

## Disease progression and evolution of the HIV-1 *env* gene in 24 infected infants<sup>☆</sup>

Antonio Carvajal-Rodríguez<sup>a,b,\*</sup>, David Posada<sup>b</sup>, Marcos Pérez-Losada<sup>a</sup>, Emily Keller<sup>a</sup>, Elaine J. Abrams<sup>c</sup>, Raphael P. Viscidi<sup>d</sup>, Keith A. Crandall<sup>a</sup>

<sup>a</sup> Department of Biology, Brigham Young University, 84602 Provo, UT, USA

<sup>b</sup> Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, 36310 Vigo, Spain

<sup>c</sup> Department of Pediatrics, Columbia University College of Physicians and Surgeons and Harlem Hospital Center, NY, USA

<sup>d</sup> Stanley Division, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Received 3 August 2007; received in revised form 23 October 2007; accepted 24 October 2007

Available online 1 November 2007

### Abstract

We have studied the relationship between disease progression and HIV-1 evolution in 24 infants classified as rapid or non-rapid progressors, during nearly the entire disease progression cycle from infection to AIDS. Specifically, we examined the temporal relationship between clinical status and changes in genetic diversity, divergence, selection and recombination at the C2V3C3 region of the *env* gene during a period of 3 years. Statistical analyses were performed using linear mixed models that are particularly well-suited for longitudinal studies in which repeated measures are taken from the same patients. We did not observe significant differences in genetic diversity or overall substitution rates between clinical categories. However, the nonsynonymous substitution rate per nonsynonymous site (dN) evolved differently between groups. Changes in dN explained the evolutionary slowdown of the dN/dS ratio in the rapid progressors, while in non-rapid progressors the dN/dS ratio continuously increased through time. The number of positively selected sites had limited power for predicting disease progression. Recombination rate estimates were different among groups, although not significantly in the linear mixed models analysis. They showed some power predicting clinical categories and, interestingly, they were significantly correlated with the frequency of positively selected sites. Overall, the results obtained confirm that viral adaptation in the C2V3C3 region of the *env* gene is related to disease progression, although the statistical characterization of such pattern seems rather difficult.

© 2008 Published by Elsevier B.V.

**Keywords:** HIV-1; Disease progression; *env*; Selection; Recombination

### 1. Introduction

The broad inpatient genetic diversity characteristic of the human immunodeficiency virus type 1 (HIV-1) infection is usually assumed to be the result of positive selection, especially at the *env* gene (Ross and Rodrigo, 2002; Seibert et al., 1995;

Williamson, 2003; Yamaguchi-Kabata and Gojobori, 2000). Early HIV-1 studies suggested that progression to AIDS results from virus adaptation to the host environment (Nowak et al., 1991, 1996; Wolinsky et al., 1996). Since then, some studies have found a positive relationship between levels of genetic (antigenic) diversity and the rate of disease progression (Strunnikova et al., 1995, 1998), whereas others have found the opposite pattern (Ganeshan et al., 1997; Wolinsky et al., 1996). These apparently contradictory results likely arose because many of these analyses failed to distinguish between adaptive and selectively neutral changes, which is key to understand the interaction between the virus and its host (Williamson, 2003). When this distinction is taken into account, and more patients are compared, it seems clear that positive selection is more prevalent in patients with slow progression rates to AIDS, because their viral population shows

<sup>☆</sup> Note: Nucleotide sequence data reported in this paper are available in the GenBank database under the accession numbers: AY823998–AY824179, AY824250–AY824290, AY824329–AY824409, and AY824472–AY824946. These sequences were submitted as part of an independent analysis of the data set.

\* Corresponding author at: Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, 36310 Vigo, Spain. Tel.: +34 986 813828; fax: +34 986 812556.

E-mail address: [acaaj@uvigo.es](mailto:acaaj@uvigo.es) (A. Carvajal-Rodríguez).

higher adaptation rates due to a stronger immune response (Ross and Rodrigo, 2002; although see Viscidi, 1999; Williamson, 2003; Zanotto et al., 1999). Moreover, it has been suggested that an evolutionary slowdown occurs in late infection, caused by the collapse of the immune system, rather than by a reduction in the viral replication rate due to cellular exhaustion (Lemey et al., 2007; Williamson et al., 2005). To further complicate this picture, some studies have shown the occurrence of convergent evolution and the persistence of selection acting at the same sites over long periods of time in patients with slow (Ross and Rodrigo, 2002) and fast progression rates (Strunnikova et al., 1995), while others have failed to observe these patterns (Williamson, 2003), perhaps reflecting the influence of anti-retroviral therapy (Ganeshan et al., 1997; Potter et al., 2006), random genetic drift (Shriner et al., 2004) and recombination (Anisimova et al., 2003) on the distribution of positive selected variants. Moreover, a high portion of the variation at the V3 region of the *env* gene seems to be neutral (Nielsen and Yang, 1998). Finally, it has been shown that selection in the HIV-1 *env* gene, though intense, can be context-dependent (Templeton et al., 2004).

Studying the evolutionary relationship of HIV-1 and its host and characterizing the distinct adaptation patterns in different parts of the HIV-1 genome that interact with the immune system will be key to elucidate how HIV-1 overwhelms the immune system and leads to AIDS (Williamson et al., 2005). In this study, we examine in detail the evolution of HIV-1 in 24 infants classified on the basis of progression to AIDS. For that purpose, we have sequenced the C2V3C3 region of the HIV-1 *env* gene in longitudinal samples (three to seven per patient) obtained from each infant during a period of 3 years. The C2V3C3 region of the *env* gene is very suitable to study viral adaptation because it is key for entrance of HIV-1 into host cells and is a target of the immune response to HIV-1. The main objective of this study was therefore to decipher the temporal relationship between disease progression and diversity, divergence, selection, and recombination in the HIV-1 envelope gene. Uniquely, because we have sampled infants, we were able to study nearly the entire disease progression cycle – from infection to AIDS – whereas most studies on adults can only begin sampling long after infection and often only shortly before AIDS.

## 2. Materials and methods

### 2.1. Study population

The study participants were a subset of infants with HIV-1 perinatal infection enrolled in The New York City Perinatal HIV Transmission Collaborative Study, which is an observational cohort study of HIV infected infants born at seven New York City health care institutions since 1986. The enrollment criteria and study protocol design are described in detail elsewhere (Abrams et al., 1995; Thomas et al., 1994). Infants had a medical history, physical examination and phlebotomy done at birth and during scheduled follow-up visits. Fifty-one infants were identified who had plasma viral RNA load measurements performed within 2 months of birth and at

intervals up to 3 years of age using the NASBA HIV-1 RNA quantification kit (Organon-Teknika, Durham, NC, USA). For 44 infants, aliquots of nucleic acid extracted from plasma specimens for viral load measurements were available for sequence analysis. Of the 44 infants, 11 of 14 infants with a diagnosis of AIDS and 13 of the 30 remaining infants, selected at random, were included in the present study. The date of birth of the study subjects fell between August 1991 and July 1994. Six infants received no anti-retroviral therapy during the period of study; 10 infants received Zidovudine and 8 infants received Zidovudine and Didanosine at some time during the study. The study subjects were divided into two categories based on a clinical diagnosis of AIDS following recommendations of the HIV Pediatric Guidelines Working Group (supplementary Table S1). Rapid progressors (RP) were defined as infants who, by 12 months of age, were clinically diagnosed with AIDS or died (subjects P2, P8, P10, P18, and P20–P25). Six of these 10 infants eventually succumbed to their disease. The remaining 14 infants were classified as non-rapid progressors (NRP; subjects P1, P3–P7, P9, P11–P16, and P19).

