Lack of temporal structure in the short term HIV-1 evolution within asymptomatic naïve patients

Gonzalo Bello a,b, Concepción Casado a, Soledad García c, Carmen Rodríguez c, Jorge del Romero c, Antonio Carvajal-Rodríguez d,e, David Posada e, Cecilio López-Galíndez a,⁎

a Centro Nacional de Microbiología (CNM), Instituto de Salud Carlos III, Majadahonda, Madrid 28220, Spain
b Laboratorio de AIDS and Molecular Immunology, Department of Immunology, Oswaldo Cruz Institute-Fiocruz, Rio de Janeiro, Brazil
c Centro Sanitario Sandoval (CSS), IMSALUD, Comunidad Autónoma de Madrid, Madrid 28010, Spain
d Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602, USA
e Facultad de Biología, Universidad de Vigo, Campus As Lagoas-Marcosende, Vigo 36310, Spain

Received 31 August 2006; returned to author for revision 28 October 2006; accepted 22 November 2006
Available online 31 January 2007

Abstract

HIV-1 evolution in the envelope gene (env) was analyzed in four asymptomatic antiretroviral therapy naïve patients with typical and slow disease progression rates. In typical progressors, viral populations were monophyletic and two distinct evolutionary patterns were observed. In one patient, HIV-1 evolution displayed a strong temporal structure similar to the consistent pattern previously described. In the other, viral evolution displayed a lack of temporal structure with no increase in genetic heterogeneity and divergence over time. In slow progressors, several clades were observed in viral populations. However, analysis within the major sub-population revealed the same two evolutionary patterns described for typical progressors. Synonymous and non-synonymous substitution rate analyses indicated that positive selection was the major force driving HIV-1 evolution in viral populations with temporal structure, while evolution in viral populations with an atemporal structure was dominated by genetic drift and purifying selection. These results support the existence of distinct patterns of env evolution in untreated HIV-1-infected patients.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Intra-host HIV evolution; Phylogeny; Temporal structure; Atemporal structure; dN/dS; Recombination; In vivo

Introduction

A consistent model of intra-host HIV-1 evolution has been previously proposed based on the analysis of the envelope (env) gene in nine patients with typical clinical disease progression (Shankarappa et al., 1999). According to this model, during most of the asymptomatic period intra-host HIV-1 evolution displays a strong temporal structure, characterized by a constant replacement of viral lineages and a linear increase in divergence and heterogeneity of viral quasispecies over time. A recent analysis of the synonymous (dS) and non-synonymous (dN) substitutions rates in these data suggests that positive selection is the primary driving force underlying this temporal evolutionary pattern (Williamson et al., 2005). A similar pattern has been reported in many other HIV-1 infected adults (Casado et al., 2001; Liu et al., 1997; Mani et al., 2002; Markham et al., 1998; Poss et al., 1998; Shankarappa et al., 1998; Shioda et al., 1997; Troyer et al., 2005; Wolinsky et al., 1996; Zhang et al., 1997) and children (Ganeshan et al., 1997; Halapi et al., 1997; Nowak et al., 2002; Salvatori et al., 1997; Strunnikowa et al., 1995), usually associated with positive selection. This temporal pattern of intra-host HIV-1 evolution might result from a process of high viral adaptation rate that occurs when virus replication coincides with a persistent and continual immune pressure (Grenfell et al., 2004).

In the present study we performed a longitudinal analysis of HIV-1 evolution at the C2–V5 region of env gene.
during the asymptomatic period in four antiretroviral therapy naïve patients with typical and slow disease progression rates. Our results suggested the existence of two distinct patterns of intra-host HIV-1 evolution in these patients, characterized by the presence or absence of a temporal structure.

Results

Patient characteristics

The short-term viral evolution, during four to six years, was studied in four HIV-1 asymptomatic naïve patients at the C2–V5 region in env gene, in samples taken 2 to 16 years after seroconversion. In general, patients maintained CD4+ cell numbers above 500 cells/mm$^3$ and plasma RNA viral load usually below 20,000 copies/ml (Fig. 1). Looking at variation in the CD4+ numbers, patient 59 exhibited a decline of CD4+ T cells, with a mean decrease rate of about 60 cells/mm$^3$/year ($P=0.0003$). Patient 60 showed a sharper decline in the first years and a mean decrease rate of 53 cells/mm$^3$/year ($P=0.0011$). The decreasing trend in CD4+ numbers in these patients is compatible with the pattern described for typical progressors (Betts et al., 1999). In contrast, the same analysis in patients 30 ($P=0.3427$) and 45 ($P=0.5428$) did not show such decline after 14 to 19 years of infection, allowing the classification of these patients as slow progressors or long term non-progressors (LTNPs).

