino acid length from the time of seroconversion (Fig. 7A). We found no or minimal change in the length of variable loops in HIV-1 subtype C infection over the first year after seroconversion. The V1 loop was predominantly reduced, which was evident by negative slopes in five out of eight subjects. In the V2 loop, three cases had a slight increase, while three subjects had no length change and two had decreases in length. The length of the V3 loop and the C3 region virtually did not change over the first year of infection. The V4 loop was reduced in two cases of infection with multiple viral variants (subjects 1811 and 5018), and was almost unchanged in the other six subjects. The V5 loop length was unchanged in three, decreased in four, and increased in one subject. Interestingly, while there were minimal or no change in length of variable loops within clusters (cases 1811 and 5018), the length of variable loops differed between clusters. For example, clusters in subject 1811 showed lengths of 26 and 22 aa in V1, 48 and 43 aa in V2, 52 and 51 aa in C3 region, and 30–31 and 26 aa in V4. Clusters in subject 5018 were different only in V4 loop, demonstrating lengths of 29 and 26 aa. No difference in the V-loop length was found between groups with slow and fast decline of viral load (data not shown).

The number of PNGS was relatively stable and the mean values fluctuated within  $\pm$ one amino acid within the first year of HIV-1 subtype C infection (Fig. 7B). The overall pattern of changes in the number of PNGS was similar to the amino acid length changes over time. The number of PNGS in the V1 loop showed negative slopes in five out of eight cases. Changes in the V2 and V3 loops were minimal. In the C3 region, an increase of the number of PNGS was seen in four out of eight subjects including subject 3505 who increased the number of PNGS gradually from 2 to 3 over the first year of infection. In the V4 loop, the number of PNGS increased in 3 subjects, stayed unchanged in one subject, and decreased in 4 cases. The number of PNGS in the V5 loop showed an increase in 3 cases, stayed unchanged in 2 cases, and decreased in 3 cases. Some differences between the number of PNGS were observed between clusters in subjects 1811 but not in subject 5018. Thus, clusters in subject 1811 had 3 and 2 PNGS in V2, 2 and 3 PNGS in C3 region, 4 and 3 PNGS in V4, and did not differ in V1, V3, and V5 loops. No significant differences in the number of PNGS between groups with slow and fast decrease of viral load were found (data not shown).

### Discussion

A better understanding of early events in HIV-1 subtype C infection may facilitate proper intervention strategies and better vaccine design. We analyzed the evolution of proviral quasispecies in the V1C5 region of gp120 during the first post-seroconversion year in eight acute cases of HIV-1 subtype C infection from Botswana. Despite the inherent limitations of a small sample set, the results of this study provide evidence that viral evolution in subtype C



infection might have some specific characteristics. First, infection with more than a single viral variant is not rare in HIV-1 subtype C infection. Second, early viral RNA set point from 50 to 200 days post-seroconversion seems to be associated, at least temporarily, with proviral diversity and diversification in subtype C gp120. Third, no pattern of increase in the length of variable loops was observed. Fourth, the increase in the number of PNGS in proviral gp120 was clearly not a dominant pattern of viral evolution over the first post-seroconversion year.

Two out of eight subjects in this study were infected with more than a single variant of subtype C viruses. Dual infections originating from multiple transmissions of non-related viruses from multiple donors (“multiple donors to a single recipient transmission”) were ruled out based on phylogenetic analysis and comparison of inter- and intra-patient pairwise distances. The most plausible explanation of cases 1811 and 5018 is a transmission of diversified viral variants from a single (per subject) donor. Multiple transmissions from the same donor occurring over time cannot be ruled out, and apparently are likely to happen. Both cases of infection with more than a single viral variant in this study had a high viral load accompanied by fast decline of CD4+ T cells, which was consistent with described previously rapid disease progression of HIV-1 superinfection (Gottlieb et al., 2004, 2007b; Grobler et al., 2004; Jost et al., 2002; Liu et al., 1997; Sagar et al., 2003). Infection with multiple viral variants dramatically increases the overall viral diversity and may lead to recombinant viral variants within a short period of time. It is likely that multiple viral variants might contribute to fast exhaustion of immune response and a reduced capacity of the immune system to contain viral replication. While elevated levels of proviral diversity and diversification in cases with multiple viral variants are readily predicted, no difference in the length of variable loops or number of PNGS was found between HIV-1 subtype C infections with single or multiple viral variants except V4 reduction in both 1811 and 5018 cases. Overall, these findings are consistent with previous reports on infections with multiple viral variants in HIV-1 subtype C (Coetzer et al., 2007; Grobler et al., 2004; Williamson et al., 2008), although most previous studies did not separate infections originating from multiple donors and infections with diversified viruses from a single donor. In Tanzania, the frequency of HIV-1 infections with multiple viral variants ranged from 9% to 19% depending on the risk (Herbinger et al., 2006), and was up to 25% in a high-risk cohort (McCutchan et al., 2004). Whether frequency of multiple infections differs between HIV-1 subtypes is still unknown, as well as the mechanistic reason for HIV-1 infections with multiple viral variants. What is important is that HIV-1 infections with multiple viral variants present an additional challenge for vaccine development.

Results of this study suggest the existence of differential pathways of viral evolution during the early phase of HIV-1 subtype C infection. Our data support the idea that viral population is relatively homogeneous at the time of seroconversion (Gottlieb et al., 2007a; Shankarappa et al., 1999; Zhu et al., 1993), unless subjects are infected with more than a single viral variant, in which case virus population is predictably heterogeneous. We report different patterns of viral evolution dependent on the decline of viral RNA and the early viral RNA set point, which is consistent with the assumption that higher levels of viral replication cause higher levels of viral diversity and diversification over time. In subjects with suboptimal suppression of viral replication after seroconversion we observed higher viral diversification. In contrast, in subjects with efficient suppression of viral replication, at least by 200 days after seroconversion, the diversification of virus was lower. Three out of four subjects with more diversified proviral gp120 dropped CD4+ T cell count below 200 within the first year after seroconversion and initiated ARV treatment, while none of four subjects with lower proviral diversity did. Whether ARV treatment alters proviral diversity in HIV-1 subtype C gp120 needs to be addressed in future studies.

We found no or minimal length variation in variable loops and the number of PNGS in gp120 during the first year after seroconversion, which is consistent with the recent report from South

Africa (Coetzer et al., 2007). In contrast, studies from Zambia suggested preferential transmission of HIV-1 subtype C with compact variable loops and a reduced number of PNGS (Derdeyn et al., 2004; Li et al., 2006a). Transmission of viruses with shorter and less glycosylated Env implies a gradual increase in variable loop length and in the number of PNGS over the course of infection, which has not been confirmed experimentally. While shorter variable loops in natural subtype C infections are more likely to induce broad NeuAb, no association between the number of PNGS and induction of NeuAb was reported (Rademeyer et al., 2007). The loop length and the number of PNGS differ between HIV-1 subtypes. For example, newly transmitted subtype B viruses have longer and more glycosylated gp120 than newly transmitted subtype C viruses (Li et al., 2006b), and V1V2 loops are shorter in subtype A infections (Chohan et al., 2005). However within HIV-1 subtype C no difference in the length of variable loops between newly transmitted and chronic viruses has been reported (Li et al., 2006b), which is consistent with our findings. The number of PNGS in variable loops of Env showed minimal fluctuations and no clear patterns of intra- and inter-subject variation of PNGS were observed over the first year of HIV-1 subtype C infection.

In summary, the results of the study suggest that the frequency of HIV-1 subtype C infections with more than a single viral variant is not rare and provide a characterization of the proviral evolution of gp120 in an early phase of HIV-1 subtype C infection. We found differential pathways of viral evolution dependent on viral replication during the first 200 days after. There was minimal evolution of the variable loop length and the number of PNGS over the first year of HIV-1 subtype C infection.