### 2.2. PCR amplification, cloning, screening and sequencing

Envelope sequences generated for this study are available from GenBank under accession numbers AY823998–AY824946, and are the same as those used by Edwards et al. (2006). Nucleic acid was extracted from plasma specimens using the protocol of the NASBA HIV-1 RNA kit. An aliquot of nucleic acid was subjected to reverse transcription and a nested PCR for an approximately 350 bp fragment of the C2V3C3 region of the HIV-1 *env* gene. By serial dilution analysis we determined that the specimens contained 40 or more amplifiable copies of HIV-1 RNA. The positions of the outer primers (ED31, PND-02) and nested primers (PND-01, PND-04) have been published previously (Liu et al., 1997; Strunnikova et al., 1995). The reverse transcription reaction was performed with 2.5 pmol of primer PND-02 and 20 U of Moloney murine leukemia virus reverse transcriptase (Boehringer Mannheim Biochemicals, Indianapolis, IN) following a previously described protocol (Strunnikova et al., 1995). PCR was performed with 0.25  $\mu$ M each of sense and antisense primers and 3.5 U of Expand High Fidelity enzyme mixture (Boehringer, Mannheim) following the manufacturer's instructions. Cycling conditions for the first and second round of PCR were as follows: 35 cycles of 94 °C for 15 s, 55 °C for 15 s, and 72 °C for 40 s; and a final incubation at 72 °C for 10 min. A 1:50 dilution of the first round products was used in the second round.

The PCR products were ligated into the plasmid pCR-Blunt (Invitrogen Corp., Carlsbad, CA) and transformed into One Shot TOP10 competent cells as recommended by the manufacturer. The transformed cells were streaked onto agar plates and colonies containing recombinant plasmids were identified by PCR with primers PND-03 and PND-04. PCR products from the clones were screened for sequence variants by heteroduplex mobility analysis (Delwart et al., 1993; Strunnikova et al., 1995). For the analysis, equal amounts of

PCR-amplified nucleic acid from a clone and from a lysate of HIV-1<sub>MN</sub> were mixed in a final volume of 4  $\mu$ l. The mixture was boiled for 5 min and then rapidly cooled on ice for 5 min. An aliquot of the reaction mixture was fractionated on a precast 12.5% polyacrylamide gel (GeneGel Excell 12.5/24 kit; Amersham Pharmacia Biotech, Uppsala, Sweden) with a GenPhor electrophoresis unit (Amersham Pharmacia Biotech) following the manufacturer's recommended conditions for heteroduplex analysis. Bands were visualized by silver staining. For each specimen, 24 clones, the number that would fit on a single polyacrylamide gel, were analyzed. The frequency of clones corresponding to each unique heteroduplex band pattern was recorded. One clonotype representing each pattern was selected arbitrarily for sequencing. Sequence analysis of 84 pairs of identical clonotypes obtained from 84 specimens from 20 subjects showed that the average nucleotide sequence difference between pairs was 0.3% (range: 0.0–1.9%). Use of heteroduplex analysis allowed us to screen more clones that would have been practical by sequence analysis alone. In addition, assuming that identical clonotypes have the same nucleotide sequence, the method provides for an estimate of gene frequencies.

From clones selected as described above, plasmid DNA was isolated from a 4 ml broth culture using the Bio101 RPM kit (Bio101 Inc., Vista, CA) in accordance with the manufacturers' instructions. DNA templates were sequenced in both directions on an ABI 377 automated DNA sequencer (Synthesis and Sequencing Facility, Department of Biological Chemistry, Johns Hopkins University School of Medicine) using Big Dye Terminator RR mix (PE Applied Biosystems Inc., Foster City, CA). A consensus sequence was formed from the sequences generated in the forward and reverse directions. Ambiguities were resolved by examination of the trace data.

### 2.3. Sequence alignment and phylogenetic estimation

We studied 24 patients, 3–7 sampling times for each patient, and 24 amplicons per sample, giving a total of about 2500 variant genomic fragments. After heteroduplex screening of variants, a dataset of 784 nucleotide sequences was generated. The *env* nucleotide sequences were aligned using the program Clustal X (Thompson et al., 1997), translated into amino acids using the universal genetic code in Hyphy (Pond et al., 2005) and then matched to the reference sequence HIVXB2 from Los Alamos database (<http://hiv-web.lanl.gov/content/hiv-db/main-page.html>) to assure that the reading frame was maintained. The *env* region (C2V3C3) studied included 119–122 codons, starting at HIVXB2 codon number 247 in all samples (subject 3 started at codon 296 and the length was 95 codons). The best-fit model of nucleotide substitution was selected under the Akaike information criteria (AIC; Akaike, 1974) with Modeltest v3.6 (Posada and Crandall, 1998), using maximum likelihood (ML) estimates from PAUP\* (Swofford, 2002). Maximum likelihood trees for each patient were inferred under the best-fit model with Phyml v.2.4.1 (Guindon and Gascuel, 2003). Before proceeding with the analysis, we explored the data for possible cross-contamination. For that purpose we estimated a ML

phylogenetic tree for all the samples combined. If a given sample did not cluster together with the other samples from the same patient, cross-contamination was inferred.

### 2.4. Genetic diversity and substitution rates

Genetic diversity ( $\theta = 4N\mu$ , where  $N$  is the effective population size and  $\mu$  is the mutation rate) was estimated for each infant, overall and for every time point, using Watterson's (1975) estimator. Maximum likelihood estimates of the substitution rates were also calculated for each patient using the dated tips model (Rambaut, 2000) as implemented in PAML (Yang, 1997), assuming a molecular clock and including dates of isolation. Effective population size to compare between rapid and non-rapids progressors was estimated using BEAST (Drummond and Rambaut, 2006).

### 2.5. Selection analyses

#### 2.5.1. Synonymous and nonsynonymous diversity

Synonymous  $\pi_S$ , and nonsynonymous  $\pi_N$ , diversities for each patient and time point were estimated with the Jukes–Cantor correction using the program DnaSP (Rozas et al., 2003).

#### 2.5.2. Divergence analysis

We studied the evolution of the dN/dS ratio across time for each patient, where dN is the rate of nonsynonymous substitution per nonsynonymous site, and dS is the rate of synonymous substitution per synonymous site. We measured dN and dS divergence following Williamson et al. (2005). Under this approach, the most recent common ancestor (MRCA) in each patient is estimated 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 this method does not assume a particular phylogenetic structure, it has been suggested (Williamson et al., 2005) that this method is more robust to the false inference of positive selection caused by recombination (Anisimova et al., 2003; Shriner et al., 2003).

