SCIENCE CHINA
Life Sciences

• RESEARCH PAPERS •

October 2010 Vol.53 No.10: 1204–1214
doi: 10.1007/s11427-010-4066-5

Complex positive selection pressures drive the evolution of HIV-1 with different co-receptor tropisms

ZHANG ChiYu[1*], DING Na[1], CHEN KePing[1] & YANG RongGe[2*]

[1] *Institute of Life Sciences, Jiangsu University, Zhenjiang 212013, China;*
[2] *HIV Molecular Epidemiology and Virology Research Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China*

Received March 17, 2010; accepted August 1, 2010

HIV-1 co-receptor tropism is central for understanding the transmission and pathogenesis of HIV-1 infection. We performed a genome-wide comparison between the adaptive evolution of R5 and X4 variants from HIV-1 subtypes B and C. The results showed that R5 and X4 variants experienced differential evolutionary patterns and different HIV-1 genes encountered various positive selection pressures, suggesting that complex selection pressures are driving HIV-1 evolution. Compared with other hypervariable regions of Gp120, significantly more positively selected sites were detected in the V3 region of subtype B X4 variants, V2 region of subtype B R5 variants, and V1 and V4 regions of subtype C X4 variants, indicating an association of positive selection with co-receptor recognition/binding. Intriguingly, a significantly higher proportion (33.3% and 55.6%, $P<0.05$) of positively selected sites were identified in the C3 region than other conserved regions of Gp120 in all the analyzed HIV-1 variants, indicating that the C3 region might be more important to HIV-1 adaptation than previously thought. Approximately half of the positively selected sites identified in the *env* gene were identical between R5 and X4 variants. There were three common positively selected sites (96, 113 and 281) identified in Gp41 of all X4 and R5 variants from subtypes B and C. These sites might not only suggest a functional importance in viral survival and adaptation, but also imply a potential cross-immunogenicity between HIV-1 R5 and X4 variants, which has important implications for AIDS vaccine development.

HIV-1 co-receptor tropism, adaptive evolution, positively selected site, R5 and X4, Gp120, epitopes

Citation: Zhang C Y, Ding N, Chen K P, *et al*. Complex positive selection pressures drive the evolution of HIV-1 with different co-receptor tropisms. Sci China Life Sci, 2010, 53: 1204–1214, doi: 10.1007/s11427-010-4066-5

The usage of different kinds of chemokine receptors (referred to as co-receptors) in the cellular entry of HIV-1 strains confers their co-receptor tropisms [1], which can further influence viral phenotypes. HIV-1 co-receptor tropism is central for understanding the transmission and pathogenesis of HIV-1 infection [2,3]. The two major HIV-1 co-receptor tropisms are R5 and X4, using CCR5 and CXCR4 as co-receptors, respectively [4]. R5 and X4 have different viral characteristics, with R5 variants dominating the viral quasispecies early in and even throughout infection [1], whereas X4 variants classically emerge late in infection. X4 exhibits rapid replication and higher virulence in peripheral blood mononuclear cells (PBMCs) than R5 [5]. The emergence of X4 strains is usually accompanied by an accelerated decrease in CD4$^+$ T cell counts, which leads to an accelerated disease progression [6]. HIV-1 co-receptor switch from R5 to X4 occurs in approximately 50% of HIV-1 subtype B infected individuals during progression to AIDS. This co-receptor switch occurs to a lower extent in HIV-1 subtype C infected individuals [2]. Therefore, an evolutionary comparison between subtypes B and C will help to understand the mechanism of HIV-1 co-receptor

*Corresponding author (email: zhangcy1999@hotmail.com; ryang@wh.iov.cn) switch.

HIV-1 has high mutation rates and is often subject to strong selective pressures from human immune responses [7]. Positive (diversifying) selection has been widely detected in whole genomes, but especially in the *env* gene of HIV-1 group M, HIV-1 group O, HIV-2 and SIV [8–14]. Furthermore, an association between positive selection and AIDS disease progression was observed in pediatric and adult HIV-1 infections [15,16]. Although only the C2V3C3 region of the *env* gene of R5 variants was used in these studies, increasing evidence showed that positive selection was more prevalent in individuals with slow HIV-1 disease progression than those with rapid disease progression [7,17]. The reason for this was likely due, in part, to a stronger immune response in slow progressors and a destroyed immune system in rapid progressors. However, the reverse was observed in one study based on HIV-1 subtype B V3 sequences, in which syncytium-inducing (SI) variants appeared to evolve faster than the non-syncytium-inducing (NSI) variants [18], implying that SI variants were subject to stronger positive selection than NSI variants. The disagreement was likely due to the usage of only part of the *env* gene in these studies. To reveal the intricate nature of selection pressures driving the evolution of HIV-1 R5 and X4 variants, adaptive evolutionary analyses based on whole *env* genes, or even whole genomic sequences are required. Therefore, a genome-wide comparison between the adaptive evolution of R5 and X4 variants from HIV-1 subtypes B and C was performed in the present study. We found that R5 and X4 variants underwent obviously different evolutionary patterns and different HIV-1 genes were subject to various selection pressures. We found that a significantly higher proportion of positively selected sites were identified in the C3 region than in other conserved regions of Gp120 in all analyzed HIV-1 variants, indicating that the C3 region might be more important to HIV-1 adaptation than previously believed. In addition, approximately half of the positively selected sites identified in *env* genes were identical between R5 and X4 variants. These common positively selected sites might not only imply functional importance in viral survival and adaptation, but also have important implications for AIDS vaccine development.

1 Materials and methods

1.1 Sequences and alignment

All of the sequences used in this study were collected from the Los Alamos National Laboratory (LANL) HIV Sequence Database (http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html). Because the full-length HIV-1 sequences with known co-receptor tropism are very limited, we primarily downloaded all 624 subtype B and 466 subtype C full-length sequences from this database. To determine the co-receptor tropism of each full-length sequence, three robust online prediction tools WebPSSM (http://indra.mullins.microbiol.washington.edu/webpssm) [19], Geno2pheno (http://coreceptor.bioinf.mpi-inf.mpg.de/index.php) [20], and HIV-1PhenoPred (http://yjxy.ujs.edu.cn/R5-X4%20pred.rar) [21] were used based on hypervariable region 3 (V3) of Gp120. Only sequences that yielded consistent prediction results by the three tools were selected and divided into CCR5 and CXCR4 datasets. Because a very short evolutionary distance is able to reduce the discriminatory power [22], sequences with closer evolutionary distances were deleted. As a consequence, 37 R5 and 33 X4 sequences from subtype B, and 28 R5 and 13 X4 sequences from subtype C were kept. These sequences were divided into four HIV-1 sub-populations for adaptive evolutionary analyses and their GenBank accession numbers are provided in Appendix Table 1.

To compare the different HIV-1 genes that play various functions in the HIV-1 life cycle, each complete genome was divided into five genes: *gag*, reverse transcriptase (RT), integrase (IN), *gp120* and *gp41*. For X4 variants of subtype C, only 13 *env* (*gp120* and *gp41*) sequences were kept for adaptive evolutionary analyses and other gene subsets were excluded due to only a few sequences being available. The sequences of each data subset were aligned using Clustal W implemented in MEGA4 [23] and manually adjusted. The phylogenetic tree of each data subset was obtained using the maximum likelihood (ML) method (PHYML v2.4.4) [24].