Phylogenetic analyses

In general, models of nucleotide substitution selected by Modeltest were quite complex (TVM+G, GTR+1+G, K81uf+1+G, and TIM+1+G for patients 30, 45, 59, and 60, respectively; see Posada and Crandall, 2001 for a detailed description) and are described in more detail in the Materials and methods section. The complexity of the models indicate that the variation of the substitution rate among sites is an important feature of HIV-1 evolution, even within patients.

In typical progressors only one major clade was identified per patient, but two distinct evolutionary patterns were observed. In patient 59, viral evolution showed a strong temporal structure (“temporal model”) characterized by the progressive diversification of the env gene over time (Fig. 2a), similar to the pattern described by Shankarappa et al. (1999). Variants isolated at earlier time points gave rise to later variants which were more distant to the MRCA (most recent common ancestor) than earlier ones. The mean genetic heterogeneity at the last time point was higher ($P<0.001$) than at the first one (Table 1); and the mean genetic divergence exhibited an increase over time, with a constant annual rate of 1.76% ($P=0.0028$) (Fig. 3a).

In contrast, in patient 60 the viral evolutionary pattern was characterized by the lack of a temporal structure (“atemporal model”). The variants recovered at a given time point were not the origin of the virus isolated immediately later, and more importantly, the viruses isolated at later time points were not
more distant to the MRCA than viruses isolated earlier (see Fig. 2b). The mean genetic heterogeneity was not significantly different in the first and last samples ($P>0.05$) (Table 1), and the mean genetic divergence did not show a significant linear increase over time ($P=0.2080$) (Fig. 3b).

In the LTNP patients, several well-supported clades were observed (Figs. 2c and d), as previously noted (Bello et al., 2004). Up to five clades (a to e) could be identified in patient 30, and three (a to c) in patient 45. According to the dating of the different clades, some clades represented actively replicating variants (a, b and c in patient 30, and a in patient 45), while others constituted ancestral non-replicating strains (d, and e in patient 30; b and c in patient 45) close in time to the sequences at seroconversion time (Bello et al., 2004). In each patient, one clade (designated a) dominated the quasispecies during most of the follow up (Figs. 2c and d). The phylogenetic analysis within these major clades revealed the same two evolutionary patterns described above for the typical progressors. Clade a of patient 30 displayed a temporal structure similar to that described for patient 59, with a significant increase of the mean heterogeneity over time ($P<0.05$) and with a mean annual increase of divergence rate of 1.2% ($P<0.05$) (see Table 1 and Fig. 3c). In contrast, clade a of patient 45 exhibited a lack of temporal structure similar to that described for patient 60, with no consistent increase of the mean heterogeneity ($P>0.05$) and divergence ($P<0.05$) over time (see Table 1 and Fig. 3d). All the subsequent analyses in patients 30 and 45 were carried out only with the dominant “a” clades.

Recombination analyses

The recombination permutation test was significant for all patients and time points, and the recombination rate estimates were similar for most patients (data not shown), suggesting that the distinct evolutionary patterns are not the result of different recombination rates.

Selective pressure analyses

In most cases (85%) the dual model of rate variation gave a lower AIC value than the non-synonymous model, while 63% of the LRTs were significant ($P$-value<0.05), indicating that there is a significant $dS$ rate variation within these data sets. The $dN/dS$ estimates were typically above 1 in patients 30 and 59; but always below 1 in patients 45 and 60 (Table 2). Positively and negatively selected sites were identified at most time points, but important differences were found between individuals. In patients 59 and 30, the mean number of sites under positive selection was generally higher than the number of sites under negative selection, whereas in patients 45 and 60 the opposite was true (Table 2). The analysis of the $dN$ and $dS$ divergence (Williamson et al., 2005) clearly showed that in patients 30 and 59, $dN$ accumulates steadily over time and at a faster rate than $dS$, whereas in patients 45 and 60, $dV$ or $dS$ did not accumulate over time (Table 3). In the tree-based analysis, $dV$ or $dS$ estimates were of smaller magnitude, although showed similar trends; except for patient 30, where no significant accumulation of $dN$ was detected over time (Table 3).