#### 2.5.3. Detection of positively selected sites

In order to detect which sites are being positively selected at the C2V3C3 region, we ran a fixed effects likelihood (FEL) analysis for each patient and time point. For those samples with significant synonymous rate variation (see below), we also ran a random effects likelihood (REL) analyses under a *Dual* model (Pond and Muse, 2005). Both FEL and REL analyses were performed at the DataMonkey server (Pond and Frost, 2005b). Both methods provide dN and dS expectations by fitting to the data a MG94 codon model (Muse and Gaut, 1994) crossed with the best-fit substitution model. While FEL uses a likelihood ratio test (LRT) to identify sites under negative or positive selection, REL performs an empirical Bayes analysis for the same purpose. Importantly, the REL method allows for rate

heterogeneity both in synonymous and nonsynonymous rates (the Dual model), reducing the chances for misidentification of positively selected sites. However because REL can suffer from high Type I error rates (Pond and Frost, 2005a), we applied the REL model only when the test for the presence of synonymous rate variation was significant. For each individual, we also stored the frequency of positively selective sites (FPSS). In addition, in samples with a high number of inferred positively selected sites and a significant recombination signal, we confirmed these results using a method implemented in DataMonkey (GARD + REL) (Pond and Frost, 2005b) that takes recombination into account (Scheffler et al., 2006). Moreover we simultaneously estimated the recombination rate and the dN/dS ratio for each patient using omegaMap (Wilson and McVean, 2006). In this case we had to pool the temporal samples in order obtain reliable estimates.

#### 2.5.4. Synonymous rate variation

We tested for the presence of synonymous rate variation within each patient and time point using the Nonsynonymous and Dual models with general discrete distributions as implemented in Hyphy (Pond and Muse, 2005; Pond et al., 2005). When using three nonsynonymous and three synonymous rate categories these two models are nested, and a  $\chi^2$  distribution with four degrees of freedom can be used to compute the *P* values for a LRT comparing these models. We also computed the AIC for each model.

#### 2.6. Recombination analysis

We estimated the population recombination rate  $\rho$  ( $4Nr$ , where  $N$  is the effective population size and  $r$  is the recombination rate per locus per generation) with the composite likelihood estimator (CLE) (McVean et al., 2002) implemented in the program pairwise of the *LDhat* package, freely 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.

#### 2.7. Linear mixed models

Linear mixed models (LMM), also known as random effect linear models or hierarchical linear models (Raudenbush and Bryk, 2002), are flexible models which generalize the general linear model to better support repeated (non-independent) measures and fixed and random effects. Fixed effects are constant across individuals, while random effects vary among individuals. In this study the fixed effects considered were the sampling time (age in months), drug treatment (treated or untreated at each time sampling time), clinical category (rapid progressors or non-progressors), and their interactions. The only random effect we considered was sampling time, which we therefore allowed to have both a constant and a varying effect between individuals. Because observations were taken longitudinally on the same subjects, making subject errors to be autocorrelated, and because these observations are not equally spaced, we used a continuous first-order autoregressive process

in the errors in which the strength of the correlation is inversely related to the units of time that separate the within-subject observations (the *corCAR1* function in R statistical package) (Team, 2006). The different outcome variables, analyzed in turn, were  $\theta$ ,  $\rho$ ,  $\pi_S$ ,  $\pi_N$ , dS, dN and FPSS.

Linear mixed models were fitted for each of the clinical categories independently and for the full dataset. Full models and up to 12 different restricted models were fitted with the function *lme* in the *nlme* library (Pinheiro and Bates, 2000) in R 2.4.1 (Team, 2006), using restricted maximum likelihood (REML). Models without random effects were fitted by REML with the generalized least squares function *gls* in R. The AIC was used to select the best-fit models for each outcome variable and dataset.

#### 2.8. Predictive classification

Finally, we also used a cross-validation test as implemented in the discriminant function analysis of the SPSS 12.0 statistical package to investigate the predictive power of the frequency of positive selected sites (FPSS) and recombination rates to assign individuals to clinical categories. We used also the logistic regression tool within the same package to take into account drug treatment. When necessary we compared the linear regression slopes using the parallel regression test as described in Sokal and Rohlf (1981).

### 3. Results

The GTR + I + G model was selected as the best-fit model of nucleotide substitution for the combined data set (all patients) ( $\pi_A = 0.4121$ ,  $\pi_C = 0.1814$ ,  $\pi_G = 0.2121$ , and  $\pi_T = 0.1943$ ;  $r_{CT} = 5.3938$ ,  $r_{CG} = 0.5097$ ,  $r_{AT} = 0.7055$ ,  $r_{AG} = 4.0379$ , and  $r_{AC} = 1.5957$ ;  $\alpha = 0.9762$  and proportion of invariable sites  $I = 0.0482$ ). This model is often selected for the HIV-1 *env* gene (Posada and Crandall, 2001). Upon inspection of the ML tree, we inferred cross-contamination for samples P14.2.3 and P14.2.15, which were eliminated from further analyses.

#### 3.1. Genetic diversity and substitution rates

Although viral populations isolated from NRP seem to be more diverse than those isolated from RP, the average genetic diversities and substitution rates were not significantly different between clinical categories (Table 1). However, the increase of diversity through time was significant for NRP (regression  $R^2 = 0.09$ ;  $P = 0.028$ ) but not for RP (regression  $R^2 = 0.03$ ;  $P = 0.27$ ). In addition, we did not detect a significant correlation between diversity or substitution rate with the CD4 count.

#### 3.2. Selection analyses

##### 3.2.1. Synonymous and nonsynonymous diversity

Levels of synonymous or nonsynonymous variation at each time point were not significantly different among clinical categories. In addition, there was not a significant increase through time.

Table 1  
Average genetic diversity, substitution rates, and frequency of positive selected sites (FPSS) for each infant across time points

Clinical category/patient	Average genetic diversity	Substitution rate	Average FPSS
<b>Non-rapid progressors (NRP)</b>			
P1	19.2	0.000489	0.004098
P3	14.5	0.000020	0.000000
P4	8.3	0.000454	0.000000
P5	25.5	0.000110	0.002066
P6	8.6	0.000799	0.000000
P7	13.0	0.001076	0.001667
P9	8.5	0.000299	0.001667
P11	23.2	0.000374	0.001667
P12	12.3	0.000014	0.000000
P13	19.3	0.000116	0.031933
P14	30.3	0.000195	0.005263
P15	46.1	0.001478	0.033333
P16	42.9	0.000926	0.024793
P19	9.2	0.000010	0.000000
Average	20.1	0.000450	0.007610
<b>Rapid progressors (RP)</b>			
P2	22.4	0.000035	0.000000
P8	10.3	0.000044	0.000000
P10	21.7	0.000026	0.004762
P18	13.5	0.000141	0.000000
P20	3.1	0.000019	0.000000
P21	11.8	0.000026	0.000000
P22	6.5	0.000334	0.000000
P23	4.6	0.000423	0.000000
P24	12.2	0.000111	0.000000
P25	13.3	0.000091	0.000000
Average	11.9	0.000125	0.000480

### 3.2.2. Synonymous and nonsynonymous divergence

Both for NRP and RP, dN and dS increased significantly with time (Table 2). However, this increase was higher for dN than for dS in NRP but similar for dN and dS in RP, which is reflected in the dN/dS ratio (Fig. 1). The slopes between NRP and RP regressions were significantly different for both dN (0.0087 for NRP and 0.0046 for RP; parallel regression test  $t = 13.29$ ,  $P < 0.0001$ ) and dS (0.0046 for NRP and 0.0062 for RP; parallel regression test  $t = 3.99$ ,  $P = 0.0001$ ). No differences between clinical categories were detected when comparing the slopes for the dN/dS values (parallel regression test  $t = 0.003$ ,  $P = 0.99$ ). In addition, only for NRP was there a significant correlation between dN/dS and CD4 count, at sampling time 1

Table 2  
Average dN and dS divergence for RP and NRP

Time	NRP		RP	
	dN	dS	dN	dS
1	0.0013	0.0020	0.0057	0.0074
2	0.0098	0.0109	0.0137	0.0145
3	0.0170	0.0109	0.0156	0.0200
4	0.0280	0.0173	0.0200	0.0260
$R^2$ ( $P$ value)	0.33 (0.000*)	0.13 (0.006*)	0.17 (0.010*)	0.15 (0.019*)

NRP: non-rapid progressors; RP: rapid progressors;  $R^2$ : regression; \* $P < 0.05$ .