1.2 Analysis of positive selection and identification of positively selected sites

Positive selection is measured by comparing the rate of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) with that of synonymous substitutions per synonymous site (d_S). The d_N/d_S ratio (ω) is traditionally used as an index to assess positive selection. The ω greater than 1 is taken as evidence of positive selection, ω equal to 1 indicates neutral selection, and ω less than 1 reflects strong negative (purifying) selection. The analyses of adaptive molecular evolution on all datasets were performed using six codon substitution models: M0 (one-ratio), M1a (nearly neutral), M2a (positive selection), M3 (discrete), M7 (beta), and M8 (beta and ω), as implemented in PAML 4.0 [25]. The details of these models were described in a previous study [26]. The likelihood ratio test (LRT) comparing three pairs of nested codon evolutionary models (M0 vs. M3, M1a vs. M2a, and M7 vs. M8) was used to test against the null hypothesis of no positive selection [27]. The null hypotheses of three pairs of nested models were rejected and positive selection was inferred when the LRT statistic was significant for a χ^2 distribution with the degrees of freedom equivalent to the difference in the number of parameters between nested models. Then the datasets were subjected to the identification of positively selected sites using three

Table 1 Phylogenetic analysis by ML estimation for the *gp120* gene of HIV-1 subtype B sequences with R5 and X4 tropisms[a)]

Tropism	Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites[b)]
R5	M0	−13809.11	ω=0.52	1998.50	None
	M3	−12809.86	p_0=0.5982, p_1=0.2658 (p_2=0.1360), ω_0=0.07, ω_1=0.85, ω_2=2.75	*P*=0.0000	Not shown[c)]
	M1a	−12963.07	p_0=0.6514 (p_1=0.3486)	302.66	Not allowed[d)]
	M2a	−12811.74	p_0=0.6169, p_1=0.2620 (p_2=0.1211), ω_2=2.99	*P*=0.0000	**19T 31T 85V 87V 164S** 169V **178K** 183P **195S 200V** 219A **232T 283T 290T 293E 308R** 333I **336A 337K 340N 343K 344Q 347S 360I** 362K **363Q 440S 442Q** 444R
	M7	−12947.00	*p*=0.1795, *q*=0.3087	306.48	Not allowed
	M8	−12793.76	p_0=0.8659 (p_1=0.1342), *p*=0.2272, *q*=0.5413, ω=2.66	*P*=0.0000	4K 10L 13W **19T 31T 85V 87V 164S 169V** 175F **178K 183P 195S 200V** 219A **232T 283T 290T** 291S **293E 308R 333I 336A 337K 340N 343K 344Q** 346A **347S 360I 362K 363Q** 389Q **440S 442Q 444R**
X4	M0	−11488.66	ω=0.67	1420.24	None
	M3	−10778.54	p_0=0.5291, p_1=0.3181 (p_2=0.1527), ω_0=0.06, ω_1=0.89, ω_2=3.11	*P*=0.0000	Not shown
	M1a	−10925.70	p_0=0.6162 (p_1=0.3838)	294.34	Not allowed
	M2a	−10778.53	p_0=0.5481, p_1=0.3272 (p_2=0.1247), ω_2=3.43	*P*=0.0000	**10L 12R** 19T 33K **49T** 85V **87V 161I 169V 170Q** 200V **232T 275V 300N 301N 302N 303T 306R 308R** 317F 328Q 333I **337K 343K 347S 354G 363Q 389Q 440S 442Q 467I**
	M7	−10922.28	*p*=0.1662, *q*=0.2439	295.74	Not allowed
	M8	−10774.41	p_0=0.8487 (p_1=0.1513), *p*=0.2050, *q*=0.3892, ω_2=2.97	*P*=0.0000	**10L 12R 19T 33K 49T 85V 87V 161I 169V 170Q** 192K **200V** 208V **232T 275V** 291S **300N 301N 302N 303T 306R 308R** 316A **317F** 318V **328Q 333I** 335R **337K** 340N **343K 347S 354G 363Q 389Q 440S 442Q 467I**

a) ln*L*, log-likelihood value. 2Δl, the likelihood ratio test statistics (2 delta lambda statistics). The *p* values represent a level of significance with a χ^2 distribution and degrees of freedom=4 (M0 vs. M3) or 2 (M1a vs. M2a and M7 vs. M8). b) Positively selected sites were identified with posterior probability *P*⩾95%; in boldface, *P*⩾99%. c) Because M3 can overestimate the number of positively selected sites, the positively selected sites identified by M3 are not shown. d) The null models do not allow for a site with ω>1.

positive selection models (M2a, M3 and M8). The sites that were detected with high posterior probabilities (*P*>0.95) within the class with ω greater than 1 by selection models were admitted as positively selected [26].

Because of the overestimate of the number of actual positively selected sites [27,28], the results under model M3 were not used to identify positively selected sites. To reduce or avoid possible false-positive results, positively selected sites identified simultaneously by models M2a and M8 in the Codeml program were defined as positively selected [28]. To confirm the results obtained using the codon substitution models in PAML, similar analyses were also performed using online DataMonkey package [29].

2 Results

2.1 Positive selection on three major genes *gag*, *pol* and *env*

HIV-1 R5 and X4 variants exhibit different phenotypic characteristics in cellular tropism and replication ability [5]. Two viral proteins Gp120 and RT are closely associated with HIV-1 phenotypes since the former determines cellular tropism by specifically recognizing co-receptors and the latter determines viral replication ability. Here, *gp120*, *gp41*, RT, IN and *gag* of R5 and X4 variants from subtypes B and C were subjected to adaptive evolutionary analyses. The results of the codon-based maximum likelihood analyses are shown in Tables 1 and 2 for HIV-1 subtype B *gp120* and *gp41* genes, and in Tables 3 and 4 for subtype C *gp120* and *gp41* genes, respectively. The results of other genes are shown in Appendix Tables 2–7.

Positive selection was detected in all analyzed genes by three positive selection models (M2a, M3 and M8), except for the IN gene of R5 variants of the B subtype. Comparison between different genes showed that stronger positive selection acted on *env* and *gag* genes than on RT and IN genes. The *env* genes appeared to be under the strongest positive selection pressure [8–11]. The numbers of positively selected sites identified in different genes further supported stronger positive selection acting on *env* and *gag* genes. These results suggested that differential HIV-1 genes suffered differential positive selection patterns.

2.2 Identification of positively selected sites in *env* genes

In HIV-1 subtype B datasets, the numbers of positively selected sites in *gp120* were 29 and 31 for R5 and X4

Table 2 Phylogenetic analysis by ML estimation for the *gp41* gene of HIV-1 subtype B sequences with R5 and X4 tropisms[a)]

