



A Comparative Study of Adaptive Molecular Evolution in Different HIV Groups and Subtypes

M. Choisy*, C. H. Woelk§, J.-F. Guégan*, D. L. Robertson#

*CNRS / IRD, Montpellier, France
 §University of California San Diego, La Jolla, USA
 #University of Manchester, Manchester, UK

GEMI, UMR CNRS-IRD 2724
 Centre IRD, 911 Avenue Agropolis, BP 64501
 34394 Montpellier Cedex 5 France
 choisy@mpl.ird.fr
 +33 4 67 41 61 80
 +33 4 67 41 62 99

ABSTRACT [1]

Molecular adaptation, as characterized by the detection of positive selection, was quantified in a number of genes from different human immunodeficiency virus type 1 (HIV-1) group M, group O, and an HIV-2 subtype using the codon-based maximum likelihood method of Yang and coworkers [8]. The *env* gene was investigated further since it exhibited the strongest signal for positive selection compared to those of the other two major HIV genes (*gag* and *pol*). In order to investigate the pattern of adaptive evolution across *env*, the location and strength of positive selection in different HIV-1 sequence alignments was compared. The number of sites having a significant probability of being positively selected varied among these different alignment data sets, ranging from 25 in HIV-1 group M subtype A to 40 in HIV-1 group O. Strikingly, there was a significant tendency for positively selected sites to be located at the same position in different HIV-1 alignments, ranging from 10 to 16 shared sites for the group M intersubtype comparisons and from 6 to 8 for the group O to M comparisons, suggesting that all HIV-1 variants are subject to similar selective forces. As the host immune response is believed to be the dominant driving force of adaptive evolution in HIV, this result would suggest that the same sites are contributing to viral persistence in diverse HIV infections. Thus, the position of the positively selected sites were investigated in reference to the inferred locations of different epitope types (antibody, T-helper, and cytotoxic T lymphocytes) and the positions of N- and O-glycosylation sites. We found a significant tendency for positively selected sites to fall outside T-helper epitopes and for positively selected sites to be strongly associated with N-glycosylation sites.

INTRODUCTION

The development of candidate HIV vaccines demands a thorough investigation into the consistency of the selective environment – presumed primarily to be due to the host immune response – between divergent lineages. This study aims to explore whether there are differences in the location and intensity of selection among the different HIV groups and subtypes. The position of positively selected sites are further related to the position of epitopes and glycosylation sites.

DATA SETS

We used alignments of DNA sequences of the *env* gene (Figure 1A & B) for HIV-1 group M, subtype A, B, C, and D, group O and HIV-2 group A and of the *gag* and *pol* genes (Figure 1A & B) for the HIV-1 group M subtype A, B, and C, and HIV-2 subtype A (Figure 1C). All the sequences were retrieved from the Los Alamos National Laboratory HIV Sequence Database [4], except those of HIV-1 group O which were retrieved from GenBank. The positions of the antibody (Ab), cytotoxic T lymphocytes (CTL) and T helper (Th) of the HIV-1 group M subtype B reference sequence were retrieved from the Los Alamos National Laboratory Immunology Database [3].

METHODS

Detection of positive selection involves the comparison of synonymous (d_s) and nonsynonymous (d_n) substitutions between protein-coding DNA sequences. The d_n/d_s ratio, ω , is then used to measure the difference between these two rates of substitution such that an $\omega < 1$ corresponds to purifying (negative) selection, an $\omega = 1$ corresponds to neutral evolution (absence of selection), an $\omega > 1$ indicates adaptive evolution (positive selection). Estimations were carried out on alignments of sequences in the maximum likelihood (ML) framework of Yang and coworkers [8] that accounts for phylogenetic structure, biases in codon usage and the transition/transversion (T_4/T_2) rate ratio. A Bayesian approach further allows statistical significance to be attached to the assignment of each site to the conserved, neutral or positively selected classes. Positive selection was tested for in *gag*, *pol*, and *env* genes sequences alignments of HIV-1 group M, subtypes A to D, group O, and HIV-2 subtype A. The location and strength of positive selection were compared between clades and analyzed with respect to epitopes and glycosylation sites. Paired Wilcoxon ranked sum test and Monte Carlo simulations were used to determine significant results. The PAUP* package [6] was used to build ML trees needed for selection analysis and glycosylation sites were predicted by the NetNGlyc and NetOGlyc programs [2]. Models of positive selection analysis were implemented using the CODEML program of the PAML package, version 3.1 [7].