## Methods

### Study subjects

Eight cases of acute HIV-1 subtype C infections were identified as part of a primary HIV-1 infection study in Botswana (Novitsky et al., 2007b), approved by the Institutional Review Boards in Botswana and the US. A positive HIV-1 RT-PCR accompanied by a negative HIV-1 serology, the RNA<sup>+</sup>Ab<sup>-</sup> status was used as inclusion criteria for enrollment. Following identification as being acutely infected with HIV, participants had weekly and bi-weekly visits for the first 4 months, and monthly visits for the following 8 months. The analyzed subjects included 2 males and 6 females whose age ranged from 20 to 53 years old. The mean value of viral RNA peak was  $6.25 \pm 0.92 \log_{10}$  copies/ml. Two patterns in evolution of viral RNA and CD4<sup>+</sup> T cells were evident suggesting slow or fast decline of viral RNA during the first two post-seroconversion months ( $5.47 \pm 0.46$  vs.  $3.72 \pm 0.65 \log_{10}$ ;  $p = 0.005$ ). While the viral RNA peaks did not differ between subjects with slow or fast decline of viral RNA ( $6.79 \pm 0.97 \log_{10}$  vs.  $5.72 \pm 0.53 \log_{10}$  for slow and fast decline, respectively; 95% CI for difference of means:  $-0.28$  to  $2.41$ ;  $p = 0.10$ ), the rates of viral RNA decline measured by slopes were different between groups ( $-0.0128 \pm 0.0058 \log_{10}$  per day vs  $-0.0275 \pm 0.0082 \log_{10}$  per day for groups with slow and fast RNA decline, respectively;  $p = 0.03$ ). The assumption was made that the peak of viral RNA should decline by day 50 after seroconversion. The early viral RNA set point defined as mean  $\pm$  standard deviation of plasma RNA measurements from 50 to 200 days differed between subjects with slow and fast decline of viral RNA (Fig. 1;  $5.25 \pm 0.22$  vs.  $3.62 \pm 0.45 \log_{10}$  for slow and fast decline of viral RNA,  $p < 0.001$ ). Subjects 1811, 2865, 3312, and 5018 showed slow decline of viral RNA and a high early viral RNA set point. Subjects 3430, 3505, 3603, and 5582 demonstrated fast decline of viral RNA and a low early set point. The distinct viral RNA phenotypes were associated with different CD4<sup>+</sup> T cell count in subjects ( $271 \pm 59$  vs.  $491 \pm 38$ ; 95% CI for difference of means  $-305$  to  $-134$ ;  $p < 0.001$ , and  $250 \pm 71$  vs.  $482 \pm 57$ ; 95% CI for difference of means  $-344$  to  $-121$ ;  $p = 0.002$  for CD4 comparisons between groups during the first 2 months and 2 to 6 months, respectively). Access to ART was free of charge if participants' CD4<sup>+</sup> T cell count dropped below 200 cells/mm<sup>3</sup>, or if they had an

AIDS-defining illness. Three out of four subjects with slow decline of viral RNA dropped their CD4+ T cell count below 200 and initiated ART within the first year after seroconversion. None of four subjects with fast decline of viral RNA dropped CD4+ T cells below 200 and did not initiate ART within the first post-seroconversion year.

### Single genome amplification and sequencing

The proviral DNA was amplified from PBMC isolated at multiple visits from seroconversion up to 440 days thereafter. In this study, time 0 corresponded to the time of seroconversion. A strategy of single-genome amplification by limiting dilutions and sequencing close to the recently published method (Salazar-Gonzalez et al., 2008) with some modifications was employed. The targeted region of proviral gp120 spanned the V1 loop to the C terminus end of gp120, V1C5, and corresponded to nucleotide positions 6615 to 7757 of HXB2 (amino acids 131 to 511 in relation to the gp120 CDS in HXB2). Amplicons were purified by Exo-SAP (Dugan et al., 2002), and sequenced directly on the ABI 3730 DNA Analyzer using the BigDye technology. About 10 viral sequences per subject per time point were analyzed. The obtained sequences were tested by HYPERMUT v.2.0 (Rose and Korber, 2000) and hypermutated sequences were excluded. Diversity of quasispecies, divergence from the most recent common ancestor (MRCA) and earliest viral quasispecies over time, length of variable loops, and number of PNGS were longitudinally characterized. Sequences from ARV-treated subjects 1811, 2865, and 3312 were included in the analysis. Time of ART initiation is indicated by arrows in Figs. 5 and 6.

### Diversity and divergence

Alignment of nucleotide sequences was performed and refined by Muscle (Edgar, 2004) followed by a manual adjustment in BioEdit (Hall, 1999). The maximum-likelihood (ML) method was used to estimate pairwise nucleotide distances. PAUP\* version 4.0 (Swofford, 2003) and Modeltest version 3.7 (Posada and Buckley, 2004; Posada and Crandall, 1998) with the Akaike information criterion were utilized to identify an appropriate substitution model. The identified substitution model was used in PAUP\* to estimate ML-corrected pairwise distances. The proviral diversity per subject per time point was computed as a mean value  $\pm$  standard deviation of pairwise distances between proviral quasispecies at the given time point. The mean values and standard deviations are shown in figures as fractions of 1, and presented as percentages in the text for convenience. Quantification of viral diversification was performed by using a mean value  $\pm$  standard deviation of pairwise distances between viral quasispecies at the given time point and a subject's MRCA or consensus sequence of the viral quasispecies from the earliest time point. The MRCA for each subject was reconstructed at the basal node of viral quasispecies collected at all time points using ML criteria in PAUP\*. The alignment was gap-stripped, and single sequences from other seven subjects were used as outgroups, as described elsewhere (Learn et al., 2002). The majority consensus sequence for the earliest quasispecies was built in BioEdit.

### Phylogenetic analysis

The genealogy reconstruction of the proviral sequences was implemented in PhyML (Guindon and Gascuel, 2003) using the HKY model of nucleotide substitution. The branching topology of generated maximum-likelihood phylogenetic tree in PhyML was comparable with NJ trees generated by MEGA4. Only maximum likelihood trees visualized by FigTree v1.1.2 (Rambaut, 2008) are presented. For HIV-1 subtyping purpose, a total of 35 reference sequences representing HIV-1 M group subtypes from A to H from Los-Alamos HIV Sequence Database (<http://www.hiv.lanl.gov/>) were included in the analysis.



## Statistical methods

Data are summarized with means±standard deviations and correlations. Comparisons between groups, including grouping of subjects with slow and fast decline of viremia, were based on *t*-tests and Mann–Whitney Rank Sum tests for continuous and binary outcomes, respectively. All reported *p*-values are 2-sided.

## Acknowledgments

We are grateful to the subjects from the *Tshedimoso* study in Botswana. We thank Gaseboloke Mothowaeng, Florence Modise, S'khatele Molehabangwe, and Sarah Masole for their dedication and outstanding work in the clinic and outreach. We express thanks to Busisiwe Mlotshwa for excellent laboratory support. We greatly appreciate the enthusiasm and strong commitment of Erin McDonald, Melissa Ketunuti, Carl Davis, Kenneth Onyait, and Mary Fran McLane in achieving the overall study goals. 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 the NIH grant D43 TW000004 (DN) and also through the AAMC FIC/Ellison Overseas Fellowships in Global Health and Clinical Research (LK and RR).

## References

- Bredell H, Martin DP, Van Harmelen J, Varsani A, Sheppard HW, Donovan R, Gray CM, Williamson C. HIV type 1 subtype C gag and nef diversity in Southern Africa. *AIDS Res Hum Retrovir* 2007;23:477–481. [PubMed: 17411382]
- Chohan B, Lang D, Sagar M, Korber B, Lavreys L, Richardson B, Overbaugh J. Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1–V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels. *J Virol* 2005;79:6528–6531. [PubMed: 15858037]
- Choisy M, Woelk CH, Guegan JF, Robertson DL. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J Virol* 2004;78:1962–1970. [PubMed: 14747561]
- Coetzer M, Cilliers T, Papathanasopoulos M, Ramjee G, Karim SA, Williamson C, Morris L. Longitudinal analysis of HIV type 1 subtype C envelope sequences from South Africa. *AIDS Res Hum Retrovir* 2007;23:316–321. [PubMed: 17331039]
- Delwart EL, Sheppard HW, Walker BD, Goudsmit J, Mullins JI. Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays. *J Virol* 1994;68:6672–6683. [PubMed: 8084001]
- Delwart E, Magierowska M, Royz M, Foley B, Peddada L, Smith R, Heldebrant C, Conrad A, Busch M. Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection. *AIDS* 2002;16:189–195. [PubMed: 11807302]
- Derdeyn CA, Decker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, Denham SA, Heil ML, Kasolo F, Musonda R, Hahn BH, Shaw GM, Korber BT, Allen S, Hunter E. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science* 2004;303:2019–2022. [PubMed: 15044802]
- 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: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5:113. [PubMed: 15318951]
- Freel SA, Fiscus SA, Pilcher CD, Menezes P, Giner J, Patrick E, Lennox JL, Hicks CBJJ Jr, Shugars DC. Envelope diversity, coreceptor usage and syncytium-inducing phenotype of HIV-1 variants in saliva and blood during primary infection. *AIDS* 2003;17:2025–2033. [PubMed: 14502005]
- Frost SD, Liu Y, Pond SL, Chappey C, Wrin T, Petropoulos CJ, Little SJ, Richman DD. Characterization of human immunodeficiency virus type 1 (HIV-1) envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B. *J Virol* 2005;79:6523–6527. [PubMed: 15858036]