Tropism	Model code	ln*L*	Estimates of parameters	$2\Delta l$	Positively selected sites
R5	M0	−10580.74	$\omega=0.57$	1518.12	None
	M3	−9821.68	$p_0=0.6478$, $p_1=0.2360$ ($p_2=0.1162$), $\omega_0=0.08$, $\omega_1=0.86$, $\omega_2=3.74$	$P=0.0000$	Not shown
	M1a	−10014.68	$p_0=0.7306$ ($p_1=0.2694$)	384.62	Not allowed
	M2a	−9822.37	$p_0=0.6731$, $p_1=0.2198$ ($p_2=0.1072$), $\omega_2=3.96$	$P=0.0000$	**7L 24M 77K 96A** 107S **113N** 125N **129S 130L 133S 137E 189A 209H** 213P **235I 236R 239N 253C 281A 306A 318V 321V 325A 326C 334R 340L 343I**
	M7	−10032.47	$p=0.2010$, $q=0.3902$	418.48	Not allowed
	M8	−9823.23	$p_0=0.8789$ ($p_1=0.12113$), $p=0.3113$, $q=0.8201$, $\omega=3.55$	$P=0.0000$	**7L 24M 32Q 77K 96A 107S 113N 119E 125N 129S 130L 133S 137E 189A 209H 213P 235I 236R 239N 245I 253C 281A 304L 306A 318V 321V 325A 326C 334R 340L 343I**
X4	M0	−8388.95	$\omega=0.63$	1183.34	None
	M3	−7797.28	$p_0=0.6907$, $p_1=0.2243$ ($p_2=0.0497$), $\omega_0=0.11$, $\omega_1=1.29$, $\omega_2=5.08$	$P=0.0000$	Not shown
	M1a	−7963.59	$p_0=0.7054$ ($p_1=0.2946$)	327.88	Not allowed
	M2a	−7799.65	$p_0=0.6497$, $p_1=0.2545$ ($p_2=0.0959$), $\omega_2=4.57$	$P=0.0000$	**24M 96A 108L 109E 110Q 113N 129S 130L 133S 163N 167N 189A** 209H 210L **212T 213P 235I 236R** 239N **281A 308A 325A 326C 343I**
	M7	−7980.27	$p=0.1591$, $q=0.2925$	352.78	Not allowed
	M8	−7803.88	$p_0=0.8908$ ($p_1=0.1092$), $p=0.2182$, $q=0.4835$, $\omega=4.18$	$P=0.0000$	**24M** 32Q **96A** 101A **108L 109E 110Q 113N 129S 130L 133S 163N 167N 189A 209H 210L 212T 213P 235I 236R 239N 247D 281A 308A** 322V **325A 326C 343I**

a) For details, see Table 1.

Table 3 Phylogenetic analysis by ML estimation for the *gp120* gene of HIV-1 subtype C sequences with R5 and X4 tropisms[a)]

Tropism	Model code	ln*L*	Estimates of parameters	$2\Delta l$	Positively selected sites
R5	M0	−10763.38	$\omega=0.47$	1581.90	None
	M3	−9972.43	$p_0=0.7142$, $p_1=0.2445$ ($p_2=0.0413$), $\omega_0=0.10$, $\omega_1=1.29$, $\omega_2=5.76$	$P=0.0000$	Not shown
	M1a	−10127.16	$p_0=0.7172$ ($p_1=0.2828$)	305.00	Not allowed
	M2a	−9974.66	$p_0=0.68058$, $p_1=0.2529$ ($p_2=0.0666$), $\omega_2=4.31$	$P=0.0000$	**7Y 10L 84V 181I** 240T **281A 295N 300N 335R 344Q 346A 350R 365A 389Q 404G 405S**
	M7	−10133.41	$p=0.1878$, $q=0.4031$		Not allowed
	M8	−9977.03	$p_0=0.9223$ ($p_1=0.0777$), $p=0.2293$, $q=0.5610$, $\omega=3.72$	312.76 $P=0.0000$	**7Y 10L 84V 181I** 240T **281A 295N 300N 335R 344Q 346A 350R 365A 389Q 404G 405S**
X4	M0	−8030.39	$\omega=0.54$	993.48	None
	M3	−7533.65	$p_0=0.6346$, $p_1=0.2910$ ($p_2=0.0744$), $\omega_0=0.10$, $\omega_1=1.30$, $\omega_2=6.87$	$P=0.0000$	Not shown
	M1a	−7647.70	$p_0=0.6544$ ($p_1=0.3456$)	224.46	Not allowed
	M2a	−7535.47	$p_0=0.5929$, $p_1=0.3111$($p_2=0.0959$), $\omega_2=5.31$	$P=0.0000$	**7Y 137D** 138T 140T **141N 173Y 186N 295N 300N 320I 335R 344Q 346A** 360I **362K** 389Q **404G 405S 406N 408T 410G** 429K **461S 500K**
	M7	−7658.08	$p=0.1654$, $q=0.2725$	242.32	Not allowed
	M8	−7536.92	$p_0=0.8857$ ($p_1=0.1144$), $p=0.2070$, $q=0.3918$, $\omega=4.49$	$P=0.0000$	**7Y** 17G 85V **87V** 132T **137D 138T 140T** 141N **169V 173Y 186N 281A 295N 300N 320I 335R** 343K **344Q 346A 360I 362K 363Q 389Q 404G 405S 406N 408T 410G 429K** 460N **461S 500K**

a) For details, see Table 1.

sub-populations, respectively. In *gp120* of HIV-1 subtype C datasets, the numbers of sites were 16 and 24 for R5 and X4 sub-populations, respectively (Table 5). The number of positively selected sites identified in X4 sub-populations was greater than in the R5 sub-populations, suggesting that *gp120* gene of X4 variants was subject to a stronger positive selection pressure. An opposite pattern was observed in *gp41* genes, where 27 and 17 positively selected sites were identified in R5 sub-populations of B and C subtypes, respectively, which was greater than the 24 and 11 sites in the X4 sub-population of B and C subtypes, respectively (Table 5). This implied that *gp41* gene of R5 variants underwent a stronger positive selection than that of X4 variants. Therefore, although as a whole there was no obvious difference in

Table 4 Phylogenetic analysis by ML estimation for the *gp41* gene of HIV-1 subtype C sequences with R5 and X4 tropisms[a)]

Tropism	Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites
R5	M0	–7964.17	ω=0.53	811.42	None
	M3	–7558.46	p_0=0.6953, p_1=0.2258 (p_2=0.0789), ω_0=0.1083, ω_1=1.1385, ω_2=3.39	*P*=0.0000	Not shown
	M1a	–7629.70	p_0=0.7019 (p_1=0.2982)	141.48	Not allowed
	M2a	–7558.76	p_0=0.6775, p_1=0.2281 (p_2=0.0944), ω_2=3.12	*P*=0.0000	**96A 101A 108L 109E 113N 129D** 144K **156A** 160N **163N** 212T **239N** 256S 266I **281A 321V** 345L
	M7	–7652.02	*p*=0.2013, *q*=0.3581	174.02	Not allowed
	M8	–7565.01	p_0=0.8674 (p_1=0.1326), *p*=0.3539, *q*=0.9639, ω_2=2.66	*P*=0.0000	**96A 101A** 107S **108L 109E 113N 129D** 137E 140N **144K** 154K **156A 160N 163N** 210L **212T** 232D **239N 256S 266I 281A 321V** 324G 326C **345L**
X4	M0	–4927.55	ω=0.57	430.82	None
	M3	–4712.14	p_0=0.6766, p_1=0.2808 (p_2=0.0427), ω_0=0.09, ω_1=1.46, ω_2=5.67	*P*=0.0000	Not shown
	M1a	–4758.74	p_0=0.6516 (p_1=0.3484)	86.78	Not allowed
	M2a	–4715.35	p_0=0.6277, p_1=0.2940 (p_2=0.0783), ω_2=3.97	*P*=0.0000	**96A 108L 109E 113N 133S** 163N 210L 258H **281A 321V 330R**
	M7	–4768.49	*p*=0.0970, *q*=0.1519	98.00	Not allowed
	M8	–4719.49	p_0=0.9026 (p_1=0.0975), *p*=0.1526, *q*=0.3012, ω=3.64	*P*=0.0000	84I **96A 108L 109E 113N 133S 163N 210L** 258H **281A 321V 330R**

a) For details, see Table 1.

