J Mol Evol (2003) 57:212–221
DOI: 10.1007/s00239-003-2467-9

JOURNAL OF MOLECULAR EVOLUTION

Widespread Adaptive Evolution in the Human Immunodeficiency Virus Type 1 Genome

Wa Yang, Joseph P. Bielawski, Ziheng Yang

Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

Received: 30 September 2002 / Accepted: 5 March 2003

Abstract. We investigated variable selective pressures among amino acid sites in HIV-1 genes. Selective pressure at the amino acid level was measured by using the nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$). To identify amino acid sites under positive selection with $\omega > 1$, we applied maximum likelihood models that allow variable ω ratios among sites to analyze genomic sequences of 26 HIV-1 lineages including subtypes A, B, and C. Likelihood ratio tests detected sites under positive selection in each of the major genes in the genome: *env*, *gag*, *pol*, *vif*, and *vpr*. Positive selection was also detected in *nef*, *tat*, and *vpu*, although those genes are very small. The majority of positive selection sites is located in gp160. Positive selection was not detected if ω was estimated as an average across all sites, indicating the lack of power of the averaging approach. Candidate positive selection sites were mapped onto the available protein tertiary structures and immunogenic epitopes. We measured the physiochemical properties of amino acids and found that those at positive selection sites were more diverse than those at variable sites. Furthermore, amino acid residues at exposed positive selection sites were more physiochemically diverse than at buried positive selection sites. Our results demonstrate genomewide diversifying selection acting on the HIV-1.

Key words: HIV-1 — Positive selection — Physiochemical properties — Epitopes

Correspondence to: Ziheng Yang; *email:* z.yang@ucl.ac.uk

Introduction

The primary human defense system against HIV-1 is a combination of humoral (antibody) and cytotoxic T-lymphocyte (CTL) responses. Antibody response is produced by B lymphocytes upon recognition of foreign antigens via immunoglobulin receptors (Bondada and Chelvarajan 1999). T helpers are responsible for activation of neutralizing antibodies and maintenance of the CTL response (Phillips et al. 2001). CTLs are a major component of T cells and are capable of recognizing and eliminating infected host cells (Gotch 1998). Several studies have shown both antibody and CTL responses following invasion of HIV-1 (reviewed by Fomsgaard 1999). Many viral proteins are targeted by antibodies, CTLs, and T helpers throughout the course of infection. However, the immune system itself is the main target of HIV-1, leading to the reduced $CD4^+$ T-lymphocyte (T-helper) count that is characteristic of HIV infection. When $CD4^+$-cell counts drop to <200 cells/μl, the infected individual becomes very susceptible to additional opportunistic infections and is characterized as an AIDS patient (Siliciano 2001).

The success of HIV-1 lies in its ability to evade recognition by CTLs and antibodies. HIV-1 specific CTL counts tend to peak in the first 3 to 4 weeks of infection. This is followed by a sharp increase in viral mutants with substantial sequence variation (Allen et al. 2000). Decline of CTL epitope-specific lymphocytes is a direct result of viral escape by means of mutation. Direct sequencing of the entire virus

genome immediately after the acute infection phase (8 weeks) showed that amino acid altering mutations accumulated within CTL epitopes (Allen et al. 2000). Presumably, there was strong positive selection acting on CTL epitopes, as neutral evolution alone could not account for the observed level of divergence in such a short period of time (Lukashov and Goudsmit 1997). Furthermore, a sharp increase in antibody titer is also observed following HIV infection. Various *in vitro* and *in vivo* studies using animal models demonstrated that selective pressure was also exerted on neutralizing antibody epitopes and was sufficient to generate escape mutations (Langedijk et al. 1995; Igarashi et al. 1996; Calarota et al. 1996; McLain et al. 2001). Apparently, maximizing variation in the surface antigen allows HIV-1 to escape immune surveillance.

The HIV-1 genome encodes seven proteins and two polyproteins. The genome is highly compact, with every gene overlapping with other genes or with long terminal repeats. The HIV-1 genome is also characterized by a high rate of evolution (with an average rate of 1.6×10^{-2} nucleotide substitutions per site per year) compared with the substitution rates of human nuclear genes, estimated to be 1.3×10^{-9} substitutions per site per year (Eyre-Walker and Keightley 1999; Fu 2000). Positive selection at sites crucial to maintaining antigenic variation could contribute to this high evolutionary rate. Positive selection is likely to operate in regions of a protein where a high level of structural specificity is not required (e.g., Walker and Goulder 2000). Thus, identification of sites with excess amino acid replacements could contribute to our understanding of positive selection and antigenic variation. In this cross-sectional statistical study, we attempted to identify sites that were under recurrent selective pressure over long evolutionary time.

Materials and Methods

Sequence Alignment and Phylogenetic Inference

The complete genomes of 26 HIV-1 isolates were obtained from GenBank. The data set comprised 5 subtype A, 7 subtype B, and 14 subtype C nonrecombinant isolates (accession Nos. AF069669–AF069673, AF042100–AF042106, and AF110959–AF110981) (Robertson et al. 1999). Due to extensive overlap with different reading frames, we excluded *tat*, *rev*, and *vpu* from most of the analysis. All overlapping regions of the remaining genes were also excluded. The data set was aligned using ClustalX (Thompson et al. 1997) and manually adjusted using GeneDoc (Nicholas et al. 1997). Regions of large indels were excluded from analysis. Due to alignment uncertainty, we also excluded hypervariable regions of gp160 (part of VI and V2 loops). Alignments are available from the authors on request. A phylogenetic tree was estimated using maximum likelihood under the model of Hasegawa et al. (1985), as implemented in the program PAUP* 4a7b (Swofford 2000).