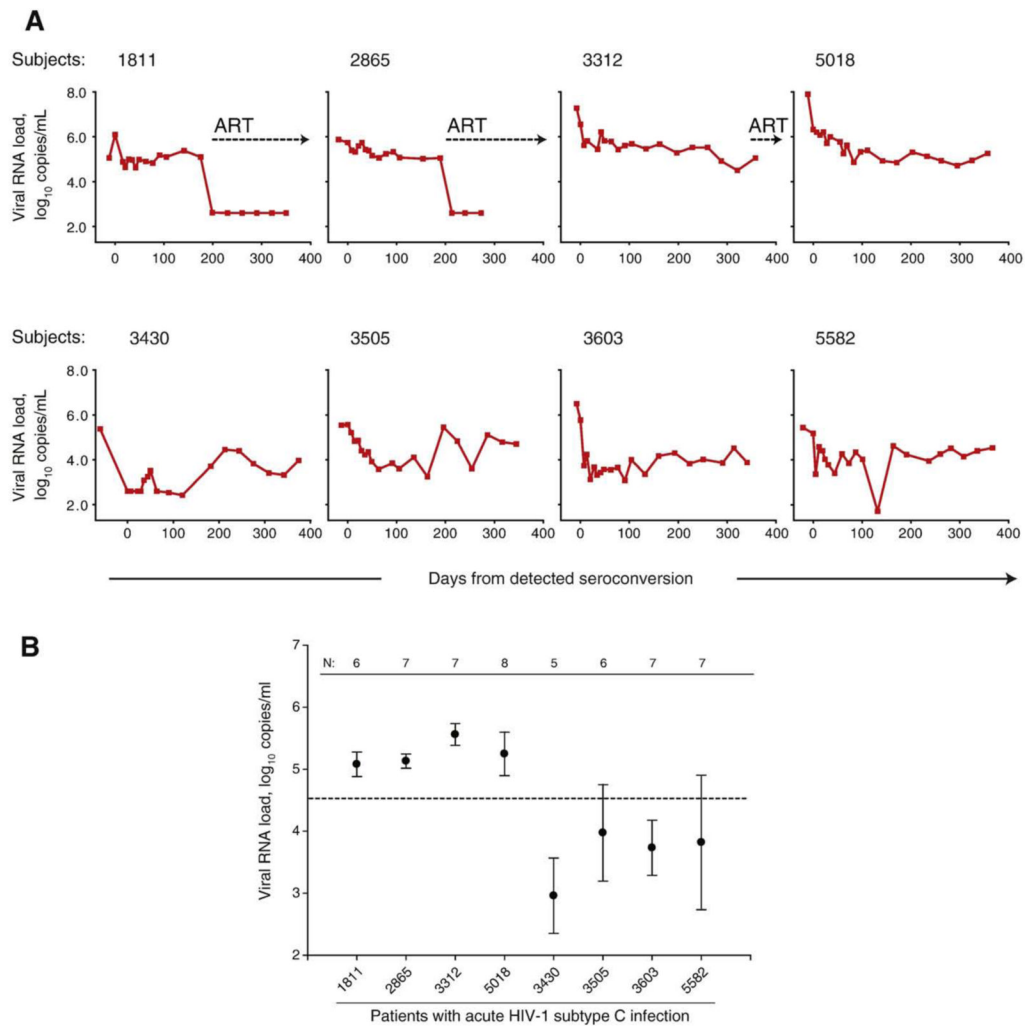
- Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T, Korber B. Diversity considerations in HIV-1 vaccine selection. *Science* 2002;296:2354–2360. [PubMed: 12089434]
- Gnanakaran S, Lang D, Daniels M, Bhattacharya T, Derdeyn CA, Korber B. Clade-specific differences between human immunodeficiency virus type 1 clades B and C: diversity and correlations in C3–V4 regions of gp120. *J Virol* 2007;81:4886–4891. [PubMed: 17166900]
- Gottlieb GS, Nickle DC, Jensen MA, Wong KG, Grobler J, Li F, Liu SL, Rademeyer C, Learn GH, Karim SS, Williamson C, Corey L, Margolick JB, Mullins JI. Dual HIV-1 infection associated with rapid disease progression. *Lancet* 2004;363:619–622. [PubMed: 14987889]
- Gottlieb GS, Heath L, Nickle DC, Wong KG, Leach S, Jacobs B, Gezahegne S, von't Woul AB, Jacobson LP, Margolick JB, Mullins JI. Complex HIV-1 Populations Prior to Seroconversion in MSM: Analysis of HIV-1 Plasma RNA Positive-Seronegative Subjects from the MACS. *CROI 2007a*;2007
- Gottlieb GS, Nickle DC, Jensen MA, Wong KG, Kaslow RA, Shepherd JC, Margolick JB, Mullins JI. HIV type 1 superinfection with a dualtropic virus and rapid progression to AIDS: a case report. *Clin Infect Dis* 2007b;45:501–509. [PubMed: 17638203]
- Gottlieb GS, Heath L, Nickle DC, Wong KG, Leach SE, Jacobs B, Gezahegne S, van 't Wout AB, Jacobson LP, Margolick JB, Mullins JI. HIV-1 variation before seroconversion in men who have sex with men: analysis of acute/early HIV infection in the multicenter AIDS cohort study. *J Infect Dis* 2008;197:1011–1015. [PubMed: 18419538]
- Gray ES, Moore PL, Choge IA, Decker JM, Bibollet-Ruche F, Li H, Leseka N, Treurnicht F, Mlisana K, Shaw GM, Karim SSA, Williamson C, Morris L. the CST. Neutralizing antibody responses in acute human immunodeficiency virus type 1 subtype C infection. *J Virol* 2007;81:6187–6196. [PubMed: 17409164]
- Grobler J, Gray CM, Rademeyer C, Seoighe C, Ramjee G, Karim SA, Morris L, Williamson C. Incidence of HIV-1 dual infection and its association with increased viral load set point in a cohort of HIV-1 subtype C-infected female sex workers. *J Infect Dis* 2004;190:1355–1359. [PubMed: 15346349]
- Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003;52:696–704. [PubMed: 14530136]
- 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.
- Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* 2006;20:W13–23. [PubMed: 17053344]
- Herbeck JT, Nickle DC, Learn GH, Gottlieb GS, Curlin ME, Heath L, Mullins JI. Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host. *J Virol* 2006;80:1637–1644. [PubMed: 16439520]
- Herbinger KH, Gerhardt M, Piyasirisilp S, Mloka D, Arroyo MA, Hoffmann O, Maboko L, Birx DL, Mmbando D, McCutchan FE, Hoelscher M. Frequency of HIV type 1 dual infection and HIV diversity: analysis of low- and high-risk populations in Mbeya Region, Tanzania. *AIDS Res Hum Retrovir* 2006;22:599–606. [PubMed: 16831083]
- Johnston MI, Fauci AS. An HIV vaccine—evolving concepts. *N Engl J Med* 2007;356:2073–2081. [PubMed: 17507706]
- Jost S, Bernard MC, Kaiser L, Yerly S, Hirschel B, Samri A, Autran B, Goh LE, Perrin L. A patient with HIV-1 superinfection. *N Engl J Med* 2002;347:731–736. [PubMed: 12213944]
- Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping LH, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH, Shaw GM. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* 2008;105:7552–7557. [PubMed: 18490657]
- Learn GH, Muthui D, Brodie SJ, Zhu T, Diem K, Mullins JI, Corey L. Virus population homogenization following acute human immunodeficiency virus type 1 infection. *J Virol* 2002;76:11953–11959. [PubMed: 12414937]