Discussion

In this study we detected the existence of two distinct patterns of short-term intra-host HIV-1 evolution in asymptomatic naïve patients, and described, for the first time, an evolutionary model without temporal structure.

Intra-host HIV-1 evolution in two patients (59 and 30), one typical progressor and one LTNP, displayed a strong temporal structure characterized by the continual replacement of variants and the progressive increase of the mean viral heterogeneity and divergence over time consistent with the evolutionary pattern described by Shankarappa et al. (1999). The mean annual divergence rate of 1.2% (LTNP 30) and 1.8% (typical progressor patient 59) was also in agreement with previous reports (Shankarappa et al., 1999).

In the other two patients (60 and 45), one typical progressor and one LTNP, intra-host HIV-1 evolution was characterized by a lack of temporal structure. The viruses detected at later time points were less heterogeneous and showed lower divergence from the MRCA than viruses obtained earlier. A similar pattern of viral evolution has been only previously described in sequences isolated from latent reservoirs (Gunthard et al., 1999) or from PBMCs (Frenkel et al., 2003) in patients under HAART. A weak temporal structure was also previously described in some naïve patients with extreme disease progression rates like rapid or non-progressors (Bello et al., 2005; Ganeshan et al., 1997; Mani et al., 2002; Shioda et al., 1997; Strunnikowa et al., 1995; Wang et al., 2002; Wolinsky et al., 1996). In those patients however, the “atemporal” pattern was associated with highly homogeneous quasispecies; in contrast to the highly heterogeneity (up to 5%) observed in quasispecies of patients 45 and 60.

We should note that the fact that virus sampled from the same patient are linked by a common evolutionary history violates one the fundamental assumptions of standard (parametric or nonparametric) statistics used to compare genetic diversity and divergence between viral samples at different time points, namely, that data points are independent. This lack of independence leads to the underestimation of standard deviations, making the t-tests or the correlation tests non-conservative. Indeed, these caveats are operating in any study describing changes in genetic diversity and divergence through time within HIV patients. The only solution would be the use of independent viral lineages, which could be impossible to obtain from individual HIV patients. In any case, the differences in diversity and divergence obtained are clear, and other lines of evidence, like the shape of the phylogenetic trees, point out in the same direction, suggesting that the patterns observed represent a real phenomenon.

We investigated different evolutionary factors which could be accountable for these two distinct patterns. Recombination analyses performed suggest that the distinct intra-host evolutionary patterns are not associated with different recombination
rates within patients; however, significant differences between patterns were observed regarding the contribution of positive and negative selection. The strong temporal structure of the HIV-1 evolution in patients 30 (clade a) and 59, was associated with positive selection operating within and across time points ($dN/dS > 1.0$), consistent with the notion that the primary driving force of the temporal evolutionary model is positive selection (Williamson et al., 2005). The signal for positive selection across time points in patients 30 and 59 was clearer with the Williamson’s method, which in theory should be less affected by recombination than tree-based approaches, because it does not assume a particular phylogenetic structure (Williamson et al., 2005). In contrast, the lack of temporal structure seen in patients 45 (clade a) and 60, was associated with negative selection operating within and across time points ($dN/dS < 1.0$). Thus, differential host-imposed immune selective pressures could be the origin of the discordant viral evolutionary patterns observed in this study.

According to the phylodynamic models described by Grenfell et al. (2004), the temporal structure and high viral adaptation rate seen in patients 30 and 59 could result from the coincidence of appreciable virus replication with a persistent medium immune selection pressure. In this situation, HIV-1 chronic infection is sustained by rapid and continuous cycles of productive infection and cell death of the activated CD4+ T cells (Ho et al., 1995; Perelson et al., 1996; Wei et al., 1995) due to the constant rise and selection of immune escape variants (adaptive model of evolution) (Lukashov and Goudsmit, 1998; Wolinsky et al., 1996).

The lack of temporal structure seen in patients 45 and 60 is more compatible with a pattern of neutral evolution as suggested by $dN/dS$ ratios ($<1$) (Table 2 and 3). This atemporal model could result from the coincidence of appreciable virus replication with low immune pressure (Grenfell et al., 2004). Such immunodeficient situation in patients 45 and 60 could correspond to the period where the clinical symptoms persist.