Likelihood Ratio Test of Positive Selection and Estimation d_N/d_S (ω) Ratios Across the Genome

We employed six codon substitution models (Yang et al. 2000)—M0 (one ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (β), and M8 (β and ω)—as implemented in PAML (Yang 2000). We used the discrete model (M3) with three ω classes. The LRT comparing M0 with M3 was a test for among-site rate variation. The comparisons of M1 with M2 and M7 with M8 were tests for positive selection. Codon usage bias is well known to affect estimation of synonymous and nonsynonymous substitution rates (Yang and Nielsen 1998). Thus we used two models to account for codon usage bias in all our analyses: (i) F3×4, which computes equilibrium codon frequencies from the nucleotide frequencies at the three codon positions; and (ii) F61, which uses empirical estimates of individual codon frequencies. We found that the results under the two models were similar, and present those under F3×4 only. Branch lengths of the phylogeny (measured as the expected number of nucleotide substitutions per codon along a branch) and the transition-to-transversion rate ratio (κ) were estimated using ML. The posterior probability that a site belonged to each ω class was estimated using Bayesian methods. Sites with a posterior probability $>95\%$ of being from the class with $\omega > 1$ were identified from M2 and are referred to as "positive selection sites" (Yang et al. 2000). Sites identified from the $\omega = 1$ rate class of M2 with a posterior probability $>95\%$ are referred to as "variable sites."

Amino Acid Acceptability, Protein 3D Structure, and Epitope Mapping

Amino acid acceptability was measured at each site relative to four physiochemical properties: (i) polarity (Grantham 1974), (ii) volume (Grantham 1974), (iii) hydropathy (Kyte and Doolittle 1982), and (iv) isoelectric point (Alff-Steinberger 1969). The means and standard deviations (SD) of polarity, volume, hydropathy, and isoelectric point were computed for amino acids at each site using DAMBE version 4.0.39 (Xia 2000). Amino acid acceptability at a site was measured for each physiochemical property as $100 \times (SD/mean)$. Amino acid acceptability is a measure of functional constraints acting on the amino acids at a site, with low acceptability indicating conservation of physiochemical properties at a site.

Tertiary structures were available from RCSB Protein Data Bank for all proteins except Vif. Buried and exposed sites were identified using the program WEBMOL (Walther 1997). We compared patterns of amino acid acceptability at positive selection sites and variable sites and also at exposed and buried residues of the proteins.

CTL, antibody, and T-helper epitopes were collected from the HIV Molecular Immunology 2000 database (Korber et al. 2000). CTL epitopes were used only if CTLs recognized the naturally processed epitopes and both the optimal epitope and the restricting HLA molecule were defined; this subset of CTL epitopes is given by Brander and Goulder (2000).

Results

Positive Selection in the HIV-1 Genome

Maximum likelihood estimates of parameters under different models of variable ω ratios among sites are presented in Table 1 for the five major genes in the HIV-1 genome. Patterns of selective pressure were

Table 1. Parameter estimates under five models of variable ω's among sites

		Parameter estimates under different models					
Gene	L_C	M0 (one-ratio)	M1 (neutral)	M2 (selection)	M3 (discrete)	M7 (β)	M8 (β & ω)
gag	420	$\omega = 0.24$	$p_0 = 0.60, \omega_0 = 0$ $p_1 = 0.40, \omega_1 = 1$	$p_0 = 0.60, \omega_0 = 0$ $p_1 = 0.38, \omega_1 = 1$ **$p_2 = 0.01, \omega_2 = 4.02$**	$p_0 = 0.72, \omega_0 = 0.05$ $p_1 = 0.23, \omega_1 = 0.59$ **$p_2 = 0.05, \omega_2 = 1.81$**	$B(0.20, 0.59)$	$B(0.29, 1.18)$ $p_0 = 0.95$ **$p_1 = 0.05, \omega = 1.79$**
Pal	907	$\omega = 0.32$	$p_0 = 0.68, \omega_0 = 0$ $p_1 = 0.32, \omega_1 = 1$	$p_0 = 0.68, \omega_0 = 0$ $p_1 = 0.31, \omega_1 = 1$ **$p_2 = 0.01, \omega_2 = 7.20$**	$p_0 = 0.81, \omega_0 = 0.05$ $p_1 = 0.18, \omega_1 = 0.52$ **$p_2 = 0.01, \omega_2 = 3.40$**	$B(0.24, 1.02)$	$B(0.33, 1.74)$ $p_0 = 0.99$ **$p_1 = 0.01, \omega = 3.49$**
vif	152	$\omega = 0.61$	$p_0 = 0.50, \omega_0 = 0$ $p_1 = 0.50, \omega_1 = 1$	$p_0 = 0.50, \omega_0 = 0$ $p_1 = 0.47, \omega_1 = 1$ **$p_2 = 0.03, \omega_2 = 4.57$**	$p_0 = 0.61, \omega_0 = 0.06$ $p_1 = 0.36, \omega_1 = 0.95$ **$p_2 = 0.03, \omega_2 = 3.74$**	$B(0.19, 0.29)$	$B(0.21, 0.31)$ $p_0 = 0.97$ **$p_1 = 0.03, \omega = 3.49$**
vpr	69	$\omega = 0.72$	$p_0 = 0.46, \omega_0 = 0$ $p_1 = 0.54, \omega_1 = 1$	$p_0 = 0.54, \omega_0 = 0$ $p_1 = 0.38, \omega_1 = 1$ **$p_2 = 0.09, \omega_2 = 3.81$**	$p_0 = 0.68, \omega_0 = 0.06$ $p_1 = 0.23, \omega_1 = 0.73$ **$p_2 = 0.09, \omega_2 = 2.73$**	$B(0.16, 0.26)$	$B(0.29, 0.78)$ $p_0 = 0.91$ **$p_1 = 0.09, \omega = 2.71$**
env	694	$\omega = 0.79$	$p_0 = 0.46, \omega_0 = 0$ $p_1 = 0.54, \omega_1 = 1$	$p_0 = 0.46, \omega_0 = 0$ $p_1 = 0.43, \omega_1 = 1$ **$p_2 = 0.11, \omega_2 = 4.67$**	$p_0 = 0.62, \omega_0 = 0.08$ $p_1 = 0.27, \omega_1 = 0.82$ **$p_2 = 0.11, \omega_2 = 3.30$**	$B(0.21, 0.36)$	$B(0.31, 0.68)$ $p_0 = 0.89$ **$p_1 = 0.11, \omega = 3.18$**
Supergene	2473	$\omega = 0.34$	$p_0 = 0.74, \omega_0 = 0$ $p_1 = 0.26, \omega_1 = 1$	$p_0 = 0.74, \omega_0 = 0$ $p_1 = 0.24, \omega_1 = 1$ **$p_2 = 0.03, \omega_2 = 5.88$**	$p_0 = 0.87, \omega_0 = 0.10$ $p_1 = 0.11, \omega_1 = 1.03$	$B(0.74, 0.26)$	$B(0.27, 0.90)$ $p_0 = 0.95$ **$p_1 = 0.05, \omega = 3.58$**