- Leitner T, Kumar S, Albert J. Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J Virol* 1997;71:4761–4770. [PubMed: 9151870]
- Li B, Decker JM, Johnson RW, Bibollet-Ruche F, Wei X, Mulenga J, Allen S, Hunter E, Hahn BH, Shaw GM, Blackwell JL, Derdeyn CA. Evidence for potent autologous neutralizing antibody titers and compact envelopes in early infection with subtype C human immunodeficiency virus type 1. *J Virol* 2006a;80:5211–5218. [PubMed: 16699001]
- Li M, Salazar-Gonzalez JF, Derdeyn CA, Morris L, Williamson C, Robinson JE, Decker JM, Li Y, Salazar MG, Polonis VR, Mlisana K, Karim SA, Hong K, Greene KM, Bilska M, Zhou J, Allen S, Chomba E, Mulenga J, Vwalika C, Gao F, Zhang M, Korber BTM, Hunter E, Hahn BH, Montefiori DC. Genetic and neutralization properties of subtype C human immunodeficiency virus type 1 molecular *env* clones from acute and early heterosexually acquired infections in Southern Africa. *J Virol* 2006b; 80:11776–11790. [PubMed: 16971434]
- Liu SL, Schacker T, Musey L, Shriner D, McElrath MJ, Corey L, Mullins JI. Divergent patterns of progression to AIDS after infection from the same source: human immunodeficiency virus type 1 evolution and antiviral responses. *J Virol* 1997;71:4284–4295. [PubMed: 9151816]
- Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* 1999;73:152–160. [PubMed: 9847317]
- Long EM, Martin HL Jr, Kreiss JK, Rainwater SM, Lavreys L, Jackson DJ, Rakwar J, Mandaliya K, Overbaugh J. Gender differences in HIV-1 diversity at time of infection. *Nat Med* 2000;6:71–75. [PubMed: 10613827]
- Markham RB, Wang WC, Weisstein AE, Wang Z, Munoz A, Templeton A, Margolick J, Vlahov D, Quinn T, Farzadegan H, Yu XF. Patterns of HIV-1 evolution in individuals with differing rates of CD4 T cell decline. *Proc Natl Acad Sci U S A* 1998;95:12568–12573. [PubMed: 9770526]
- McCutchan FE. Global epidemiology of HIV. *J Med Virol* 2006;78(Suppl 1):S7–S12. [PubMed: 16622870]
- McCutchan, F.; Piyasirisilp, S.; Hoffmann, O.; Sanders-Buell, E.; Wilson, G.; Tovanabutra, S.; Bix, DL.; Hoelscher, M.; Maboko, L. Detecting HIV-1 dual infections in a high-risk cohort in Tanzania. Abstract number TuOrA1139. *Int. Conf. AIDS; Bangkok; Thailand. Jul 11–16, 2004; 2004.*
- Moore PL, Gray ES, Choge IA, Ranchope N, Mlisana K, Karim SS, Williamson C, Morris L. The C3–V4 region is a major target of autologous neutralizing antibodies in hiv-1 subtype C infection. *J Virol* 2008;82:1860–1869. [PubMed: 18057243]
- Ndung'u T, Renjifo B, Essex M. Construction and analysis of an infectious human immunodeficiency virus type 1 subtype C molecular clone. *J Virol* 2001;75:4964–4972. [PubMed: 11333875]
- Ndung'u T, Renjifo B, Novitsky VA, McLane MF, Gaolekwe S, Essex M. Molecular cloning and biological characterization of full-length HIV-1 subtype C from Botswana. *Virology* 2000;278:390–399. [PubMed: 11118362]
- Ndung'u T, Sepako E, McLane MF, Chand F, Bedi K, Gaseitsiwe S, Doualla-Bell F, Peter T, Thior I, Moyo SM, Gilbert PB, Novitsky VA, Essex M. HIV-1 subtype C in vitro growth and coreceptor utilization. *Virology* 2006;347:247–260. [PubMed: 16406460]
- Novitsky V, Flores-Villanueva PO, Chigwedere P, Gaolekwe S, Bussman H, Sebetso G, Marlink R, Yunis EJ, Essex M. Identification of most frequent HLA class I antigen specificities in Botswana: relevance for HIV vaccine design. *Hum Immunol* 2001a;62:146–156. [PubMed: 11182225]
- Novitsky V, Rybak N, McLane MF, Gilbert P, Chigwedere P, Klein I, Gaolekwe S, Chang SY, Peter T, Thior I, Ndung'u T, Vannberg F, Foley BT, Marlink R, Lee TH, Essex M. Identification of human immunodeficiency virus type 1 subtype C Gag-, Tat-, Rev-, and Nef-specific elispot-based cytotoxic T-lymphocyte responses for AIDS vaccine design. *J Virol* 2001b;75:9210–9228. [PubMed: 11533184]
- Novitsky V, Cao H, Rybak N, Gilbert P, McLane MF, Gaolekwe S, Peter T, Thior I, Ndung'u T, Marlink R, Lee TH, Essex M. Magnitude and frequency of cytotoxic T-lymphocyte responses: identification of immunodominant regions of human immunodeficiency virus type 1 subtype C. *J Virol* 2002a; 76:10155–10168. [PubMed: 12239290]