RESULTS

Positive selection was strongest in the envelope gene, particularly in the *gp120* subunit, irrespective of the clade examined (Figure 1D). In *env*, the number of positively selected sites was 22 for HIV-2 subtype A, between 30 and 35 for HIV-1 group M subtypes, and 40 for HIV-1 group O (Table 1). Moreover, these positively selected sites tend to occur at the position in different clades (Table 2). Assuming that positive selection on this gene is primarily due to immune pressure, these results suggest that the immune response tends to target the same sites on the HIV envelope glycoprotein, irrespective of the HIV clade. Areas involved in CD4 and chemokine binding are apparently not under positive selection between individuals, presumably to maintain efficient interactions with these cellular molecules. However, sites under positive selection significantly tend to include N-glycosylation sites (Table 4), which are not necessarily associated with predicted CTL and Ab sites (Table 3).

CONCLUSION

On the assumption that the immune response provides the evolutionary pressure for amino acid change at the positively selected sites, the result that positively selected sites are shared between divergent HIV-1 lineages suggests that the immune response may be targeting the same viral regions in the different groups and subtypes, thus raising the possibility of cross-subtype or group immunogenicity. Moreover the fact that positively selected sites tends to coincide with glycosylation sites and not with epitopes tends to validate the *Cytosin Shield Model* of Kwong and coworkers [8] which suggests that epitopes in regions essential for viral fitness and that are thus unable to tolerate mutation can be protected from neutralizing antibodies by carbohydrates bound to glycosylation sites. Mutations at these sites would cause the permanent rearrangement of the carbohydrates, thus creating a moving protecting shield around epitopes that are unable to tolerate mutation, as they are in functionally conserved regions (Figure 2). These results have implications in understanding HIV evasion mechanisms and potential practical applications for vaccine design.

Table 2 – H_0 : no match between positively selected sites and epitopes
 H_1 : match between positively selected sites

HIV data set	HIV-1-MA	HIV-1-MB	HIV-1-MC	HIV-1-MD	HIV-1-O
HIV-1-MB	E: 2.018				
	O: 13				
	P: 0.001				
HIV-1-MC	E: 1.990	2.015			
	O: 16	15			
	P: 0.001	0.001			
HIV-1-MD	E: 1.760	1.793	1.741		
	O: 10	14	15		
	P: 0.001	0.001	0.001		
HIV-O	E: 1.016	0.936	0.889	0.848	
	O: 7	7	8	6	
	P: 0.001	0.001	0.001	0.001	
HIV-2-A	E: 0.633	0.722	0.571	0.511	0.729
	O: 3	1	2	2	1
	P: 0.024	0.535	0.034	0.091	0.539

Monte Carlo simulations testing the association of sites of positive selection between lineages. E, expected value from a random distribution; O, observed value; and P, level of significance at which H_0 is different from H_1 . Significant results (P-values) are in bold.

Table 3 – H_0 : random repartition of positively selected sites with respect to epitopes
 H_1 : anti-match between positively selected sites and epitopes
 H_2 : match between positively selected sites and epitopes

Epitopes	H_1 vs H_0			H_2 vs H_0			
	N_{obs}	N_{exp}	P_{sig}	N_{obs}	N_{exp}	P_{sig}	
Ab	370	18	22.17	0.846	208	17	12.53
CTL	394	19	23.82	0.976	184	16	11.12
Th	499	26	30.16	0.989	79	9	4.78
Ab and CTL	567	30	30.64	0.726	71	5	4.17
Ab and Th	537	33	32.40	0.696	43	2	1.80
CTL and Th	524	27	31.67	0.999	54	8	2.02

The first 3 rows correspond to the 3 epitope types analyzed separately (Ab, antibody; CTL, cytotoxic T-cell; and Th, T-helper response), and the remaining rows refer to combinations of these epitope types analyzed together. N_{obs} , number of sites targeted by epitopes from the HIV Immunology Database [3]; N_{exp} , observed number of identified positively selected sites that fall inside the epitope regions; N_{exp} , expected number of positively selected sites in the epitope regions as calculated by Monte Carlo simulations; and P_{sig} , significance level at which H_0 differs from H_1 . N_{exp} , number of sites targeted by epitopes from the HIV Immunology Database [3]; N_{obs} , observed number of identified positively selected sites that fall outside the epitope regions; N_{exp} , expected number of positively selected sites out of the epitope regions as calculated by Monte Carlo simulations; and P_{sig} , significance level at which H_0 differs from H_1 . Significant values ($P_{sig} < 0.05$) are in bold.