Note: L_C is the number of codons after the removal of alignment gaps. p_i is the proportion of sites assigned to an ω category or to a β distribution with parameters p and q. ω ratios greater than 1.0 and corresponding proportions of sites are in bold face. The supergene was constructed by concatenating all five protein-coding genes.

Table 2. Likelihood ratio statistics ($2\Delta l$) for comparing models of variable ω's among sites

Gene	M0 vs. M3 ($\chi^2_{0.01,4} = 13.28$)	M1 vs. M2 ($\chi^2_{0.01,2} = 9.21$)	M7 vs. M8 ($\chi^2_{0.01,2} = 9.21$)
gag	361.11	20.13	18.36
pol	524.86	82.54	84.29
vif	160.80	22.79	15.91
vpr	88.66	13.73	14.66
env	1666.43	550.63	382.46
Supergene	1137.28	1047.95	252.42

Table 3. Sites identified as evolving by positive selection under M2

Gene	Combined data set	Consensus selection sites from separate subtypes
gag	15R, **28K**, 54S, 62G, 69Q, 79Y, 84T, **91R**, 138I, 146A, 215V, 252N, 280T, 357G	28K, 62G, 69G, 84V, **91R**, 146A, 280T
pol	119L, **278D**, 317S, 328K, **362Q**, 366R, **400V**, 441T, 489Q, **531T**, 532T, 590V, **623T**, 638Y, **709A**, 834S, 839A	119L, 278D, **328K**, 362Q, 623T, 709A
vif	31V, 33G, 36R, 37G, **39F**, 61D, 63R, 92K, 101E, **127H**, 132R, **167T**	33G, 36R, **39F**, 92K, 127H, **167T**
vpr	28N, **37I**, 41G, 48E, 55A, 60I, 77R, **84T**	28N, **37I**, 41G, 60I, **77R**, 84T
env	62D, **85V**, **87V**, **92N**, **130K**, **132T**, 173Y, 178K, 183P, **187D**, 188P, 190S, 200V, 230N, 231K, **232T**, 238P, 240T, 275V, 277F, **281A**, 283T, 289N, 291S, **295N**, **308R**, 321G, **336A**, **337K**, 279D, **340N**, **343K**, **344Q**, 345A, **346A**, 351E, 350R, **362K**, **389Q**, **440S**, **442Q**, 446S, **460N**, **461S**, 467I, 500K, 607A, **612A**, **619L**, **620E**, **621Q**, **624N**, **640S**, 641L, **644S**, 815L, 817A, 832V, 833V, 836A, 851L	**85V**, 87V, **130K**, 173Y, 187D, 200V, **240T**, 232T, **281A**, **291S**, 308R, **336A**, 337K, 340N, 344Q, **346A**, **350R**, 362K, **440S**, **442Q**, 446S, 460N, 612A, 640S, **644S**, 815L, 817A, 851L

Note: Sites are numbered according to the reference sequence HXB2 (GenBank accession number K03455). Positive selection sites were identified using the empirical Bayes approach with posterior probability $P \geq 90\%$, with those at $P \geq 95\%$ in boldface. Consensus selection sites were positive selection sites identified in separate analyses of all three subtypes.

similar among all five genes. Here we present the *gag* gene as an example of our findings. Estimation of ω as an average across all sites and evolutionary history (M0) suggested that *gag* was evolving by purifying selection. The average ω of 0.24 indicated that a replacement mutation had approximately one-fourth the chance of being fixed compared to a silent mutation (Table 1). However, analyses using models that allowed variable selective pressures indicated that sites of the *gag* gene were subjected to different selection intensities (Table 1). Under the neutral model (M1), maximum likelihood estimation suggested that 60% of the sites were conserved with $\omega_0 = 0$, while the remaining 40% evolve "neutrally" with $\omega_0 = 1$. The addition of an extra rate class in M2 indicated that 1% of the sites are under positive selection, with $\omega_2 = 4$ (Table 1). Under M3, 72% of sites had an ω ratio of 0.05 (Table 1), 23% of sites had an ω ratio of 0.59, and 5% had an ω ratio of 1.81. Under M8, the estimates suggest that ~5% of sites are evolving by positive selection with $\omega = 1.79$. The likelihood ratio test suggested rejection of the null model in each of the three comparisons: M0 (one-ratio) against M3 (discrete), M1 (neutral) against M2 (selection), and M7 (β) against M8 (β and ω) (Table 2).