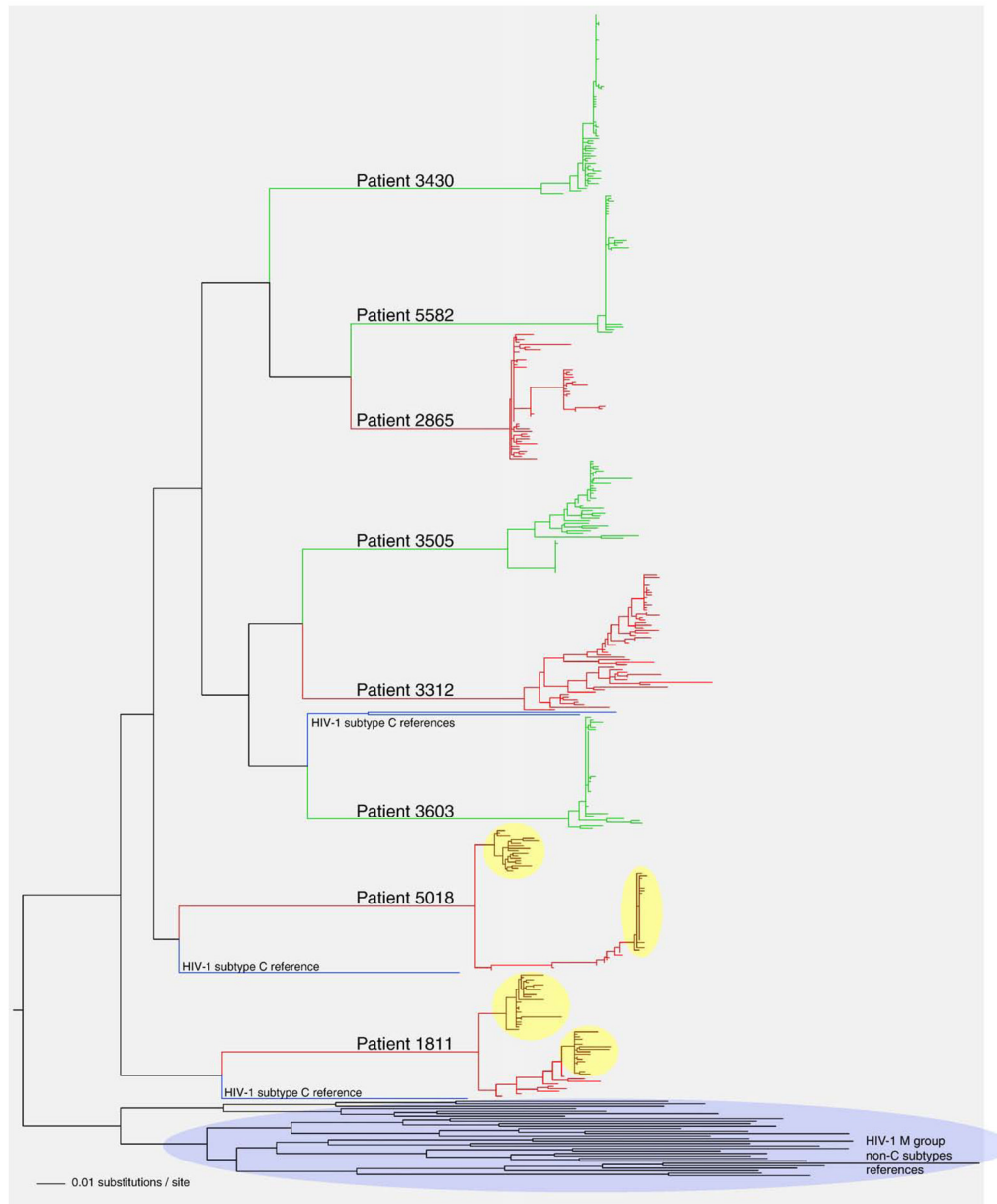
- Novitsky V, Smith UR, Gilbert P, McLane MF, Chigwedere P, Williamson C, Ndung'u T, Klein I, Chang SY, Peter T, Thior I, Foley BT, Gaolekwe S, Rybak N, Gaseitsiwe S, Vannberg F, Marlink R, Lee TH, Essex M. HIV-1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? *J Virol* 2002b;76:5435–5451. [PubMed: 11991972]
- Novitsky V, Gilbert P, Peter T, McLane MF, Gaolekwe S, Rybak N, Thior I, Ndung'u T, Marlink R, Lee TH, Essex M. Association between virus-specific T-cell responses and plasma viral load in HIV-1 subtype C infection. *J Virol* 2003;77:882–890. [PubMed: 12502804]
- Novitsky V, Wester CW, DeGruttola V, Bussmann H, Gaseitsiwe S, Thomas A, Moyo S, Musonda R, Widenfelt E, Marlink RG, Essex M. The reverse transcriptase 67N 70R 215Y genotype is the predominant TAM pathway associated with virologic failure among HIV-1C-infected adults treated with ZDV/ddI-containing HAART in Southern Africa. *AIDS Res Hum Retrovir* 2007a;23:868–878. [PubMed: 17678469]
- Novitsky V, Woldegabriel E, Wester C, McDonald E, Rossenkhan R, Ketunuti M, Makhema J, Seage GR III, Essex M. Identification of primary HIV-1C infection in Botswana. *AIDS Care* 2007b;20(7): 806–811. [PubMed: 18608056]
- Osmanov S, Pattou C, Walker N, Schwardlander B, Esparza J. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. *J Acquir Immune Defic Syndr Human Retrovirol* 2002;29:184–190.
- Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 1998;14:817–818. [PubMed: 9918953]
- Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol* 2004;53:793–808. [PubMed: 15545256]
- Poss M, Martin HL, Kreiss JK, Granville L, Chohan B, Nyange P, Mandaliya K, Overbaugh J. Diversity in virus populations from genital secretions and peripheral blood from women recently infected with human immunodeficiency virus type 1. *J Virol* 1995;69:8118–8122. [PubMed: 7494333]
- Rademeyer C, van Harmelen JH, Ramjee G, Karim SS, Williamson C. Heterosexual transmission of multiple highly conserved viral variants in HIV-1 subtype C-infected seronegative women. *AIDS* 2004;18:2096–2098. [PubMed: 15577636]
- Rademeyer C, Moore PL, Taylor N, Martin DP, Choge IA, Gray ES, Sheppard HW, Gray C, Morris L, Williamson C. Genetic characteristics of HIV-1 subtype C envelopes inducing cross-neutralizing antibodies. *Virology* 2007;368:172–181. [PubMed: 17632196]
- Rambaut, A. FigTree v1.1.2. 2008. <http://tree.bio.ed.ac.uk/software/figtree>
- Ritola K, Pilcher CD, Fiscus SA, Hoffman NG, Nelson JAE, Kitrinos KM, Hicks CB, Eron JJ Jr, Swanstrom R. Multiple V1/V2 env variants are frequently present during primary infection with human immunodeficiency virus type 1. *J Virol* 2004;78:11208–11218. [PubMed: 15452240]
- Rong R, Bibollet-Ruche F, Mulenga J, Allen S, Blackwell JL, Derdeyn CA. Role of V1V2 and other human immunodeficiency virus type 1 envelope domains in resistance to autologous neutralization during clade C infection. *J Virol* 2007a;81:1350–1359. [PubMed: 17079307]
- Rong R, Gnanakaran S, Decker JM, Bibollet-Ruche F, Taylor J, Sfakianos JN, Mokili JL, Muldoon M, Mulenga J, Allen S, Hahn BH, Shaw GM, Blackwell JL, Korber BT, Hunter E, Derdeyn CA. Unique mutational patterns in the envelope {alpha}2 amphipathic helix and acquisition of length in gp120 hypervariable domains are associated with resistance to autologous neutralization of subtype C human immunodeficiency virus type 1. *J Virol* 2007b;81:5658–5668. [PubMed: 17360739]
- Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G→A hypermutation. *Bioinformatics* 2000;16:400–401. [PubMed: 10869039]
- Rousseau CM, Daniels MG, Carlson JM, Kadie C, Crawford H, Prendergast A, Matthews P, Payne R, Rolland M, Raugi DN, Maust BS, Learn GH, Nickle DC, Coovadia H, Ndung'u T, Frahm N, Brander C, Walker BD, Goulder PJR, Bhattacharya T, Heckerman DE, Korber BT, Mullins JI. HLA class-I driven evolution of human immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. *J Virol* 2008;82:6434–6446. [PubMed: 18434400]
- Sagar M, Lavreys L, Baeten JM, Richardson BA, Mandaliya K, Chohan BH, Kreiss JK, Overbaugh J. Infection with multiple human immunodeficiency virus type 1 variants is associated with faster disease progression. *J Virol* 2003;77:12921–12926. [PubMed: 14610215]

- Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, Derdeyn CA, Farmer P, Hunter E, Allen S, Manigart O, Mulenga J, Anderson JA, Swanstrom R, Haynes BF, Athreya GS, Korber BTM, Sharp PM, Shaw GM, Hahn BH. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* 2008;82:3952–3970. [PubMed: 18256145]
- 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]
- Swofford, DL. PAUP\*. Phylogenetic Analysis Using Parsimony (3and Other Methods). Version 4.0. Sinauer Associates; Sunderland, Massachusetts: 2003.
- van Harmelen J, Wood R, Lambrick M, Rybicki EP, Williamson AL, Williamson C. An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. *AIDS* 1997;11:81–87. [PubMed: 9110079]
- van Harmelen, J.; Bredell, H.; Morris, L.; van der Ryst, E.; Lyons, S.; York, D.; Williamson, C. 12th World AIDS Conference; Geneva, Switzerland. 1998.
- van Harmelen JH, van der Ryst E, Loubser AS, York D, Madurai S, Lyons S, Wood R, Williamson C. A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. *AIDS Res Hum Retrovir* 1999;15:395–398. [PubMed: 10082124]
- van Harmelen J, Williamson C, Kim B, Morris L, Carr J, Karim SS, McCutchan F. Characterization of full-length HIV type 1 subtype C sequences from South Africa. *AIDS Res Hum Retrovir* 2001;17:1527–1531. [PubMed: 11709097]
- 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]
- Williamson C, Engelbrecht S, Lambrick M, van Rensburg E, Wood R, Bredell W, Williamson AL. HIV-1 subtypes in different risk groups in South Africa. *Lancet* 1995;346:782. [PubMed: 7658903]
- Williamson C, Morris L, Maughan MF, Ping LH, Dryga SA, Thomas R, Reap EA, Cilliers T, van Harmelen J, Pascual A, Ramjee G, Gray G, Johnston R, Karim SA, Swanstrom R. Characterization and selection of HIV-1 subtype C isolates for use in vaccine development. *AIDS Res Hum Retrovir* 2003;19:133–144. [PubMed: 12639249]
- Williamson, C.; Abrahams, M.; Treurnicht, F.; Seioighe, C.; Passmore, JA.; Wood, N.; Mlisana, K.; Hahn, B.; Abdool Karim, S. the CAPRISA 002 study and the Center for HIV-AIDS Vaccine Immunology Consortium. Single variant transmission predominates in HIV-1 subtype C infection, with multiple variant transmission associated with increased genital inflammatory cytokines. Abstract 284 15th CROI; Boston, MA. 2008.
- Wolfs TF, Zwart G, Bakker M, Goudsmit J. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. *Virology* 1992;189:103–110. [PubMed: 1376536]
- Zhang LQ, MacKenzie P, Cleland A, Holmes EC, Brown AJ, Simmonds P. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. *J Virol* 1993;67:3345–3356. [PubMed: 8497055]
- Zhang H, Hoffmann F, He J, He X, Kankasa C, Ruprecht R, West JT, Orti G, Wood C. Evolution of subtype C HIV-1 Env in a slowly progressing Zambian infant. *Retrovirology* 2005;2:67. [PubMed: 16274482]
- Zhang H, Hoffmann F, He J, He X, Kankasa C, West JT, Mitchell CD, Ruprecht RM, Orti G, Wood C. Characterization of HIV-1 subtype C envelope glycoproteins from perinatally infected children with different courses of disease. *Retrovirology* 2006;3:73. [PubMed: 17054795]
- Zhu T, Mo H, Wang N, Nam DS, Cao Y, Koup RA, Ho DD. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* 1993;261:1179–1181. [PubMed: 8356453]