**Fig. 2.** Maximum likelihood phylogenetic trees for the env gene. Partial HIV-1 PBMC-derived gp120 env sequences of patients 59 (a), 60 (b), 30 (c), and 45 (d), along with reference strains (italic). Sequences are represented with an arbitrary symbol corresponding to the sampling year, as indicated in the key. Numbers at branch nodes indicate bootstrap support (only values greater than 50% are shown). Branch lengths are drawn to scale.
arise or AIDS. However, this is unlikely because patients 45 and 60 were away from the AIDS period and they did not show any clinical symptom of disease progression, maintaining low RNA viral loads and remaining therapy naïve during the follow-up (Fig. 1). Instead, this “atemporal” structure might result from the coincidence of appreciable virus replication with a persistent high immune pressure able to reduce the rise and selection of immune escape variants. Under this viral replication model the intra-host HIV-1 evolution will be driven principally by genetic drift (Frost et al., 2001; Plikat et al., 1997). In this situation, the persistent virus replication would be mainly sustained by the constant activation of new variants from the latently infected CD4+ T cell reservoir (Proximal Activation and Transmission, or PAT model) (Grossman et al., 1998; Grossman et al., 1999). According to the PAT model, sequences derived from later points should not necessarily replicate through more generations and accumulate more changes than variants isolated earlier, leading to a lack of temporal structure and no consistent increase in dS divergence.

We did not find any association between the evolutionary pattern, time since seroconversion and disease progression rate. The same pattern of viral evolution was observed in patients with different disease progression rates and at different times.

Fig. 2 (continued).
after seroconversion (patients 30 and 59, or patients 45 and 60); while distinct patterns were detected in patients with similar disease stages and at similar times since seroconversion (patients 59 and 60). However, the viral load observed in patients 30 and 59 at the last time points could indicate an association of temporal structure with higher risk of failure of viral replication control (Fig. 1), resulting from the rapid and constant selection of immune escape variants.

It is important to indicate that both patterns of viral evolution were observed in the global population in patients 59 and 60, but only in the major clades in patients 30 and 45. This could result from the co-existence of viral variants originated from different compartments (organs, tissues, and cell types), and/or latent viral reservoirs, which would produce large fluctuations in the genetic distance estimates (Nickle et al., 2003). The different viral dating of clades for patients 30 and 45 (Bello et al., 2004) is fully consistent with this hypothesis. Discontinuous HIV-1 intra-host evolutionary patterns associated to the co-evolution of several distinct lineages (Casado et al., 2001; Holmes et al., 1992; Simmonds et al., 1991), or to the reappearance of founder ancestral viral variants (Karlsson et al., 1999; Nowak et al., 2002; Ostrowski et al., 1998), has been also previously observed in other patients.

In conclusion, our data suggests that, apart from the consistent intra-host HIV-1 evolutionary pattern previously described, a second pattern without temporal structure could be

<table>
<thead>
<tr>
<th>Patient</th>
<th>Heterogeneity a</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>2.4±1.3</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>60</td>
<td>5.5±2.1</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>30 (clade a)</td>
<td>3.2±1.6</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>45 (clade a)</td>
<td>2.4±1.3</td>
<td>&gt;0.05</td>
</tr>
</tbody>
</table>

a Mean nucleotide distance between all pair-wise comparisons of sequences within each patient (%±S.D.), assessed at the first and last time point of the follow-up.

Table 2
dV and dS diversities within time points

<table>
<thead>
<tr>
<th>Time point</th>
<th>REL estimates of dV/dS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P30 (clade a)</td>
</tr>
<tr>
<td>1</td>
<td>1.20</td>
</tr>
<tr>
<td>2</td>
<td>1.51</td>
</tr>
<tr>
<td>3</td>
<td>1.02</td>
</tr>
<tr>
<td>4</td>
<td>1.31</td>
</tr>
<tr>
<td>5</td>
<td>0.61</td>
</tr>
<tr>
<td>6</td>
<td>0.88</td>
</tr>
<tr>
<td>Mean</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Table 1
Nucleotide heterogeneity analyses at the first and last time points of the follow-up