Table 4 – H_0 : no match between positively selected sites and N-glycosylation sites
 H_1 : match between positively selected sites and N-glycosylation sites

HIV-1 data set	Sites of N-glyc.	Conserved N-glyc.	Observed	Expected	P-value
HIV-1-MA	28	5	11	1.64	0.001
HIV-1-MB	27	2	5	1.62	0.009
HIV-1-MC	30	2	7	1.69	0.002
HIV-1-MD	22	5	4	1.17	0.023
HIV-1-O	39	6	13	2.46	0.001

Sites of N-glyc., total number of N-glycosylation sites in the data set; Conserved N-glyc., number of N-glycosylation sites that are conserved across all sequences of the data set; Observed, observed number of association between positively selected sites and sites of N-glycosylation; Expected, expected number of associations between positively selected sites and those of N-glycosylation, as calculated from the mean of the Monte Carlo simulated distribution; P-value, significance level at which Observed differs from Expected; significant values (P-value < 0.05) are indicated in boldface type.

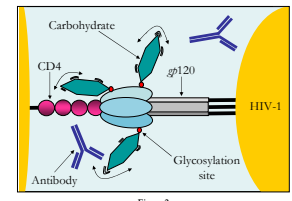


Figure 2 – The Cytosin Shield Model of Kwong and coworkers [8]. The glycosylation sites of the envelope protein anchor big carbohydrates which act as a shield by protecting the epitopes from neutralizing antibodies. Mutation at the glycosylation sites change the orientation of carbohydrates and thus makes this shield a moving structure.

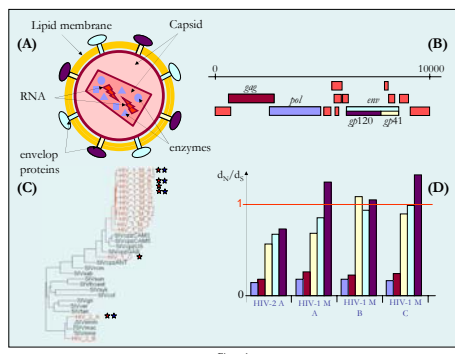


Figure 1 – (A) structure of the HIV B₁ gene map of the virus, showing the ORFs analyzed for positive selection. (B) ML tree of HIV-1/2 gene highlighting groups and subtypes analyzed. Blue stars indicate groups and subtypes which were analyzed for the *gag*, *pol*, and *env* genes (see D) and red stars indicate groups and subtypes which could have been analyzed only for the *env* gene (see Tables 1 to 4). (C) Mean ω ratios in *gag*, *pol*, *env*, *env*gp120, and *env*gp130 for HIV-1 group M subtypes A, B, and C, and HIV-2 subtype A, to A, B, and D, each ORF analyzed is symbolized by one green color (see blue for *env*).

Table 1 – Results of selection analysis

HIV-1 data set	No. of seq.	No. of codons	Mean ω	Highest ω	No. of sites
HIV-1-MA	16	578	0.690	4.70204	33
HIV-1-MB	30	578	0.623	4.08588	35
HIV-1-MC	30	578	0.610	4.46262	33
HIV-1-MD	15	578	0.568	3.82134	30
HIV-1-O	30	621	0.590	3.92448	40
HIV-2-A	22	679	0.444	3.56765	25

No. of seq. is the number of sequences in the alignments analyzed, and No. of codons is the number of codons in each sequence of the alignments analyzed. The mean ω is calculated by averaging over all the sites and all the selection classes. Highest ω is the value of ω in the positively selected class, and No. of sites is the number of sites in the positively selected class.

REFERENCES

- Choisy et al. 2004 Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *Journal of Virology* 87 (in press).
- Hansen et al. 1998 NetOGlyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycosylation Journal* 15: 115-130.
- Korber et al. 2000 HIV-1 molecular immunology. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA.
- Kulkarni et al. 2000 HIV-1 sequence compendium. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA.
- Kwong et al. 2002 HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* 420: 678-682.
- Swofford 2000 Phylogenetic analysis using parsimony (* and other methods). Version 4.06b. Sinauer Associates, Sunderland, MA, USA.
- Hansen 1997 PAML: a program package for the phylogenetic analysis by maximum likelihood. *CABIOS* 13: 555-556.
- Yang et al. 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155: 431-449.