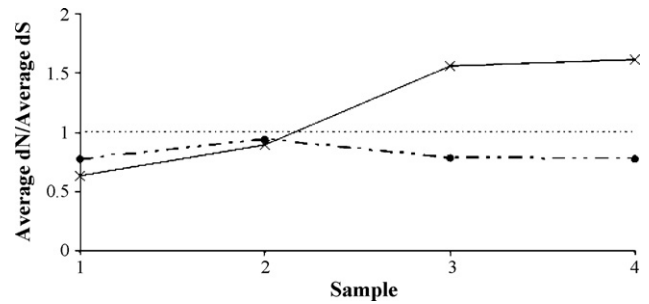


Fig. 1. Evolution of nonsynonymous and synonymous divergence in NRP (continuous line) and RP (dashed line). In this analysis, only the first four samples for each patient were used, as RP were sampled only four times.

( $r = 0.67$ ,  $P < 0.05$ ) and sampling time 2 ( $r = 0.63$ ,  $P < 0.05$ ). Finally, a logistic regression on clinical categories for the dN/dS ratio estimated with omegaMap in the presence of recombination was not significant.

### 3.2.3. Synonymous rate variation across sites

There was a significant synonymous rate variation in 8 out of the 24 patients. However, these cases were uniformly distributed among the two clinical categories: 5 out of 14 for NRP (P12–P16) and 3 out of 10 for RP (P2, P23 and P24). In all these cases, the Dual model resulted in a higher number of positively selected sites than the Nonsynonymous model.

### 3.2.4. Positively selected sites

The mean number of positively selected sites by individual (Table 1) was significantly higher in NRP than in RP (Mann–Whitney test,  $P = 0.01$ ). The number of positively selected sites had significant power to classify individuals in NRP or RP when drug therapy was considered as a covariate (71% of cases were correctly classified; Wallis test for group means,  $P = 0.03$ ). The logistic regressions using clinical category as the dependent variable and FPSS and drug therapy (yes/no) as independent variables resulted in improved classifications (75% correctly classified) albeit not significant ( $P = 0.17$  and 0.08, respectively). The number of positively selected sites at each patient and time point was not significantly correlated with the CD4 count.

The distribution of sites under positive selection was widespread and distinct between the two clinical categories (Fig. 2). Considering all patients together, 30 positively selected codons were identified from a total of 120 analyzed. Twelve of these were located in the C2 region and four in the V3 loop. At the C3 region a long interval of positively selected sites was identified stretching from codons 331 to 352. Codon 334 was found as significant four times in three different individuals, being the most consistent selected site in our samples.

The number of selective sites detected was practically constant through time for RPs (2, 2, 2 and 3 at the first four time points) and clearly increased for NRPs (2, 2, 17 and 39). These trends were significantly different according to the LMM analyses (see below). In fact, more than 90% of the positively selected sites were detected in subjects with 20 months of age or older.

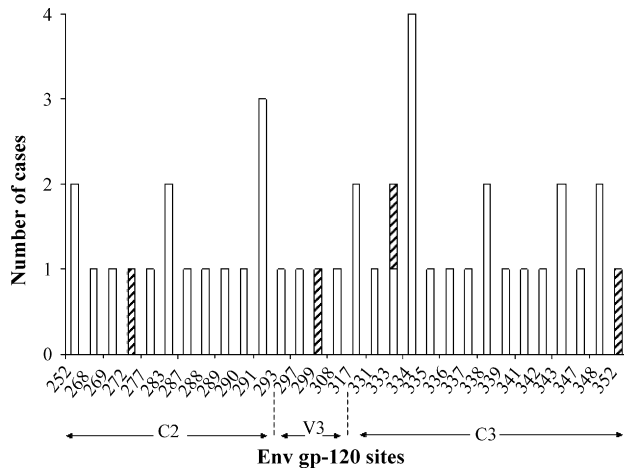


Fig. 2. Distribution of positive selected sites in NRP (white bars) and RP (hatched bars). Numbers refer to corresponding codons in the gp120 region of the HIVXB2 *env* gene.

### 3.3. Recombination analyses

The mean recombination rate and the percentage of significant recombination tests,  $32.45 \pm 11$  and 57%, respectively, were significantly higher for NRP than for RP,  $8.59 \pm 1$  and 20% (Mann–Whitney one-tailed test,  $P = 0.03$ ). Recombination rates had not much power predicting clinical categories (58.3% of the individuals were correctly classified) and the group means were not significantly different (Wallis test  $P = 0.1$ ). Logistic regressions on clinical categories including drug therapy were not significant including when the recombination rates were estimated simultaneously with the  $dN/dS$  ratio in omegaMap. Interestingly, recombination rate estimates were significantly correlated with the FPSS (Pearson correlation = 0.81,  $P = 0.000$ ), but not with the CD4 count.

### 3.4. Linear mixed models

The linear mixed models analysis indicated that time (age) had a remarkable effect on many different parameters, like genetic diversity, recombination, frequency of positive selected sites and synonymous and nonsynonymous substitution rates. Interestingly, time had a significant effect on nonsynonymous diversity,  $dN/dS$  and frequency of positive selected sites only for NRP (Table 3). Conversely, there was no increase of the recombination rate with time for either group. When the clinical category was included as a fixed effect, only nonsynonymous divergence seems to be significantly different between NRP and RP (see the ‘progress’ column at Table 4).

## 4. Discussion

The study of infected infants provides a comprehensive view of how HIV-1 genetic diversity fluctuates over the entire course of infection leading to AIDS. Most HIV-1 studies are on adults and they only sample at the onset of AIDS, which is well down the timeline from initial infection. Our data have the unique aspect of exploring early infection stages, since the infant’s

initial samplings were taken close to birth. Similar with other studies (Ganeshan et al., 1997; Shankarappa et al., 1999; Wolinsky et al., 1996), we observed higher values of genetic diversity and substitution rates in NRP than in RP; however, these differences were not significant, a fact that could be explained either by lack of power or, alternatively, by the predominance of random genetic drift of neutral mutations (Shriner et al., 2004). However, note that Edwards et al. (2006) using the same data we analyze here, conclude that intra-host HIV-1 evolution in envelope is rather dominated by purifying selection against low frequency deleterious mutations that do not reach fixation. Although Shankarappa et al. (1999) proposed that genetic diversity is positively associated with disease progression, Ross and Rodrigo (2002) reanalyzed the same data and did not find significant differences for substitution rates between long-term and slow-term progressors. Our results are consistent with this latter conclusion, but on an independent data set with a different approach to measuring genetic diversity and substitution rates. Indeed, different results relating disease progression and polymorphisms have been also reported for other HIV-1 genes, like *nef* (Chakraborty et al., 2006; Walker et al., 2007).