Table 5 Comparison of positively selected sites between different regions of Gp120 and Gp41

Subtype	Co-receptor tropism	Amino acid sites under	Gp120									Gp41		
			Variable regions[a)]						C3[a)]	C1–C5	Total	HR1+HR2	Other regions	Total
			V1	V2	V3	V4	V5	Total[b)]						
B	R5	Positive selection	0	**5**	1	0	0	6	**10**	23	29	6	21	27
		Non-positive selection	26	**34**	35	34	12	141	**43**	341	482	78	240	318
	X4	Positive selection	0	3	**8**	1	1	**13**	**6**	18	31	3	21	24
		Non-positive selection	26	36	**28**	33	11	**134**	**47**	346	480	81	240	321
C	R5	Positive selection	0	1	1	3	0	5	**5**	11	16	2	15	17
		Non-positive selection	26	38	35	31	12	143	**48**	353	485	82	246	328
	X4	Positive selection	4	2	2	6	1	**15**	**5**	9	24	1	10	11
		Non-positive selection	22	37	34	28	11	**132**	**48**	355	487	83	251	334

a) The numbers in boldface represent the statistically significant differences (*P*<0.05) between one variable region and all other variable regions or between variable (V1–V5) and conserved regions (C1–C5) using Fisher's exact test. b) The numbers in boldface represent the statistically significant differences (*P*<0.05) between C3 and all other conserved regions (C1, C2, C4, and C5) using Fisher's exact test.

the number of positively selected sites on *env* genes between R5 and X4 sub-populations, the *gp120* and *gp41* genes underwent different evolutionary pathways in R5 and X4 variants.

When taking subtypes into account, the number of positively selected sites in the *env* genes of subtype B was greater than in subtype C. For the subtype B *gp120* genes, 29 and 31 sites were identified in R5 and X4 sub-populations, respectively. There were 16 and 24 sites in the R5 and X4 sub-populations of subtype C, respectively. For the *gp41* genes, 27 and 24 sites were identified in R5 and X4 sub-populations of subtype B, respectively, which was greater than the 17 and 11 sites in R5 and X4 sub-populations of subtype C, respectively. These results suggest that *env* genes of subtype B underwent stronger positive selection than that of subtype C.

2.3 Identification of positively selected sites in *gag* and *pol* genes

The HIV-1 *gag* gene encodes four structural proteins. The *gag* products are crucial targets, recognized by the human immune system. In *gag* genes, 9 and 10 positively selected sites were detected in R5 sub-populations of subtypes B and C, respectively, obviously greater than the six sites in the X4 sub-population of subtype B (Appendix Tables 2 and 3), however, four of these sites (91, 138, 280 and 374) were all detected in both the R5 and X4 sub-populations of subtype B. Two (91 and 138) of the four common sites were also identified in the R5 sub-population of subtype C, possibly implying importance of HIV-1 adaptation.

HIV-1 RT and IN are key enzymes in the HIV life cycle. A comparison of the selection pressures for the two genes in the R5 and X4 sub-populations could help to distinguish the difference in replication rates between R5 and X4 variants. In HIV-1 subtype B datasets (Appendix Table 4), two (162 and 376) sites in the R5 sub-population and one (211) site in the X4 sub-population were detected under positive selection in RT genes. Two sites (118 and 123) were detected under positive selection for the IN gene in the X4 sub-population, whereas no positive selection was detected in the R5 sub-population (Appendix Table 6). In the R5 sub-population of subtype C dataset, three sites (123, 344, and 377) in RT and three sites (50, 72, and 125) in IN were identified under positive selection (Appendix Tables 5 and 7).

2.4 Comparison of the locations of positively selected sites in HIV-1 Env

Gp120 is the HIV-1 surface glycoprotein that not only determines viral tropism, but also is the most important target for the host immune response. Gp120 contains five conserved (C1–C5) and five hypervariable regions (V1–V5). Comparing the location of positively selected sites in both conserved and hypervariable regions showed that significantly more positively selected sites occurred in hypervariable regions in the X4 sub-populations of subtypes B (41.9%, P=0.041) and C (62.5%, P=0.0002) than in conserved regions relative to the proportion (28.8%) of hypervariable regions in whole Gp120 (Table 5). This result implied that Gp120 hypervariable regions of X4 variants were subject to stronger positive selection pressures. We further analyzed the distribution of positively selected sites in five hypervariable regions. In X4 sub-population of subtype B, significantly more sites (61.5%, P=0.0011) appeared in the V3 region that was a critical determinant of co-receptor tropism and the main epitope for eliciting neutralizing antibody [30,31] compared with other hypervariable regions relative to the proportion (24.5%) of V3 in whole hypervariable regions. This implied a stronger positive selection pressure on V3, consistent with previous observations [7,18]. In X4 sub-population of subtype C, however, higher proportions (66.7%, P=0.0316) of positively selected sites were located in the V1 and V4 regions, both of which account for 34% of hypervariable regions (Table 5). Additionally, in R5 sub-population of subtype B, significantly higher proportions (83.3%, P=0.0013) of positively selected sites were located in the V2 region that accounts for only 26.5% of the hypervariable region (Table 5).

When taking conserved regions into account, all four sub-populations exhibited significantly higher proportions of positively selected sites in the C3 region than in other conserved regions (R5 sub-population of subtype B: 43.5%, P<0.0001; X4 sub-population of subtype B: 33.3%, P= 0.0206; R5 sub-population of subtype C: 45.5%, P=0.0032; X4 sub-population of subtype C: 55.6%, P=0.0004; Table 5). These results suggested that C3 might be more important in Gp120 evolution than previously thought. With regard to Gp41, we found that relatively few positively selected sites (9.1%–22.2%) occurred in the two heptad repeat (HR) regions when compared with the proportion (24.3%) of HR1 and HR2 in whole Gp41 (Table 5 and Figure 1).