We note that models M1 and M2 are unrealistic for most genes. As a result, they tend to be conservative for the purposes of testing and identifying positively selected sites (Yang et al. 2000; Anisimova et al. 2001, 2002). Because M2 has a site class with $\omega = 1$, some sites evolving under weaker positive selection (with ω slightly >1) tend to be lumped into this class, so that M2 often identifies a subset of the sites identified by M3 (discrete) or M8 (β and ω). This is the case for the *gag* gene, although parameter estimates for other genes are very similar between the models (Table 1). Thus, in the analysis of chemical properties of amino acids at positive selection sites (see below), we use the conservative set of sites identified under M2 (Table 3).

Parameter estimates under all models indicated that all five genes (*gag*, *pol*, *vif*, *vpr*, and *env*) were subject to similar patterns of selective pressure. In each gene, the majority of sites was subjected to strong functional constraints, with ω close to zero. Among the remaining sites a large fraction was evolving under weak functional constraints, with ω ratios between 0.5 and 0.95 (M3; Table 1). Most importantly, a small fraction of sites in each gene was evolving by positive selection, with $\omega > 1$ (Table 1).

Table 4. Partition of sites within different class of immunogenic epitopes

Epitope type	Total sites	Positive selection sites
CTL		
Epitope	1057	44
Nonepitope	1134	68
Antibody		
Epitope	960	43
Nonepitope	1231	69
T helper		
Epitope	1151	79
Nonepitope	1040	33
Total		
Epitope	1679	98
Nonepitope	512	14

Table 5. Acceptability of amino acid substitutions at exposed and buried sites

	Exposed sites	Buried sites
Polarity		
Positive selection sites	16.91	8.49
Variable sites	5.16	5.13
Volume		
Positive selection sites	41.34	25.38
Variable sites	14.54	12.37
Hydropathy		
Positive selection sites	44.68	32.88
Variable sites	13.22	19.80
Isoelectric point		
Positive selection sites	24.56	20.01
Variable sites	8.31	1.39

Note. Acceptability was calculated as $100 \times (\text{SD}/\text{mean})$. There are 75 exposed and 37 buried positive selection sites and 163 exposed and 30 buried variable sites.

In each gene, LRTs indicated that some variation in selective pressure was due to positive selection (Table 2). Parameter estimates under M2 and M8 were consistent with M3 in indicating evolution by positive selection at a small fraction of sites (Table 1).

We also analyzed a ''supergene'' sequence constructed by concatenating all five protein-coding genes. Results of this analysis were virtually the same as in the separate analysis (Tables 1 and 2). Similar parameter estimates and likelihood values were obtained using different trees reconstructed using different phylogenetic methods (data not shown); hence, our analysis was robust to tree topology.

At a threshold posterior probability of 90%, M2, M3, and M8 identified 112 positive selection sites (Table 3). We also conducted a separate analysis for each of the A, B, and C subtypes represented in our sample. Sites identified to be under positive selection (with posterior probabilities ≥90%) in every subtype were designated consensus selection sites (CSS). Because this criterion is much more stringent, the CSS was a subset of sites identified in the combined analysis of all subtypes (Table 3).

Amino Acid Diversity, Protein Tertiary Structure, and Immunogenic Epitopes

We mapped known immunogenic epitopes (CTL, antibody, and T helper) onto our sample of DNA sequences and partitioned all the codon sites into epitope and nonepitope sites (Table 4). We used a χ^2 test to determine whether there was a significant excess of the 112 positive selection sites in any one class of epitope by comparing the frequency of positive selection sites at epitope and nonepitope sites to the expected frequencies if such sites had been drawn at random from the genome. At CTL and antibody epitopes the observed and expected proportions did not differ significantly. However, there was a highly significant excess of positive selection sites at T-helper epitopes ($p < 0.0001$). In total, 98 of the 112 positive selection sites were located at at least one of the three types of epitope site. Our results, however, suggest that the dominant pattern is an association with T-helper epitopes.

We selected sites from M2 with a posterior probability of 95% and estimated the acceptability of amino acids at those sites (see Materials and Methods). Acceptability was measured in terms of polarity, volume, hydropathy, and isoelectric point. As a test of robustness, we also estimated acceptability for positive selection sites inferred under M3. Results under M3 were highly similar to those obtained under M2; results under M2 are listed in Table 5. Note that at the threshold of 95%, M8 inferred the same set of positive selection sites as M3.

To investigate differences between positive selection and relaxed functional constraints, we compared the acceptability of positive selection sites with that of variable sites ($\omega_2 = 1$ rate class, under M2, with a posterior probability ≥95%) (Fig. 1). Amino acids at positive selection sites had a higher physiochemical diversity than those at variable sites. Furthermore, the mean acceptabilities at exposed positive selection sites, in terms of polarity, volume, and hydropathy, were substantially higher than both buried positive selection sites and exposed variable sites (Table 5). Differences between acceptability of exposed and that of buried variable sites depended on the physiochemical property, with greater acceptability in volume and isoelectric point, nearly equal acceptability in polarity, and lower acceptability in hydropathy at exposed sites (Table 5).