<table>
<thead>
<tr>
<th>Patient</th>
<th>Heterogeneity a</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>2.4±1.3</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>60</td>
<td>5.5±2.1</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>30 (clade a)</td>
<td>3.2±1.6</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>45 (clade a)</td>
<td>2.4±1.3</td>
<td>&gt;0.05</td>
</tr>
</tbody>
</table>

Fig. 3. Nucleotide divergence regression analysis. The mean nucleotide distance from a given time point to the reconstructed MRCA sequence (see Materials and methods) was calculated for each sampling year in the four patients. Error bars indicate standard errors of the mean. Subject identification numbers, slope, coefficient of regression, and P values, are indicated in each graph.
as described in Casado et al. (2001) and viral DNA was obtained from 1 × 10^7 cells by a standard phenol extraction method. PCR amplification conditions were those described in Casado et al. (2001). DNA viral load was determined as described in Rodrigo et al. (1997) in the same nested PCR used to amplify the C2–C5 gp120 region. To avoid genetic bottleneeking, first PCR included at least 20 copies of viral DNA which used primers 91ECU (5′-CTTAGGCATCTCCTATGGC-3′, 5956–5974 HXB2 position) and 22ED (Casado et al., 2001). Second PCR included 1 μl of the first PCR products and primers 27EU (Casado et al., 2001) and 96ED (5′-AGACAAATAATTGCTTGGCCTGTACCGT-3′, 7862–7836 HXB2 position). When a low DNA viral load did not allow the simultaneous amplification of at least 20 copies, two to five first-round PCRs were pooled according to the DNA load to prevent the loss of sequence heterogeneity, and 5 μl of this mixture were used in the second PCR.

Cloning and sequencing

Two microliters of the nested PCR products were ligated into the PCR 2.1 plasmid and cloned according to the TA Cloning kit instructions (Invitrogen, Carlsbad, Calif.). An average of 17 clones (range 11–20 clones) per time point were sequenced with primer 27EU using the ABI PRISM Dye Terminator reaction kit (Perkin-Elmer, Norwalk, Conn.) according to manufacturer’s instructions in an ABI PRISM 377 automated sequencer (Perkin-Elmer, Norwalk, Conn.).

Phylogenetic analyses

All nucleotide sequences included a fragment of 614-bp of the env gene, spanning from the distal portion of C2 to the proximal portion of C5 (nucleotides 7068 to 7682 in HXB2 clone). Nucleotide sequences were aligned using CLUSTAL X (Thompson et al., 1994) and revised by eye. All positions with gaps were excluded from the analysis. For each patient, best-fit models of nucleotide substitution were selected according to the Akaike Information Criterion (AIC) (Akaike, 1974) with Modeltest 3.6 (Posada and Buckley, 2004; Posada and Crandall, 1998). The model selected were TVM+G (transversional model with a rate variation among sites) for patient 30, GTR+G+I model (general time reversible plus invariable sites and rate variation among sites) for patient 45, K81uf+I+G model (two transversions—parameter model 1 with unequal frequencies plus invariable sites and rate variation among sites) for patient 59 and TIM+I+G (transitional model plus invariable sites and rate variation among sites) for patient 60 (see Posada and Crandall, 2001 for details on these models).

Maximum likelihood (ML) trees were estimated under the best-fit model of nucleotide substitution using the searching algorithm implemented in Phylm v.2.4.1 (Guindon et al., 2005), starting the search from a BIONJ tree (Gascuel, 1997), Phylogenetic confidence was assessed by bootstrap (Felsenstein, 1988) with 1000 replicates. Distinct clades were defined for patients 30 and 45, usually supported by high bootstrap values (>75%). The mean inter-clades nucleotide distances were high (5% to 12%). These clades were also confirmed with the help of the PAQ program as described in Baccam et al. (2001).

Genetic distances analyses

Genetic distances within each patient were estimated by ML under the best-fit model of nucleotide substitution. To estimate

<table>
<thead>
<tr>
<th>Table 3</th>
<th>dN and dS divergence from the MRCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time point</td>
<td>dN</td>
</tr>
<tr>
<td>P30 (clade a)</td>
<td>0.0172</td>
</tr>
<tr>
<td>P45 (clade a)</td>
<td>0.0365</td>
</tr>
<tr>
<td>P59</td>
<td>0.0407</td>
</tr>
<tr>
<td>P60</td>
<td>0.0353</td>
</tr>
</tbody>
</table>

Materials and methods

Study subjects

Four HIV-1-infected individuals, three homosexual and one intravenous drug user patients (patient 59) from an outpatient clinic (Centro Sanitario Sandoval, IMSALUD, Madrid) were selected because they remained asymptomatic without antiretroviral therapy during the analyzed period. CD4+ and vRNA loads are summarized in Fig. 1. Patients 59 and 60 were typical non-progressors, whereas patients 30 and 45 were considered long-term non-progressors (LTNPs).