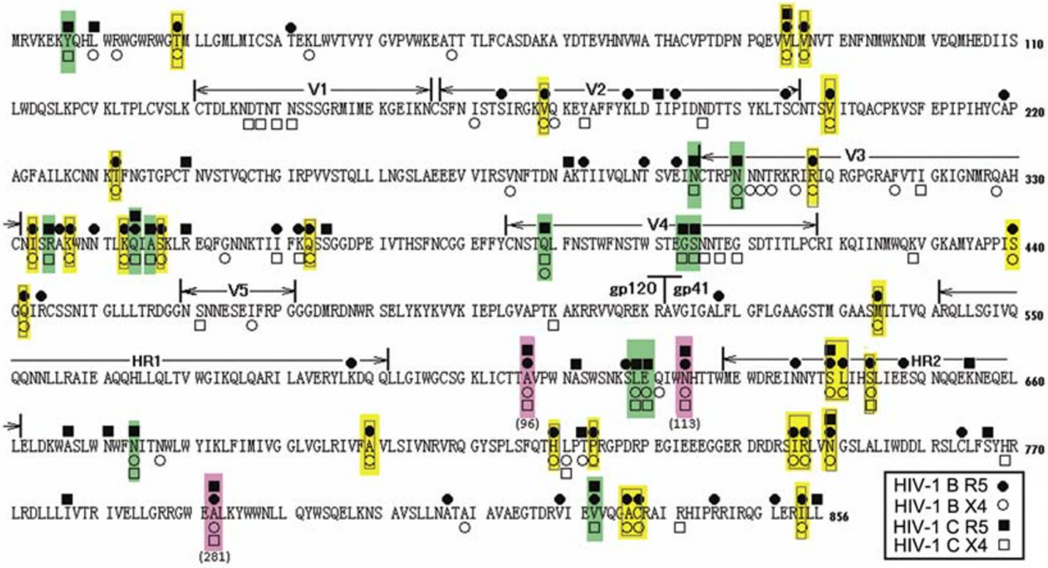
MRVKEKYQHL WRWGWRWGTM LLGMLMICSA TEKLWVTVYY GVPVWKEATT TLFCASDAKA YDTEVHNVWA THACVPTDPN PQEVVLVNVT ENFNMWKNDM VEQMHEDIIS 110
V1
V2
LWDQSLKPCV KLTPLCVSLK CTDLKNDTNT NSSSGRMIME KGEIKNCSFN ISTSIRGKVQ KEYAFFYKLD IIPIDNDTTS YKLTSCNTSV ITQACPKVSF EPIPIHYCAP 220
V3
AGFAILKCNN KTFNGTGPCT NVSTVQCTHG IRPVVSTQLL LNGSLAEEEV VIRSVNFTDN AKTIIVQLNT SVEINCTRPN NNTRKRIRIQ RGPGRAFVTI GKIGNMRQAH 330
V4
CNISRAKWNN TLKQIASKLR EQFGNNKTII FKQSSGGDPE IVTHSFNCGG EFFYCNSTQL FNSTWFNSTW STEGSNNTEG SDTITLPCRI KQIINMWQKV GKAMYAPPIS 440
V5
gp120
gp41
GQIRCSSNIT GLLLTRDGGN SNNESEIFRP GGGDMRDNWR SELYKYKVVK IEPLGVAPTK AKRRVVQREK RAVGIGALFL GFLGAAGSTM GAASMTLTVQ ARQLLSGIVQ 550
HR1
HR2
QQNNLLRAIE AQQHLLQLTV WGIKQLQARI LAVERYLKDQ QLLGIWGCSG KLICTTAVPW NASWSNKSLE QIWNHTTWME WDREINNYTS LIHSLIEESQ NQQEKNEQEL 660
(96)
(113)
LELDKWASLW NWFNITNWLW YIKLFIMIVG GLVGLRIVFA VLSIVNRVRQ GYSPLSFQTH LPTPRGPDRP EGIEEEGGER DRDRSIRLVN GSLALIWDDL RSLCLFSYHR 770
LRDLLLIVTR IVELLGRRGW EALKYWWNLL QYWSQELKNS AVSLLNATAI AVAEGTDRVI EVVQGACRAI RHIPRRIRQG LERILL 856
(281)
HIV-1 B R5 ●
HIV-1 B X4 ○
HIV-1 C R5 ■
HIV-1 C X4 □

Figure 1 Mapping of positively selected sites across *env* for different HIV-1 subtypes with different co-receptor tropisms. Positively selected sites are detected with high posterior probabilities greater than 95% within the class with ω greater than 1 by at least two of three selection models (M2, M3 and M8). The positions of the five variable regions (V1–V5) in Gp120 and the two HR regions in Gp41 are indicated. Positively selected sites identified in various HIV-1 sub-populations with different subtypes and different co-receptor tropisms are highlighted by different symbols. The solid and blank circles represent R5 and X4 variants of subtype B, respectively. The shaded and unshaded squares represent R5 and X4 variants of subtype C, respectively. The positively selected sites shared between R5 and X4 variants are highlighted with yellow and green for subtypes B and C, respectively. The positively selected sites shared by all four sub-populations are highlighted in pink.

We observed that approximately half of the positively selected sites identified in the *env* genes were the same between R5 and X4 sub-populations regardless of whether they were subtypes B or C. As an example, in subtype B, 30 of 56 positively selected sites identified in the R5 sub-population were also detected in X4. In subtype C, 16 positively selected sites were common between the R5 and X4 sub-populations (Table 5 and Figure 1). Of additional note were three positively selected sites (96, 113 and 281) in *gp41* genes that were common among all four sub-populations (Figure 1).

3 Discussion

By comparing selection pressures acting on several key genes of HIV-1 from subtypes B and C and from R5 and X4 sub-populations, we found that *env* (*gp120* and *gp41*) and *gag* genes underwent higher selection pressures than other genes. These results suggested that certain HIV-1 genes were subject to different selection pressures [8]. Comparison of positively selected sites identified in *env* of R5 and X4 sub-populations showed that both variants experienced obviously different evolutionary patterns. For the *gp120* genes, more positively selected sites were identified in the X4 than in the R5 sub-population (Table 5), similar to previous observations of highly positive selection in the V3 region of SI compared with NSI variants [18]. However, this pattern was reversed for Gp41, which underwent stronger positive selection in R5 compared with X4 variants (Table 5). These results suggested R5 and X4 had different evolutionary patterns.

At least two kinds of potentially positive selection pressures, the host immune system and the target cell range, can drive HIV-1 evolution in treatment-naïve HIV-1-infected individuals [2]. The host immune responses including humoral and cellular immune responses were generally assumed to be the most important evolutionary pressures for adaptive evolution observed in the HIV-1 genome [7]. The *gag* gene encodes viral structural proteins that are less involved in viral replication and co-receptor recognition/binding. A total of 19 positively selected sites were identified in *gag* genes of three analyzed HIV-1 sub-populations (R5 and X4 sub-populations from subtype B and R5 sub-population from subtype C) (Appendix Tables 2 and 3). All these sites were found to be associated with at least one of three kinds of epitopes (Ab, CTL and T-helper). In particular, 83.3% and 66.7% of these sites were associated with CTL and T-helper epitopes, respectively. These results indicated that the positive selection pressures on *gag* genes were primarily imposed by the host immune response. Among these positively selected sites, two sites at 91 and 138 were detected in all three sub-populations, possibly implying an additional importance for HIV-1 adaptation. A previous study demonstrated that a residue change in site 30 of Gag was able to confer a species specific replication advantage in HIV or SIV to adapt to their hosts [32]. The potential roles of sites 91 and 138 in HIV-1 adaptation need to be assessed by site-directed mutagenesis analyses.

RT is a key enzyme responsible for HIV-1 replication. HIV-1 X4 viruses usually exhibit higher replication rates than R5 viruses [2]. A total of six positively selected sites were identified in RT of three sub-populations. All these sites were located in the DNA polymerase domain of RT [33], and the sites identified in R5 and X4 variants were different, implying that these positively selected sites might be associated with specific replication characteristics of R5 or X4 variants. All these sites were associated with CTL-specific epitopes. Therefore, cellular immune responses were also likely to drive the evolution of RT.

The envelope (Env) glycoprotein of HIV-1 is exposed on the surface of the virus particle and HIV-1 infected cells, playing an important role in viral survival. It not only determines the co-receptor tropism of HIV-1, but is also the major determinant of immunogenicity for humoral and cellular immune responses. Moreover, the highest mutation rate of the *env* gene in HIV-1 genome confers a potential ability to escape host immune responses [34,35]. Therefore, the adaptive evolution of *env* genes was thought to be complex and might involve multiple selection factors such as cell source, host immune responses and the virus itself [2,8].