Discussion

Evidence of Positive Selection in the HIV-1 Genome

Many studies on HIV-1 evolution have focused primarily on single-gene analysis, with separate studies

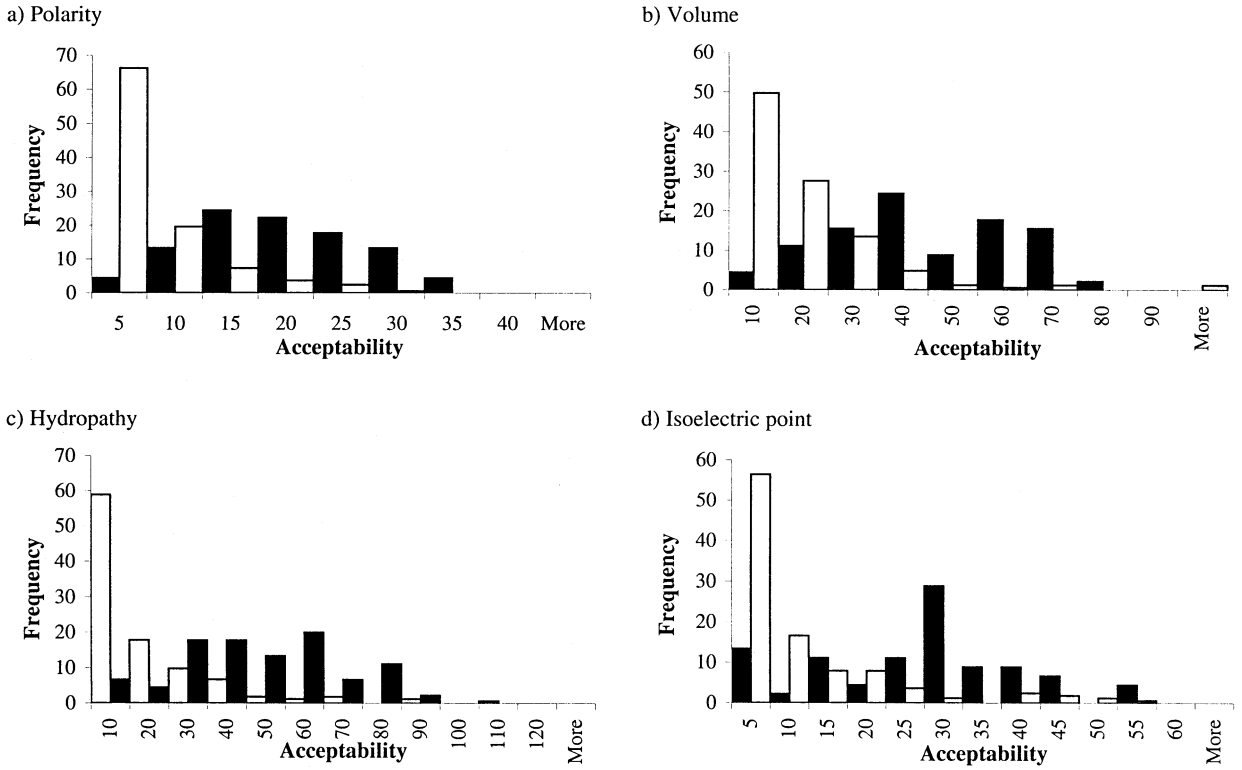

Fig. 1. Relative frequency of positive selection sites (*black bars*) and variable sites (*white bars*) at different amino acid acceptabilities, with the acceptability measure: (**a**) polarity, (**b**) volume, (**c**) hydropathy, and (**d**) isoelectric point.

on *pol*, *vif*, *env*, and *nef* indicating a role for positive selection (Zanotto et al. 1999; Yamaguchi-Kabata and Gojobori 2000; Yang et al. 2000). We expanded upon previous studies by analyzing all genes in the entire HIV-1 genome. We sampled three of the most prevalent subtypes (A, B, and C) to identify sites that were under continual selective pressure. The likelihood analysis suggested positive selection in all five major genes, *gag*, *pol*, *vif*, *vpr*, and *env*, in all three subtypes. We also analyzed *nef* for subtypes A and C and found that it was undergoing positive selection; it was excluded in the combined analysis because it was incomplete in subtype B. The two small genes, *tat* and *vpu*, were also found to be under positive selection by the likelihood ratio test. However, we excluded them from further analysis because they had only 39 and 58 nonoverlapping codons, respectively. We were unable to analyze *rev*, as its reading frame is completely overlapped by *tat* and *env*. In sum, our study is the first to suggest that essentially every gene in the HIV-1 genome is evolving under positive selection.

In viral evolution, ω is commonly used as a measure of selective constraints on a protein (e.g., Crandall et al. 1999; Zanotto et al. 1999; Yamaguchi-Kabata and Gojobori 2000). Early analyses estimated ω as an average across all sites between a pair of lineages and indicated no role for positive selection in HIV-1 evolution (Seibert et al. 1995; Leigh Brown 1997; Plikat et al. 1997). Recent studies, however, showed that a subset of sites in *vif*, *env*, and *nef* is evolving by positive Darwinian selection (Nielsen and Yang 1998; Yang et al. 2000; Zanotto et al. 1999). Traditional approaches failed to detect positive selection in *env* and *nef* because most sites in those genes were evolving under purifying selection, yielding an average d_N/d_S over all sites of $\ll 1$ (Sharp 1997; Yang et al. 2000). Our findings indicate that this pattern is characteristic of all HIV-1 genes examined; most sites are evolving under purifying selection, with positive Darwinian selection acting on only a small set of sites. Furthermore, when we estimated ω as an average over all sites, no positive selection was indicated in any HIV-1 gene. This situation is not unique to HIV-1, as methods that average d_N/d_S over sites had a low power to detect positive selection in other viruses (e.g., Gojobori et al. 1990).