Quantification of RNA plasma viral load

Plasma HIV-1 RNA was quantified with the Amplicor HIV Monitor test kit with a detection limit of 500 copies/ml for early (before 1998) and 50 copies/ml for late samples (Roche Diagnostics Systems, Somerville, NJ) following manufacturer’s instructions.

Purification, PCR amplification and quantification of viral DNA

Peripheral blood mononuclear cells (PBMCs) were obtained as described in Casado et al. (2001) and viral DNA was obtained from 1 × 10^7 cells by a standard phenol extraction method. PCR amplification conditions were those described in Casado et al. (2001). DNA viral load was determined as described in Rodrigo et al. (1997) in the same nested PCR used to amplify the C2–C5 gp120 region. To avoid genetic bottleneeking, first PCR included at least 20 copies of viral DNA which used primers 91ECU (5′-CTTAGGCATCTCCTATGGC-3′, 5956–5974 HXB2 position) and 22ED (Casado et al., 2001). Second PCR included 1 μl of the first PCR products and primers 27EU (Casado et al., 2001) and 96ED (5′-AGACAAATAATTGCTTGGCCTGTACCGT-3′, 7862–7836 HXB2 position). When a low DNA viral load did not allow the simultaneous amplification of at least 20 copies, two to five first-round PCRs were pooled according to the DNA load to prevent the loss of sequence heterogeneity, and 5 μl of this mixture were used in the second PCR.

Cloning and sequencing

Two microliters of the nested PCR products were ligated into the PCR 2.1 plasmid and cloned according to the TA Cloning kit instructions (Invitrogen, Carlsbad, Calif.). An average of 17 clones (range 11–20 clones) per time point were sequenced with primer 27EU using the ABI PRISM Dye Terminator reaction kit (Perkin-Elmer, Norwalk, Conn.) according to manufacturer’s instructions in an ABI PRISM 377 automated sequencer (Perkin-Elmer, Norwalk, Conn.).

Phylogenetic analyses

All nucleotide sequences included a fragment of 614-bp of the env gene, spanning from the distal portion of C2 to the proximal portion of C5 (nucleotides 7068 to 7682 in HXB2 clone). Nucleotide sequences were aligned using CLUSTAL X (Thompson et al., 1994) and revised by eye. All positions with gaps were excluded from the analysis. For each patient, best-fit models of nucleotide substitution were selected according to the Akaike Information Criterion (AIC) (Akaike, 1974) with Modeltest 3.6 (Posada and Buckley, 2004; Posada and Crandall, 1998). The model selected were TVM+G (transversional model with a rate variation among sites) for patient 30, GTR+G+I model (general time reversible plus invariable sites and rate variation among sites) for patient 45, K81uf+I+G model (two transversions—parameter model 1 with unequal frequencies plus invariable sites and rate variation among sites) for patient 59 and TIM+I+G (transitional model plus invariable sites and rate variation among sites) for patient 60 (see Posada and Crandall, 2001 for details on these models).

Maximum likelihood (ML) trees were estimated under the best-fit model of nucleotide substitution using the searching algorithm implemented in Phylm v.2.4.1 (Guindon et al., 2005), starting the search from a BIONJ tree (Gascuel, 1997), Phylogenetic confidence was assessed by bootstrap (Felsenstein, 1988) with 1000 replicates. Distinct clades were defined for patients 30 and 45, usually supported by high bootstrap values (>75%). The mean inter-clades nucleotide distances were high (5% to 12%). These clades were also confirmed with the help of the PAQ program as described in Baccam et al. (2001).

Genetic distances analyses

Genetic distances within each patient were estimated by ML under the best-fit model of nucleotide substitution. To estimate
viral heterogeneity at a given time point, we determined the mean and standard error of the pairwise nucleotide distances between all sequences. For viral divergence, we determined the mean and standard deviation of the pairwise nucleotide distances between all sequences from this time point and the most recent common ancestor sequence (MRCA) of the patient, calculated as the consensus sequence of the earliest sample. In viral quasispecies with multiple divergent clades, the MRCA sequence was inferred using only the nucleotide sequences at the earliest sample included in the major clade.