The hypervariable rather than conserved regions were demonstrated to determine HIV-1 co-receptor tropism. The V3 region plays the most important role in the determination of HIV-1 co-receptor tropism [30,36,37] and other hypervariable regions, such as V1V2 and V4 regions, affect the co-receptor usage [38–44]. Comparing the location of positively selected sites in Gp120 showed that in X4 sub-populations significantly more sites were located in the hypervariable regions (P<0.05). However, a similar pattern was not observed in R5 sub-populations. Further analyses showed that significantly more positively selected sites were in the V3 region (61.5%, P=0.0011) of subtype B X4 variants, the V2 region (83.3%, P=0.0013) of subtype B R5 variants, and V1 and V4 regions (66.7%, P=0.0316) of subtype C X4 variants compared with other hypervariable regions of Gp120 (Table 5). These results distinctly indicated that positive selection acting on the Gp120 hypervariable regions was closely associated with the function of co-receptor recognition/binding. The V1V2, V4 and V5 regions have been demonstrated to contribute to autologous neutralization [45]. This implied that humoral immune response-imposed positive selection also contributed to the evolution of Gp120.

A higher proportion (33.3%–55.6%, P<0.05) of positively selected sites were located in the C3 region than in other conserved regions of Gp120 in all four HIV-1 sub-populations (Table 5). This result indicated that the C3 region might be more important to the function of HIV-1 Gp120 than previously believed. Moreover, the C3 region of the subtype C virus was able to elicit early autologous neutralizing response to HIV-1 infection by forming an important structural motif together with the V4 region [45]. The results observed in the C3 region also supported that humoral immune response-imposed positive selection might play a role in the evolution of Gp120.

Like the S2 domain of SARS-CoV spike (S) protein, HIV-1 Gp41 contains two HR regions, which have been shown to be important in virus membrane fusion [46]. We found that low proportions (9.1%–22.2%) of positively selected sites occurred in two HR regions of Gp41 (Figure 1), possibly arguing against membrane fusion as a major selection factor for the evolution of HIV-1 Gp41 [47]. However, three sites (96, 113 and 281) were detected in Gp41 of all four HIV-1 sub-populations, and two of these sites (96 and 113) were located in the middle region between two HR regions. This finding suggested that the three mutual sites might play some role in the membrane fusion function of Gp41, supporting membrane fusion as a minor selection factor for the evolution of HIV-1 Gp41. Furthermore, we found that approximately half of the positively selected sites identified in *env* genes were identical between R5 and X4 variants (Figure 1). These common positively selected sites not only indicated that they were functionally important for the survival of both HIV-1 R5 and X4 variants, but also suggested that immune responses might be targeting the same viral region in both variants. These positively selected sites shared by all X4 and R5 variants might have important implications for AIDS vaccine development.

This work was supported by the National Natural Science Foundation of China (*Grant No. 30600352*), *the "Top-notch Personnel" Project of Jiangsu University, the National Basic Research Program of China* (*Grant No. 2006CB504200*), *and the Open Research Fund Program of the State Key Laboratory of Virology of China* (*Grant No. 2009008*).

1 Berger E A, Murphy P M, Farber J M. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol, 1999, 17: 657–700

2 Regoes R R, Bonhoeffer S. The HIV coreceptor switch: a population dynamical perspective. Trends Microbiol, 2005, 13: 269–277

3 Moore J P, Kitchen S G, Pugach P, *et al.* The CCR5 and CXCR4 coreceptors—central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. AIDS Res Hum Retroviruses, 2004, 20: 111–126

4 Berger E A, Doms R W, Fenyo E M, *et al.* A new classification for HIV-1. Nature, 1998, 391: 240

5 Bjorndal A, Deng H, Jansson M, *et al.* Coreceptor usage of primary human immunodeficiency virus type 1 isolates varies according to biological phenotype. J Virol, 1997, 71: 7478–7487

6 Connor R I, Sheridan K E, Ceradini D, *et al.* Change in coreceptor use correlates with disease progression in HIV-1-infected individuals. J Exp Med, 1997, 185: 621–628

7 Ross H A, Rodrigo A G. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J Virol, 2002, 76: 11715–11720

8 Choisy M, Woelk C H, Guegan J F, *et al.* Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol, 2004, 78: 1962–1970

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

10 Yang W, Bielawski J P, Yang Z. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J Mol Evol, 2003, 57: 212–221

11 Travers S A, O'Connell M J, McCormack G P, *et al.* Evidence for heterogeneous selective pressures in the evolution of the *env* gene in different human immunodeficiency virus type 1 subtypes. J Virol, 2005, 79: 1836–1841

12 Zanotto P M, Kallas E G, de Souza R F, *et al.* Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics, 1999, 153: 1077–1089

13 de Oliveira T, Salemi M, Gordon M, *et al.* Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? Genetics, 2004, 167: 1047–1058

14 Soares A E, Soares M A, Schrago C G. Positive selection on HIV accessory proteins and the analysis of molecular adaptation after interspecies transmission. J Mol Evol, 2008, 66: 598–604

15 Carvajal-Rodriguez A, Posada D, Perez-Losada M, *et al.* Disease progression and evolution of the HIV-1 *env* gene in 24 infected infants. Infect Genet Evol, 2008, 8: 110–120

16 Leal E, Janini M, Diaz R S. Selective pressures of human immunodeficiency virus type 1 (HIV-1) during pediatric infection. Infect Genet Evol, 2007, 7: 694–707

17 Williamson S. Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol, 2003, 20: 1318–1325

18 Yamaguchi Y, Gojobori T. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc Natl Acad Sci USA, 1997, 94: 1264–1269

19 Jensen M A, Li F S, van't Wout A B, *et al.* Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. J Virol, 2003, 77: 13376–13388

20 Sing T, Low A J, Beerenwinkel N, *et al.* Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. Antivir Ther, 2007, 12: 1097–1106

21 Xu S, Huang X, Xu H, *et al.* Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loop sequence using random forest. J Microbiol, 2007, 45: 441–446

22 Nozawa M, Suzuki Y, Nei M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci USA, 2009, 106: 6700–6705

23 Tamura K, Dudley J, Nei M, *et al.* MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol, 2007, 24: 1596–1599

24 Guindon S, Lethiec F, Duroux P, *et al.* PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res, 2005, 33: W557–W559

25 Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol, 2007, 24: 1586–1591

26 Yang Z, Nielsen R, Goldman N, *et al.* Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics, 2000, 155: 431–449

27 Anisimova M, Bielawski J P, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol, 2001, 18: 1585–92

28 Anisimova M, Bielawski J P, Yang Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol,

2002, 19: 950–958

29 Pond S L, Frost S D. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics, 2005, 21: 2531–2533

30 Hwang S S, Boyle T J, Lyerly H K, *et al.* Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science, 1991, 253: 71–74

31 Javaherian K, Langlois A J, LaRosa G J, *et al.* Broadly neutralizing antibodies elicited by the hypervariable neutralizing determinant of HIV-1. Science, 1990, 250: 1590–1593

32 Wain L V, Bailes E, Bibollet-Ruche F, *et al.* Adaptation of HIV-1 to its human host. Mol Biol Evol, 2007, 24: 1853–1860

33 Sarafianos S G, Das K, Tantillo C, *et al.* Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. EMBO J, 2001, 20: 1449–1461

34 Parren P W, Moore J P, Burton D R, *et al.* The neutralizing antibody response to HIV-1: viral evasion and escape from humoral immunity. AIDS, 1999, 13 Suppl A: S137–S162