An inverse relation between adaptation rates and disease progression has been reported when diversity was disentangled into adaptive and selectively neutral changes (Ross and Rodrigo, 2002; Williamson, 2003). This seemed to be the case here as well. Positively selected sites accumulated through time only in NRPs. We observed higher, albeit non-significant, nonsynonymous variation in NRP than in RP. This pattern was confirmed by a significant increase in NRP, but not in RP, of nonsynonymous variation through time (Table 3). We hypothesize that this relationship is the result of a longer dynamic between the individual virus population and the host immune system and not a difference in starting levels of nonsynonymous variation ( $dN$ ). Indeed, NRP start with a lower average  $dN$  (0.0013) relative to RP (0.0057) (Table 1). Thus, initial levels of nonsynonymous variation do not appear to be predictors of disease progression. On the contrary, Strunnikova et al. (1995) obtained a positive relationship between diversity and substitution rates with disease progression in children (Strunnikova et al., 1998). Indeed, different sample sizes and/or the effect of drug therapy (reducing the effective population size and increasing genetic drift) can alter the apparent relationship between disease progression and synonymous and nonsynonymous variation.

To better understand the putative inverse relationship between disease progression status and positive selection, we measured the rate of adaptive and neutral divergence through time. Importantly, the method we have used is fairly permissive in terms of recombination and natural selection, assuming only that the underlying rate of substitution is approximately constant through time. We detected a faster increase of  $dN$  with respect to  $dS$  in NRP, and a slowdown of  $dN/dS$  in RP due to a slower increase through time of  $dN$  with respect to  $dS$ . In any case the  $dN/dS$  was stabilized at the end. Importantly, these results were confirmed by the LMM analyses. Such a pattern is expected under the hypothesis of relaxation of immune

Table 3  
Linear mixed-effects model fit for each progression category by maximum likelihood (REML)

	Fixed effects				Random effects			d.f.
	Intercept	Age	Drugs	Age × drugs	Age	CAR1	AIC	
$\theta$								
NRP-full	0.0261*	0.0292*	0.4637	0.3622	<i>I + S</i>	5.93E–09	457.35	9
NRP-AIC	0.0066**	0.0236*	–	–	<i>S</i>	–	451.22	4
RP-full	0.0027**	0.8080	0.4992	0.1180	<i>I + S</i>	4.98E–15	227.16	9
RP-AIC	0.0085**	0.0088**	–	–	<i>I</i>	–	221.25	4
$\rho$								
NRP-full	0.0383*	0.7195	0.6022	0.5608	<i>I + S</i>	2.69E–09	542.41	9
NRP-AIC	0.0241*	0.4418	–	–	–	–	535.89	3
RP-full	0.0197*	0.3106	0.5673	0.2817	<i>I + S</i>	–	320.95	9
RP-AIC	0.0117*	0.8754	–	–	–	–	319.03	4
$\pi_S$								
NRP-full	0.0465*	0.1067	0.9804	0.8062	<i>I + S</i>	0	–265.45	9
NRP-AIC	0.0234**	0.0526	–	–	<i>S</i>	–	–295.09	4
RP-full	0.0016**	0.4827	0.9917	0.2797	<i>I + S</i>	0	–191.77	9
RP-AIC	0.0009***	0.1951	–	–	–	–	–222.29	3
$\pi_N$								
NRP-full	0.0749	0.0597	0.7382	0.6192	<i>I + S</i>	0.6818	–299.00	9
NRP-AIC	0.0233*	0.0241*	–	–	<i>S</i>	–	–329.46	4
RP-full	0.0001***	0.2223	0.0459*	0.0169*	<i>I + S</i>	0	–229.41	9
RP-AIC	0.0001***	0.3634	–	–	–	–	–257.39	3
$dS$								
NRP-full	0.2948	0.0830	0.8726	0.8600	<i>I + S</i>	0	–307.05	9
NRP-AIC	0.2813	0.0317*	–	–	<i>I + S</i>	–	–334.48	6
RP-full	0.1136	0.0501	0.4154	0.6747	<i>I + S</i>	0.0859	–192.34	9
RP-AIC	0.0806	0.0026**	–	–	<i>I + S</i>	–	–219.34	6
$dN$								
NRP-full	0.7632	0.0000***	0.8402	0.5172	<i>I + S</i>	0.7056	–375.05	9
NRP-AIC	0.9647	0.0000***	–	–	<i>S</i>	0.7217	–404.95	5
RP-full	0.0823	0.0007***	0.8630	0.3239	<i>I + S</i>	0.4827	–213.86	9
RP-AIC	0.0222	0.0002***	–	–	<i>I + S</i>	–	–239.80	6
$dN/dS$								
NRP-full	0.5180	0.0099**	0.6062	0.1231	<i>I + S</i>	0.7609	377.46	9
NRP-AIC	0.9210	0.0357*	0.1971	–	<i>I + S</i>	0.7673	376.48	8
RP-full	0.0134*	0.1510	0.7227	0.2994	<i>I + S</i>	0.8582	172.70	9
RP-AIC	0.0026**	0.6266	–	–	<i>I + S</i>	–	171.01	4
FPSS								
NRP-full	0.9049	0.3301	0.9856	0.6650	<i>I + S</i>	0.3486	–252.80	9
NRP-AIC	0.7734	0.0432*	–	–	<i>S</i>	–	–281.70	4
RP-full	0.7879	0.7120	0.2013	0.3675	<i>I + S</i>	0.2609	–297.94	9
RP-AIC	0.9272	0.3855	–	–	<i>S</i>	–	–333.01	4

Note: The parameter estimates evaluated in turn for the non-progressors (NRP) and rapid progressors data (RP) were  $\theta$ : genetic diversity;  $\rho$ : population recombination rate;  $\pi_S$ : synonymous diversity;  $\pi_N$ : nonsynonymous diversity;  $dS$ : rate of synonymous substitutions per synonymous site;  $dN$ : rate of nonsynonymous substitution per nonsynonymous site; FPSS: frequency of positive selected sites. We evaluated the full model with all parameters included (full) and the best-fit model according to the Akaike information criterion (AIC). Only age was considered as a random effect, but in some models only the intercepts (*I*) or the slopes (*S*) vary among individuals. The autocorrelation was described as a continuous first-order autoregressive function (CAR1). The symbol “–” indicates that the parameter is not included in the particular model. d.f. are the degrees of freedom for each model. The values in the fixed effects cells are the *P* values. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001.

selective pressure (Williamson et al., 2005). Furthermore, only in NRP did we observed a  $dN/dS > 1$ . This might suggest that in RP the virus is actively trying to escape from the immune system, which is not able to maintain the selective pressure. Thus, RP should correspond to the first or second category in the context of phylodynamics (Grenfell et al., 2004), i.e. RPs have large viral population reflected in weak host immune response resulting in low net viral adaptation. In contrast, in NRP viral adaptation might be less effective and the immune system would be able to

hold the selection pressure longer, reducing the viral population size resulting in higher viral adaptation rate. This corresponds to category 4 in Grenfell et al. (2004). However, we obtained some estimates of  $N_e$  under different growth models in BEAST (Drummond and Rambaut, 2006) and found no significant differences between the two groups.