The empirical Bayes approach has become a popular method for inferring positive selection sites (Haydon et al. 2001; Swanson et al. 2001; Fares et al. 2001), however, Haydon et al. (2001) recently expressed concerns about type I error rates. Recent simulation studies suggested that the Bayes approach was reliable provided that the sample size was not too small and the sequence divergence was not too low (Anisimova et al. 2002). In simulations, identification of positively selected sites was reliable for a data set

with as few as 17 taxa and 0.07 expected substitution per branch (Anisimova et al. 2002). With approximately 0.16 expected substitution per branch under M2, M3, and M8, our data sets appear to be well within the window of sequence divergence required for reliable identification of positively selected sites. Two additional lines of evidence support this notion. First, under all three models positive selection sites with a posterior probability >90% were almost identical. Second, sites we identified in *env* were highly consistent with those identified previously (Yamaguchi-Kabata and Gojobori 2000; Yang 2001). Yamaguchi-Kabata and Gojobori (2000) identified 33 positive selection sites, 16 of which were located in regions excluded from our analysis due to alignment uncertainty. We identified the remaining 17 sites as being under positive selection. We also identified seven additional positive selection sites, four of which were also identified in Yang's (2001) analysis. Our analysis differed from the previous two studies of *env* by only three additional sites, although our data set included sequences from two additional subtypes of HIV-1.

The Influence of Recombination

The HIV-1 genome is characterized by a high recombination rate that is estimated to be as high as three crossovers per genome per replication event (Jetzt et al. 2000). There are two types of recombination, intersubtype and intrasubtype. Intersubtype recombination gives rise to circular recombinants but does not affect this analysis because we sampled only those sequences classified to be nonrecombinant at the subtype level (see Materials and Methods). However, intrasubtype recombination is much more difficult to detect, limiting our ability to define a completely recombination-free data set. With recombination, different sites may have different phylogenetic histories. Hence assuming one phylogeny may lead to false detection of sites under selection because the phylogeny is incorrect for those sites. To investigate the impact of assuming an incorrect phylogeny, we analyzed these data assuming a star tree, which should be wrong for all sites. Results obtained from the star tree were highly similar to those obtained using the estimated gene tree. Furthermore, one expects false positives due to recombination to be clustered along the primary sequence, as a recombination event is expected to affect a set of contiguous bases in the primary sequence. However, in our analysis, positive selection sites were dispersed in the primary sequence and clustered on the tertiary structure. Those results suggest that our results concerning positive selection may not be seriously affected by possible recombination events.

Amino Acid Substitution Patterns at Positive Selection Sites

There are potentially contrasting views regarding amino acid substitution pattern and positive selection. One suggests that positive selection promotes change among residues with large differences in physiochemical properties (Hughes et al. 1990; McClellan and McCracken 2001). The other suggests that positive selection must operate within certain structural constraints and amino acid substitution should be physiochemically constrained (Haydon et al. 1998, 2001). Our findings suggest that in HIV, substitution patterns promoted by positive selection are related to tertiary structure. Positive selection appears to promote physiochemically diverse substitutions at externally located positive selection sites, whereas less radical amino acid substitutions appear to be favored at buried positive selection sites. Rather than competing hypotheses, these two views appear to reflect two different aspects of adaptive evolution in HIV.

The tertiary structure of a protein could have a profound effect on all amino acid substitutions, not just those at positive selection sites. It is expected that residues in regions such as the core of the protein should be subjected to strong functional constraint, with conserved patterns of amino acid evolution. This notion was first linked to HIV-1 sequence evolution when Yamaguchi-Kabata and Gojobori (2000) noted that endodomains of the most variable HIV-1 protein, gp120, are more conserved than the ectodomains. We expand on this by noting that buried residues were subjected to more intense functional constraints relative to isoelectric point. However, for polarity, volume, and hydropathy, both exposed and buried variable sites were subjected to similar levels of selective constraints. Positive selection sites exhibited much less constraint on physiochemical properties than variable sites. Moreover, acceptability was always much lower at buried positive selection sites than at exposed positive selection sites. Our results suggest that even positive selection must operate within the functional constraints associated with internal residues. Interestingly, this is not unique to viruses, as a study of *fimA* from *Escherichia coli* yielded similar conclusions (Peek et al. 2001).

It is important to note that the majority of positive selection sites was located in gp120, and this subunit undergoes many conformational changes that are not stable enough to undergo structural assessment. Hence, our knowledge of what are the "constant" core residues is limited. However, regardless of the structural information, our findings support the notion that there are at least two classes of positive selection sites: physiochemically radical and conser-

vative. Moreover, the fact that these two classes of sites are related to the inferred tertiary structure is consistent with the notion that the available 3D structure, although problematic, provides a reasonable model of the true 3D structure.

Diversifying Selection, Antigenic Variation, and Epitope Evolution

Adaptive evolution at the DNA level could be either directional or diversifying. Directional positive selection promotes a specific type of substitution that is selectively advantageous, leading to an increase in a certain phenotype, whereas diversifying selection maximizes the variation of a population. Directional positive selection is less easily detected using the $\omega > 1$ threshold. Hence it is more likely that the sites inferred in our analysis reflect the impact of diversifying selection rather than directional positive selection. This notion is supported by the observation that the pattern of amino acid substitution is physiochemically most diverse at external positive selection sites (Hughes et al. 1990).

As the immune response is a mixture of antibody and CTL response, either antibody or CTL epitopes might experience more intense selective pressure compared to the rest of the protein. If evolution at these epitopes strongly favors immune escape, they are expected to be evolving by positive selection for escape mutants. This notion was confirmed in experimental longitudinal studies of CTL escape, where the frequency of such escape mutations increased over time (Evans et al. 1999; Allen et al. 2000). Although we found 44 positive selection sites at known CTL epitopes, this number was slightly less than expected if we had sampled sites at random from the genome. Experimental longitudinal studies also indicate an accumulation over time of escape mutations for neutralizing antibodies (Zhang et al. 1999; Beaumont et al. 2001). Again, we found a large number of positive selection sites at antibody epitopes (43), but this number was less than expected under a random sample of sites. It is important to note that the approach used in this paper detects only sites that have been subject to recurrent positive selection over long periods of time. Our results do not indicate that these epitopes are not under selection for escape mutation: rather, they indicate that selection for escape mutation at either CTL or antibody epitopes has not persisted at any one site for long periods of time.