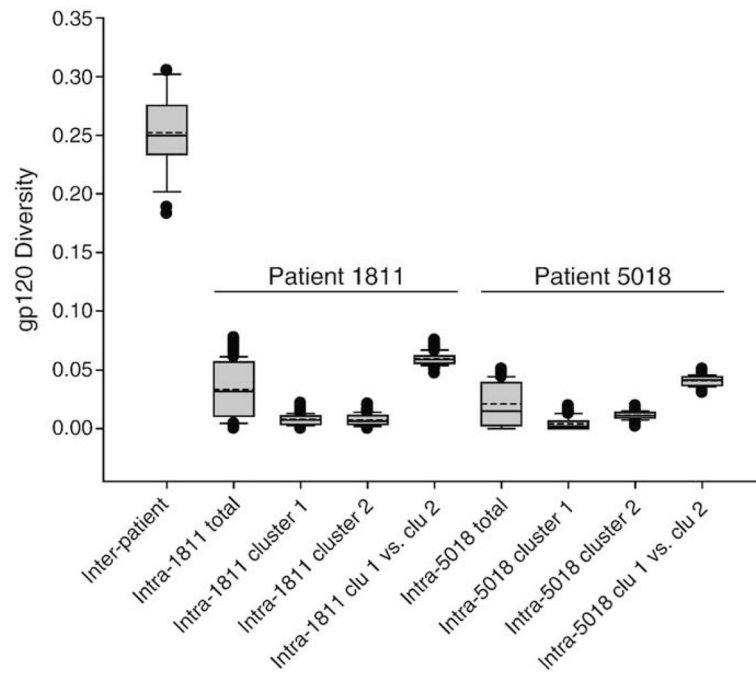




**Fig. 1.** Viral RNA load in acute HIV-1 subtype C infection. (A) Individual curves of viral RNA load in eight cases of acute infection. The timeline shows days from detected seroconversion. Time 0 corresponds to the first seropositive test. Plasma viral RNA load is expressed as  $\log_{10}$  copies per ml of plasma. Measurements of viral RNA before time 0 are pre-seroconversion. Numbers at top of boxes correspond to subject cases. Dotted lines with arrows indicate initiation of ART. (B) Early viral RNA set point in acute HIV-1 subtype C infection. The level of viral RNA at early set point was defined as a mean  $\pm$  standard deviation of measurements from 50 to 200 days from detected seroconversion (after assuming reduction of viral RNA peak). N shows the number of viral RNA measurements for the period from 50 to 200 days per subject. Dashed line shows a median of early viral RNA set point computed for all eight subjects at 4.53  $\log_{10}$  copies/ml. The first four subjects 1811, 2865, 3312, and 5018 correspond to the group of slow decline of viral RNA and high early viral RNA set point. Subjects 3430, 3505, 3603, and 5582 correspond to the group of fast decline of viral RNA and low early viral RNA set point.