Interestingly, and at least for this study, the rate of change in  $dN$  seems to be a more reliable predictor of progression than the  $dN/dS$  ratio.

Table 4

Linear mixed-effects model fit by maximum likelihood (REML) with progression (P) considered as a fixed effect in the models

	Fixed								Random		
	Intercept	Age (A)	Progress (P)	Drugs (D)	A × P	A × D	P × D	A × P × D	Age	AIC	d.f.
$\theta$											
Full	0.0082**	0.0103*	0.7369	0.3798	0.1598	0.2759	0.3209	0.1607	I + S	698	13
AIC	0.0002***	0.0071**	–	–	–	–	–	–	S	690	4
PM	0.0003***	0.0066**	0.3446	–	–	–	–	–	I + S	695	8
$\rho$											
Full	0.0263*	0.7018	0.4860	0.5783	0.3353	0.5350	0.9461	0.5914	I + S	857	13
AIC	0.0016**	0.4624	–	–	–	–	–	–	–	853	3
PM	0.0111*	0.4455	0.7525	–	–	–	–	–	I + S	859	8
$\pi_S$											
Full	0.0202*	0.0637	0.5774	0.9787	0.1729	0.7773	0.8026	0.3187	I + S	–452	13
AIC	0.0007***	0.0146*	–	–	–	–	–	–	S	–531	4
PM	0.0026**	0.0292*	0.7428	–	–	–	–	–	I + S	–515	8
$\pi_N$											
Full	0.0176*	0.0150*	0.4053	0.7043	0.1345	0.5304	0.2931	0.1879	I + S	–510	13
AIC	0.0004***	0.0174*	–	–	–	–	–	–	S	–591	4
PM	0.0007***	0.0183*	0.4534	–	–	–	–	–	I + S	–574	8
dS											
Full	0.3347	0.0481*	0.3185	0.9927	0.8809	0.9721	0.8320	0.9867	I + S	–494	13
AIC	0.0188*	0.0006***	–	–	–	–	–	–	S	–571	5
PM	0.3254	0.0010**	0.1179	–	–	–	–	–	I + S	–559	8
dN											
Full	0.8528	0.0000***	0.0744	0.6380	0.9184	0.4232	0.6562	0.7765	I + S	–592	13
AIC	0.0783	0.0000***	–	–	–	–	–	–	S	–668	5
PM	0.7910	0.0000***	0.0465*	–	–	–	–	–	I + S	–657	8
dN/dS											
Full	0.4696	0.0020**	0.0564	0.5743	0.0220*	0.0762	0.6490	0.1548	I + S	565	13
AIC	0.4731	0.0764	0.6912	0.2605	–	–	–	–	I + S	564	9
PM	0.4088	0.1518	0.8320	–	–	–	–	–	I + S	566	8
FPSS											
Full	0.8986	0.2546	0.9097	0.9748	0.6475	0.5806	0.8975	0.8596	S + I	–449	13
AIC	0.6733	0.0330*	–	–	–	–	–	–	S	–525	4
PM	0.9241	0.0464	0.3759	–	–	–	–	–	I + S	–510	8

Note: The parameter estimates evaluated were  $\theta$ : genetic diversity;  $\rho$ : population recombination rate;  $\pi_S$ : synonymous diversity;  $\pi_N$ : nonsynonymous diversity; dS: rate of synonymous substitutions per synonymous site; dN: rate of nonsynonymous substitution per nonsynonymous site; FPSS: frequency of positive selected sites. We evaluated the full model with all parameters included (full) and the best-fit model according to the Akaike information criterion (AIC). Only age was considered as a random effect, but in some models only the intercepts (I) or the slopes (S) vary among individuals. The symbol “–” indicates that the parameter is not included in the particular model. d.f. are the degrees of freedom for each model. The values in the fixed effects cells are the *P* values. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001.

Our results seem to be in contrast with those obtained by Lemey et al. (2007), who were able to measure *absolute* synonymous divergence rate and unexpectedly found that it correlated negatively with disease progression. Indeed, we could not find the same relationship here as we have measured instead the *relative* synonymous divergence, implicitly assuming viral generation times are equal across patients. In addition, we have not explicitly removed the potential confounding effect of deleterious mutations by not concentrating our estimates on particular (backbone) branches of the tree. In both regards, their approach is more powerful. On the other hand, we have explicitly taken into account recombination, autocorrelation between different time measures and site-to-site variation in dS, and our results are consistent with previous studies suggesting a relationship between neutralizing antibodies and disease progression (Ross and Rodrigo, 2002; Williamson, 2003).

There have been many attempts in the past to identify particular sites undergoing positive selection in the HIV-1 *env* gene (Crandall et al., 1999; Ganeshan et al., 1997; Huelsenbeck et al., 2006; Nielsen and Yang, 1998; Suzuki and Gojobori, 1999; Yamaguchi-Kabata and Gojobori, 2000). The persistence of selection acting at the same sites over long periods of time (Ross and Rodrigo, 2002) has been postulated in some cases, but this pattern does not seem to always hold (Ganeshan et al., 1997; Williamson, 2003). In the V3 loop, several positions between codons 303 and 327 have been reported as being under selection (Ganeshan et al., 1997; Nielsen and Yang, 1998; Suzuki and Gojobori, 1999; Yamaguchi-Kabata and Gojobori, 2000), and among these, codon 308 seems to be predominant. As in previous studies (Ross and Rodrigo, 2002; Williamson, 2003), we have detected more persistently positively selected sites in NRP. We have also identified codons 308 and 317 as being under significant positive selection (Fig. 2). Regarding the C3 region,

a short interval including codons 333–347 has been reported as being under positive selection (Huelsenbeck et al., 2006; Nielsen and Yang, 1998; Yamaguchi-Kabata and Gojobori, 2000). Here we have identified a similar region including this same interval. However, the distribution of these sites does not show any consistent pattern within patient groups. A similar lack of convergence has been previously reported for another study in children (Ganeshan et al., 1997), which might be due to the individual variation in the selective regimen.

We also found that a third of the individuals exhibited synonymous substitution rate variation along the C2V3C3 region, although those samples were distributed uniformly among patient groups. Importantly, in all of these samples, the use of the Dual model under a REL approach to account for this variation allowed the detection of more positively selected sites than the use of a Nonsynonymous model under a FEL approach. The frequency of positively selected sites had only marginal power to predict the clinical category. However, the percentage of correct classification was better (although not significant) when the presence of therapy was taken into account. However, neither the effect nor the interactions held after the LMM analysis.