In contrast to CTL and antibody epitopes, we found a significant excess of sites under positive selection at T-helper epitopes ($p = 0.0001$). Although involved in the maintenance of antibody and CTL response, the exact role of T helpers in viral clearance is much less well understood. It has been difficult to obtain clear evidence supporting directional T-helper escape in an experimental longitudinal study (Harcourt et al. 1998). However, Harcourt et al. (1998) have isolated viral mutants that were not recognized by T helpers, implicating involvement of T helpers in viral control. Moreover, mounting experimental evidence indicates that T-helper identification plays a critical role in controlling HIV-1 infection (Rosenberg et al. 2000; Altfeld et al. 2001). In addition to this cross-sectional study, a statistical longitudinal study of HIV-1 has identified an excess of positive selection sites at T-helper epitopes in long-term progressors (Ross and Rodrigo 2002), further emphasizing the importance of T helpers in viral clearance. Our findings indicate that selective pressure at T-helper epitopes differs in a fundamental way from that at CTL and antibody epitopes in that it is stable at some sites over long periods of time. This finding supports the notion that T-helper epitopes should play a more important role in vaccine design (Norris et al. 2001; Altfeld et al. 2001).

From experimental studies, it is evident that escape mutations at T-cell or antibody epitopes occur frequently as a consequence of selective pressure (Allen et al. 2000; Beaumont et al. 2001). As many CTL, antibody, and T-helper epitopes overlap, it is difficult to assess the major target of the immune system. This matter is further complicated in this study by the long divergence time of the three subtypes sampled, where the virus experiences selective pressure from individuals with different HLA loci. Hence, sites that were once selected may drift once the selective pressure is off. Clearly, longitudinal studies with well-defined epitopes and known HLA loci are ideal for identifying directional escape mutations (such as Harcourt et al. 1998). However, statistical cross-sectional studies have the power of identifying sites under long-term recurrent selective pressure, as sites that only briefly experience selection would not have been detected. Hence, regardless of the underlying selective mechanism, sites with high posterior probabilities of evolving under positive selection identified in this study must be highly immunogenic.

Acknowledgments. We thank Stephane Aris-Brosou and Maria Anisimova for many useful discussions and suggestions. This study was supported by a Biotechnology and Biological Sciences Research Council grant to Z.Y. W.Y. was supported by a Biotechnology and Biological Sciences Research Council studentship. We thank the two anonymous referees for their suggestions and comments.

References

Alff-Steinberger C (1969) The genetic code and error transmission. Proc Natl Acad Sci USA 64:584–591

Allen TM, O'Connor DH, Jing P, et al. (2000) Tat-specific cytotoxic T lymphocytes selected for SIV escape variants during resolution of primary viraemia. Nature 407:386–390

Altfeld M, Rosenberg ES, Shankarappa R, et al. (2001) Cellular immune responses and viral diversity in individuals treated during acute and early HIV-1 infection. J Exp Med 193:169–180

Anisimova M, Bielawski JP, Yang Z (2001) The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. Mol Biol Evol 18:1585–1592

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayesian prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958

Beaumont T, van Nuenen A, Broersen S, Blattner WA, Lukashov VV, Schuitemaker H (2001) Reversal of human immunodeficiency virus type 1 IIIB to a neutralization-resistant phenotype in an accidentally infected laboratory worker with a progressive clinical course. J Virol 75:2246–2252

Bondada S, Chelvarajan RL (1999) B lymphocytes. In: Encyclopedia of science 2001. Nature Publishing Group, London

Brander C, Goulder PJR (2000) The evolving field of HIV CTL epitope mapping: New approaches to the identification of novel epitopes. In: Korber B, Brander C, Haynes BF, Koup R, Kuiken C, Moure JP, Walker BD, Watkins DI (eds) HIV molecular immunology 2001. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, NM.

Calarota S, Jansson M, Levi M, Broliden K, Libonatti O, Wigzell H, Wahren B (1996) Immunodominant glycoprotein 41 epitope identified by seroreactivity in HIV type 1-infected individuals. AIDS Res Hum Retroviruses 12:705–713

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP (1999) Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol Biol Evol 16:372–382

Domingo E, Holland JJ (1997) RNA virus mutations and fitness for survival. Annu Rev Microbiol 51:151–178

Evans DT, O'Connor DH, Jing PC, et al. (1999) Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. Nature Med 5:1270–1276

Eyre-Walker A, Keightley PD (1999) High genomic deleterious mutation rates in hominids. Nature 397:344–347

Fares MA, Moya A, Escarmis C, Baranowski E, Domingo E, Barrio E (2001) Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol Biol Evol 18:10–21

Fomsgaard A (1999) HIV-1 DNA vaccines. Immunol Lett 65: 127–131

Fu YX (2001) Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol Biol Evol 18: 620–626

Gojobori T, Moriyama EN, Kimura M (1990) Molecular clock of viral evolution and the neutral theory. Proc Natl Acad Sci USA 87:10015–10018

Goldman N, Yang Z (1994) A codon based model of nucleotide substitution for protein coding DNA sequences. Mol Biol Evol 11:725–736

Gotch FM (1998) T lymphocytes: Cytotoxic. In: Encyclopedia of science 2001. Nature Publishing Group, London

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Harcourt GC, Garrard S, Davenport MP, Edwards A, Philips RE (1998) HIV-1 variation diminishes CD4 T lymphocyte recognition. J Exp Med 188:1785–1793