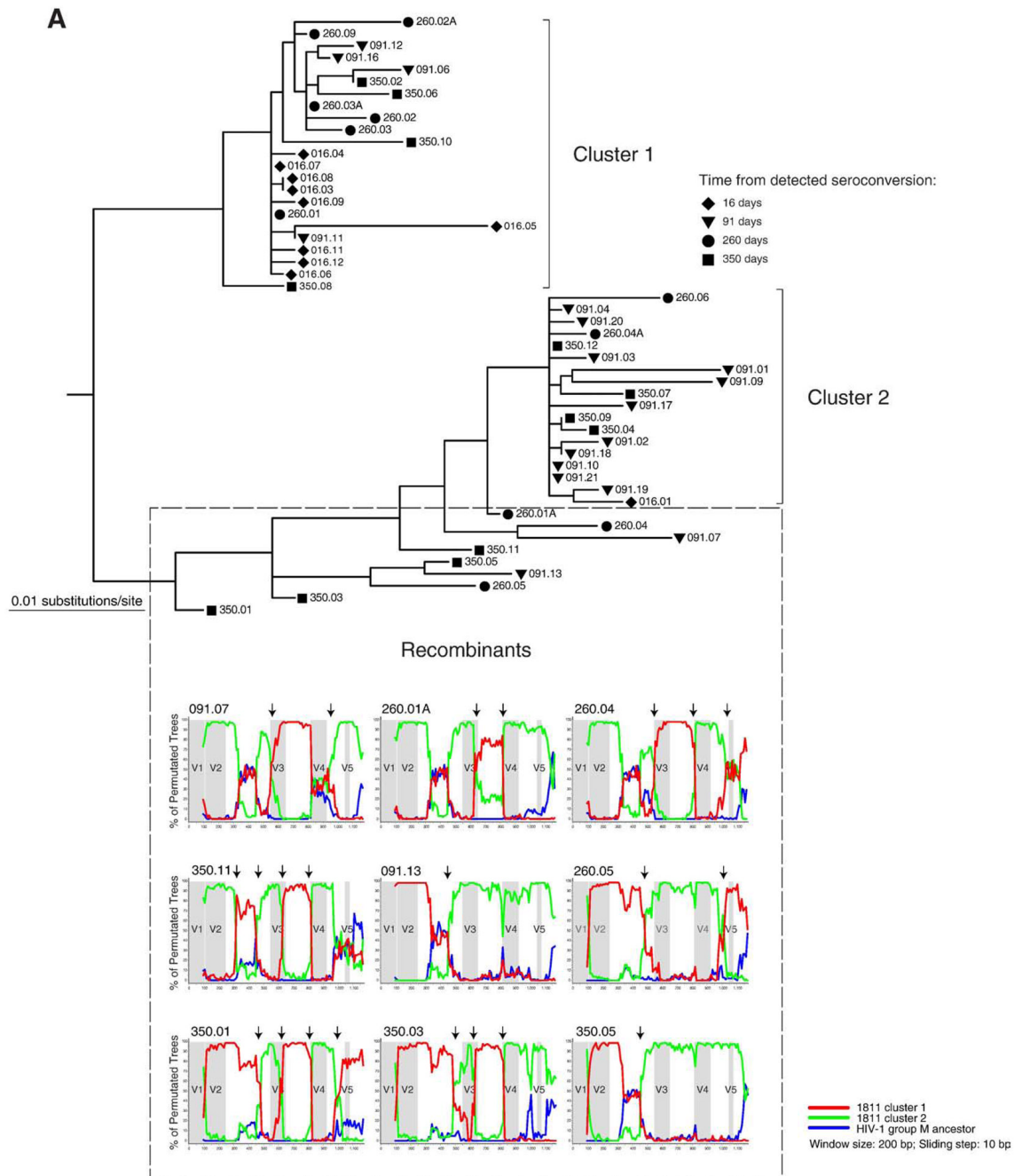


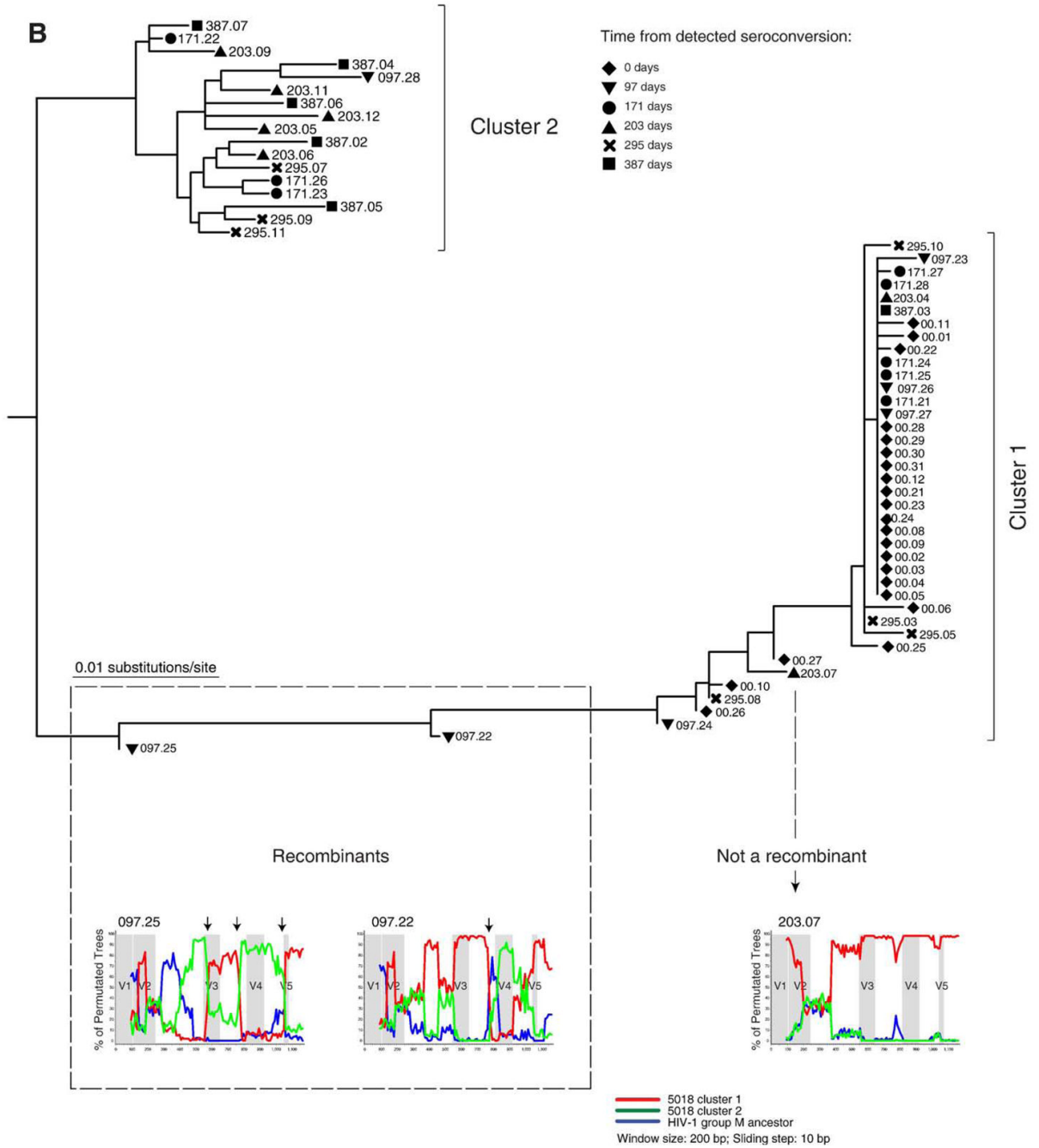
**Fig. 2.** Phylogenetic relationships of gp120 sequences in acute HIV-1 subtype C infection. A phylogenetic tree was generated by PhyML (Guindon and Gascuel, 2003) using the HKY model for nucleotide substitution, and visualized in FigTree v.1.1.2 (Rambaut, 2008). Main branches are labeled with subjects' numbers. Coloring corresponds to the viral RNA phenotypes: subjects with slow decline of viral RNA and a high early viral RNA set point are shown in red, and subjects with fast decline of viral RNA and a low early viral RNA set point are shown in green. Clusters are highlighted for subjects 1811 and 5018. HIV-1 subtype C reference sequences are shown in blue. The cluster of all HIV-1 M group non-subtype C references is shown at the bottom.



**Fig. 3.**

A comparison of inter-patient viral diversity in 8 subjects and intra-patient diversity in subjects 1811 and 5018. Pairwise maximum likelihood distances are shown. The inter-patient viral diversity was quantified as pairwise distances between MRCA of 8 subjects. The intra-patient diversity was measured as pairwise distances between viral quasispecies within each subject or each cluster. Distance between clusters were compared by quantifying pairwise distances between sequences belonging to each cluster. In the box plots: the boundary of the box closest to zero indicates the 25th percentile, a solid line within the box marks the median value, a dashed line within the box shows the mean, and the boundary of the box farthest from zero indicates the 75th percentile. *Whiskers* above and below the box indicate the 10th and 90th percentiles. Points above and below the whiskers indicate outliers outside the 10th and 90th percentiles.