Recombination analyses

We estimated the recombination rate (4Nr, where N is the effective population size and r is the recombination rate per gene per generation) with the composite likelihood estimator (CLE) (McVean et al., 2002), implemented in Pairwise program included in the LDhat package, available at http://www.stats.ox.ac.uk/~mcvean/LDhat/. The statistical significance of these estimates was assessed through a likelihood permutation test with 1000 replicates (McVean et al., 2002).

Selection analyses

We used several strategies in order to study the selective pressures acting on these patients.

Synonymous rate variation

We applied a test for the presence of variation in the rate of synonymous substitution per synonymous site (dS) within each patient and time point using the non-synonymous and dual models with general discrete distributions as implemented in the web server DataMonkey (Pond and Frost, 2005). In short, the non-synonymous model allows only for variation in the rate of non-synonymous substitution per non-synonymous site (dN) across sites, while the dual model allows for independent rate distributions both for dN and dS across sites (Pond and Muse, 2005). Using 3 non-synonymous and 3 synonymous rate categories these two models are nested and a $\chi^2$ distribution with 4 degrees of freedom can be used to compute the $P$-values for a likelihood ratio test (LRT) contrasting these two models. We also computed the AIC for each model.

Non-synonymous and synonymous rates within time points

For each patient and time point we ran a random effect likelihood (REL) analysis (Pond and Frost, 2005) using the DataMonkey server. This analysis provides dN and dS expectations by fitting codon models by ML. In addition, this analysis also identifies sites under positive and negative selection through an empirical Bayesian procedure. This method allows for rate heterogeneity both in synonymous and non-synonymous rates, reducing the chances for misidentification of positively selected sites (Pond and Muse, 2005).

Non-synonymous and synonymous rates across time points

To analyze the evolution of the dN/dS ratio across time for each patient we used two independent approaches. First we measured dS and dN divergence through time as described by Williamson et al. (2005). The MRCA in each patient was approximated as the consensus sequence at the earliest sample. Conveniently, this method uses the temporal information contained in the longitudinal samples to correct for saturation and to identify the evolutionary path between codons that differ at more than one position. In addition, because it does not assume a particular phylogenetic structure, it has been suggested to be less affected by recombination than other methods (Williamson et al., 2005). Second, we performed a phylogenetic-based analysis of dS and dN divergence. First we estimated a ML haplotype tree for each patient with Phylm using the best-fit model of nucleotide substitution selected by Modeltest. We then estimated the dN and dS rates for each branch under a Muse and Gaut codon model (Muse, 1996) crossed with a time reversible model (MG95 ×REV) in HYPHY (Pond et al., 2005). Finally, we used HYPHY to calculate the mean dN or dS path (the sum of branch lengths) from the sequences sampled at a given time point to the root of the patient tree, resulting in maximum likelihood estimates of mean dN and dS divergence from the MRCA at each time point.

Statistical analyses

Unpaired Student’s $t$-tests were used to compare group means. All statistical analyses were carried out with GraphPad Prism version 2.01 programs (GraphPad Software Incorporated).

Acknowledgments

We want to thank Sergei L. Kosakovsky Pond for his generous help with HYPHY, and to Scott Williamson for providing us with his code for the divergence analysis. This work was supported by grants SAF 2002/626, SAF 2003/04987, SAF 2005/03833, by the Plan Nacional del SIDA and in part by the “Red Tematica Cooperativa de Investigación en SIDA (Red de grupos 173) del FISss” G.B. was supported by a fellowship from AECI (MUTIS program) and from Instituto Carlos III. A. C-R and D.P were supported by grant R01-GM66276 from the US NIH. DP is also supported by grant BFU2004-02700 of the Spanish Ministry of Education and Science and by the “Ramón y Cajal” program of the Spanish government.


References

subset of human immunodeficiency virus type 1 long-term non-progressors is characterized by the unique presence of ancestral sequences in the viral population. J. Gen. Virol. 86 (2), 355–364.


progression of human immunodeficiency virus type 1 infection. J. Virol. 73 (12), 10489–10502.