It is well-known that the recombination rate in HIV-1 is one of the highest of all organisms (Rambaut et al., 2004). HIV-1 recombination has been associated with disease progression (Liu et al., 2002), resistance to drug therapy, and immune escape (Kellam and Larder, 1995; Morris et al., 1999; Nájera et al., 2002). In this study, we detected significant differences in the recombination estimates between NRP and RP with the “standard” analysis, but not with the LMMs. We observed a much lower percentage of individuals with significant recombination rates within RP. Such a result might be explained in terms of the interaction of selection and recombination, in which the signal for recombination would have been obscured or lost in RP because of rapid selective sweeps (Carvajal-Rodríguez et al., 2006). In such a situation, we would expect to see lower substitution rates in RP, which was in fact the case. Furthermore, in our data we observed a positive correlation between recombination and the number of positive selected sites. There are several explanations for this. One is that recombination might be simply inflating the signal for selection (Anisimova et al., 2003; Shriner et al., 2003), but this does not seem to be the case here, as suggested by the analyses performed taking recombination into account. An alternative explanation is that the immune pressure could be favoring at the same time the fixation of nonsynonymous substitutions and the increase of the recombination rate in the population, as this might put together beneficial CTL escape mutations, as it has been shown for drug-resistance (Althaus and Bonhoeffer, 2005; Carvajal-Rodríguez et al., 2007; Kellam and Larder, 1995; Rouzine and Coffin, 2005). Finally, recombinants might accumulate in the population and/or become more detectable as diversity increases. Indeed, more data are needed to clarify these questions.

Overall, the results obtained seem to confirm that viral adaptation in the C2V3C3 region of the *env* gene is related to disease progression. Importantly, the statistical characterization

of such a relationship appears to be difficult to demonstrate. This difficulty is probably due to lack of statistical power. The latter could be the cause of the absence of any significant difference allowing the segregation between progressors and non-progressors reported in a similar study in which diversity and  $dN/dS$  were studied in seven HIV-1 subtype C infected infants (Zhang et al., 2006). In our study, we have 24 infants but although we can perceive different trends, often these are non-significant. Although not confirmed by the LMM analysis, recombination could play an important role in disease progression, due either to its adaptive value or to its potential confounding effects. In any case, recombination should be taken into account when studying the relationship between HIV-1 evolution and progression to AIDS. In addition, it will be very interesting to try to measure absolute  $dN$  and  $dS$  rates in different data sets (Lemey et al., 2007).

In this study we have performed two different types of analyses. The first type, linear regression, is very common in longitudinal studies of HIV-1 in which repeated measures are taken from the same patient. However, longitudinal data violate several assumptions of the linear regression model, particularly that data points are independent. On the other hand, linear mixed models (LMM) provide a powerful statistical framework to analyze longitudinal data. They are regression models in which the regression coefficients are allowed to vary across the subjects. Linear mixed models enable us to describe the trend over time while taking into account the correlation that exists between successive measurements. Moreover, they allow us to describe the variation in the baseline measurement and in the rate of change over time. Importantly, subjects are not assumed to be measured on the same number of time points, time points do not need to be equally spaced, and the analyses can be conducted for subjects who may miss one or more of the measurement occasions, or who may be lost to follow-up at some point during study. Here, the linear regression and the LMM analyses disagree on several occasions. For example, the relationship between recombination rate and progression status, detected by the regression analyses was not significant under the LMM. The same was true for the higher diversity and FPSS in NRP. Given that these data violate several of the assumptions of the linear regression analysis, the LMM seems more reliable and therefore, these relationships might be false positives. The main result of this study, that  $dN$  increases faster in NRP than in RP, was however suggested by both types of analyses.

## Acknowledgements

This work was supported by NIH grants R01-HD34350 (KAC, RPV) and R01-GM66276 (KAC, DP) and Brigham Young University (EK). DP was supported grant BFU2004-02700 of the Spanish Ministry of Education and Science and by the “Ramon y Cajal” programme of the Spanish government. ACR is currently funded by an Isidro Parga Pondal research fellowship from Xunta de Galicia (Spain). We thank Jing Gu, Samantha Crowe, James Demma and Imran Fatani for generation of sequence data and Emilio Rolán-Alvarez for statistical advice.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.meegid.2007.10.009](https://doi.org/10.1016/j.meegid.2007.10.009).

## References

- Abrams, E.J., Matheson, P.B., Thomas, P.A., Thea, D.M., Krasinski, K., Lambert, G., Shaffer, N., Bamji, M., Hutson, D., Grimm, K., et al., 1995. Neonatal predictors of infection status and early death among 332 infants at risk of HIV-1 infection monitored prospectively from birth. New York City Perinatal HIV Transmission Collaborative Study Group. *Pediatrics* 96, 451–458.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723.
- Althaus, C.L., Bonhoeffer, S., 2005. Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus type 1. *J. Virol.* 79, 13572–13578.
- Anisimova, M., Nielsen, R., Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229–1236.
- Carvajal-Rodríguez, A., Crandall, K.A., Posada, D., 2006. Recombination estimation under complex evolutionary models with the coalescent composite likelihood method. *Mol. Biol. Evol.* 23, 817–827.
- Carvajal-Rodríguez, A., Crandall, K.A., Posada, D., 2007. Recombination favors the evolution of drug resistance in HIV-1 during antiretroviral therapy. *Infect. Genet. Evol.* 7, 476–483.
- Chakraborty, R., Reinis, M., Rostron, T., Philpott, S., Dong, T., D'Agostino, A., Musoke, R., Silva, E., Stumpf, M., Weiser, B., Burger, H., Rowland-Jones, S.L., 2006. Nef gene sequence variation among HIV-1-infected African children. *HIV Med.* 7, 75–84.
- Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, C.H., Salzman, N.P., 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* 16, 372–382.
- Delwart, E.L., Shpaer, E.G., Louwagie, J., McCutchen, F.E., Grez, M., Rubsamen-Waigmann, H., Mullins, J.I., 1993. Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes. *Science* 262, 1257–1261.
- Drummond, A.J., Rambaut, A., 2006. BEAST v1.4.
- Edwards, C.T., Holmes, E.C., Wilson, D.J., Viscidi, R.P., Abrams, E.J., Phillips, R.E., Drummond, A.J., 2006. Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evol. Biol.* 6, 28.
- Ganeshan, S., Dickover, R.E., Korber, B.M., Bryson, Y.J., Wolinsky, S.M., 1997. Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease. *J. Virol.* 71, 663–677.
- Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L., Daly, J.M., Mumford, J.A., Holmes, E.C., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303, 327–332.
- Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704.
- Huelsenbeck, J.P., Jain, S., Frost, S.W., Pond, S.L., 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6263–6268.
- Kellam, P., Larder, B.A., 1995. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *J. Virol.* 69, 669–674.
- Lemey, P., Pond, S.L.K., Drummond, A.J., Pybus, O.G., Shapiro, B., Barroso, H., Taveira, N., Rambaut, A., 2007. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol.* 3, e29.
- Liu, S.L., Schacker, T., Musey, L., Shriner, D., McElrath, M.J., Corey, L., Mullins, J.I., 1997. Divergent patterns of progression to AIDS after infection from the same source: human immunodeficiency virus type 1 evolution and antiviral responses. *J. Virol.* 71, 4284–4295.
- Liu, S.L., Mittler, J.E., Nickle, D.C., Mulvania, T.M., Shriner, D., Rodrigo, A.G., Kosloff, B., He, X., Corey, L., Mullins, J.I., 2002. Selection for human immunodeficiency virus type 1 recombinants in a patient with rapid progression to AIDS. *J. Virol.* 76, 10674–10684.
- McVean, G.A.T., Awadalla, P., Fearnhead, P., 2002. A coalescent based-method for detecting and estimating recombination from gene sequences. *Genetics* 160, 1231–1241.
- Morris, A., Marsden, M., Halcrow, K., Hughes, E.S., Brettle, R.P., Bell, J.E., Simmonds, P., 1999. Mosaic structure of the human immunodeficiency virus type 1 genome infecting lymphoid cells and the brain: evidence for frequent in vivo recombination events in the evolution of regional populations. *J. Virol.* 73, 8720–8731.
- Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724.
- Nájera, R., Delgado, E., Pérez-Alvarez, L., Thomson, M.M., 2002. Genetic recombination and its role in the development of the HIV-1 pandemic. *AIDS* 16 (Suppl. 4), S3–S16.
- Nielsen, R., Yang, Z., 1998. Likelihood methods for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936.
- Nowak, M.A., Anderson, R.M., McLean, A.R., Wolfs, T.F.W., Goudsmit, J., May, R.M., 1991. Antigenic diversity threshold and the development of AIDS. *Science* 254, 963–969.
- Nowak, M.A., Anderson, R.M., Boerlijst, M.C., Bonhoeffer, S., May, R.M., McMichael, A.J., 1996. HIV-1 evolution and disease progression. *Science* 274, 1008–1010.
- Pinheiro, J.C., Bates, D.M., 2000. *Mixed-effects Models in S and S-PLUS*. Springer, New York.
- Pond, S.K., Muse, S.V., 2005. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* 22, 2375–2385.
- Pond, S.L.K., Frost, S.D., 2005a. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22, 1208–1222.
- Pond, S.L.K., Frost, S.D.W., 2005b. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21, 2531–2533.
- Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679.
- Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Posada, D., Crandall, K.A., 2001. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.* 18, 897–906.
- Potter, S.J., Lemey, P., Dyer, W.B., Sullivan, J.S., Chew, C.B., Vandamme, A.M., Dwyer, D.E., Saksena, N.K., 2006. Genetic analyses reveal structured HIV-1 populations in serially sampled T lymphocytes of patients receiving HAART. *Virology* 348, 35–46.
- Rambaut, A., 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16, 395–399.
- Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C., 2004. The causes and consequences of HIV evolution. *Nat. Rev. Genet.* 5, 52–61.
- Raudenbush, S.W., Bryk, A.S., 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Sage, Thousand Oaks, CA.
- Ross, H.A., Rodrigo, A.G., 2002. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J. Virol.* 76, 11715–11720.
- Rouzine, I.M., Coffin, J.M., 2005. Evolution of human immunodeficiency virus under selection and weak recombination. *Genetics* 170, 7–18.
- Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., Rozas, R., 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19, 2496–2497.
- Scheffler, K., Martin, D.P., Seoighe, C., 2006. Robust inference of positive selection from recombining coding sequences. *Bioinformatics* 22, 2493–2499.