35 Klenerman P, Wu Y, Phillips R. HIV: current opinion in escapology. Curr Opin Microbiol, 2002, 5: 408–413

36 Fouchier R A, Groenink M, Kootstra N A, *et al.* Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol, 1992, 66: 3183–3187

37 Xiao L, Owen S M, Goldman I, *et al.* CCR5 coreceptor usage of non-syncytium-inducing primary HIV-1 is independent of phylogenetically distinct global HIV-1 isolates: delineation of consensus motif in the V3 domain that predicts CCR-5 usage. Virology, 1998, 240: 83–92

38 Boyd M T, Simpson G R, Cann A J, *et al.* A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. J Virol, 1993, 67: 3649–3652

39 Smyth R J, Yi Y, Singh A, *et al.* Determinants of entry cofactor utilization and tropism in a dualtropic human immunodeficiency virus type 1 primary isolate. J Virol, 1998, 72: 4478–4484

40 Groenink M, Fouchier R A, Broersen S, *et al.* Relation of phenotype evolution of HIV-1 to envelope V2 configuration. Science, 1993, 260: 1513–1516

41 Pastore C, Nedellec R, Ramos A, *et al.* Human immunodeficiency virus type 1 coreceptor switching: V1/V2 gain-of-fitness mutations compensate for V3 loss-of-fitness mutations. J Virol, 2006, 80: 750–758

42 Sullivan N, Thali M, Furman C, *et al.* Effect of amino acid changes in the V1/V2 region of the human immunodeficiency virus type 1 gp120 glycoprotein on subunit association, syncytium formation, and recognition by a neutralizing antibody. J Virol, 1993, 67: 3674–3679

43 Ghaffari G, Tuttle D L, Briggs D, *et al.* Complex determinants in human immunodeficiency virus type 1 envelope gp120 mediate CXCR4-dependent infection of macrophages. J Virol, 2005, 79: 13250–13261

44 Cho M W, Lee M K, Carney M C, *et al.* Identification of determinants on a dualtropic human immunodeficiency virus type 1 envelope glycoprotein that confer usage of CXCR4. J Virol, 1998, 72: 2509–2515

45 Bunnik E M, Pisas L, van Nuenen A C, *et al.* Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. J Virol, 2008, 82: 7932–7941

46 Chambers P, Pringle C R, Easton A J. Heptad repeat sequences are located adjacent to hydrophobic regions in several types of virus fusion glycoproteins. J Gen Virol, 1990, 71: 3075–3080

47 Zhang C Y, Wei J F, He S H. Adaptive evolution of the spike gene of SARS coronavirus: changes in positively selected sites in different epidemic groups. BMC Microbiol, 2006, 6: 88

Appendix Table 1 List of GenBank accession numbers for HIV-1 genomic sequences analyzed in the text

Subtype	Co-receptor tropism	Sequence number	GenBank accession numbers
HIV-1 B	R5	37	AB286956, AB253432, AF003888, AF042101, AF224507, AY173952, AY037282, EU576191, AY586543, AY713412, AY835748, AY713411, EU574998, AY561236, AY970946, AY839827, AY857022, D10112, DQ854714, DQ837381, DQ886031, FJ469746, EF363124, EF637046, EF514699, EF637049, EU786675, FJ460501, FJ469770, FJ495937, FJ469703, FJ469731, M93258, U23487, U63632, FJ496085, FJ496150, K02007
	X4	33	EU281726, K02013, AB287363, AB287365, AF049494, AF086817, AF146728, AY037268, AY736821, AY173956, AY180905, AY560108, AY835767, AY835768, D86068, DQ127534, DQ396398, DQ823363, EF514712, FJ469686, FJ469692, FJ469736, FJ469737, FJ469739, FJ469748, FJ469753, FJ469759, L02317, L31963, M17449, M26727, FJ496166
HIV-1 C	R5	28	AB254141, DQ369991, AY734550, DQ275642, EU786673, AY878054, AF286227,AY945738, FJ496185, U46016, AY713414, AF110978, AF110981, AF286224, AF286231, AY444800, AY463217, AY563170, AF286233, AF286234, AF290027, AY043176, AF391231, AY118165, AY253303, AY228556, AY228557, AY253321
	X4	13	FJ846637, FJ846642, AY878064, DQ093600, AY529666, FJ846647, AY529678, DQ382362, DQ382372, DQ382378, AY529677, AY529673, AF411966

Appendix Table 2 Phylogenetic analysis by ML estimation for *gag* gene of HIV-1 subtype B variants with different coreceptor tropisms[a)]

Co-receptor tropism	Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites
R5	M0(one-ratio)	−10957.60	ω=0.24	1329.12 P=0.0000	None
	M3(discrete)	−10293.04	p_0=0.7101, p_1=0.2228 (p_2=0.0671), ω_0=0.03, ω_1=0.56, ω_2=2.58		**12E 30K 67S 84T 91R 138I 146A 215V 223I** 252N **280T 374A 375T 389I 425D 441Y 473P 478P 483L** 487T
	M1a(Nearly neutral)	−10382.84	p_0=0.7869 (p_1=0.2131)	145.4 P=0.0000	Not allowed
	M2a(Positive selection)	−10310.14	p_0=0.7719, p_1=0.1899 (p_2=0.0381), ω_2=3.82		**84T 91R 138I 223I 280T 374A 389I 473P 483L**
	M7(beta)	−10364.21	p=0.1217, q=0.4289	157.08 P=0.0000	Not allowed
	M8(beta&ω)	−10285.67	p_0=0.9509 (p_1=0.0491), p=0.1539, q=0.6937, ω=3.06		67S **84T 91R 138I** 146A **223I 280T 374A** 375T **389I** 425D **473P** 478P **483L**
X4	M0(one-ratio)	−8991.79	ω=0.25	850.6 P=0.0000	None
	M3(discrete)	−8566.49	p_0=0.6919, p_1=0.2246 (p_2=0.08350), ω_0=0.02, ω_1=0.49, ω_2=2.06		12E 15R **30K 67S 79Y 84T 91R 93E 102D** 119D **125S** 124N **138I** 219H **223I 252N 280T 374A** 389I 403R 418K **441Y 473P** 478P **473L** 477T
	M1a(Nearly neutral)	−8607.63	p_0=0.7827 (p_1=0.2173)	51.28 P=0.0000	Not allowed
	M2a(Positive selection)	−8581.99	p_0=0.7777, p_1=0.0448 (p_2=0.1775), ω_2=2.86		67S 91R 138I **280T 374A** 478P
	M7(beta)	−8603.18	p=0.1197, q=0.4075	73.92 P=0.0000	Not allowed
	M8(beta&ω)	−8566.22	p_0=0.9279 (p_1=0.0720), p=0.1844, q=0.9994, ω=2.20		67S 84T 91R 125S **138I** 223I **280T 374A** 478P 473L

a) Positively selected sites were identified with posterior probability P≥95%; in boldface, P≥99%.