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Haydon DT, Lea S, Fry L, Knowles N, Samuel AR, Stuart D, Woolhouse MEJ (1998) Characterising sequence variation in the VP1 capsid proteins of foot and mouth disease virus (serotype 0) with respect to virion structure. J Mol Evol 46: 465–475

Haydon DT, Bastos AD, Knowles NJ, Samuel AR (2001) Evidence of positive selection in foot-and-mouth disease virus capsid genes from field isolates. Genetics 157:7–15

Hughes AL (1992) Positive selection and interallelic recombination at the Merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. Mol Biol Evol 9:381–393

Hughes AL, Ota T, Nei M (1990) Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol Biol Evol 7:515–524

Igarashi T, Kuwata T, Takehisa J, Ibuki K, Shibata R, Mukai R, Komatsu T, Adachi A, Ido E, Hayami M (1996) Genomic and biological alteration of a human immunodeficiency virus type 1 (HIV-1)-simian immunodeficiency virus strain mac chimera, with HIV-1 Env, recovered from a long-term carrier monkey. J Gen Virol 77:1649–1658

Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP (2000) High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol 74:1234–1240

Korber B, Brander C, Haynes B, Koup R, Kuiken C, Moore JP, Walker BD, Watkins DI (eds) (2000) HIV molecular immunology 2000. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, NM

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Langedijk JPM, Zwart G, Goudsmit J, Meloen RH (1995) Fine specificity of antibody recognition may predict amino acid substitution in the 3rd variable region of GP120 during HIV type-1 infection. AIDS Res Hum Retroviruses: 11:1153–1162

Leigh Brown AJ (1997) Analysis of HIV-1 *env* gene reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci USA 94:1862–1865

Lukashov VV, Goudsmit J (1997) Evolution of the human immunodeficiency virus type 1 subtype-specific V3 domain is confined to a sequence space with a fixed distance to the subtype consensus. J Virol 71:6332–6338

McClellan DA, McCracken KG (2001) Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. Mol Biol Evol 18:917–925

McLain L, Brown JL, Cheung L, Reading SA, Parry C, Jones TD, Cleveland SM, Dimmock NJ (2001) Different effects of a single amino acid substitution on three adjacent epitopes in the gp41 C-terminal tail of a neutralizing antibody escape mutant of human immunodeficiency virus type 1. Arch Virol 146:157–166

Metzgar D, Wills C (2000) Evolutionary changes in mutation rates and spectra and their influence on the adaptation of pathogens. Microbes Infect 2:1513–1522

Nicholas KB, Nicholas HB, Deerfield DW (1997) GeneDoc: Analysis and visualization of genetic variation. Version 2.5. EMBnew News 4:14

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Norris PJ, Sumaroka M, Brander C, Moffett HF, Boswell SL, Nguyen T, Sykulev Y, Walker BD, Rosenberg ES (2001) Multiple effector functions mediated by human immunodeficiency virus-specific CD4(+) T-cell clones. J Virol 75:9771–9779

Peek AS, Souza V, Eguiarte LE, Gaut BS (2001) The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (*fimA*) from *Escherichia coli*. J Mol Evol 52:193–204

Phillips RE, Harcourt GC, Price DA (2001) $CD4^+$ T cells: The great escape. Nature Med 7:777–778

Plikat U, Nieselt-Struwe K, Meyerhans A (1997) Genetic drift can dominate short-term human immunodeficiency virus type 1 *nef* quasispecies evolution *in vivo*. J Virol 71:4233–4240

Robertson DL, Anderson JP, Bradac JA, et al. (1999) HIV-1 nomenclature proposal. In: Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA, Mellors JW, Mullins JI, Sodroski J, Wolinksy S (eds) Human retroviruses and AIDS 1999. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, pp 492–505

Rosenberg ES, Altfeld M, Poon SH, Phillips MN, Wilkes BM, Eldridge RL, Robbins GK, D'Aquila RT, Goulder PJR, Walker BD (2000) Immune control of HIV-1 after early treatment of acute infection. Nature 407:523–526

Ross HA, Rodrigo AG (2002) Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J Virol 76:11715–11720

Seibert SA, Howell CY, Hughes MK, Hughes AL (1995) Natural selection on the *gag*, *pol* and *env* genes of human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 12:803–813

Siliciano RF (2001) Acquired immune deficiency syndrome (AIDS). In: Encyclopedia of life science 2001. Nature Publishing Group, London

Sharp PM (1997) In search of molecular Darwinism. Nature 385:111–112

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci USA 98:2509–2514

Swofford DL (2000) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Walker BD, Goulder PJR (2000) AIDS—Escape from the immune system. Nature 407:313–314

Walther D (1997) WEBMOL, a Java based PDB viewer. Trends Biochem Sci 22:274–275

Williams EJB, Pal C, Hurst LD (2000) The molecular evolution of signal peptides. Gene 253:313–322

Xia X (2000) Data analysis in molecular biology and evolution. Kluwer Academic, Dordrecht

Yamaguchi-Kabata Y, Gojobori T (2000) Reevaluation of amino acid variability of the HIV-1 gp120 envelop glycoprotein and prediction of new discontinuous epitopes. J Virol 74:4335–4350

Yang Z (2000) Phylogenetic analysis by maximum likelihood (PAML). Version 3. University College London, London

Yang Z (2001) Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. Pacific Symp Biocomput pp 226–237

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon substitution models for heterogeneous selection pressure at amino acid sites. Genetics 15:1600–1611

Zanotto PMde A, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics 153:1077–1089

Zhang PF, Chen X, Fu DW, Margolick JB, Quinnan GV (1999) Primary virus envelope cross-reactivity of the broadening neutralizing antibody response during early chronic human immunodeficiency virus type 1 infection. J Virol 73:5225–5230