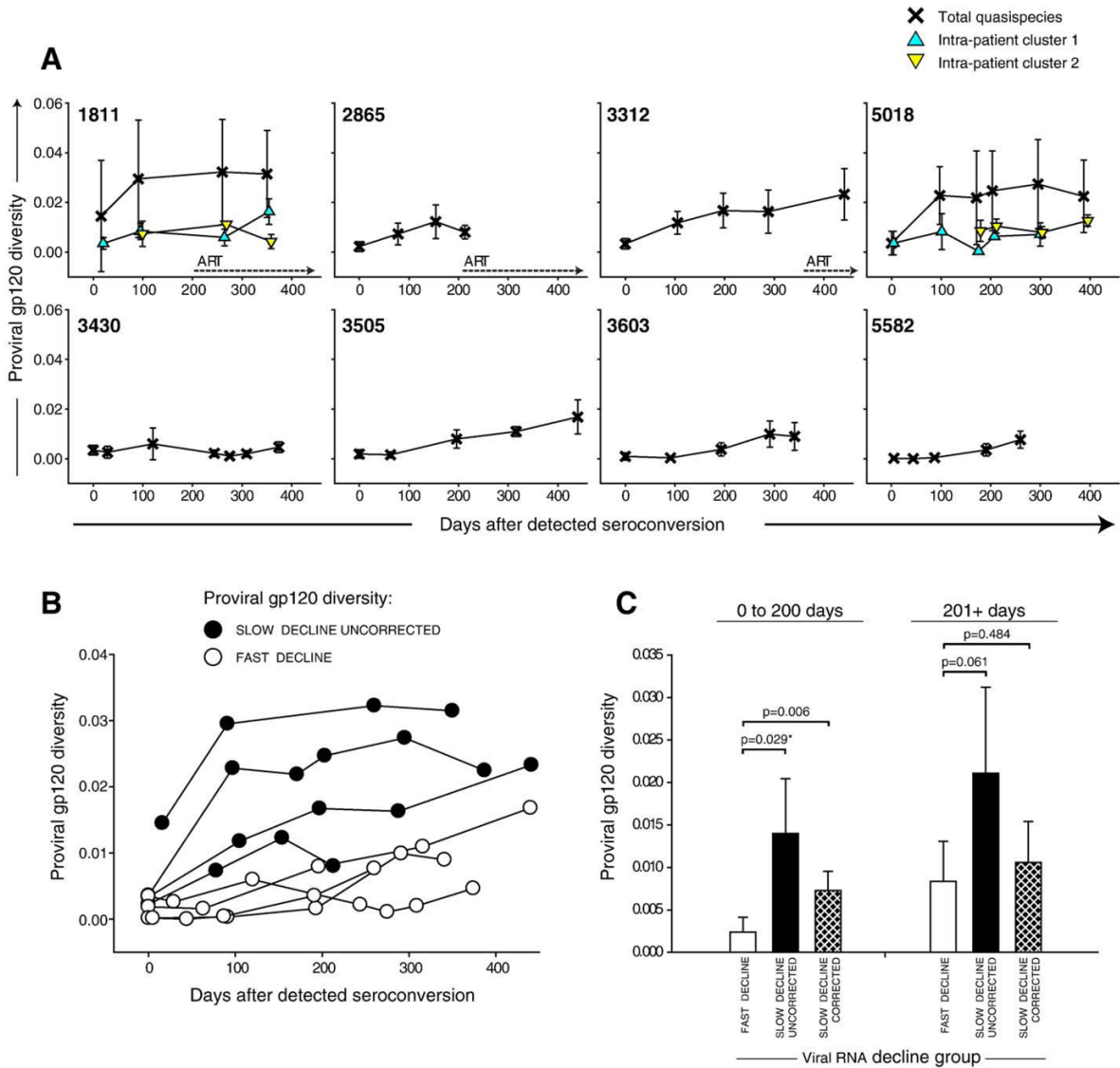




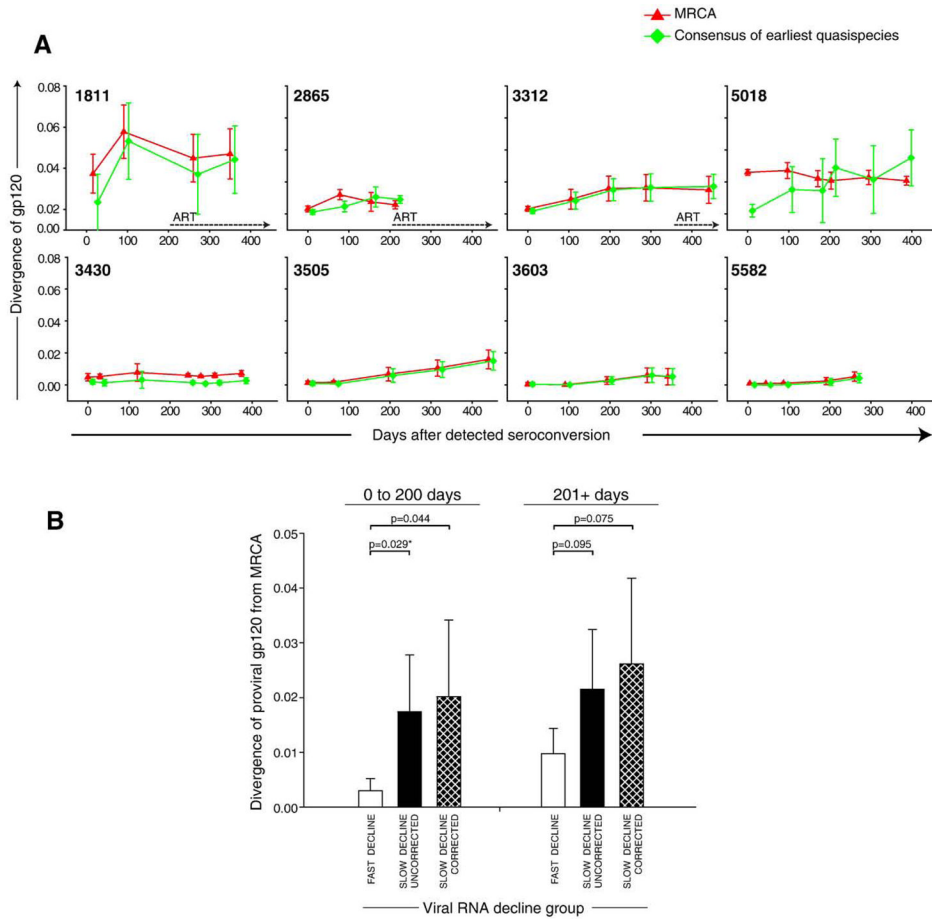
**Fig. 4.** Analysis for potential recombination in subjects 1811 and 5018. Distinct clusters and potential recombinants are defined. Each sequence is shown by a symbol corresponding to time points of sampling followed by number of days from detected seroconversion and ID of viral quasispecies. The recombinant search was performed by bootstrap analysis in the SimPlot program (Lole et al., 1999). The HIV-1 group M ancestor sequence was used as outlier (blue line). Arrows on the top of bootstrap graphs designate potential breakpoints of recombination between viral variants. The location of variable loops in gp120 is shaded. (A) Subject 1811. Time points of sampling are shown by a diamond at 16 days, a triangle to the bottom at 91 days, a circle at 260 days, and a square at 350 days. The 16 day sequences in cluster 1 and



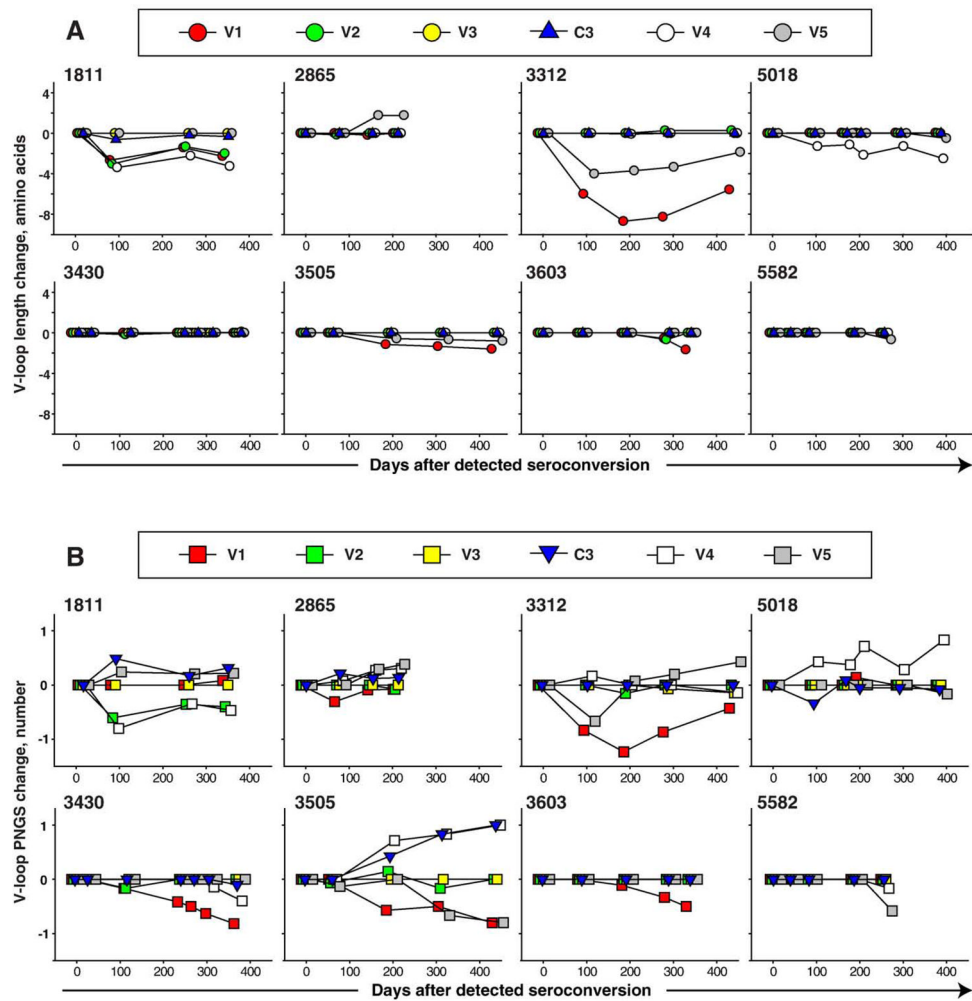
sequence 016.01 were used as references for the first (red) and second (green) clusters. The red and green lines correspond to bootstrap values per sliding window. (B) Subject 5018. Time points of sampling are designated by a diamond at day 0, a triangle to the bottom at 97 days, a circle at 171 days, a triangle to the top at 203 days, a cross at 295 days, and a square at 387 days. The 0 time point sequences in cluster 1 and majority consensus sequence for the entire cluster 2 were used as references for the first and second cluster (red and green lines correspond to bootstrap values, respectively).



**Fig. 5.** Intra-patient viral diversity in gp120. Mean values±standard deviation of pairwise maximum likelihood nucleotide distances are shown. The timeline shows days from detected seroconversion. (A) Individual trajectories of viral diversity in gp120. Numbers within boxes correspond to subject cases. The extent of total viral diversity per time point per subject is shown by crosses. The intra-cluster diversity in subjects 1811 and 5018 is designated by triangles (recombinant sequences are removed). Dotted lines with arrows indicate initiation of ART. (B) Levels of proviral diversity in gp120 between subjects with slow (filled circles) and fast (open circles) decline of viral RNA. (C) Viral diversity within groups defined on viral RNA decline, fast and slow decline of viral RNA. The slow decline group is presented by uncorrected distances (defined as total intra-patient pairwise distances) and cluster-corrected distances (sequences within clusters used separately, no recombinants included). Analysis was performed for two time intervals: 0 to 200 days and 201+ days after seroconversion. Comparison between groups is performed by *t*-test or Mann–Whitney Rank Sum Test (shown with \*).



**Fig. 6.** Diversification of viral gp120 in acute HIV-1 subtype C infection. Mean values±standard deviation are shown. Viral diversification at each time point was compared to the per-subject MRCA sequence reconstructed from viral quasispecies at all time points (red triangle), and consensus sequence reconstructed from the earliest available quasispecies (green square). The timeline shows days from detected seroconversion. (A) Individual trajectories of viral diversification. Numbers within boxes correspond to subject cases. Dotted lines with arrows indicate initiation of ART. (B) Viral diversification from MRCA that was reconstructed from viral quasispecies is compared between groups defined on viral RNA phenotype, fast and slow decline of viral RNA. The slow decline group is presented in two ways: by uncorrected distances (defined as total intra-patient pairwise distances to MRCA) and cluster-corrected distances (defined as intra-cluster pairwise distances to MRCA, no recombinants were included). Analysis was performed for two time intervals: 0 to 200 days and 201+ days after seroconversion. Comparison between groups is performed by *t*-test or Mann–Whitney Rank Sum Test (shown with \*).



**Fig. 7.** Changes in the length of V-loops and the number of PNGS in acute HIV-1 subtype C infection are presented as a difference between mean value at a given time point and earliest available time point. Variable loops are shown by red (V1), green (V2), yellow (V3), white (V4), and gray (V5) colors. The timeline shows days from detected seroconversion. Numbers above boxes correspond to subject cases. (A) Evolution of length of variable loops in gp120. (B) Evolution in number of PNGS within V-loops in gp120.