Appendix Table 3 Phylogenetic analysis by ML estimation for *gag* gene of HIV-1 subtype C variants with R5 tropism[a)]

Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites
M0 (one-ratio)	−9276.18	ω=0.27	1131.26 P=0.0000	None
M3 (discrete)	−8710.55	p_0=0.7768, p_1=0.1893 (p_2=0.0338), ω_0=0.05, ω_1=0.93, ω_2=3.99		**28K** 54S **79Y 90Q 91R 138I 146A 241S 371T** 440S **458P**
M1a (Nearly neutral)	−8778.63	p_0=0.7926 (p_1=0.2074)	135.54	Not allowed
M2a (Positive selection)	−8710.86	p_0=0.7825, p_1=0.1852(p_2=0.0324), ω_2=4.14	P=0.0000	**28K 79Y 90Q 91R 138I 146A 241S 371T** 440S **458P**
M7 (beta)	−8782.96	p=0.1194, q=0.4021	147.9	Not allowed
M8 (beta&ω)	−8709.01	p_0=0.9639 (p_1=0.0360), p=0.1426, q=0.5506, ω=3.71	P=0.0000	**28K** 54S **79Y 90Q 91R 138I 146A** 223V **241S 371T 440S 458P**

a) Positively selected sites were identified with posterior probability P≥95%; in boldface, P≥99%.

Appendix Table 4 Phylogenetic analysis by ML estimation for RT gene of HIV-1 subtype B variants with different coreceptor tropisms[a)]

Co-receptor tropism	Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites
R5	M0(one-ratio)	−7803.46	ω=0.14	602.52 P=0.0000	None
	M3(discrete)	−7502.20	p_0=0.7218, p_1=0.2236 (p_2=0.0546), ω_0=0.02, ω_1=0.27, ω_2=1.46		Not shown
	M1a(Nearly neutral)	−7534.98	p_0=0.8974 (p_1=0.1026)	22.66 P=0.0000	Not allowed
	M2a(Positive selection)	−7523.65	p_0=0.8971, p_1=0.0897 (p_2=0.0132), ω_2=3.05		162S **376T**
	M7(beta)	−7525.67	p=0.1371, q=0.7393	49.5 P=0.0000	Not allowed
	M8(beta&ω)	−7500.92	p_0=0.9604 (p_1=0.0396), p=0.2362, q=2.1903, ω=1.75		**162S** 211R 245V 297E 332Q 360A **376T** 386T
X4	M0(one-ratio)	−6749.04	ω=0.17	488.7 P=0.0000	None
	M3(discrete)	−6504.69	p_0=0.6640, p_1=0.2606 (p_2=0.0753), ω_0=0.01, ω_1=0.23, ω_2=1.51		Not shown
	M1a(Nearly neutral)	−6522.31	p_0=0.8821 (p_1=0.1179)	18.96 P=0.0001	Not allowed
	M2a(Positive selection)	−6512.83	p_0=0.8834, p_1=0.0940 (p_2=0.0226), ω_2=2.65		211R
	M7(beta)	−6526.67	p=0.1286, q=0.6409	46.92 P=0.0000	Not allowed
	M8(beta&ω)	−6503.21	p_0=0.9479 (p_1=0.0521), p=0.2461, q=2.1554, ω_2=1.86		207Q **211R** 215T 245V **357M** 376T

a) Positively selected sites were identified with posterior probability P≥95%; in boldface, P≥99%.

Appendix Table 5 Phylogenetic analysis by ML estimation for RT gene of HIV-1 subtype C variants with R5 tropism[a)]

Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites
M0(one-ratio)	−7159.79	ω=0.15	573.44 P=0.0000	None
M3(discrete)	−6873.07	p_0=0.8749, p_1=0.1176 (p_2=0.0075), ω_0=0.05, ω_1=0.90, ω_2=5.24		**123D 334Q 377T**
M1a(Nearly neutral)	−6900.15	p_0=0.8841 (p_1=0.1159)	53.04 P=0.0000	Not allowed
M2a(Positive selection)	−6873.63	p_0=0.8817, p_1=0.1109 (p_2=0.0073), ω_2=5.41		**123D 334Q 377T**
M7(beta)	−6909.81	p=0.1378, q=0.6736	65.26 P=0.0000	Not allowed
M8(beta&ω)	−6877.18	p_0=0.9919 (p_1=0.0081), p=0.1647, q=0.9165, ω=4.91		**123D 334Q 377T**

a) Positively selected sites were identified with posterior probability $P\geqslant95\%$; in boldface, $P\geqslant99\%$.

Appendix Table 6 Phylogenetic analysis by ML estimation for IN gene of HIV-1 subtype B variants with different coreceptor tropisms[a)]

Co-receptor tropism	Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites
R5	M0(one-ratio)	−4567.46	ω=0.12	304.34 P=0.0000	None
	M3(discrete)	−4415.29	p_0=0.82278, p_1=0.1239 (p_2=0.0533), ω_0=0.03, ω_1=0.34, ω_2=1.62		**10E 16S** 27L **38S** 44L **71L 100L** 111T **121T** 124T 199I
	M1a(Nearly neutral)	−4427.40	p_0=0.8994 (p_1=0.1006)	6.9 P=0.0317	Not allowed
	M2a(Positive selection)	−4423.95	p_0=0.9011, p_1=0.0816 (p_2=0.0173), ω_2=2.64		None
	M7(beta)	−4435.26	p=0.1511, q=0.8502	37.46 P=0.0000	Not allowed
	M8(beta&ω)	−4416.53	p_0=0.9444 (p_1=0.0556), p=0.3879, q=4.7314, ω=1.58		16S 71I 100L 124T
X4	M0(one-ratio)	−3994.45	ω=0.14	198.36 P=0.0000	None
	M3(discrete)	−3895.27	p_0=0.7992, p_1=0.1833 (p_2=0.0175), ω_0=0.04, ω_1=0.51, ω_2=3.19		100L 118S 123A 124T
	M1a(Nearly neutral)	−3909.48	p_0=0.8738 (p_1=0.1262)	12.02 P=0.0025	Not allowed
	M2a(Positive selection)	−3903.47	p_0=0.8744, p_1=0.1123 (p_2=0.0133), ω_2=3.95		118S 123A
	M7(beta)	−3910.05	p=0.1933, q=0.9867	26.96 P=0.0000	Not allowed
	M8(beta&ω)	−3896.57	p_0=0.9821 (p_1=0.0179), p=0.2779, q=1.8653, ω=3.13		100L **118S 123A** 124T

a) Positively selected sites were identified with posterior probability $P\geqslant95\%$; in boldface, $P\geqslant99\%$.

Appendix Table 7 Phylogenetic analysis by ML estimation for IN gene of HIV-1 subtype C variants with R5 tropism[a)]

Model code	ln*L*	Estimates of parameters	2Δl	Positively selected sites
M0(one-ratio)	−4143.50	ω=0.16	292.76 P=0.0000	None
M3(discrete)	−3997.12	p_0=0.7776, p_1=0.1910 (p_2=0.0314), ω_0=0.04, ω_1=0.44, ω_2=2.70		**11E 50M 72V 125T** 269R
M1a(Nearly neutral)	−4016.89	p_0=0.8952 (p_1=0.1048)	26.58 P=0.0000	Not allowed
M2a(Positive selection)	−4003.60	p_0=0.8940, p_1=0.0851 (p_2=0.0209), ω_2=3.32		50M 72V **125T**
M7(beta)	−4021.05	p=0.1546, q=0.7387	48.34 P=0.0000	Not allowed
M8(beta&ω)	−3996.88	p_0=0.9696 (p_1=0.0304), p=0.2734, q=1.9860, ω=2.74		**11E 50M 72V 125T**

a) Positively selected sites were identified with posterior probability $P\geqslant95\%$; in boldface, $P\geqslant99\%$.