- Seibert, S.A., Howell, C.Y., Hughes, M.K., Hughes, A.L., 1995. Natural selection on the *gag*, *pol*, and *env* genes of human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.* 12, 803–813.
- Shankarappa, R., Margolick, J.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.-L., Mullins, J.I., 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* 73, 10489–10502.
- Shriner, D., Nickle, D.C., Jensen, M.A., Mullins, J.I., 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet. Res.* 81, 115–121.
- Shriner, D., Shankarappa, R., Jensen, M.A., Nickle, D.C., Mittler, J.E., Margolick, J.B., Mullins, J.I., 2004. Influence of random genetic drift on human immunodeficiency virus type 1 *env* evolution during chronic infection. *Genetics* 166, 1155–1164.
- Sokal, R.R., Rohlf, F.J., 1981. *Biometry*, 2nd ed. W.H. Freeman and Co., New York.
- Strunnikova, N., Ray, S.C., Livingston, R.A., Rubalcaba, E., Viscidi, R.P., 1995. Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. *J. Virol.* 69, 7548–7558.
- Strunnikova, N., Ray, S.C., Lancioni, C., Nguyen, M., Viscidi, R.P., 1998. Evolution of human immunodeficiency virus type 1 in relation to disease progression in children. *J. Hum. Virol.* 1, 224–239.
- Suzuki, Y., Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16, 1315–1328.
- Swofford, D.L., 2002. *PAUP\**. Phylogenetic Analysis Using Parsimony (\*and Other Methods), 4th ed. Sinauer Associates, Sunderland, MA.
- Team, R.D.C., 2006. R: a language and environment for statistical computing. <http://www.R-project.org> (online).
- Templeton, A.R., Reichert, R.A., Weisstein, A.E., Yu, X.-F., Markham, R.B., 2004. Selection in context: patterns of natural selection in the glycoprotein 120 region of human immunodeficiency virus 1 within infected individuals. *Genetics* 167, 1547–1561.
- Thomas, P.A., Weedon, J., Krasinski, K., Abrams, E., Shaffer, N., Matheson, P., Bamji, M., Kaul, A., Hutson, D., Grimm, K.T., et al., 1994. Maternal predictors of perinatal human immunodeficiency virus transmission. The New York City Perinatal HIV Transmission Collaborative Study Group. *Pediatr. Infect. Dis. J.* 13, 489–495.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24, 4876–4882.
- Viscidi, R., 1999. The evolution of HIV. HIV evolution and disease progression via longitudinal studies. In: Crandall, K.A. (Ed.), *The Evolution of HIV*. The Johns Hopkins University Press, pp. 346–389.
- Walker, P.R., Ketunuti, M., Choge, I.A., Meyers, T., Gray, G., Holmes, E.C., Morris, L., 2007. Polymorphisms in Nef associated with different clinical outcomes in HIV type 1 subtype C-infected children. *AIDS Res. Hum. Retrovirus.* 23, 204–215.
- Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276.
- Williamson, S., 2003. Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. *Mol. Biol. Evol.* 20, 1318–1325.
- Williamson, S., Perry, S.M., Bustamante, C.D., Orive, M.E., Stearns, M.N., Kelly, J.K., 2005. A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients. *Mol. Biol. Evol.* 22, 456–468.
- Wilson, D.J., McVean, G., 2006. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 172, 1411–1425.
- Wolinsky, S.M., Korber, B.T.M., Neumann, A.U., Daniels, M., Kunstman, K.J., Whetsell, A.J., Furtado, M.R., Cao, Y., Ho, D.D., Safrit, J.T., Koup, R.A., 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science* 272, 537–542.
- Yamaguchi-Kabata, Y., Gojobori, T., 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* 74, 4335–4350.
- Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556.
- Zanotto, P.M.D.A., Kallas, E.G., de Souza, R.F., Holmes, E.C., 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* 153, 1077–1089.
- Zhang, H., Hoffmann, F., He, J., He, X., Kankasa, C., West, J.T., Mitchell, C.D., Ruprecht, R.M., Orti, G., Wood, C., 2006. Characterization of HIV-1 subtype C envelope glycoproteins from perinatally infected children with different courses of disease. *Retrovirology* 3, 73.